The Pennsylvania State University

The Graduate School

College of Engineering

SYSTEMS ENGINEERING BASED APPROACHES FOR BIOLOGICAL NETWORK INFERENCE, ANALYSIS AND REDESIGN

A Thesis in

Chemical Engineering

by

Madhukar S. Dasika

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2007

The thesis of Madhukar S. Dasika was reviewed and approved* by the following:

Costas D. Maranas
Donald B. Broughton Professor of Chemical Engineering
Thesis Advisor
Chair of Committee

Patrick C. Cirino
Assistant Professor of Chemical Engineering

Antonios Armaou
Assistant Professor of Chemical Engineering

Réka Albert
Assistant Professor of Physics

Andrew L. Zydney
Walter L. Robb Chair and Professor of Chemical Engineering
Head of the Department of Chemical Engineering

*Signatures are on file in the Graduate School

ABSTRACT

Nature has created a wide range of diversity in the form of microorganisms with sometimes less than 200 genes to highly complex biological entities such as humans. All these biological systems are driven by an underlying complex cascade of biochemical reactions. These reactions are often represented as “networks”; these networks come in different flavors such as Metabolic, Signaling and Gene Regulatory. It is now clear that development of systematic procedures to analyze biological networks will lead to advances in biotechnology and therapeutics. To this end, in this thesis we develop computational approaches which can serve as powerful tools in various stages of biological research and discovery. Specifically, this thesis is composed of four parts that address inference, topological analysis and redesign of biological networks. While complementary these approaches are not necessarily interlinked.

The first part of this thesis focuses on development of a mathematical programming based framework to extract the underlying layer of complex component interaction networks using high-throughput biological data. Specifically, a mixed-integer linear programming based modeling and solution framework for inferring time delay in gene regulatory networks is developed. Solution of the model with real microarray data indicates that (i) The model predicts considerable number of interactions with a non zero value of time delay suggesting that time delay is prominent in gene regulation and (ii) Predicts a network that is more sparse and less sensitive to random fluctuations in gene expression, when time delay is accounted for.

Subsequently, optimization based frameworks are introduced for elucidating the input-output structure and redesigning large-scale cell signaling networks for pinpointing targeted disruptions leading to the silencing of undesirable outputs in the context of therapeutic interventions. The Min-Input framework is used to exhaustively identify all input-output connections implied by the signaling network structure. Results reveal that there exists two distinct types of outputs in the signaling network that either can be elicited by many different input combinations (i.e., degenerate) or are highly specific requiring dedicated inputs. The Min-Interference framework is next used to precisely pinpoint key disruptions that negate undesirable outputs while leaving unaffected necessary ones. In addition to identifying disruptions of terminal steps, we also identify complex disruption combinations in upstream pathways that indirectly negate the targeted output by propagating their action through the signaling cascades.

Next, we focus our attention to the development of computational tools construct and validate biological networks. Specifically, a discrete event based mechanistic simulation platform DEMSIM was developed for testing and validating putative regulatory interactions. The proposed framework models the main processes in gene expression, which are transcription, translation and decay processes, as stand-alone modules while superimposing the regulatory circuitry to obtain an accurate time evolution of the system. The stochasticity inherent to gene expression and regulation processes is captured using Monte Carlo based sampling. Overall, the results demonstrate the simulation framework's ability to make accurate predictions about system behavior in

response to perturbations and distinguish between different plausible regulatory mechanisms postulated to explain observed gene expression profiles.

Finally, we explore how optimization based approaches can be used to construct synthetic genetic networks that exhibit a specific function. Specifically, we introduce OptCircuit, an optimization based framework that automatically identifies the circuit components from a list and connectivity that brings about the desired functionality. The dynamics that govern the interactions between the elements of the genetic circuit are currently modeled using deterministic ODE's but the framework is general enough to accommodate stochastic simulations. The desired circuit response is abstracted as the maximization/minimization of an appropriately constructed objective function. The optimization framework is applied on a variety of applications ranging from the design of circuits that exhibit a specific time course response to circuits that discriminate between the presence, absence and level of external stimuli (e.g., genetic decoder). The results for the demonstrate the ability of the framework to (i) generate the complete list of circuit designs of varying complexity that exhibit the desired response; (ii) rectify a non-functional biological circuit and restore functionality by modifying an existing component and/or identifying additional components to append to the circuit.

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgements

I express my deep gratitude to my advisor Dr. Costas D. Maranas for his encouragement, support and guidance through out this course work. My experience working with him has had a profound impact on me and I will carry the many words of wisdom he has kindly shared with me throughout my life. I cannot overstate the amount of respect I have for him. Next, I would like to extend a special thanks to Dr. Antonios Armaou, Dr. Réka Albert and Dr. Patrick Cirino for agreeing to serve on my committee. I would like to express my sincerest gratitude to Dr. Anshuman Gupta who is more a mentor to me than a fellow worker. I am extremely thankful to him for all I have achieved during this course work. I thank the past and present members of my research group especially Dr. Anthony Burgard for their invaluable advice and guidance. I would like to also acknowledge both personal and professional help extended by all my friends (esp. Nitin Kumar, Vamsi Salaka, Dheeban Kannan, Sujit Nair, Amit Varshney and Praveen Depa) for making my stay in State College a memorable one. I wish them success in their endeavors. I would like to acknowledge the unconditional support extended by my girlfriend Savitha, my brother Sridhar and my sister Mrudula. I simply cannot put to words the amount of respect I have for my parents Leela and Suryanarayana for their unwavering support and encouragement they have provided and to them, I dedicate this thesis.

CHAPTER 1

Introduction

1.1 Background

Gene expression is the primary method through which a living organism processes the information stored in its DNA to form all functional cellular components. The underlying processes governing gene expression are the transcription and translation. Briefly the transcription process copies the information stored in the DNA duplex to an intermediate mRNA transcript. Subsequently, the translation process transfers this information into a functionally active protein. The proteins then perform key cellular functions such as transport, house keeping, signal processing etc. The complexity of this process is enhanced by the fact that a typical genome encodes for a number of proteins and subsequently the response of the organism to the extracellular and intracellular stimuli is driven by the complex cascade of underlying protein, gene and metabolite interactions. It is now clear that extracting and analyzing the interactions underpinning biological systems is an essential first step towards realizing the goal of understanding the response of cellular systems to stimuli with potential applications ranging from biotechnology to therapeutics.

Understanding whole-cell physiology has remained a daunting task for several decades. However, the research community has begun to take small but sure steps towards achieving this goal in the past decade. These accomplishments were made possible by numerous developments in experimental protocols that enable high-throughput biological data generation. Sequencing a genome, which only a few years ago was a tremendous feat, is now routine. Many examples of transcript array analysis can be found in the literature [1-6]; an exponential increase since the ground breaking work of Brown and co-workers on *S. cerevisae* [7]. Genomics, proteomics and metabolomics are now the cutting edge of physiological analysis [8-13]. These technologies when married with physiological measurements, genome-wide transcription measurements and genetic sequence generate huge tracts of integrated data. All this astounding complexity is being housed in numerous databases that characterize every step in the flow of information from the genome to the proteome. For example, Stanford microarray database supports the research of over 43 organisms and stores data generated from more than 60,000 microarray experiments [14]. Similarly, the KEGG database provides information regarding the genomes of over 512 organisms [15]. This data explosion currently overwhelming biology presents a challenging paradox: “you can see everything, but understand very little”. In response to this challenge the research community has recognized the need to develop efficient computational approaches to reconcile these huge tracts of biological information and provide testable hypotheses to *in vivo/in vitro* experimental protocols in order to accelerate research and discovery across life sciences.

These computational approaches can be broadly classified into two distinct categories depending on the nature of challenge they address. Specifically, the first class of frameworks may be classified as “network inference” based approaches. These approaches are typically “top-down” in nature as they use high level “snap shot” biological data to infer the bottom-level inner working of the cellular system. For example, Yeung et al. have used a singular value decomposition based approach to reverse engineer gene networks from microarray data. More recently Faith et al. has develop an algorithm named CLR (context likelihood of relatedness) to extract putative regulatory interactions in *E.coli* [16]. Other prominent examples of network inference approaches include Bayesian, Clustering, etc.

The computational approaches described in the previous paragraph provide network representation of cellular interactions. In addition to alleviating the seemingly intractable cellular complexity to a certain extent, a network representation of cellular interactions makes them amenable to computational frameworks that enable a systems-level perspective of the cellular system. Several examples of genome-wide cellular networks are now available across the literature. For example, the database RegulonDB provides a compilation of all known gene regulatory interactions in *E. coli* [17]. The *Reaction* entries in the TRANSPATH database [18, 19] allow the query of the upstream and downstream connectivity of signaling molecules by providing directionality and stoichiometry information for each interaction. The integration of TRANSPATH with TRANSFAC [20], a database for transcription factors and their DNA binding sites, provides the means to obtain complete signaling pathways from the binding of a ligand to the set of affected genes. The Alliance for Cellular Signaling (AfCS) [21], has brought forward the Molecule Pages database [22] which contains extensive information about more than 3,700 signaling proteins present in cellular signaling. The increasing availability of genome-scale cellular networks motivates the development of a second class of computational approaches termed as “network analysis based frameworks”. Several successful examples of network analysis approaches have been developed in the past to examine the organizational principles of cellular networks. For example, Albert has established that several naturally occurring networks obey the power law degree distribution [23]. Similarly, Milo et al. has established that large-scale cellular networks are build from more simple motifs such as feed forward loop, single input module etc. [24]. In addition to examining the organizational principles network analysis approaches also lend themselves in performing engineering interventions within cellular systems to accomplish a biotechnological/therapeutic objective. An outstanding example of this approach is the OptKnock framework developed by Burgard et al. who have implemented an optimization based approach to identify gene knockouts for overproduction of biochemicals [25].

The previous paragraph has focused on system identification and analysis approaches for biological networks. In contrast a complementary procedure embraces a "bottom-up" approach by deploying mechanistic level information of cellular components ("inner working") to construct modeling and simulation platforms that are parameterized to comply with experimental observations. The reductionist approaches in biology continue to provide a multitude of information regarding fundamental cellular components and their respective functionalities. For example, researchers now have a molecular level understanding of the regulation mechanism of several cellular components such as *lac* promoter and the light-switch mechanism of *arac* promoter [26]. Other prominent examples include *B subtilis* stress response [27]; Drosophila segment polarity [28] and *E. coli* SOS response [29]. By taking into account the fundamental cellular processes the "mechanistic simulation" based approaches enable the construction of models which may be employed to predict the behavior of cellular systems in response to novel perturbations *in silico*. These predictive capabilities confer upon these platforms the ability to test and validate alternative hypotheses regarding cellular processes.

The development of such platforms assumes an even greater significance with the advent of new discipline of synthetic biology which is defined as development of new biological components, parts and systems that are capable of exhibiting novel functions. An important consideration associated with the design and fabrication of genetic circuits is the proper matching of kinetic rates of individual elements of the circuits [30, 31]. Mechanistic simulation platforms provide a comprehensive description of the underlying circuit dynamics and hence can serve as an important tool to enable rational design of genetic circuit. The design of genetic circuits is further complicated by the number of ways in which one can choose and interconnect the basic components to accomplish a particular response. Hence there is a need to develop accurate integration schemes that choose and optimize circuit components to improve circuit performance. To date, several small synthetic gene networks that accomplish a specific functionality have been constructed and several researchers have employed synthetic circuits to investigate the dynamics and inner workings of more complex natural genetic networks. In addition to uncovering the design principles of natural genetic networks, synthetic genetic networks are now increasingly finding roles in applications ranging from biotechnology, medicine and bio-sensing. Hence the development of such platforms represents an important step in bringing to fruition the promise of synthetic biology.

The central theme of this thesis is the development of efficient computational and in particular optimization based approaches to address the questions and challenges raised in the previous paragraphs (see Figure 1.1). Specifically we explore the development of systems engineering based approaches for the following challenges:

(i) Inference of gene regulatory networks from high-throughput biological data.

(ii) Developing computational frameworks to perform topological analysis of large-scale cell-signaling networks and identify sites for targeted interventions to accomplish a therapeutic objective.

(iii) Developing mechanistic simulation platform to test/validate alternative gene regulatory hypothesis

(iv) Develop accurate circuit integration schemes that couple these platforms with optimization algorithms to aid the progress of synthetic biology.

1.2 Thesis Overview

The following chapters are introduced in this thesis are focused in development of computational and modeling frameworks to address challenges involved in biological network inference, analysis/redesign and validation.

The last few years has witnessed the development of a number of computational approaches aimed at unraveling the regulatory circuitry from microarray data. In spite of the progress made these approaches do not account of several key biological features such as time delay. From a biological viewpoint, time delay in gene regulation arises from the delays characterizing the various underlying processes such as transcription, translation and transport processes. Consequently, accounting for this key attribute of the regulatory structure is essential to ensure that the proposed inference model accurately captures the dynamics of the system. This issue constitutes the core of Chapter 2. Specifically, we introduce an optimization based modeling and solution framework for inferring gene regulatory networks accounting for time delay. Computational experiments are performed for both *in numerous* and real microarray data sets and results reveal that considerable number of interactions have a non-zero value of time delay suggesting that time delay is ubiquitous in gene regulation leading to a networks that are more sparse and less sensitive to random fluctuations in gene expression. The complete description of the mathematical frameworks and the obtained results can be found in

- Dasika, M S., A. Gupta and C.D. Maranas (2004). "A Mixed Integer Linear Programming (MILP) Framework for Inferring Time Delay in Gene Regulatory Networks" Pac. Symp. Biocomp., 9, 474-486.

In chapter 3, we shift our focus to development of network analysis approaches. Specifically, motivated by the observation that the knowledge regarding the topology and connectivity information of signaling networks far outweighs our knowledge regarding their kinetics, we raise the question, "How much information can we extract from signaling networks using their topology information alone ?". We introduce optimization based frameworks for elucidation of input-output structure of signaling networks and for pinpointing targeted disruptions leading to silencing of undesirable outputs in therapeutic interventions. Key challenges associated with modeling the activating and inhibiting interactions embedded in signaling networks are discussed and computational experiments are performed on a large-scale cell signaling network implicated in Prostate cancer. First a mixed-integer linear programming framework termed the Min-Input framework is used to exhaustively identify all input-output connections implied by the signaling network structure. Next a bi-level optimization procedure, the Min-Interference framework is developed to precisely pinpoint key disruptions that negate undesirable outputs while leaving unaffected necessary ones. Overall, our results demonstrate that the developed computational frameworks can help elucidate the input/output relationships of signaling networks and help guide the systematic design of interference strategies. A discussion of the proposed methods and the obtained results can be found in

- Dasika, M S., A. Burgard and C.D. Maranas (2006). “A Computational Framework for Topological Analysis and Targeted Disruption of Signal Transduction Networks” Biophy. Jour., 91, 382-398.

In chapter 4, we turn our attention to the development of mechanistic simulation platforms to test and validate regulatory hypotheses. Motivated by numerous parallels between manufacturing processes and gene expression and regulation, we describe the development of a discrete event based mechanistic simulation platform DEMSIM. The event based modeling of fundamental processes underlying gene expression, which are transcription, translation and decay processes; its integration with the regulatory circuitry as well as accounting for inherent stochasticity is discussed. Overall, the results demonstrate the simulation framework’s ability to make accurate predictions about system behavior in response to perturbations and distinguish between different plausible regulatory mechanisms postulated to explain observed experimental behavior. The discussion of the simulation platform can be found in

- Dasika, M S., A. Gupta and C.D. Maranas (2005). "DEMSIM: A discrete event based mechanistic simulation platform for gene expression and regulation dynamics", Jour. of. Theor. Biol., 232(1), 55-69.

In recent years significant progress has been made in the field of synthetic biology leading to advances on both experimental and computational fronts. For example, the research community has been moving towards standardization by creating the Registry of Standard Biological parts (http://parts.mit.edu/). This registry provides a compilation of well-defined elements of a genetic circuit such as promoters, ribosome binding sites, transcripts, inducer molecules, terminator sites and plasmids among others. The impetus is that these spare parts registries will help usher the development of more rational engineering approaches for designing such circuits. The potential of using modeling and computations to better understand the function of these circuits has already been recognized [32-35]. The key observation on which this part of the thesis is based on is that the task of constructing biological circuits to meet multiple inducer specific requirements is a challenging one and accurate matching of kinetic rates of individual circuit elements is essential to ensure their functionality. We describe OptCircuit, an optimization based framework that automatically identifies the circuit components from a list and connectivity that brings about the desired functionality. The key features of the framework are accounting for underlying mechanistic detail to ensure accurate parameter matching and abstraction of the targeted circuit response as an appropriately constructed objective function. Overall, our computational results reveal the ability of the framework to synthesize circuits that exhibit a wide array of responses (inducer presence/absence, inducer concentration dependent). Further, our results also demonstrate that OptCircuit can suggest circuit configurations that go beyond the ones compatible with digital logic-based design principles as well as pinpoint parameters for modification to rectify an existing (non-functional) biological circuit and restore functionality. A discussion of the proposed methods and a description of the results can be found in

- Dasika, M S and C.D. Maranas (2007). " OptCircuit: An Optimization based method for computational design of synthetic genetic circuits", (Under review)

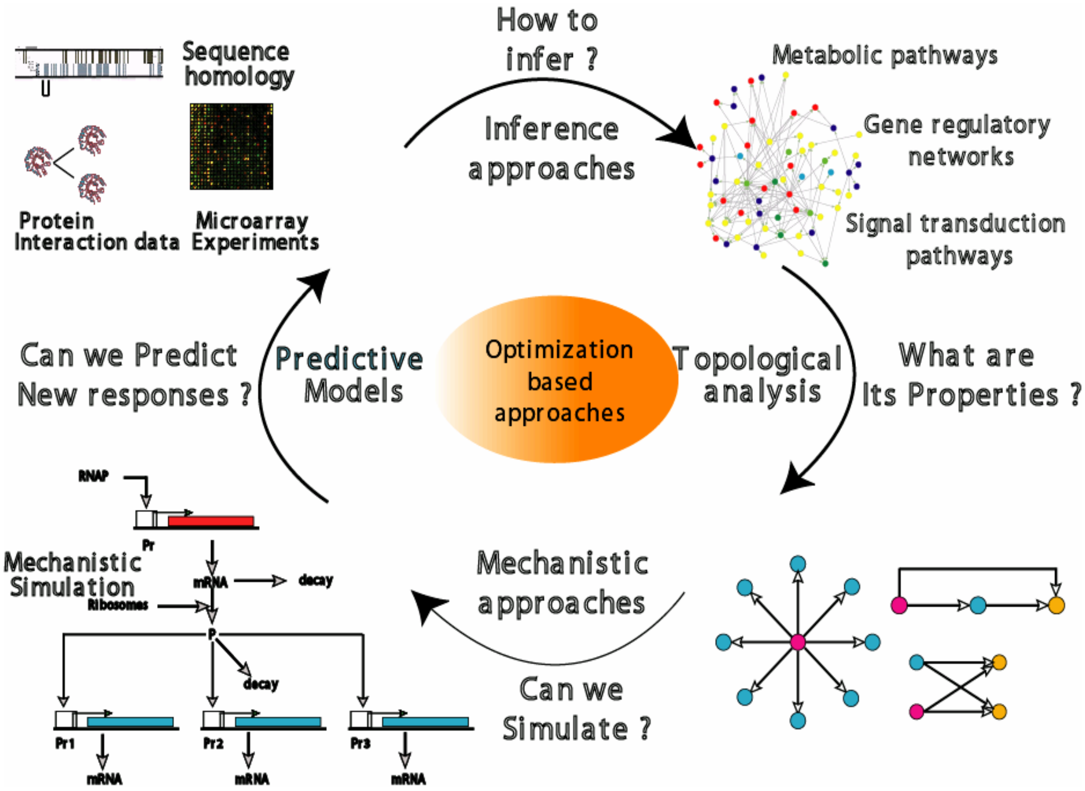


Figure 1.1: A pictorial overview of challenges addressed in this research. Computational and in particular optimization based approaches are developed to infer, analyze/redesign and validate biological networks.

CHAPTER 2

Identifying Gene Regulatory Networks Using Microarray data

2.1 Background

The advent of microarray technology has made it possible to gather genome-wide expression data. In addition to experimentally quantifying system-wide responses of biological systems, these technologies have provided a major impetus for developing computational approaches for deciphering gene regulatory networks that control the response of these systems to cellular and environmental stimuli. A complete understanding of the organization and dynamics of gene regulatory networks is an essential first step towards realizing this goal [36, 37]. To date, many computational/algorithmic frameworks have been proposed for inferring regulatory relationships from microarray data. Initial efforts primarily relied on the clustering of genes based on similarity in their expression profiles [38]. This was motivated by the hypothesis that genes with similar expression profiles are likely to be co-regulated. Hwang *et.al* [39] and Stephanopoulos *et.al* [40] extended these clustering approaches to classify distinct physiological states. However, clustering approaches alone cannot extract any causal relationship among the genes. Many researchers have attempted to explain the regulatory network structure by modeling them as Boolean networks [41, 42]. These networks model the state of the gene as either ON or OFF and the input-output relationships are postulated as logical functions. Measures of transcript levels, however, vary in a continuous manner implying that the idealizations underlying the Boolean networks may not be appropriate and more general models are required [43].

Recently, there have been many attempts to develop approaches that can uncover the extent and directionality of the interactions among the genes, rather than simply grouping genes based on the expression profiles. These approaches include the modeling of genetic expression using differential equations [44-46], Bayesian networks [47] and neural networks [48]. Even though a lot of progress has been made, key biological features such as time delay have been left largely unaddressed in the context of inferring regulatory networks. Experimentally measured time delay in gene expression has been widely reported in literature [49-51]. However, on the computational front, the fact that gene expression regulation might be asynchronous in nature (*i.e.*, the expression profile of all the genes in the system may not be regulated simultaneously), has largely been left unexplored.

From a biological viewpoint, time delay in gene regulation arises from the delays characterizing the various underlying processes such as transcription, translation and transport processes. For example, time delay in regulation may result due to the time taken for the transport of a regulatory protein to its site of action. Consequently, accounting for this key attribute of the regulatory structure is essential to ensure that the proposed inference model accurately captures the dynamics of the system. Prominent among the initial efforts made to incorporate time delay is the framework developed by Yildirim and Mackey [52]. The authors examined the effect of time delay in a previously developed mechanistic model of gene expression, in the *Lac* operon [53]. Chen *et. al* [44] proposed a general mathematical framework to incorporate time delay but did not apply it to any gene expression data to produce verifiable results. While interesting, these methods are not scalable to large expression data sets where the mechanistic details are often absent. Quin *et. al* [54] have proposed a time-shifted correlation based approach to infer time delay using dynamic programming. Since this approach relies on pair-wise comparisons, it fails to recognize the potential existence of multiple regulatory inputs with different time delays.

In this chapter, we describe an optimization based modeling and solution framework for inferring gene regulatory relationships while accounting for time delays in these interactions using mixed-integer linear programming (MILP). We compare the proposed model, both in terms of its capability to uncover a target network that exhibits time delays for a test example, as well as computational requirements with a model that does not account for time delay. The rest of the chapter is organized as follows. In the following subsection, a detailed description of the proposed model formulation is provided. Subsequently, the performance of the proposed model is evaluated on two data sets (one *in numero*, one real). Finally, concluding remarks are provided and the work is summarized.

2.2 Methods

Here, an inference method is described for extracting the regulatory inputs for each gene in a genetic regulatory network, while accounting for time delays in the system. To this end, the linear model of network inference [55-57] is adopted as a benchmark and modified to account for time delay as shown in Eq 2.1.

$$\dot{Z}_i(t)=\frac{Z_i(t+1)-Z_i(t)}{\Delta t}=\sum_{\tau=0}^{\tau^{\max}}\sum_{j=1}^{N}\omega_{ji\tau}Z_j(t-\tau)\ \forall\ i=1,2,...\mathrm{N},t=1,2,...\mathrm{T} \qquad (2.1)$$

In Eq 2.1, $Z_i(t)$ is the expression level of gene i at time point t and $\omega_{ji\tau}$ is the regulatory coefficient that captures the regulatory effect of gene j on gene i. The index τ indicates that this regulation has a time delay of τ associated with it while the integer parameter $\tau^{\max}$ denotes the longest time delay accounted for. Note that the frequency at which gene expression is sampled through the microarray experiment determines the maximum amount of *biologically relevant* time delay that can be inferred. For example, if the time points are separated by seconds/minutes then a higher value of $\tau^{\max}$ can be used. Subsequently, if $\omega_{ji\tau}$ >0 then gene j activates gene i with a time delay τ ,while if $\omega_{ji\tau}$ <0 then it inhibits the expression of gene i. If $\omega_{ji\tau}$= 0 for some i, j, τ, then no regulatory connection is implied between the genes j and i with a time delay τ.

In a typical microarray time course expression data set, the expression levels for N genes are measured at T time points where $N >> T$. In order to uniquely determine all regulatory coefficients, $N^2(\tau^{\max}+1)$ equations are needed. However, only NT equations are available implying that the system is typically underdetermined and consequently there exists a family of solutions that fit the microarray data equally well. To reduce the dimensionality of the solution space we assume a single time delay τ for every regulatory interaction. Furthermore, we limit the maximum number of regulatory inputs to each gene. In order to impose both these constraints, boolean variables $Y_{ji\tau}$ are defined as follows.

$$Y_{ji\tau} = \begin{cases} 1 & \text{if gene } j \text{ regulates gene } i \text{ with a time delay } \tau \\ 0 & \text{otherwise} \end{cases}$$

Subsequently, the network inference model with time delay is formulated as the following mixed integer linear programming (MILP) model.

$$Minimize \quad E = \frac{1}{N \cdot T}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[e_i^{+}(t) + e_i^{-}(t)\right] \tag{2.2}$$

subject to

$$\dot{Z}_i(t) - \sum_{\tau=0}^{\tau^{\max}}\sum_{j=1}^{N}\omega_{ji\tau} Z_j(t-\tau) = e_i^{+}(t) - e_i^{-}(t) \quad \forall i = 1,2,\ldots,N; t = 1,2,\ldots,T \tag{2.3}$$

$$\Omega_{ji}^{\min} \cdot Y_{ji\tau} \le \omega_{ji\tau} \le \Omega_{ji}^{\max} \cdot Y_{ji\tau} \quad \forall i,j = 1,2,\ldots,N; \tau = 0,1,\ldots,\tau^{\max} \tag{2.4}$$

$$\sum_{\tau=0}^{\tau^{\max}} Y_{ji\tau} \le 1 \quad \forall i,j = 1,2,\ldots,N \tag{2.5}$$

$$\sum_{\tau=0}^{\tau^{\max}}\sum_{j=1}^{N} Y_{ji\tau} \le N_i \quad \forall i = 1,2,\ldots,N \tag{2.6}$$

$$Y_{ji\tau} \in \{0,1\} \quad \forall i,j = 1,2,\ldots,N; \tau = 0,1,\ldots,\tau^{\max} \tag{2.7}$$

$$e_i^{+}(t) \ge 0,\; e_i^{-}(t) \ge 0 \quad \forall i = 1,2,\ldots,N; t = 1,\ldots,T \tag{2.8}$$

The objective function (Eq 2.2) minimizes the total (over all genes and time points) absolute error *E* between the predicted and the experimental expression values. The absolute value of the error is determined from Eq 2.3 through the positive and negative error variables $e_i^{+}(t)$ and $e_i^{-}(t)$ respectively. For a given gene *i* and time point *t*, only one of these variables can be non-zero. Specifically, if the error is positive then $e_i^{+}(t)$ is non-zero while if the error is negative then $e_i^{-}(t)$ is non-zero. This property arises from the fact that when the constraints of the model are placed in matrix form, the columns associated with these two variables are linearly dependent. Consequently, the linear programming (LP) theory principle that states that the columns of the basic variables (variables that are non-zero at the optimal solution) are linearly independent ensures the above property. Eq 2.4 ensures that the coefficients for all regulatory relationships *not* present in the network are forced to zero. In this constraint, $\Omega_{ji}^{\min}$ and $\Omega_{ji}^{\max}$ are the lower and upper bounds respectively on the values of regulatory coefficients. Eq 2.5 imposes the constraint that each regulatory interaction, if it exists, may assume only a single value of time delay associated with it while Eq 2.6 limits N_i, the maximum number of regulatory inputs to gene *i*.

The proposed framework has a number of key advantages. The basic linear model with no time delay is a special case of the proposed model. It can be recovered by including the following constraints.

$$Y_{ji\tau} = 0 \; \forall \, i, j = 1,2,..., N, \tau > 0 \qquad (2.9)$$

Additional environmental stimuli may be incorporated by introducing an additional node that describes the influence of the stimulus into the network. Furthermore, various biologically relevant hypotheses can be tested by translating them into either additional/alternative constraints or objective functions. For example, one of the hypotheses recently proposed, concerns the robustness of gene regulatory networks, defined as the ability of these networks to effectively tolerate random fluctuations in gene expression levels [58, 59]. Within the context of the linear model, this translates into having small values of the regulatory coefficients $\omega_{ji\tau}$ so that small variations in the expression levels of gene *j* have a small impact on the rate of change of expression of gene *i*.

From a statistical perspective, the proposed framework can be used to capture the trade-off between degree of model fit and the number of model parameters. By systematically varying the number of maximum regulatory inputs to a particular gene and computing the resulting minimum error, a trade-off curve between accuracy and model complexity can be generated. This curve provides an appropriate means for determining the critical number of regulatory inputs above which the model is tending towards over-fitting of data.

In a system with N genes, there will be $N^2(\tau^{\max}+1)$ binary variables implying a total of $2^{N^2(\tau^{\max}+1)}$ possible alternatives for the network connectivity. Even for a relatively small network inference setting it is computationally expensive to conduct an exhaustive search through these alternatives. The computational requirements can be reduced, to a certain extent, by exploiting the decomposable structure of the proposed model. This is achieved by recognizing that the model can be solved for each gene *i* separately without any loss of generality. Note, however, that this model structure is lost if an overall maximum connectivity constraint is imposed in the same spirit as the individual gene maximum connectivity constraint (Eq 2.6). In addition to improved computational performance, another key advantage of the decomposable property is that it limits the amount of computational resources that need to be expended if only a sub-network involving a sub-set of the genes is to be inferred.

The key parameters that determine the computational complexity of the proposed model are the bounds $\Omega_{ji}^{\min}, \Omega_{ji}^{\max}$ imposed on the regulatory coefficients in Eq 2.4. While in certain special application settings, there are pre-specified upper and lower bounds that are part of the model, in contrast, in our proposed model, these bounds are not known a priori. For such cases, typically the "Big-M" approach is utilized whereby arbitrarily large/small bounds are imposed [60]. Such a simplistic approach circumvents the need to determine tight valid bounds, although, at the expense of much higher computational requirements. On the other hand, if tight invalid bounds are specified, the computational gains realized will be off-set by the inability to attain the global optimal solution. In light of this trade-off between computational requirements and quality of optimal solution, a sequential bound relaxation procedure is developed and described next. As a starting point for this procedure, for a given gene i^*, both the upper and lower bounds are fixed such that $|\Omega_{ji^*}^{\min}| = |\Omega_{ji^*}^{\max}| = \Omega^0_{ji^*}$. The initial value of the bound is selected based on the scaling of the expression values. Specifically, for gene *j*, this initial bound value is determined as a value proportional to the ratio of the order of magnitude of the derivative values and that of the expression values. Subsequently, given these bounds, the inference model is solved to obtain the optimal values of the regulatory coefficient $\omega_{ji^*\tau}(\Omega^0_{ji^*})$ and the absolute error $E_{i^*}(\Omega^0_{ji^*}) = \sum_{t=1}^{T}\left(e^+_{i^*}(t) + e^-_{i^*}(t)\right)$. Next, the bounds are relaxed such that $\Omega^1_{ji^*} = (1+\delta_{i^*})\cdot\Omega^0_{ji^*}$ where $0 < \delta_{i^*} \le 1$ followed by re-optimization of the model with these updated bounds. Since the relaxation of bounds leads to a larger feasible solution, it is guaranteed that $E_{i^*}(\Omega^1_{ji^*}) \le E_{i^*}(\Omega^0_{ji^*})$. These two steps of bound relaxation and optimization are repeated until the total absolute error for gene i^* reduces to/below the desired tolerance level. This procedure is then repeated for all the genes in the network until the entire (or a sub-set) network topology has been inferred.

2.3 Computational Results

To highlight and probe the inference capabilities of the proposed model, it is applied to 3 different data-sets as listed in Table 2.1. Two artificially generated data-sets are utilized to test whether the proposed model can infer back a network of know connectivity. Data set 1 is generated by assuming known time delay in the system dynamics. The computational performance of the model is studied by scaling-up the system in data set 2. Finally, a real microarray data-set is analyzed to highlight the applicability of the inference procedure to real biological systems.

2.3.1 Data Set 1

A 10 gene network is constructed and time delay was incorporated into the interactions. The network consists of 5 regulatory interactions with time delay zero, 4 regulatory interactions with time delay one and 3 regulatory interactions with a time delay two (see Figure 2.1). Gene 4 auto regulates it self with time delay one. Genes 1 and 5 have two regulatory inputs while the rest of the genes have one regulatory input. The initial expression values are assumed and the expression data is generated based on the weight and time delay values for the various regulatory relationships. Expression and derivative values are computed at 6 time points. Figure 2.2 shows the generated time series profiles for the 10 genes under consideration.

The results predicted by the models with and without time delay indicate that, while the model with time delay recaptures the assumed network of interactions, the model without time delay requires large number of parameters to characterize the system dynamics. Figure 2.3 shows the network connectivity and the corresponding weighs inferred by the model without time delay. Further the magnitude of the observed weights is much higher than the assumed weight values indicating that the predicted network is more sensitive to random fluctuations in gene expression. This example clearly reflects the need to account for time delay to effectively describe the system.

2.3.2 Data Set 2

To examine the model behavior under more number of genes, we constructed a 40 gene network which has time delay in some of the interactions. Six genes were had three regulatory inputs, 10 genes had two regulatory inputs, while the rest had one regulatory input. 33 regulatory interactions occurred with a time delay of zero, 20 with a time delay of one and nine with a time delay of two. Expression values were computed for each of the 40 genes at 8 time points. The derivatives were computed by employing forward difference. The starting value of the bound for each gene found to be around 1.0 and a bound increment value $\delta_i = 1.0$ was used for computation. Figure 2.4(a) shows the computational requirement for the models with and without time delay. Notice that there is no considerable difference in the computational time requirement for the models with and without time delay for most of the genes. We have plotted in Figure 2.4(b), the iterations required by the *sequential bound relaxation procedure* to terminate. Observe that when time delay is not accounted for, the number of iterations is higher, and hence resulting in a network that has low robustness. The progression of the relaxation procedure can be seen in Figure 2.5. Notice that as the number of iterations increase, the magnitude of the bound imposed on the regulatory interactions and the maximum number of regulatory inputs into the gene also increase. Hence the realized error falls as the algorithm progresses. As in the previous example the model with time delay was able to recapture the underlying network efficiently while the model without time delay requires predicts large number of network connections in order to concur with the expression data.

All the computations were performed using CPLEX 6.5[61]accessed through the modeling environment GAMS [62] on an IBM RS6000-270 workstation.

2.3.3 Data Set 3

The second microarray data set analyzed consisted of time course expression profiles of 24 genes of *Bacillus subtilis* subjected to an amino-acid pulse in minimal media. Gene expression is measured using Affymetrix GeneChip® arrays at 0, 8, 13, 18, 28, 38, 68, 118 and 178 minutes. The amino-acid pulse is introduced for 8 minutes at the start of the experiment. Subsequently, cubic splines are used to interpolate the expression data and the derivatives are computed by employing a local finite difference approximation at each of the time points. The model with time delay is used to infer the regulatory network. The trade-off curve between error and the maximum number of parents is shown in Figures 2.6(a) and 2.6(b) for both the model with and without time delay. Note that the maximum number of parents determines the number of parameters available for fitting. In accordance with the results obtained for data set 1, Figure 2.6

highlights the fact that for any imposed threshold error tolerance value, the model with time delay infers a network which is sparser.

The inferred regulatory relationships are shown in Figures 2.7 and 2.8. The proposed model is able to identify a number of regulatory relationships that have been previously reported in literature. Jin *et.al* [63] have hypothesized the existence of regulatory relationship between *citH* and genes involved in aspartate production (*nadB* and *purA*). The inferred regulatory network identifies a potential indirect mechanism for these regulations mediated by *pycA* and *odhB* (Figure 2.7). Miller *et.al* [64] have reported that genes *sdhA* and *citG* might share a common regulatory mechanism. The inferred network indicates that genes involved in glycine, serine and threonine (*yqhIJ*) metabolism regulate both *citG* and *sdhC,* which is a part of the *sdhCAB* operon. These results highlight the capability of the proposed inference framework to capture biologically plausible regulatory interactions.

From a statistical perspective, in addition to the relative error, a metric that is widely used to determine how well a regression fits is the coefficient of determination (or multiple correlation coefficient) R^2 [65]. This metric quantifies the fraction of variability in the response variable that can be explained by the variability in the input variables. In the context of our current setting, the average R^2 value is given by

$$R^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Var\left[\dot{Z}_i(t)\right] - Var\left[E_i(t)\right]}{Var\left[\dot{Z}_i(t)\right]}\right] \qquad (2.10)$$

where the $Var[\cdot]$ operator determines the variance of a particular quantity over time and $E_i(t) = e_i^+(t) - e_i^-(t)$ is the computed error for gene *i* at time point *t*. Given this metric, the *additional* variance explained by the model with time delay is determined as

$$\text{Add. Variance Explained} = R^2[\text{Time delay}] - R^2[\text{Without time delay}] \quad (2.11)$$

Figure 2.9 shows the additional variance explained for data-set 2 as the number of parents is varied. In addition to the real data set, the additional variance explained for random data is also shown in Figure 2.9. Specifically, the randomized data is obtained by permuting the rows and columns of the expression matrix such that any underlying structure of the data is lost while the scaling of the data is retained. The results of Figure 2.9 indicate that the model with time delay is able to discriminate between real and randomized data only when the maximum number of inputs allowed is either 4 or 5. For relatively small number of inputs (1,2 and 3), the model is unable to capture the underlying structure of the real data due to lack of sufficient number of parameters. Similarly, at the other extreme, when too many parameters are made available (6 and 7), the model starts tending towards over-fitting leading to the overlap between real and randomized data. A clear separation between the two data sets is realized only in the intermediate range of inputs (4 and 5). These results highlight the capabilities of the proposed modeling and solution framework in not only accounting for key system dynamics such as time delay but also gaining deeper insights into the topological features of regulatory networks.

2.4 Summary

In this chapter, an optimization based modeling and solution framework, for incorporating time delay in transcriptional regulations was proposed. The proposed model used the existing linear model as a benchmark and employed boolean variables to incorporate discrete time delay into the interactions. Since, the system of equations describing the interactions is underdetermined and consequently has a family of solutions that fit the data equally well, various properties of biological networks such as sparseness, and uniqueness of time delay were employed to search through the solution space. A number of key advantages of the model in terms of examining the impact of alternative objective functions, incorporating known biological interactions and including environmental stimuli were discussed. On the computational front, however, the proposed model formulation was NP-hard implying that the computational requirements increase exponentially with the model size. To alleviate this problem, a sequential bound relaxation procedure was proposed. The inferential potential of the proposed methodology was determined by applying it to an in numero data set and a real expression data set. Results for the in numero data set confirmed the fact that neglecting time delay in a system a priori known to be characterized with it results in a significant increase in the number of parameters needed to describe the system dynamics. Subsequently, application of the model to real microarray data uncovered numerous regulatory relationships with time delay suggesting that time delay is ubiquitous in gene regulation. In the spirit of the results obtained for the first data set, inclusion of time delay resulted in inferred networks that were sparser. In addition, analysis of the amount of variance in the data explained by the model revealed that the proposed methodology explained more variance in real data as compared to randomized data.

Table 2.1: Data sets utilized for testing the proposed inference procedure

Data Set	# Genes	# Time Points	Data Type	Time Delay
1	10	6	*In numero*	Yes
2	40	8	*In numero*	Yes
3	24	9	Real	?

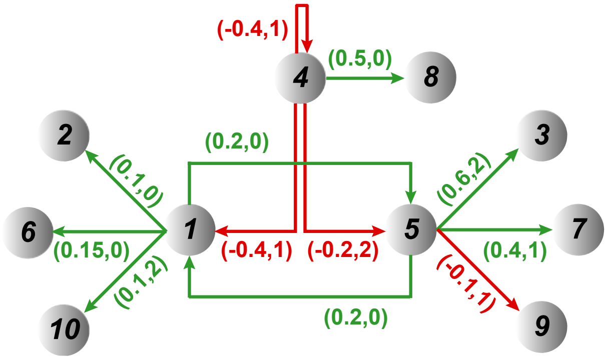


Figure 2.1: Network connectivity used for generating data set 1. The weight coefficients and time delays associated with each regulatory relationship are indicated. The red arcs imply inhibition or suppression of the activity of regulated gene, while the green arcs imply activation.

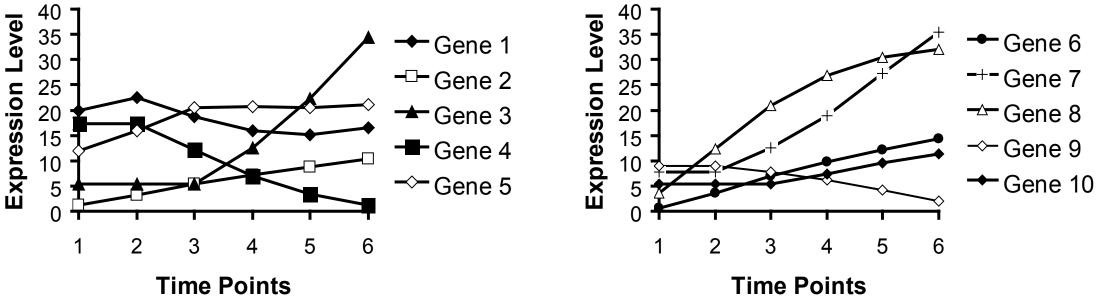


Figure 2.2: The expression levels of genes for data set 1.

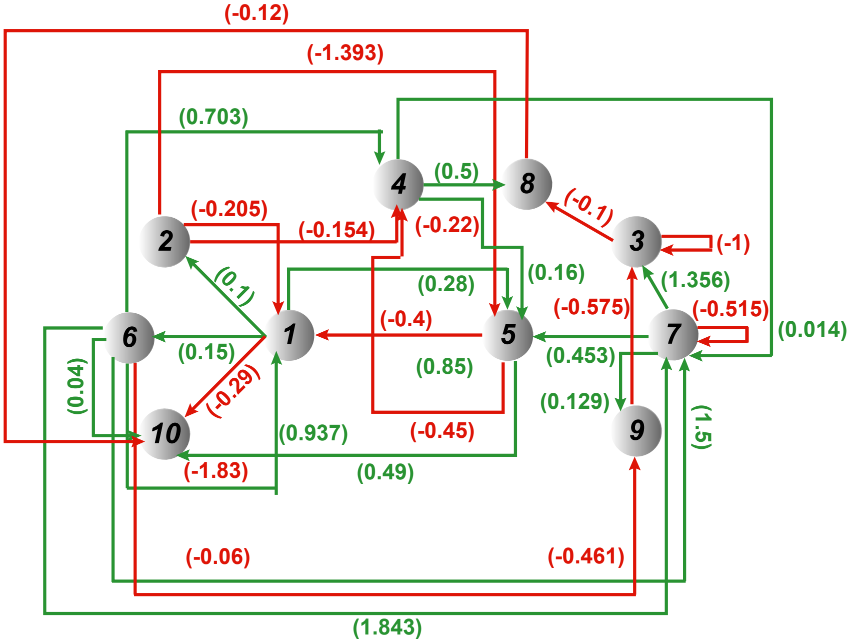
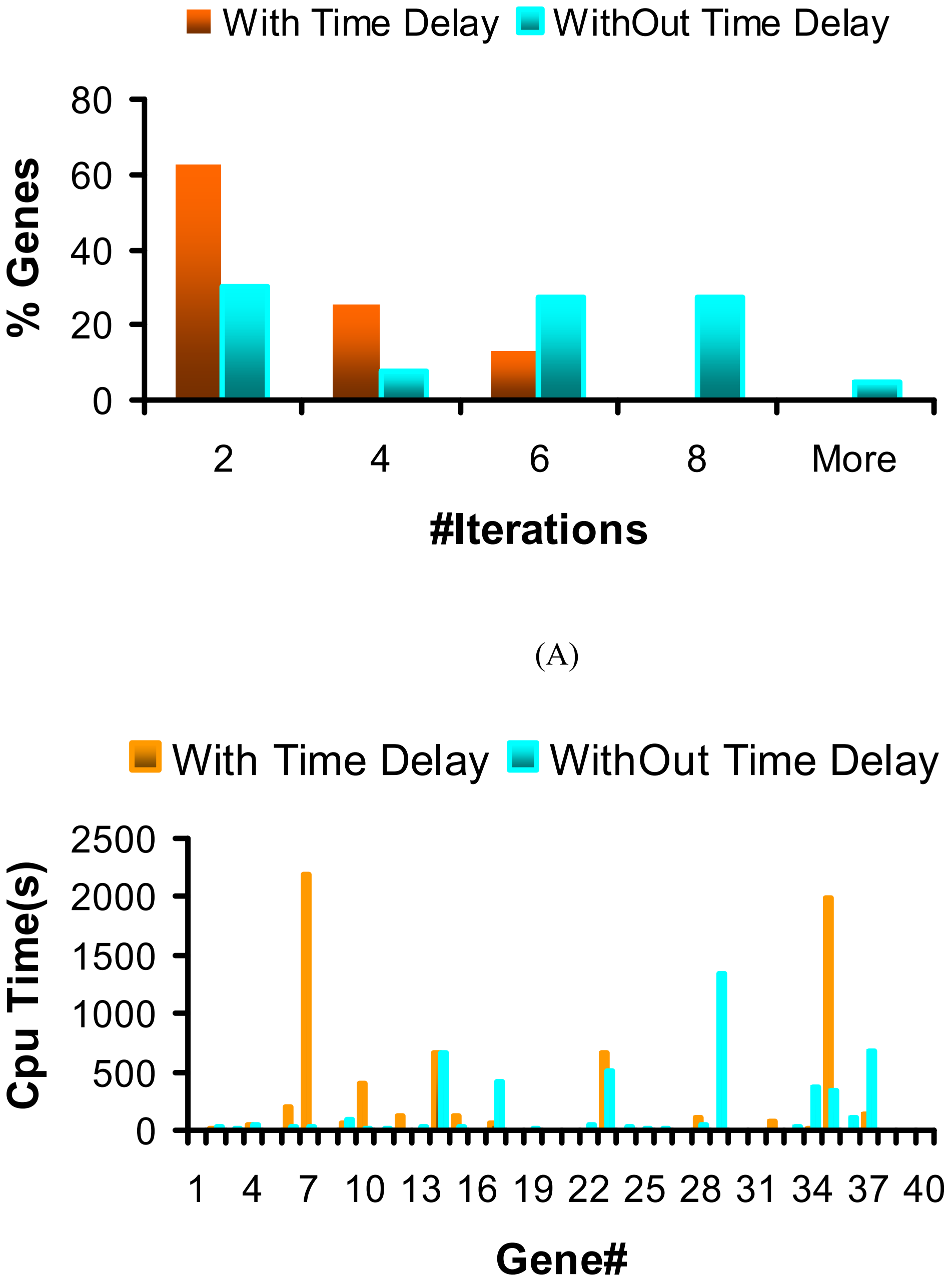


Figure 2.3: Network connectivity inferred by the model without time delay.



(A)

(B)

Figure 2.4: Comparison of computational performance for the model with and without time delay. (A) Distribution of the total number of sequential bound relaxation procedure iterations. (B) Total CPU time required for each of the 40 genes in the network

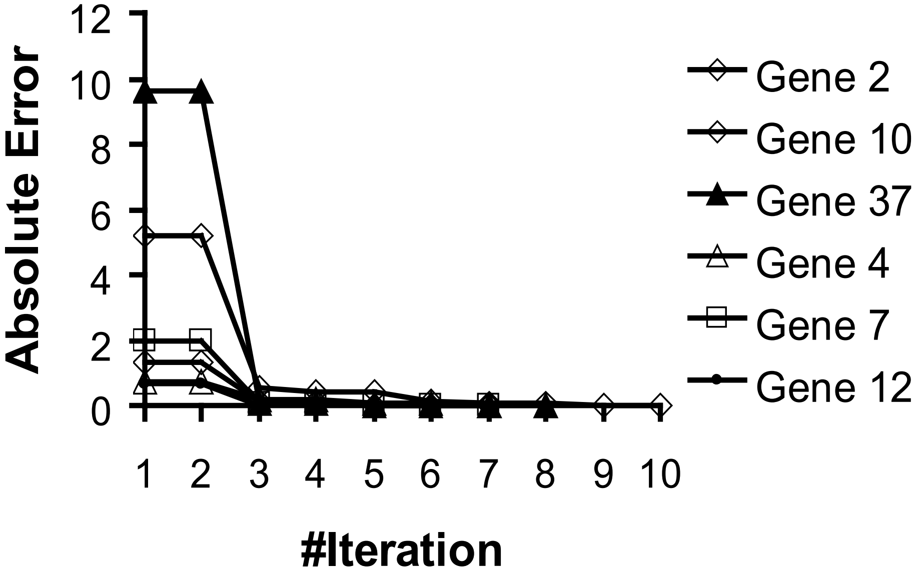
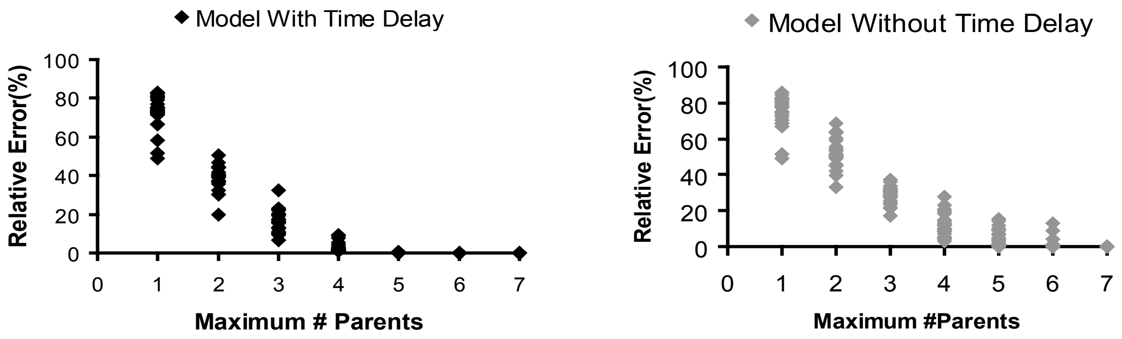


Figure 2.5: Progression of the sequential bound relaxation algorithm in terms of the reduction in error for a selected set of genes.



(A)

(B)

Figure 2.6: Trade-off between number of model parameters and quality of fit (A) Model with time delay (B) Model without time delay.

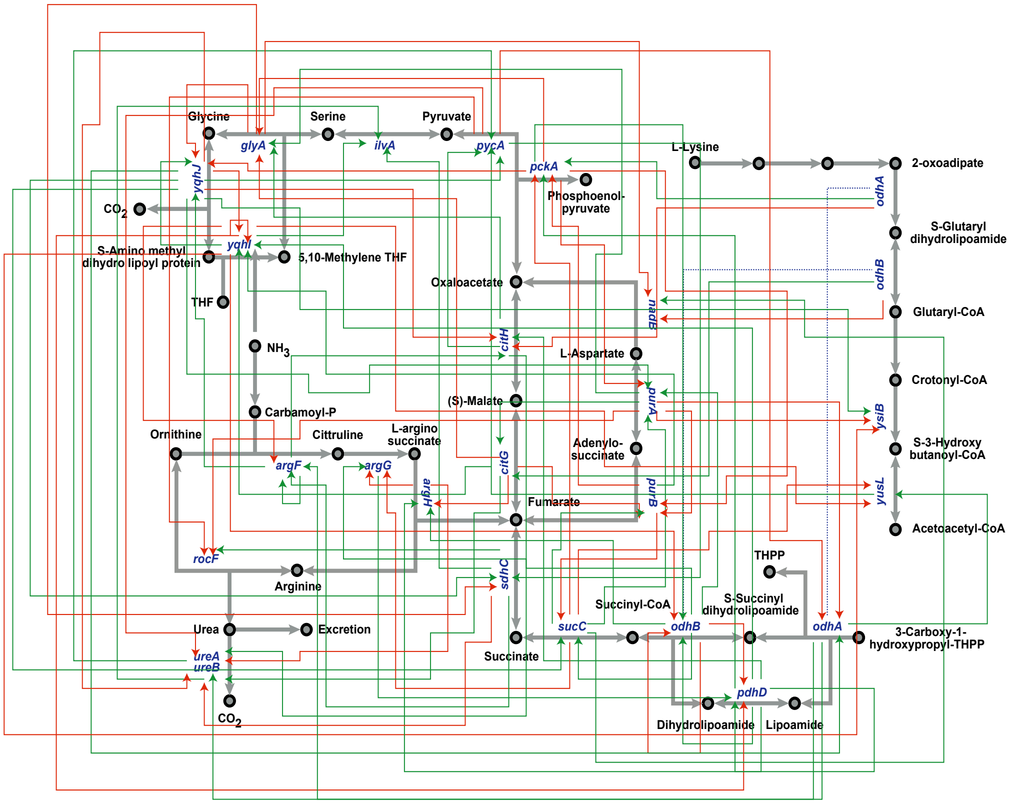


Figure 2.7: Regulatory network inferred by the model with time delay for the selected 24 genes of Bacillus subtilis (green: activation; red: inhibition; Threshold Error Tolerance of 10.0%)

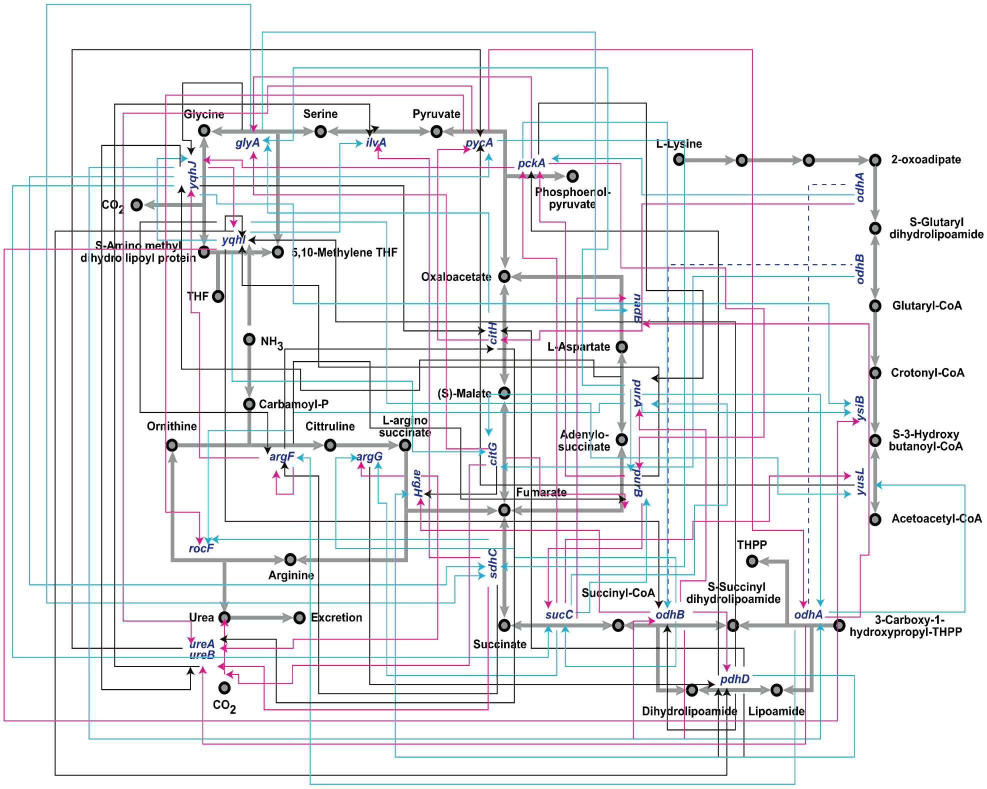


Figure 2.8: Time delays associated with the inferred regulatory relationships (black: τ = 0, pink: τ = 1, blue: τ = 2)

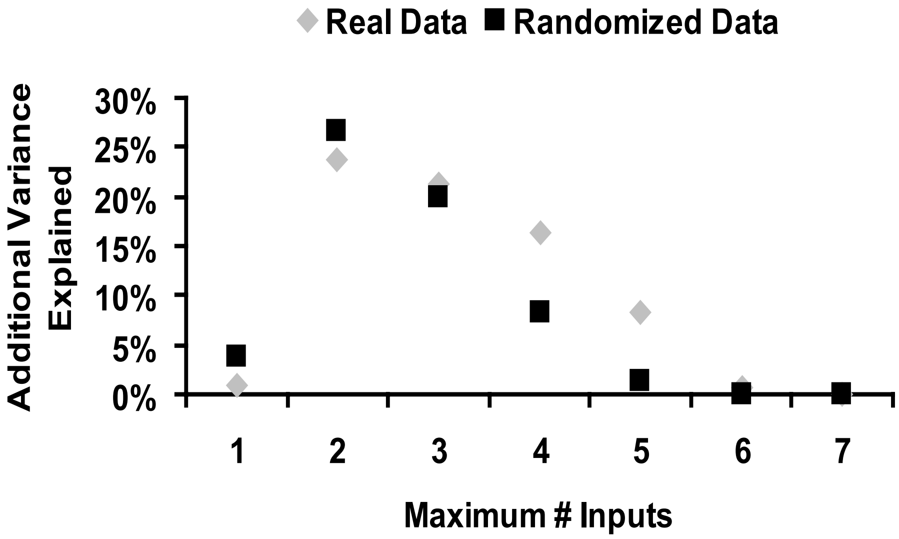


Figure 2.9: Additional variance explained by including time delay for real and randomized data.

CHAPTER 3

Topological Analysis and Targeted Disruption of Signaling Networks

3.1 Background

Recent years has witnessed an increasing interest in the study of cell signaling cascades as the critical role of these networks in various cellular events is becoming better understood. A typical signaling pathway involves the capture of extracellular signals and the subsequent transduction inward to control target proteins or gene expression [66]. For example, in response to stimulation by specific ligands the receptor tyrosine kinases regulate a great diversity of cellular processes including cell migration, cell proliferation and differentiation [67]. Similarly, the vascular endothelial growth factor (VEGF) family of ligands and receptors have been implicated in vascular development and neovascularization [68]. The connectivity of signaling networks is being unraveled at an ever increasing pace [69-71]. This brings to fore-front the challenge of devising novel strategies for systematically deducing the stimuli capable of eliciting a particular cellular response and deciphering how to "shape" their connectivity to negate undesirable outputs (e.g., *P70S6K* a suppressor of apoptosis) without affecting necessary ones (e.g., glycogen synthesis) [72, 73]. This chapter describes an integrated computational base for addressing these questions for large-scale signaling network reconstructions using a stoichiometric description of molecular transformations and boolean description of activations and inhibitions.

Genomic advances have provided a major impetus to the large-scale reconstruction of signaling pathways. Numerous databases are under development to catalogue the astounding complexity associated with cell signaling networks. For example, the *Reaction* entries in the TRANSPATH database [18, 19] allow the query of the upstream and downstream connectivity of signaling molecules by providing directionality and stoichiometry information for each interaction. The integration of TRANSPATH with TRANSFAC [20], a database for transcription factors and their DNA binding sites, provides the means to obtain complete signaling pathways from the binding of a ligand to the set of affected genes. The Alliance for Cellular Signaling (AfCS) [21], has brought forward the Molecule Pages database [22] which contains extensive information about more than 3,700 signaling proteins present in cellular signaling. Each entry, contributed by invited experts and peer-reviewed, contains information on a protein's known states including a list of sequence, kinetic, and thermodynamic parameters when available. Both the AfCS and TRANSPATH programs have the ultimate goal of providing the kinetic parameters necessary for the quantitative simulation of large signaling networks to aid in drug target discovery and evaluation. The Biomolecular Interaction Network Database (BIND) [74, 75] and the Database of Interacting Proteins (DIP) [76] store protein-protein interaction data representing approximately 15,000 and 11,000 interactions, respectively. Finally, the PANTHER^{TM} [77] database is a repository for cell signaling pathways and includes interactive resources for associating protein families with their biological pathways, as well as new tools for analyzing gene expression data in relation to molecular functions, biological processes, and pathways.

Signaling cascades were originally thought to function via linear routes where a single extracellular signal (i.e. input) would trigger a linear chain of reactions resulting in a single well-defined response (i.e. output) [78]. However nowadays, it is unanimously accepted that biological responses to external stimuli are much more complicated and the result of multiple interacting pathways containing many common molecules [79-82]. Many researchers have attempted to model and simulate the signaling cascades. These include modeling studies conducted on localized aspects of the cell signaling process such as the kinetic and spatial analysis of cell surface receptor mechanisms [83, 84], analyses of the cascades [85, 86], and analyses of specific signaling system modules [87-90]. Specifically, the Database of Quantitative Cellular Signalling (DOQCS) provides a repository of modules (i.e., less than 100 reactions) of signaling pathways containing approximately one-third of all published kinetic models of signaling pathways [69]. To date, the most detailed modular-type kinetic analysis involves the construction of a dynamic model of the MAP kinase cascade activated by epidermal growth factor (EGF) receptors [91]. This model describes the temporal concentration profiles of 94 compounds participating in 62 biochemical transformations triggered by EGF stimulation. Though impressive, a total of 94 compounds is still only a small fraction of the 2,503 unique chemical species that have so far been identified in humans and catalogued in the TRANSPATH database [18].

Faced with the paucity of accurate and comprehensive kinetic data, the key question is whether only the topology and connectivity alone of signaling networks can provide sufficient information to qualitatively predict their allowable states and responses to stimuli. Interestingly, a number of studies have shown that signaling networks are quite robust with respect to variations in kinetic parameters implying that their key properties may be largely established by their network architecture. For example, it has recently been deduced that the core topology of the interactions of the *Drosophila* segment polarity genes in differentiation was sufficient to deduce the properties expected of a developmental module, irrespective of the exact values of the kinetic parameters or initial conditions [92-94]. Furthermore, a boolean model of the segment polarity genes based solely on binary (0-1) representation of transcript and protein levels, was able to reproduce wild-type gene expression experiments along with expression patterns in various mutants and over-expression experiments [95]. In their study of the EGF signaling system, Schoeberl et al. (2002) concluded that the EGF induced responses were remarkably stable over a 100-fold range of ligand concentrations and were unexpectedly robust to variations in kinetic parameters and initial conditions. In *Escherichia coli* chemotaxis, the precision of tumbling frequency adaptation to external stimulant concentrations was found to be quite robust despite substantial variations in network-protein concentrations. [96, 97]. The local responses at each level of a signaling cascade have been shown to amplify enabling the total response of the cascade to operate almost as a switch where the target is activated in response to a given signal [98]. Lastly, it has been found that engineering the topology of signaling networks alone was able to change response specificity in *Saccharomyces cerevisiae* resulting in cells eliciting an osmolarity response to a mating signal [99, 100].

Therefore, despite the lack of detailed kinetic representations, the newly available large-scale signaling network reconstructions and their inherent parametric robustness motivates the need to explore computationally their signal transfer properties and possible redesigns. Specifically, the question of how many signaling inputs are required to elicit a particular cellular response has already drawn attention [70]. The examination of alternative sets of input signaling molecules that are capable of triggering the same response provides insight into the degeneracy built into signaling networks and their organizational principles. In another context, degeneracy has been shown to play an important role in the robust behavior exhibited by the cellular, metabolic and regulatory networks [101]. To address this need, we put forward an optimization based framework, (Min-Input problem) that exhaustively identifies all sets of input signaling molecules that are required to elicit a particular cellular outcome (See Figure 3.1A) in the context of large-scale signaling networks.

Dysfunctions in the signaling architecture have often been implicated in a wide range of diseases. For example, deregulation of ETS transcription factors results in formation of malignant cells leading to tumorous growth [102]. Similarly, dysfunctions in the activity of the receptor tyrosine kinases and corresponding signaling pathways have been linked to diabetes and cancer [67]. Several drug development studies focus on identifying therapeutic agents that are capable of disrupting a targeted set of chemical transformations within the signaling pathways through competitive binding.

Unfortunately, due to the complexity of the networks, the unintended consequences of these disruptions to desired outputs are not systematically explored. However, by considering system-wide reconstructions of signaling pathways the far reaching effects of these disruptions could be traced over the entire signaling cascades. To this end, we introduce the optimization-based framework (i.e., Min-Interference problem) that pinpoints the minimal combinations of chemical transformations that need to be disrupted to prevent an undesirable cellular response while preserving desired ones (See Figure 3. 1B).

The proposed computational base is demonstrated on nine human signaling pathways that have been implicated in the growth and development of prostate cancer (Table 1). This network, extracted from the PANTHER^{TM} [77] database of signaling networks, involves 322 chemical transformations and 526 distinct chemical entities. A description of the procedure used to download and process all pathway data is provided in the next section. Subsequently, the adopted mathematical description of the signaling network is highlighted followed by a detailed presentation of the computational frameworks for the Min-Input and Min-Interference problems including results and comparisons to data from open literature.

3.2 Mathematical Modeling

3.2.1. Pathway Data

Table 3.1 lists all nine pathways considered in this study to highlight the proposed computational frameworks. We used implication to prostate cancer as a selection criterion. Prostate cancer is the second highest cause of cancer related deaths in the US and many research efforts are currently directed towards elucidating the pathways whose upregulation or downregulation promotes malignant behavior [103].

Many of the chemical transformations in the signaling pathways are either activated or repressed by chemical entities present in the system. For example, protein tyrosine hydroxylase (*TH*) activates the transformation of *tyrosine* to 3,4-Dihydroxyphenylalanine (*DOPA*) in the adrenaline synthesis pathway. Similarly, the presence of *Akt* suppresses the recruitment of *capsase 9* in the angiogenesis pathway, which plays an important role in the blood vessel formation. Therefore, in addition to stoichiometry, representation of the network topology requires identification of the activation and inhibition agents and interactions. As explained before, the pathways investigated in this work were downloaded from the PANTHER^{TM} database of signaling networks in SMBL format [104]. PANTHER^{TM} is publicly available without restriction at http://panther.appliedbiosystems.com. We developed customized scripts using Perl [105] to mine the chemical transformations, chemical entities and activating and inhibiting interactions and convert them into a format readable by the GAMS [62] optimization environment. The final dataset consists of 322 chemical transformations, 526 chemical entities, 198 activation interactions and 38 inhibition interactions.

3.2.2. Basic Definitions

Signaling pathways are represented using a stoichiometric formalism which has been extensively used to model metabolic networks [106]. The key features of this formalism include explicit accounting of every chemical transformation such as binding, dimerization, and phosphorylation, and balancing around every chemical entity. The component balances governing a signal transduction network involving N = {1…n} chemical transformations and M = {1…m} chemical entities are as follows:

$$\frac{dC_i}{dt} = \sum_{j=1}^{n} S_{ij} \cdot r_j \text{ , } \forall \ i \in M \qquad (3.1)$$

Here C_i denotes the concentration of chemical entity i, S_{ij} is the stoichiometric coefficient of chemical entity i in chemical transformation j, and r_j is the corresponding flux of transformation j. The rate limiting steps in cell signaling processes are typically either receptor internalization or transcriptional regulation, both with time constants on the order of 10^2 seconds. The time constants for the signaling transformations are on the order of 1-10 seconds allowing a steady-state assumption to be invoked:

$$\sum_{j=1}^{n} S_{ij} \cdot r_j = 0, \ \ \forall \, i \in M \qquad (3.2)$$

Our reaction set considers the transcription factors as the endpoints and do not take into account the subsequent transcriptional regulation of targeted genes

3.2.3. Modeling activating interactions

Activators are chemical entities that act as catalysts and enable specific chemical transformations. In such situations, the corresponding chemical transformation can take place only if the requisite activator is present subject to the availability of the reactants. The following simple example explains how we model activation using only a stoichiometric description of chemical transformations.

Consider the chemical transformation A⟶B which is activated by a chemical entity X (Figure 3.2 (A)). Even though the presence of X is necessary to carry out the transformation there is no net change in the amount of X and thus an unambiguous stoichiometric coefficient value cannot be assigned to it. To overcome this dilemma, we duplicate the X chemical entity into X^R and X^P depending on whether X is a “reactant” or “product” species with respect to the reaction at hand (i.e., A⟶B). Accordingly, activation by X is modeled using the following simple reaction steps (see Figure 3.2 (B)).

$$\text{(production or input of X)} \longrightarrow X^R$$

$$A + X^R \longrightarrow B + X^P$$

$$X^R \longrightarrow X^P$$

$$X^P \longrightarrow \text{(consumption or output of X)}$$

The first reaction step ensures that all X present as input or generated through chemical transformations is denoted as X^R. This defines a pool X^R of “reactant” X which, only if available, could be used in the second step to carry out the A⟶B transformation which also converts X^R to X^P. The third step allows for X to directly flow from its “reactant” X^R to its “product” X^P form without having to necessarily participate in reaction A⟶B. Finally, the last step enables X to be consumed or become an output to prevent accumulation. This representation of activation enables a non-zero flux through the reaction A⟶ B if and only if activator X is available in the system. All the transformations modeling activating interactions are irreversible. It is important to note based on the above definitions the extent of the A⟶B reaction is constrained by the amount of X^R. However, this is not a problem because we are examining network properties of signaling pathways which are dependent upon the presence or absence of flow rather than exact values.

This formalism for modeling activation within a stoichiometric framework can be generalized for any chemical transformation. Based on the above definitions and by duplicating all activators into corresponding "reactant" and "product" pools, any reaction requiring activation by a single or multiple species can be expressed as the combination of the elementary steps described above.

3.2.4. Modeling inhibiting interactions

Inhibition interactions are ubiquitous in signaling pathways. Despite the conceptual similarity of inhibitions to activations we did not find an equivalent way to express them in a purely stoichiometric fashion. Therefore, we had to make use of binary variables Y_j (i.e., acting as ON/OFF switches) to model inhibited chemical transformations depending on the presence or absence of the inhibitor. The binary variable Y_j is defined as follows.

$$Y_j = \begin{cases} 1, \text{ implies reaction } j \text{ is active} \\ 0, \text{ implies reaction } j \text{ is disrupted} \end{cases}$$

The above condition for setting the values of Y_j along with the constraint

$$0 \le r_j \le U \cdot Y_j$$

ensure that the flux r_j is set to zero if $Y_j = 0$ and it can assume any value between 0 and U if $Y_j = 1$. The magnitude of parameter U was fixed at 10^3 for all the computational studies conducted in this work. We also define the set M_{ihb}^{j} as the set of inhibitors for transformation j.

The presence or absence of an inhibiting chemical species $i \in M_{ihb}^{j}$ is determined by examining if at least one chemical transformation leads to production of inhibitor *i* or *i* is supplied as an input. The amount of *i* produced by or supplied as input to the system is given by the term $\sum_{j \in J_P^i} S_{ij'} . r_{j'}$, where J_P^i is the set of chemical transformations (including input reactions) leading to the production of *i*.

If $\sum_{j'} S_{ij'} \cdot r_{j'} > 0,$ then inhibitor *i* is present and the flux through transformation j is set to zero. This condition is described mathematically as the following set of constraints:

$$Y_j \leq 1 - (\sum_{j'} S_{ij'} \cdot r_{j'}) / L, \ \forall \quad j \in N, i \in M_{ihb}^{j}, j' \in J_P^i$$

Parameter *L* is chosen so that the term $(\sum_{j'} S_{ij'} \cdot r_{j'}) / L$ is always less than one. Given the scale of flows in the network a value of L = 10^6 was sufficient for all computational studies presented in this chapter. Therefore, through the application of duplicated chemical species and binary variables both activating and inhibiting interactions are properly described.

3.2.5. Identifying and eliminating loops

Signaling networks are often characterized by the existence of cycles (i.e., loops) in the flow of information. Specifically, a chemical entity *i* participates in a loop if there exists a finite number of chemical transformations that starting from *i* can lead back to the formation of the chemical entity *i*. Cyclic motifs or loops lead to the formation of disjoint subnetworks which can have non-zero flows, at steady-state, even in the absence of required input signaling molecules. A practical manifestation of this is the incomplete identification of all inputs needed for an output. To illustrate this point consider the pathway posed by Papin and Palsson (2004a) for the generation of the STAT1 homodimer output from the input signaling molecules, *rIFNγ*, *JAK2*, *IFNγ*, *STAT1* and *ATP* (Figure 3.3(A)). Also shown in Figure 3a is one possible flux distribution that recruits input signaling molecules *ATP* and *STAT1* alone to produce *STAT1* homodimer. Hence, the presence of the loop at ***Interferon-γ JAK2 receptor ligand complex*** allows the assignment of non-zero flows towards the production of *STAT1* homodimer in the presence of *ATP* and *STAT1* even if the required inputs *rIFNγ*, *JAK2* and *IFNγ* are absent. To overcome this problem, we have developed a loop-breaking procedure that first identifies all linearly independent loops and subsequently breaks them by duplicating chemical species forming the junction points of the identified loops. Briefly, the procedure involves an iterative algorithm that selects each chemical entity and subsequently traces the path from the selected chemical entity to the input entities. If the same chemical entity is encountered again in the traced path, then a loop exists at the specific chemical entity. The chemical entity is subsequently duplicated to eliminate the loop and the algorithm is applied to the new network to find additional loops. This procedure is repeated until no loops are identified. For example, by applying the loop-breaking procedure to the example shown in Figure 3.3(A), a topologically equivalent loop-free network is obtained (see Figure 3.3(B)). It can be seen from Figure 3.3(B) that the network flow balance condition (Eq 2) now ensures the recruitment of all the five inputs towards the production of *STAT1* homodimer.

A composite block diagram indicating the important steps in the network modeling is shown in Figure 3.4. The statistics of the resulting network are summarized in Table 3.2. As shown in the Table, the curated network involves 1,338 chemical transformations of which 249 are inputs and 75 are the outputs and consists of 1,063 chemical entities. This network is used as the basis for all subsequent computational studies.

3.3 Computational Studies

3.3.1. Min-Input Problem

The identification of all combinations of input signals that could lead to a desired cellular response represents a significant challenge for large and highly interconnected signaling networks. The proposed approach exhaustively identifies the smallest non-decomposable sets of inputs that could elicit a particular cellular response (see Figure 3.1(A)) using an optimization based framework. The mathematical description of the optimization problem requires the definition of a number of sets that identify chemical entities that serve only as inputs M_{in} or outputs M_{out} respectively in the signaling pathways. Specifically, the stoichiometric coefficients S_{ij} for all inputs must be non-positive for every chemical transformation j. Similarly, the stoichiometric coefficients S_{ij} are non-negative for all outputs in all chemical transformations j.

Transport reactions provide input and output species with a way to enter and leave respectively the signaling pathways ensuring balanceability under the quasi steady-state assumption. Transport reactions form sets N_{in} and N_{out} respectively.

$N_{\text{in}} = \{j \in N \mid j \text{ is a source of an input}\}$

$N_{\text{out}} = \{j \in N \mid j \text{ is a sink for an output}\}$

Based on the above variable and set definitions the problem of identifying all minimal inputs capable of eliciting a desired output i^*, where $i^* \in M_{out}$, is posed as the following mixed-integer linear programming (MILP) problem.

$$Minimize \sum_{j \in N_{in}} Y_j \tag{3.3}$$

subject to

$$\sum_{j=1}^{N} S_{ij} \cdot r_j = 0 \qquad \forall \qquad i \in M \tag{3.4}$$

$$r_{Ou^{i^*}} \geq 1 \tag{3.5}$$

$$Y_j \leq 1 - \frac{(\sum_{j'} S_{ij'} . r_{j'})}{L}, \quad \forall \quad j \in N, i \in M_{ihb}^{j}, j' \in J_P^i \tag{3.6}$$

$$0 \leq r_j \leq U \cdot Y_j \qquad \forall \qquad j \in N \tag{3.7}$$

$$Y_j \in \{0,1\} \qquad \forall \qquad j \in N \tag{3.8}$$

The objective function minimizes the number of inputs required to allow a particular response (output i^*). Constraint (3.4) imposes the quasi-steady state condition. Constraint (3.5) ensures that the flux to the output transformation corresponding to the desired output (i.e. i^*) is non zero. Constraint set (3.6) accounts for the inhibition interactions. Finally constraint (3.7) ensures that the reaction flux r_j is set to zero if Y_j is equal to zero. Alternatively, if Y_j is equal to one then r_j can assume any value between 0 and U as described previously. The above formulation is solved sequentially for every chemical entity that has been characterized to be an output of the signaling network (i.e. for every $i \in M_{out}$) to extract the minimal sets of inputs for every output of the signaling network. Often times several non-decomposable sets of inputs exist that could elicit a particular cellular response. Exhaustively identifying all sets of inputs requires utilizing the above formulation in an iterative procedure while successively implementing constraints known as integer cuts. Specifically, we impose the constraint (3.9)

$$\sum_{j \in N_{in} | Y_j^{iter}=1} Y_j \leq \sum_{j \in N_{in} | Y_j^{iter}=1} Y_j^{iter} - 1 \qquad (3.9)$$

where, $Y_j^{iter}, j \in N_{in}$ corresponds to the values of binary variables obtained at a particular iteration. The constraints at each successive iteration are accumulated to exclude previously found solutions. If the problem becomes infeasible, then no other sets of inputs that can elicit the formation of the desired outcome remain and the procedure terminates.

Convex analysis based methods such as extreme pathway [107] and elementary mode [108, 109] analysis represent other alternatives for obtaining an exhaustive identification of all input/output structures. These approaches require the computation of all convex basis vectors that can represent every possible combination of reactions rates that are feasible to the network. However, convex analysis based methods have been found to have poor scalability when applied to large networks [110]. In contrast, the proposed formulation is based on linear programming (LP) and mixed-integer linear programming (MILP) principles and is scalable to thousands of chemical transformations. Similar LP and MILP based procedures have been successfully demonstrated on genome-scale metabolic networks in various microorganisms containing thousands of metabolic reactions [111-114].

In addition to solution tractability to large networks, the proposed formulation can be readily modified to address a number of biologically relevant questions. For example, the minimal sets of inputs that are required for attaining not just a single but multiple outputs can be identified by simply setting the flows through all the desired outputs greater than one. Similarly, the formulation can be modified to conduct an input/output feasibility analysis as in [70]. A feasible input/output relationship implies that given a set of signaling inputs, there exists a set of chemical transformations that lead to the production of the desired output. This is accomplished by replacing the objective function with $Maximize \quad r^{desired}$ and replacing constraint (3.5) with $r_{In^i} \geq 1 \quad \forall \quad i \in M_{In}^{ava}$

where $r^{desired}$ represents the flux on the desired output transformation and the set M_{In}^{ava} is the set of available inputs.

3.3.1.1 Computational Results

The breadth of questions that can be answered by solving the Min-Input problem and the biological insights obtained are highlighted by applying the procedure to the large-scale network model constructed from the pathways characterizing the growth and development of prostate cancer. Specifically, we address the following three key challenges in the context of the signaling network described in the Table 3.2.

(i) Identify the minimal number of inputs required to elicit a particular outcome.

(ii) Identify the degeneracy of a particular output by exhaustively enumerating all possible sets of input signaling molecules that lead to a particular outcome.

(iii) Analyze the interconnection between the inferred input/output structures.

By iteratively solving the Min Input problem once for each output we generated the distribution of the minimum number of inputs required to realize a particular output (see Figure 3.5). As shown in the Figure 3.5, the minimum number of required inputs to elicit an output ranges from as low as one input to as high as 15. This is a manifestation of the highly varied topologies of the identified input/output structures. We observe single linear paths to highly interconnected cascades (see Figure 3.6) depending on the output. Specifically, in the case of apoptosis, a single intracellular input of *capsase 3* protein needs to be provided (see Figure 3.6(A)). Alternatively, as shown in Figure 3.6(B), the input/output structure characterizing the formation of protein *Survivin* resembles a simple linear cascade. Finally, we find that input/output structure of phosphorylation of *BAD* is much more complex and is formed by multiple interacting linear cascades (see Figure 3.6(C)).

Next, the degeneracy of the outputs is examined by exhaustively identifying all sets of input molecules capable of triggering the response of a particular outcome. The distribution of the number of alternative sets of input molecules capable of eliciting a given response is shown in Figure 3.7. Interestingly, the probability distribution of output degeneracy is a convex function with a minimum in the middle and two maxima at the two extremes. This suggests that the examined signaling pathways are characterized by the existence of two distinct sets of outcomes (i.e., outputs) that are either highly degenerate or highly specific. For example, we find that nine alternative sets of input signaling molecules are capable of triggering the apoptotic machinery whereas there is only a single way of triggering the deregulation of the apoptotic machinery by enabling the activation of *NF-κB*. The presence of alternative strategies to realize an outcome can be rationalized as an evolutionary adaptation to protect against failure, thus improving response robustness [101]. Therefore, a high degree of degeneracy for a particular outcome may allude to the importance of the role played by that component in the cell.

In the previous paragraph we examined output degeneracy. Next, we quantify input degeneracy by identifying whether there exists any input signaling molecules that are highly recruited. The distribution of the number of input/output structures that require the participation of a particular input is shown in Figure 3.8. It can be seen that the number of input/output structures that recruit a particular input can be as low as one to as high as 130. This clearly demonstrates that most inputs are narrowly recruited to trigger only a handful of outputs, although, a few key inputs are implicated in triggering a large number of outputs. Specifically, 73% of inputs signaling molecules were found to be narrowly recruited (i.e. recruited by 10 or less input/output structures) whereas 10% of input signaling molecules were found to be highly recruited (i.e. recruited by 50 or more input/output structures). The complete list of highly recruited input signaling molecules is provided in Table 3.3. As expected energy transfer metabolites such as *ATP* and *GTP* and proteins such as *GRB2* and *SOS* were found to be highly recruited. This is in agreement with experimental observations that report that the protein *GRB2* is a crucial component linking the receptor tyrosine kinase pathways (eg. *VEGF*, *EGF*) with downstream proteins such as *RAS* and *SOS* [115]. The identification of highly/narrowly recruited input signaling molecules is important for developing targeted therapeutic interventions in signaling pathways. Specifically, interfering with a highly recruited signaling molecule is more likely to lead to side-effects by negating many possibly desirable outputs.

In summary, the described computational results demonstrate that the Min-Input framework is able to extract many important topological properties of signaling networks. We find that the topology of the input/output structures varies widely from simple linear paths to highly connected cascades. Output degeneracy tends to be either very high or very low while input degeneracy is very low for most inputs and very high for a few key inputs (exponential distribution). Clearly, input and output degeneracy plays a key role in understanding the organizational principles of signaling networks and crafting therapeutic interventions by blocking key transformations. In the next section, we describe how to systematically identify which transformations to disrupt in order to deny and/or enable different outcomes.

3.3.2. Min-Interference Problem

The results for the Min-Input problem indicate that cellular outputs can be stimulated by several different signaling molecules hinting at the enormous complexity associated with disrupting signal transduction. Given a set of input signaling molecules

(M_{in}), the Min-Interference problem pinpoints the minimal disruption strategies needed to prevent an undesirable cellular outcome while preserving the desired ones (see Figure 3.1(B)). At the core of the search algorithm is the bilevel optimization problem depicted pictorially in Figure 3.9. Bilevel programming problems are hierarchical optimization problems where the constraints of one problem (outer problem) are defined in part by a second parametric optimization problem (inner problem). Specifically, in the case of the Min-Interference problem, the inner level problem identifies the worst-case scenario response of the network by maximizing the flow to the undesirable response. The outer problem then guarantees that the solution of the inner problem is equal to zero by systematically disrupting a minimal number of transformations. A similar framework has been proposed before and successfully implemented for identifying gene knock-outs in metabolic networks leading to the overproduction of a particular metabolite [116].

It is important to emphasize that the presence of an inhibitor molecule leads to the disruption of the corresponding chemical transformation(s) without the need for any further action. Inhibitor molecules can either be inputs to the signaling network whose presence can be controlled or they can be produced through a set of chemical transformations. This implies that both the set of inputs present and the underlying chemical transformations in tandem determine the presence or absence of inhibiting species in the network. For example, in the context of the small pathway shown in Figure 3.10, recruitment of inputs *A*, *B*, and *D* implies that the formation of output *E* is blocked due to the production of inhibitor molecule *C* for the transformation $D \longrightarrow E$. Alternatively, assuming that only the input signaling molecule *D* is present in the system enables the production of output *E*. In the results described in this section we assume that all of the input signaling molecules (M_{in}) are present in the network. We also postulate that all inhibitor molecules that can be produced from the set of input signaling molecules (M_{in}) are present in the system. Therefore, the disruption targets identified by the Min-Interference problem are in addition to those chemical transformations that are not achievable due to presence of inhibitors. This assumption that all "reachable" inhibitors are present is described mathematically as follows,

$$\sum_{j \in N} S_{ij} \cdot r_j \geq 1, \forall\, i \in I^R$$

where the set I^R is identified by employing an input/output feasibility analysis for every inhibitor molecule [71]. This constraint forces a net production of each inhibitor *i* that is constitutively available to the system. The set of chemical transformations that are unreachable or disrupted by the presence of inhibitors N_{ihb}^{dis} is found by examining if at least one inhibitor *i* for that transformation is a member of set I^R.

The conceptual optimization model shown in Figure 3.9 is fleshed-out in full detail as follows. Given the set of input transformations (N_{in}), an undesirable output $i^* \in M_{out}$, a subset of desirable outputs $M_{out}^{des} \subset M_{out}$ to be preserved, and the set of inhibitor molecules (I^R), the bilevel optimization problem for identifying a disruption strategy to prevent an undesirable output while preserving the desired outputs is posed as follows.

$$\underset{\text{(by varying } Y_j\text{)}}{Maximize} \quad -r_{Ou^{i^*}} \qquad (3.10)$$

Subject to

$$\left(\begin{array}{llll}
\underset{\text{(by varying } r_i\text{)}}{Maximize} & r_{Ou^{i^*}} & & (3.11) \\
\text{Subject to} & & & \\
\sum_{j=1}^{N} S_{ij} \cdot r_j \ge 0 & \forall & i \in M - I^R & (3.12) \\
\sum_{j=1}^{N} S_{ij} \cdot r_j \ge 1 & \forall & i \in I^R & (3.13) \\
r_j \ge 1 & \forall & j \in N_{in} & (3.14) \\
Y_j \le 1 - \dfrac{(\sum_{j'} S_{ij'} . r_{j'})}{L} & \forall & j \in N, i \in M_{ihb}^{j}, j' \in J_P^{i} & (3.15) \\
r_{Ou^i} \ge 1 & \forall & i \in M_{out}^{des} & (3.16) \\
0 \le r_j \le U \cdot Y_j & \forall & j \in N & (3.17) \\
Y_j \in \{0,1\} & \forall & j \in N & (3.18)
\end{array}\right)$$

$$\sum_{j} (1 - Y_j) \le K \qquad \forall \qquad j \in N, j \notin N_{ihb}^{dis} \qquad (3.19)$$

The objective function (3.10) for the outer problem minimizes the flow to the undesirable outcome (i.e. i^*) while the objective function for the inner problem (3.11) maximizes the flow to the undesirable outcome. This is because the solution of the inner problem establishes the worst-case scenario for the system while the outer problem drives this worst-case flow to the undesirable output to zero by disrupting reaction steps. Disruption of chemical transformations either by inhibitor action or by targeted disruption eliminates reactions that consume the reactants involved in the disrupted chemical transformations. Consequently, this leads to the accumulation of some of these reactant species. This is allowed through constraint (3.12) which ensures that no deficit in the mass balance of any chemical species is present although a surplus or accumulation is allowed. Constraint set (3.13) ensures that, as discussed earlier, all inhibitor molecules that can be derived from the current input signaling molecules are present in the system. The inflow of input signaling molecules is switched on by setting them greater than or equal to one (constraint 3.14). Constraint (3.15) disrupts inhibited transformations (by setting Y_j = 0) if the corresponding inhibitor is present. Constraint (3.16) preserves the desired outputs (M_{out}^{des}) by ensuring that the flow to these outputs is possible. Constraint (3.17) forces the reaction flux corresponding to all disrupted chemical transformations in the network to zero and finally constraint (3.19) places an upper limit of *K* on the number of allowable interferences.

A mathematically valid disruption strategy is identified if the value of the objective function reaches zero implying that the transmission of the extracellular signal to the undesirable output is blocked. As in the case of the Min-Input problem, alternative interference strategies (i.e., multiple optima) are identified by implementing the above optimization problem within an iterative procedure where previously found solutions are excluded at each iteration by employing integer cut constraints. First, single disruptions are investigated by setting *K* equal to one. Multiple disruption strategies are investigated by successively increasing the value of *K* by one after all single disruption strategies are found.

3.3.2.1 Computational Results

The following studies were conducted to test the ability of the Min-Interference problem to elucidate targeted disruptions:

(i) Identify the minimal set of transformations that need to be disrupted to prevent each output separately (see Figure 3.11(A)).

(ii) Identify the minimal set of transformations that need to be disrupted to prevent each output separately while preserving the flow to a set of desirable outputs (see Figure 3.11(B)).

By iteratively solving the Min-Interference problem once for each output we generate the distribution of minimum interference strategies for disrupting a particular output (see Figure 3.12). Following from our assumptions stated in previous section, the bar for zero interference corresponds to outputs that are already inaccessible due to the presence of inhibitor molecules in the signaling network. Most of the outputs (i.e. 47) require a minimum of a single disruption to be blocked and only two outputs require a minimum of two disruptions.

As expected there exist multiple interference strategies that can block the formation of an undesirable outcome. We find that the Min-Interference framework is able to suggest both straightforward strategies involving the disruption of the final transformation(s) leading to the outcome and relatively less intuitive strategies that target transformations far upstream of an undesirable outcome. For example, consider the interference strategies to block the formation of complex *cJun-cFos*, a major component of the transcription factor *AP-1* which has been implicated for its role in tumor growth [117]. A straightforward strategy to block the formation of the complex involves simple disrupting the heterodimeration of transcription factors *cJun* and *cFos* (see Figure

3.13(A)). However we find a number of less intuitive strategies such as targeting the *MEKK1* dependent activation of protein *JNKK1* (see Figure 3.13(B)).

Overall a total of ten distinct disruption strategies (four single and six double) were found to block the formation of complex *cJun-cFos*. While the single disruption strategies were found to focus on transformations downstream of the *RAS-MAP* kinase cascade, the double disruption strategies target transformations within the *RAS-MAP* kinase cascade. Interestingly, *RAS-MAP* kinase cascades are known to participate in a diverse array of cellular programs including growth, proliferation and survival and several drug molecules have been developed to target these cascades as a means to eliminate undesirable outcomes [118]. For example, by employing the drugs U0126 and PD98059 it is possible to inhibit the phosphorylation of *MEK* [118]. Table 3.4 lists drug molecules that can carry out the identified disruption strategies demonstrating the relevance of identified targets.

In the second study we impose constraints that ensure that while a specific output is disrupted a set of desirable ones is left unaffected. This modification attempts to identify disruptions that are less likely to interfere with necessary biological processes. The set of desirable outputs here is listed in Table 3.5. The identified distribution of the minimum required number of disruptions for different outputs, while preserving the set of desirable outcomes (indicated by grey bars in Figure 3.12) is almost identical to the previous case. However, the number of alternative disruption strategies identified is found to be substantially decreased. For example, consider the disruption strategy identified previously to block the formation of complex *cJun-cFos*. The total number of interference strategies decreased from ten (four single, six double) to four (four single) when the flow to desirable outputs is preserved. Furthermore, the interference strategies are found to exclusively target the terminal transformations located downstream of the *RAS-MAP* kinase cascade rather than disrupting the initial steps governing the MAP kinase cascade as shown in Figure 3.14. In addition to eliminating complex *cJun-cFos*, we find that the disruption strategies that target the *MAP* kinase cascades also block the activation of *$cPLA_2$* and the expression of *Ets* transcription factors which play an important role in cell differentiation, cell proliferation, tissue remodeling and apoptosis [102] (see Table 3.5). These results indicate that when the flow to desirable outcomes is preserved, the number of alternative interference strategies decreases and the suggested strategies are found to predominantly target the terminal transformations of the signaling pathways.

The hypothesis that by preserving the flow to desirable outputs the likelihood of side-effects is reduced is next tested by considering two separate examples. First we explore blocking the formation of Endothelial nitric oxide synthase (*eNOS*) an endothelial-cell-specific isoform of NO producing enzyme. *eNOS* has been implicated in both angiogenesis and vasculogenesis suggesting that the modulation of *eNOS* may be a potent new strategy for the control of pathological neovascularization [119, 120]. As illustrated in Figure 3.15(A), one strategy for eliminating *eNOS* activity is to disrupt the transport of Ca^{2+} ion from endoplasmic reticulum. However this also results in loss of *$cPLA_2$* activity which is implicated in reduced fertility [121]. Alternatively, we find that targeting the Ca^{2+} dependent activation of *eNOS* within the plasma membrane as shown in Figure 3.15(B) preserves *cPLA$_2$* activity. Interestingly, this disruption strategy is identical to the action of Cavtratin, a cell-permeable peptide molecule which by inhibiting *eNOS* activity was shown to exhibit anti-tumor properties [119].

In the previous study we demonstrated that the identified disruption target can be accomplished by an existing drug molecule. Next, we describe an example where a disruption strategy is identified that has never been explored before. Research has implicated *Src* in the progression of tumor angiogenesis [122], qualifying *Src* as an attractive target for disruption. By disrupting both *VEGF* and *FGF* receptor ligand binding, the drug Thalidomide blocks *Src* activity (see Figure 3.16(A)) [123]. However, as shown in Figure 16a, disrupting *VEGF* and *FGF* receptor-ligand binding also interferes with the activation of proteins *VRAP*, *Sck* and *HSP27*. Experimental studies have shown that *VRAP* plays an important role in the progression of normal angiogenesis [124] and *HSP27* is known to aid in the survival and recovery of cells exposed to stressful conditions [125]. As predicted, we find that administering Thalidomide as a means of eliminating *Src* activity results in compromised wound healing and stop the normal reproductive cycle in women among other side-effects [126]. Alternatively, by imposing as a restriction the preservation of desired output in Min-Interference we identify, among others, a previously unexplored target involving the disruption of both *VEGF-VEGFR2* and *FGF-FGFR* mediated activation of protein *Src* (see Figure 3.16(B)) which preserves *VRAP*, *Sck* and *HSP27* activity.

3.4 Summary and Discussion

In this chapter, a computational base was introduced for the systematic analysis and targeted disruption of signal transduction networks. A stoichiometric formalism was adopted to model the complex network of interacting molecules in signaling pathways as a network of chemical transformations. The cellular stimuli to the signaling pathways were described as inputs to the signaling network while cellular responses as outputs. The developed frameworks were benchmarked by applying them to a large-scale signaling network constructed from nine signaling pathways known to play an active role in the growth and progression of prostate cancer.

First, we introduced the Min-Input framework to identify all cellular stimuli that can elicit the formation of a particular response. By exhaustively identifying all input/output structures Min-Input was able to extract a number of important topological properties of signaling networks. Specifically, we found that the outputs can be classified into two distinct sets, highly degenerate or highly specific depending on whether they can be elicited by many different input combinations or a few dedicated ones. This classification has important implications for guiding the development of therapeutic strategies. For example, interfering with highly recruited inputs molecules (e.g. *SOS*) is likely to impact many network functions whereas affecting inputs with dedicated participation is more likely to cause only a specific event. Similarly, blocking the formation of a highly degenerate outcome (for e.g. *cyclinD*) is hard to accomplish because it requires the disruption of multiple steps. Given a set of input signaling molecules the Min-Interference framework identifies the minimal set of disruptions needed to eliminate an undesirable outcome. Computational results for the prostate cancer study, indicated that the framework was able to suggest multiple disruption strategies that were biologically relevant as several drug molecules exist to carry out the identified disruptions. Furthermore, by proactively preserving desirable outputs disruption strategies were identified that were less likely to involve side-effects by contrasting them against the action and reported side-effects of existing drug molecules.

The Min-Interference framework can be used to study the signal transfer properties of signaling networks upon the action of existing drug molecules. Databases such as the Therapeutic Target database provide extensive information about the drug molecules and their known sites of action [127]. By disrupting transformations targeted by the drug molecule Min-Interference can pinpoint which outputs are blocked. For example, we found that Thalidomide in addition to eliminating *Src* activity, it also blocks the formation of *VRAP, Sck* and *HSP27*. Min-Interference can also be used to examine if a particular combination of drug molecules are effective when used in combination and not alone (i.e., exhibit drug synergy). This is particularly important for pathologies such as cancer, where multiple pathways may be dysfunctional requiring a combination of several drug molecules [103] for effective treatment.

The reconstruction of signaling networks is progressing with a fast pace [71]. Efforts are currently under way to identify “signature networks” that are highly specific descriptors of many diseases (e.g. renal cell carcinoma)[128, 129]. Whenever available, the Min-Interference can be used to identify ways to negate the occurrence of these “signature networks”. Nevertheless, it is important to emphasize that existing signaling reconstructions are inherently incomplete. Therefore, input-output structure (Min-Input) or disruption results (Min-Interference) are bound to, for some cases, reflect these missing links. However, results obtained that are inconsistent with biological knowledge or experiment can be used to come up with hypotheses for “filling in” gaps in signaling reconstructions. Furthermore, the lack of any kinetic information in the signaling network description can lead to an overestimation of the number of viable input-output structures embedded within a signaling network. However, this overestimation of the signaling network functionalities ensures that all identified disruption strategies will be valid for the true signaling network. Despite these limitations this work represents an important first step towards constructing an integrated computational base for elucidating the input/output structure and subsequently redesigning signaling networks.

Table 3.1: The signaling pathways investigated in this study. The pathways were obtained from the PANTHER ^{TM} database.

Pathways involved in Prostate Cancer cells
1) Angiogenesis 2) Apoptosis_signaling_pathway 3) Cell_cycle 4) EGF_receptor_signaling_pathway 5) Hypoxia_response_via_HIF_activation 6) Insulin_IGF_pathway_MAP_kinase_cascade 7) JAK_STAT_signaling_pathway 8) P53 pathway 9) PI3_kinase pathway

Table 3.2 Number of chemical transformations, chemical entities, input transformations and output transformations.

Statistics of the Network Model	
Chemical Transformations	1014
Chemical Species	1063
Input Transformations	249
Output Transformations	75

Table 3.3: Highly recruited input signaling molecules (50 or more input/output structures).

Input Signaling Molecules	# Input/Output Structures
p101_cytoplasm	50
FOXO_cytoplasmINAC	50
GPCRligand_	50
ComplexGPCRG_sub__alpha_G_sub__beta__gamma__cytoplasm	50
ComplexRasGDP_cytoplasm	63
ComplexGDPRas_cytoplasm	65
IRligand_	71
IR_cytoplasm	71
IRS_cytoplasm	71
PKB_cytoplasm	72
PI4,5P2_cytoplasm	120
p85_cytoplasm	123
p110_cytoplasm	123
SOS_cytoplasmINAC	128
GTP_cytoplasm	128
ATP_cytoplasm	130
Grb2_cytoplasm	131

Table 3.4: A list of the identified interference strategies to block the formation of *cJun-cFos* along with the list of available drug molecules reported to be able to block the targeted transformations.

#	**Type**	**Disrupted Transformation(s)**	**Outputs Blocked**	**DrugMolecule(s)**	**References**
1	Single	*R*7	*cJun-cFos*	CNI-1493/JIP-1	[118]
2	Single	*R*8	*cJun-cFos*	Retenoid Acid	[130]
3	Single	*R*9	*cJun-cFos*	52R	[131]
4	Single	*R*6	*cJun-cFos*	CEP1347	[118]
5	Double	*R*4	*cJun-cFos*	Azathioprine	[132]
		*R*3	*Ets*		
			$cPlA_2$		
6	Double	*R*4	*cJun-cFos*	Azathioprine	[132]
		*R*2	*Ets*	U0126/PD98059	[133]
			$cPLA_2$		
7	Double	*R1*	*cJun-cFos*	RKIP	[134]
		*R*4	*Ets*	Azathioprine	[132]
			$cPLA_2$		
8	Double	*R*5	*cJun-cFo*s	PN7051	[135]
		*R*1	*Ets*	RKIP	[134]
			$cPLA_2$		
9	Double	*R*2	*cJun-cFos*	U0126/PD98059	[133]
		*R*5	*Ets*	PN7051	[135]
			$cPLA_2$		
10	Double	*R*5	*cJun-cFo*s	PN7051	[135]
		*R*3	*Ets*	Azathioprine	[132]
			$cPlA_2$		

Table 3.5: The list of desirable outputs. The flow to these outputs is preserved while devising interference strategies for blocking the formation of complex *cJun-cFos*.

Desirable Outputs
Transcription_br_cellcycleprogression_Cytosol
Anti-apoptosis_Nucleus
ComplexI_kappa_BNF_kappa_B_Intracellular
a127_degraded_Intracellular
CSL_Nuclearmembrane
Ets_Nuclearmembrane
STAT3_Nuclearmembrane
STAT1_Nuclearmembrane
TCF_Nuclearmembrane
PLD_PlasmaMembrane
PLA_sub_2_endsub_PlasmaMembrane
Src_PlasmaMembrane
Survivin_PlasmaMembrane
IGFBP1_nucleus
Rb_LateG_sub_1_endsub_
a102_degraded_Mitosis
ComplexeNOSCa_super_2+_endsuper_PlasmaMembrane
ComplexcPLA_sub_2_endsubCa_super_2+_endsuper_PlasmaMembrane
VRAP_PlasmaMembrane
Paxillin_PlasmaMembrane
HSP27_PlasmaMembrane
Tumorsuppression_
Survival_
Gene_space_transcription_Cytosol
Genetranscription_nucleus
S6K_cytoplasm
GSK3_cytoplasm
Caspase-9_cytoplasm
Bad_cytoplasm
NOS_cytoplasm
ComplexFOXO14-3-3_cytoplasm
Cyclind_nucleus
GADD45_nucleus
scl-1_nucleus

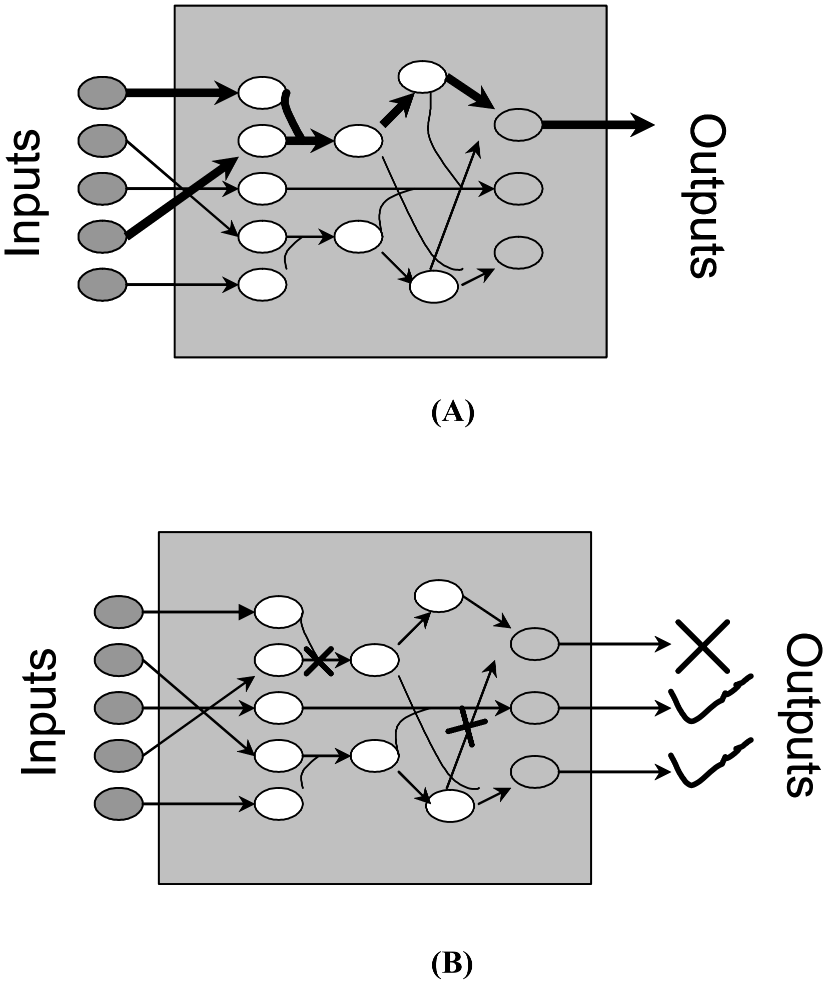


Figure 3.1: Pictorial representation of the problems and solution strategies proposed in this chapter. The Min-Input framework **(A)** identifies the minimal sets of input signaling molecules that are capable of eliciting a particular cellular outcome. The Min-Interference framework **(B)** identifies the minimal combinations of disruptions to prevent an undesirable outcome while preserving a set of the desired outputs.

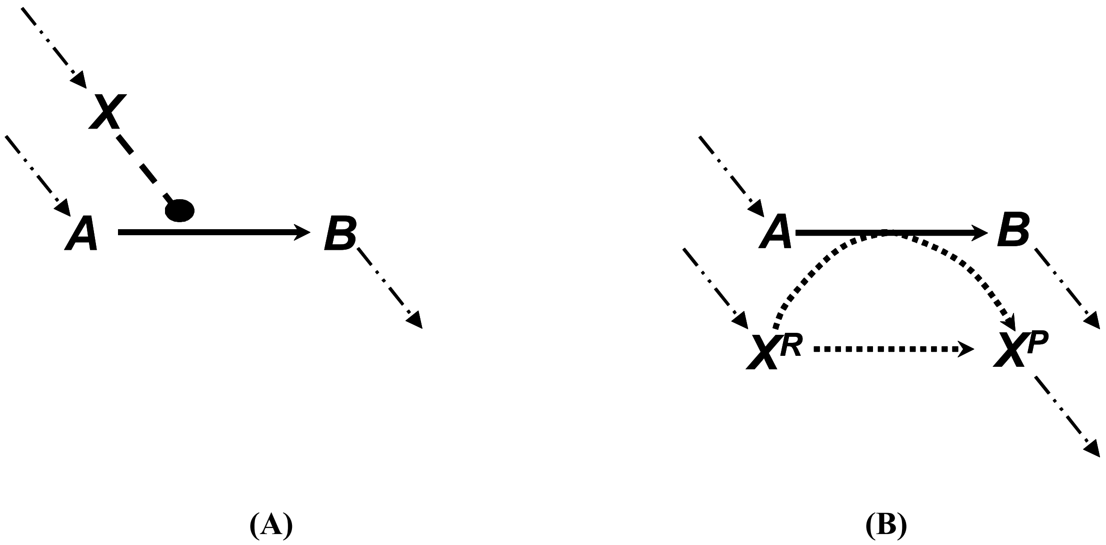


Figure 3.2: Modeling activating interactions. Shown in **(A)** is the chemical transformation A⟶B which is activated by the chemical entity X. Activation interaction is indicated by an arrow with a dot at its tail. Shown in **(B)** is the pathway resulting after accounting for activation. The activator X is duplicated as X^R and X^P. The new transformations introduced are represented by dotted arrows. In both the figures, the dashed arrows account for the production and transfer of chemical entities.

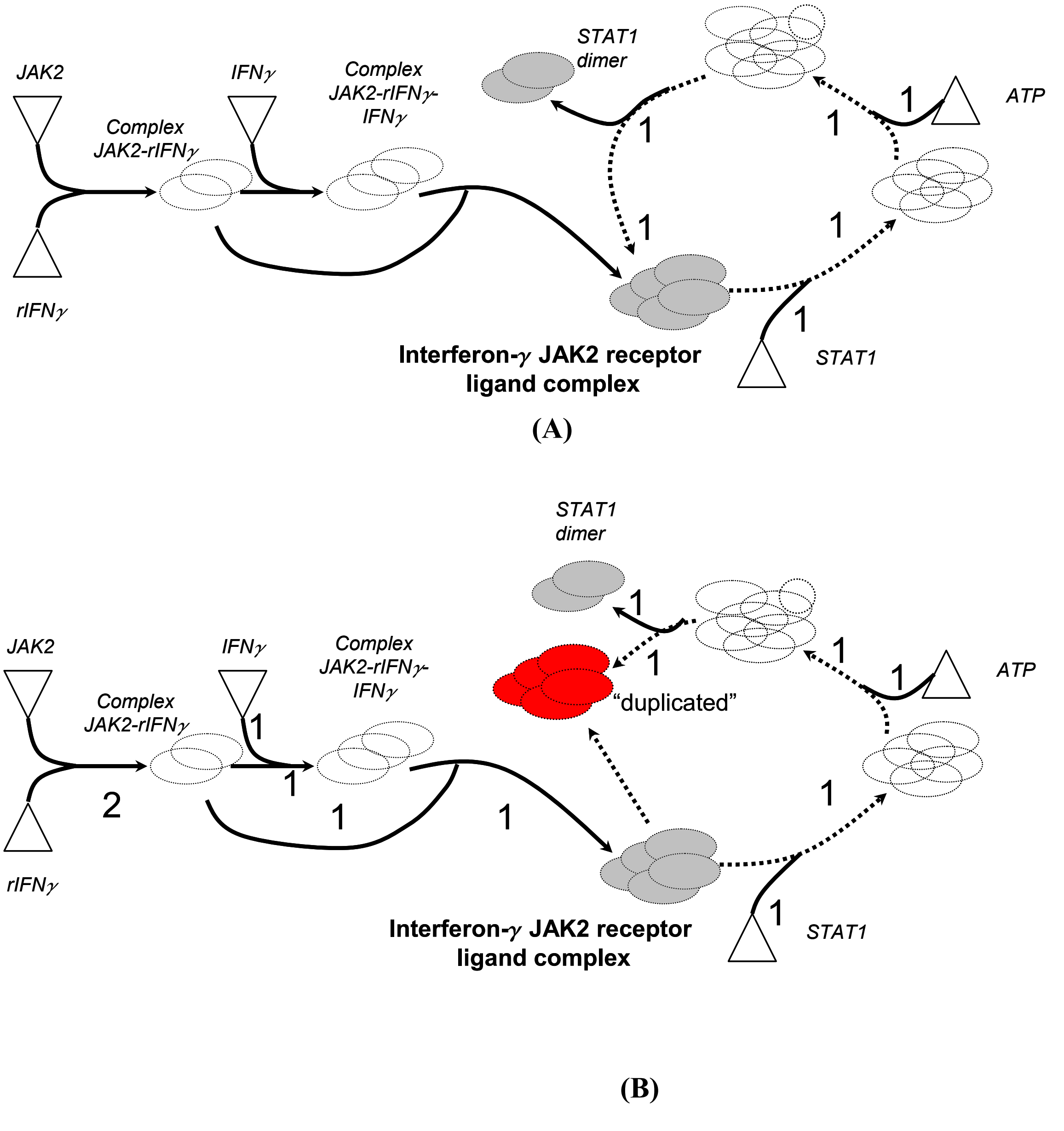


Figure 3.3: Application of loop breaking procedure Figure **(A)** shows the pathway for generating the *STAT1* homodimer prior to application of loop breaking procedure. The pathway is characterized by the existence of loop at Interferon-γ-JAK2 receptor ligand complex. The loop is represented by dotted arrows in the figure **(A)**. Figure **(B)** shows the topologically equivalent loop free pathway is obtained by duplicating the complex (represented in red color). The input molecules are shown with triangles.

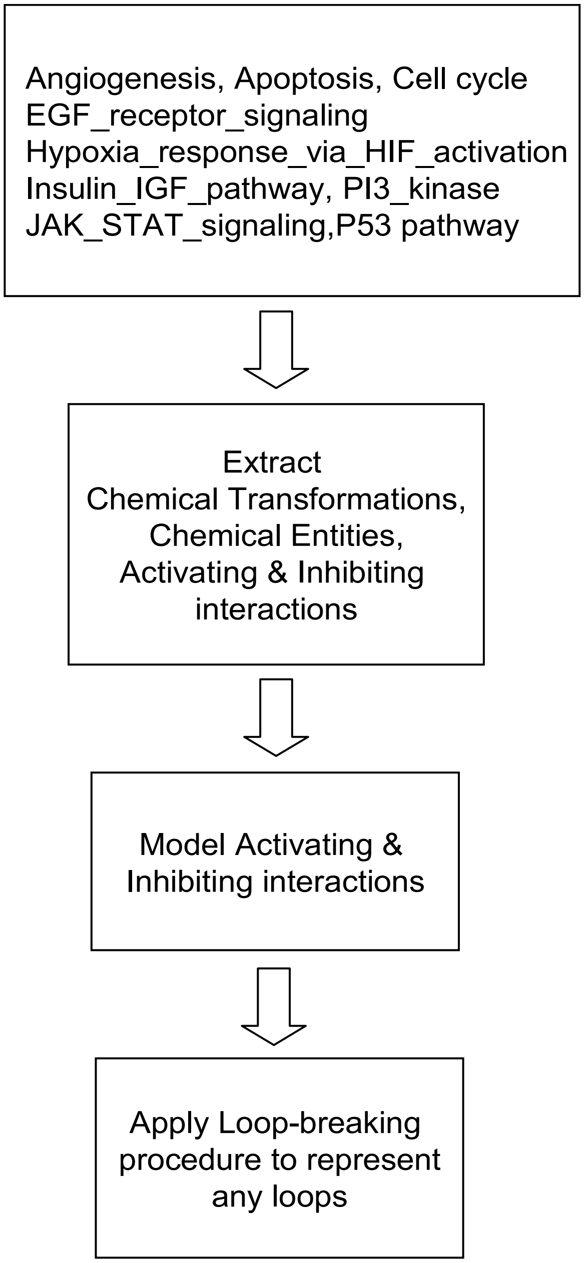


Figure 3.4: Composite block diagram illustrating the important steps in the network modeling procedure. First, the pathway data from PANTHER database is downloaded and subsequently curated to convert the data into a spreadsheet readable format. Next, we identify chemical transformations, chemical entities, activating and inhibiting interactions using customized PERL [105] scripts. Next, the identified activating and inhibiting interactions are modeled as described and finally the loop breaking procedure is employed to represent any loops embedded in the network.

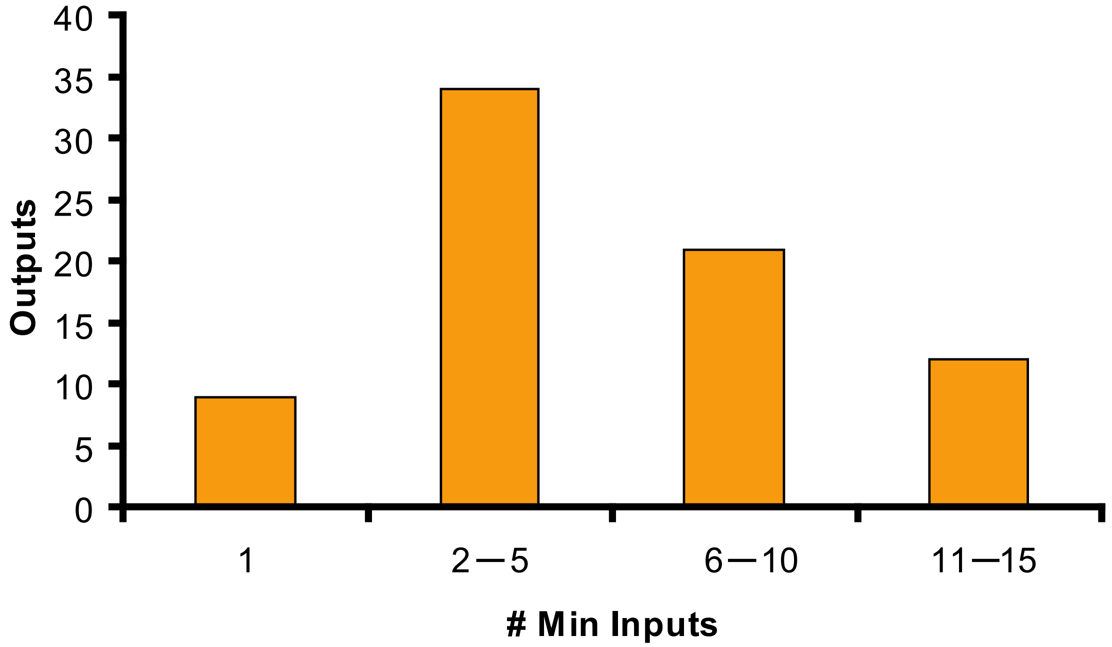


Figure 3.5: The graph depicts the distribution of minimal number of inputs required to realize a particular output. The broadness of the distribution suggests that the input/output structures span a wide spectrum in terms of their complexity.

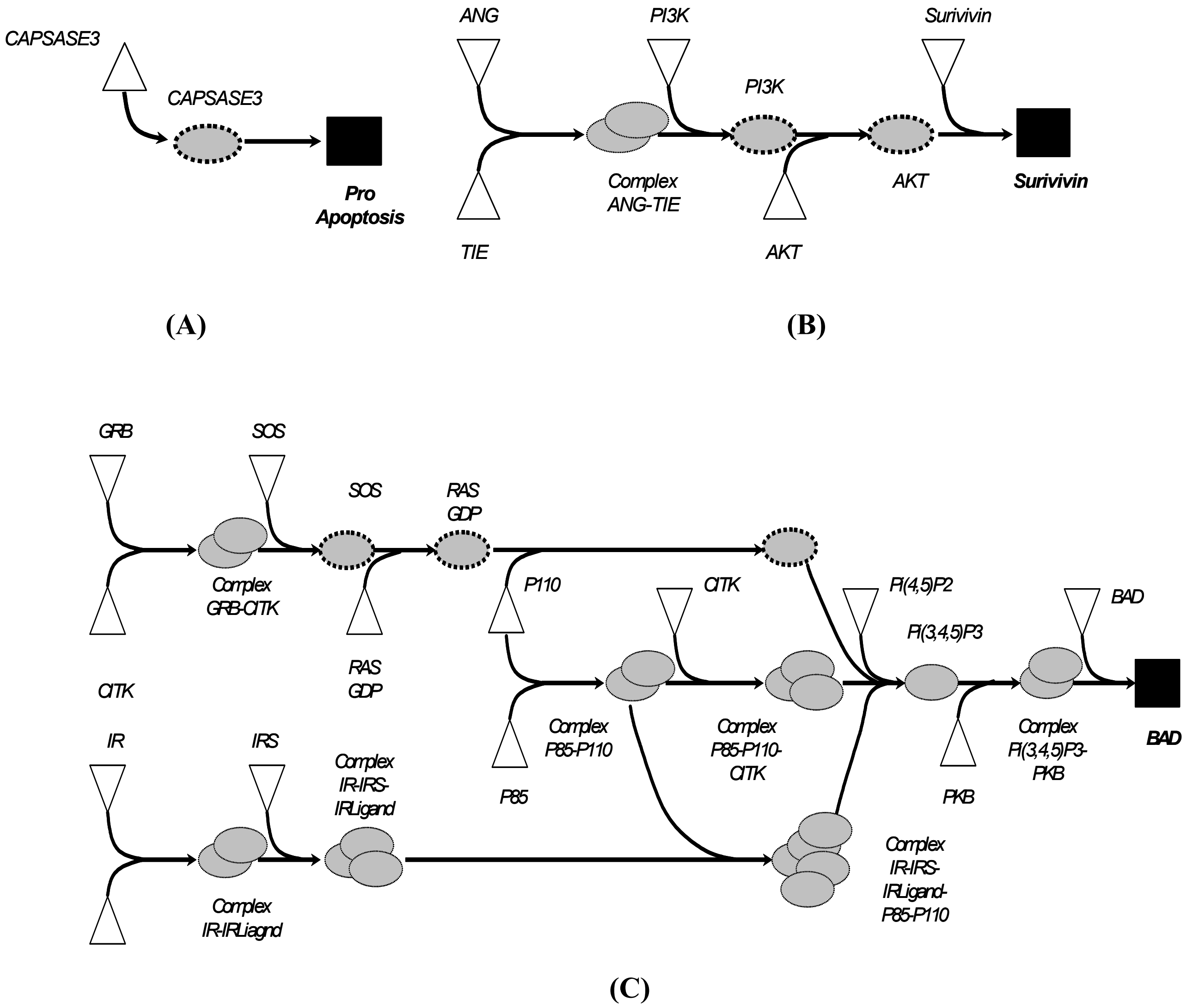


Figure 3.6: The complexity of input/output structure varies from single linear paths to highly interconnected linear cascades. In Figure **(A)** the input/output structure for *capsase 3* resembles a simple linear path requiring just one input. In **(B)** the input/output structure for *Survivin* represents a linear cascade requiring a minimum of five inputs. In **(C)** the input/output structure for *BAD* is much more complex and has many interacting linear cascades.

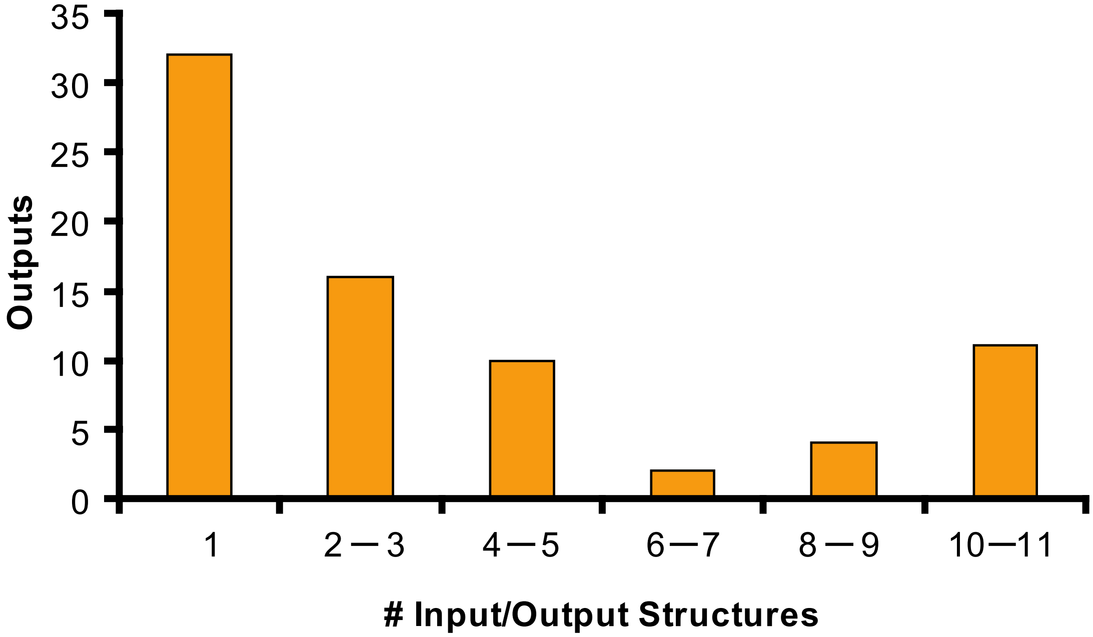


Figure 3.7: The graph shows the number alternative input/output structures realized for each output. The distribution is a convex function with a minimum in the middle and maxima at two extremes implying the existence of two distinct set of outcomes (highly degenerate or highly specific outputs).

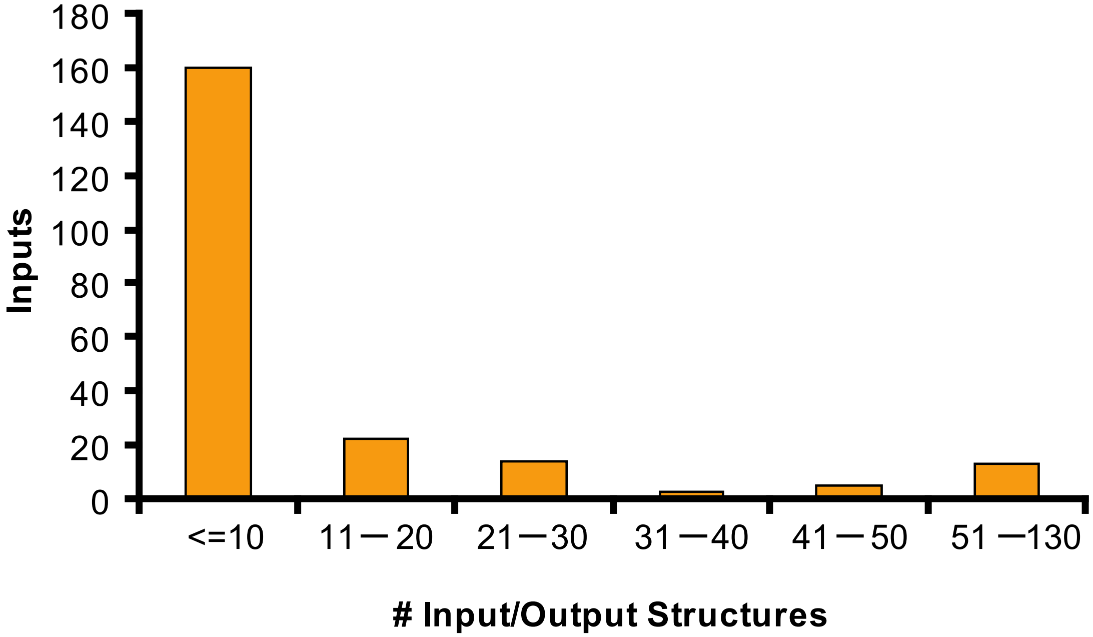
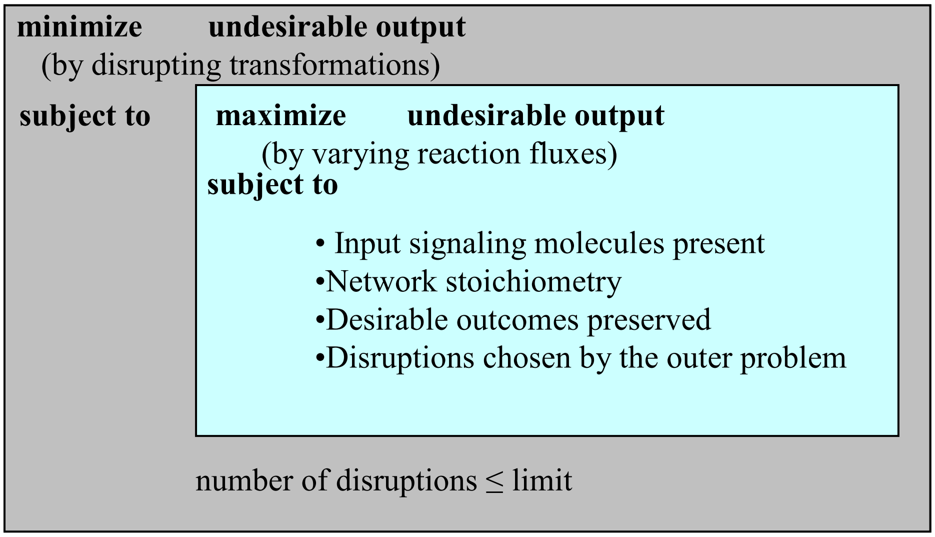


Figure 3.8: The graph depicts the number of times a particular input is recruited by an input/output structure. Highly recruited inputs include common currency such as *ATP GTP* along proteins such as *GRB2* and *SOS*.



minimize **undesirable output**
(by disrupting transformations)

subject to

maximize **undesirable output**
(by varying reaction fluxes)
subject to

- Input signaling molecules present
- Network stoichiometry
- Desirable outcomes preserved
- Disruptions chosen by the outer problem

number of disruptions ≤ limit

Figure 3.9: The bilevel optimization structure for suggesting disruption targets. The inner problem allocates the fluxes through the signaling reactions to maximize the formation of an undesirable output (i.e., worst-case scenario). The outer problem then minimizes the flow to the undesirable outcomes by restricting access (i.e., disrupting) to key transformations available to the optimization of the inner problem.

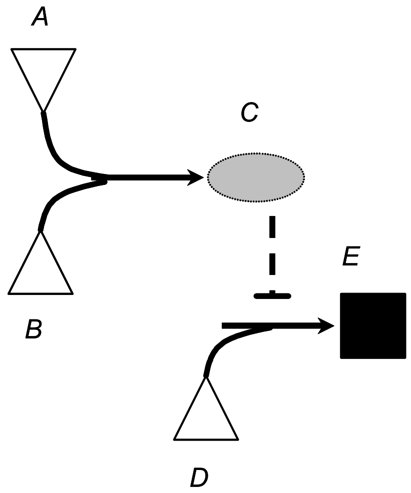
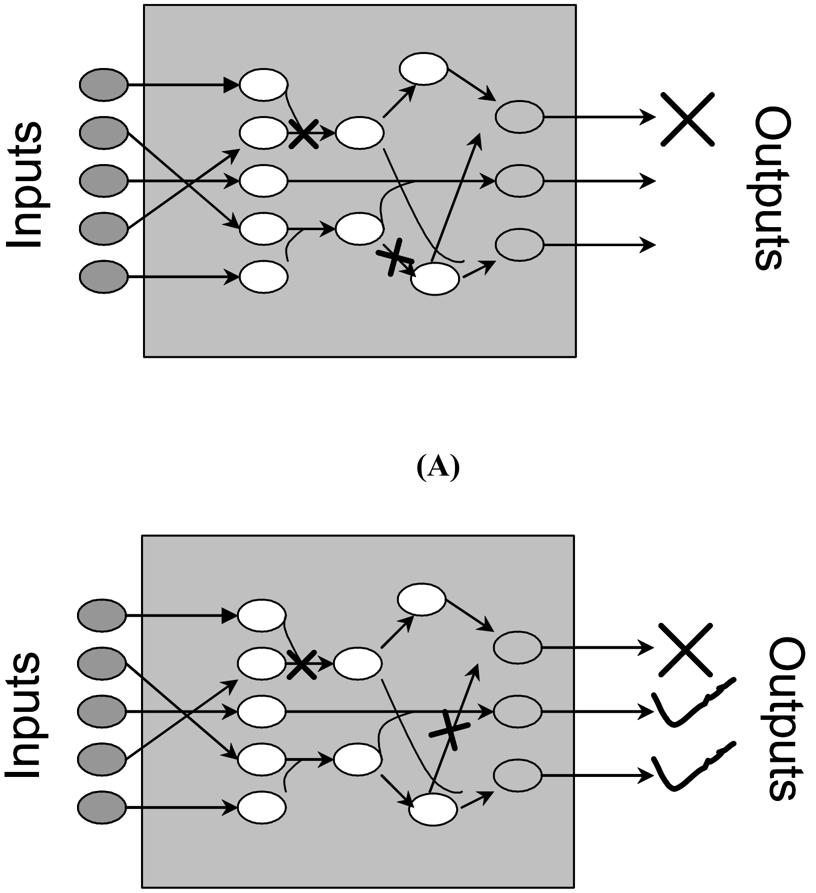


Figure 3.10: The set of input signaling molecules present determines the set of inhibitor molecules that can be formed. Recruitment of input molecules *A*, *B* and *D* blocks the production of output *E* by disruption of the transformation *D*⟶*E* by the inhibitor molecule *C*. Alternatively, if input molecule *D* alone is present production of *E* is preserved. The inhibitor action is indicated by dotted line.



(A)

(B)

Figure 3.11: Pictorial representation of the two different problems solved within the Min-Interference framework. In **(A)** we identify disruption strategies to prevent an undesirable outcome whereas in **(B)** we identify disruption strategies to prevent an undesirable outcome while preserving the formation of desirable outcomes.

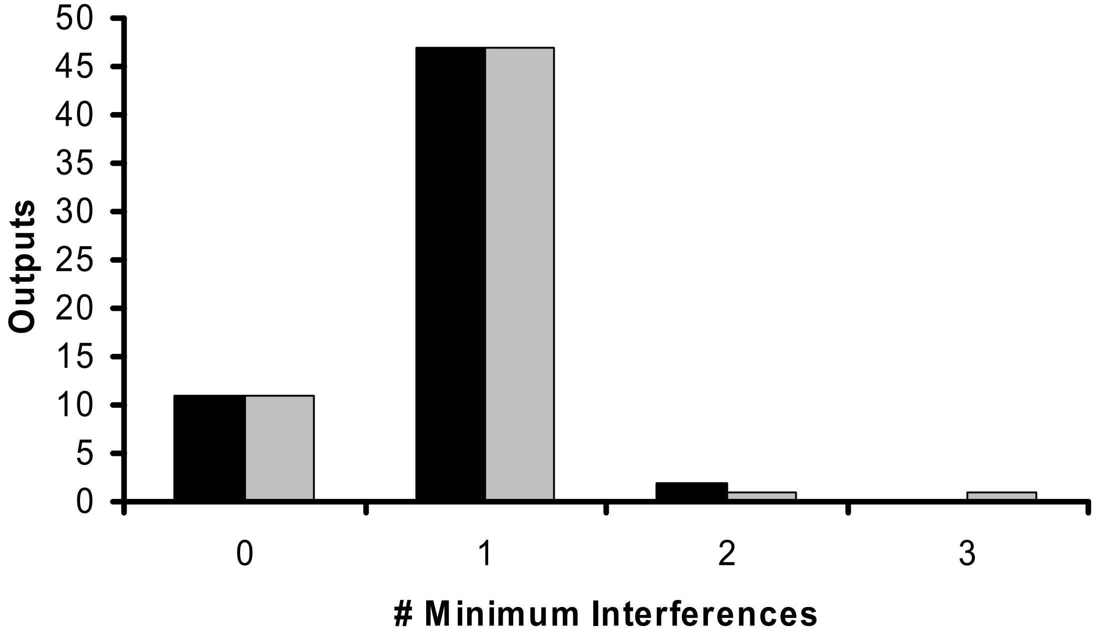


Figure 3.12: The distribution of the minimum number of interferences required to disrupt the production of a particular cellular outcome. The black bars represent the distribution of minimum number of interferences identified to block a cellular outcome alone. The grey bars correspond to the distribution of minimum number of interferences identified to block a cellular outcome while preserving the formation of desirable outcomes.

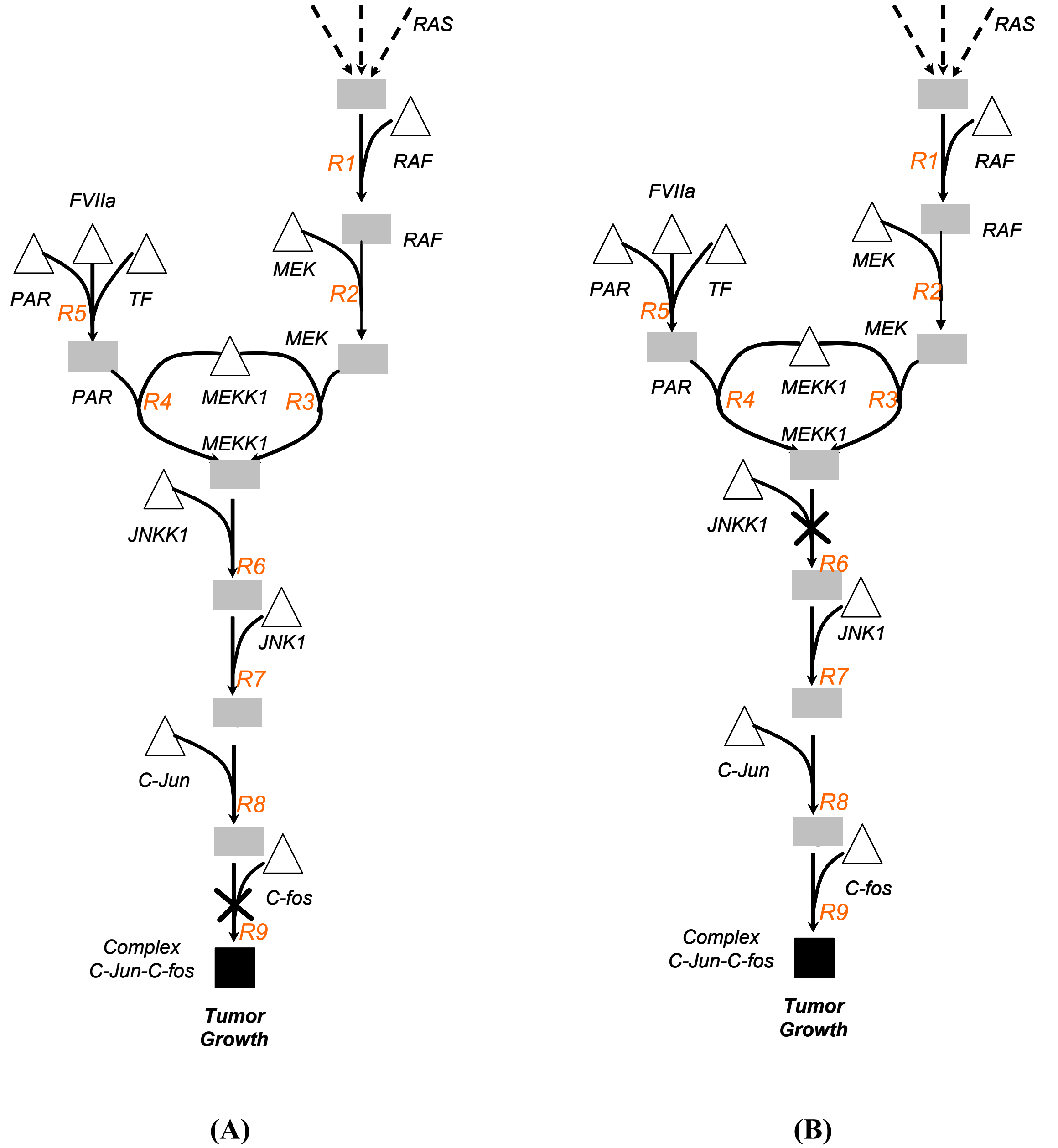


Figure 3.13: In **(A)** a straightforward strategy to block the formation of complex *cJun-cfos* involves disrupting the final transformation (*R*9) leading to the formation of the complex. A less intuitive strategy shown in **(B)** targets *MEKK1* mediated activation of *JNKK1* (*R*6) which is located far upstream to the complex.

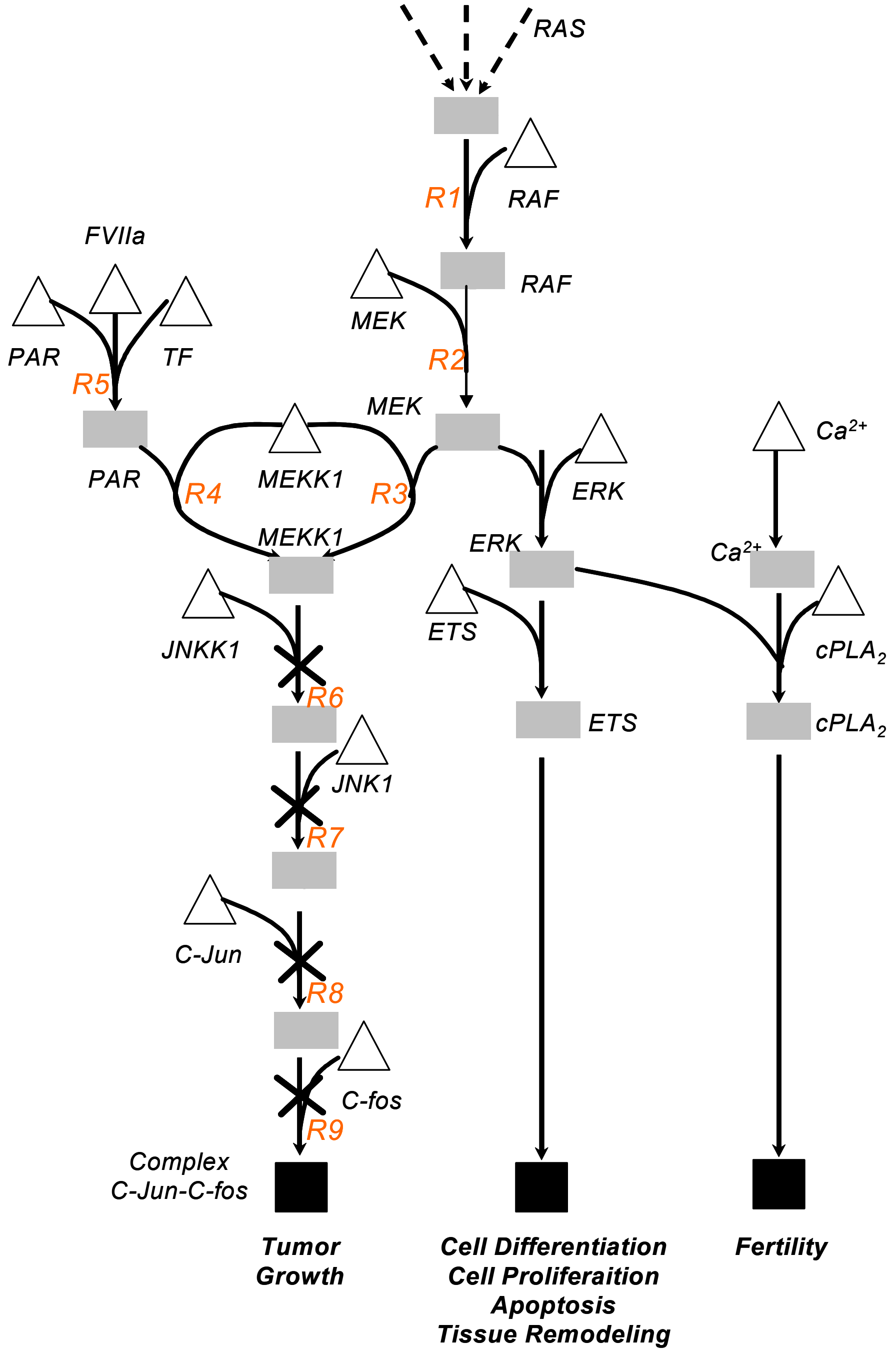
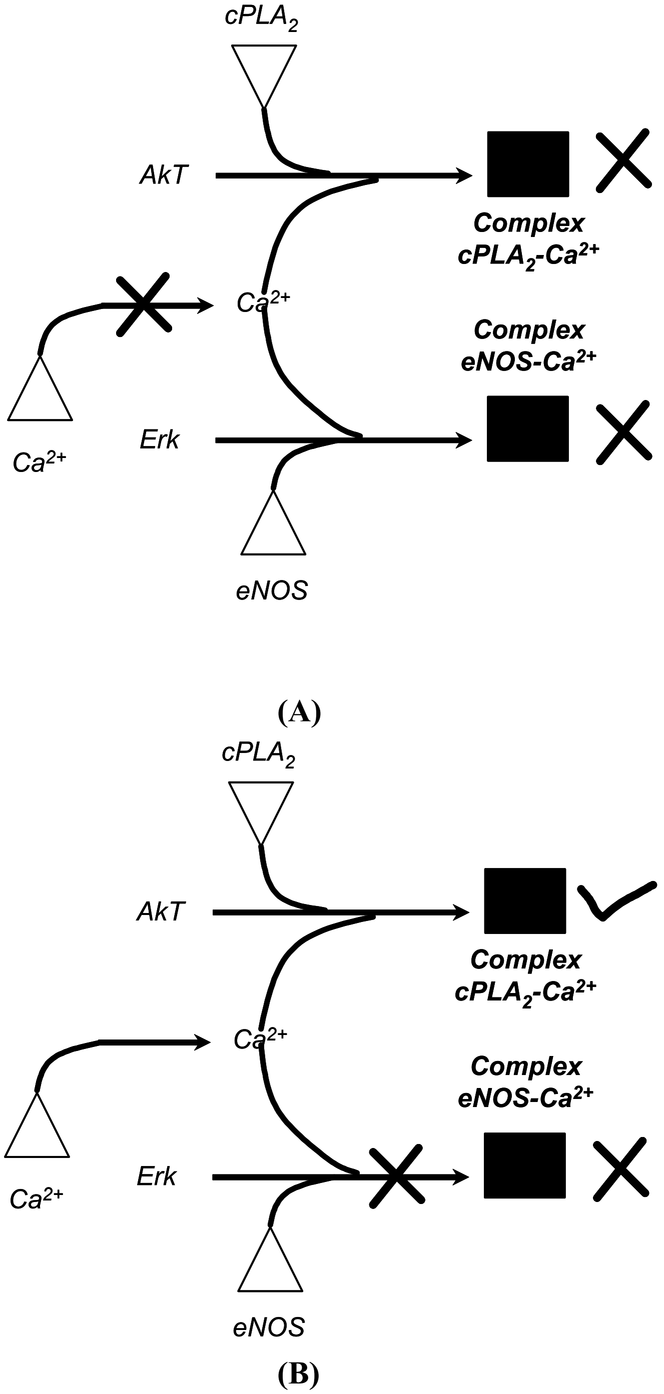


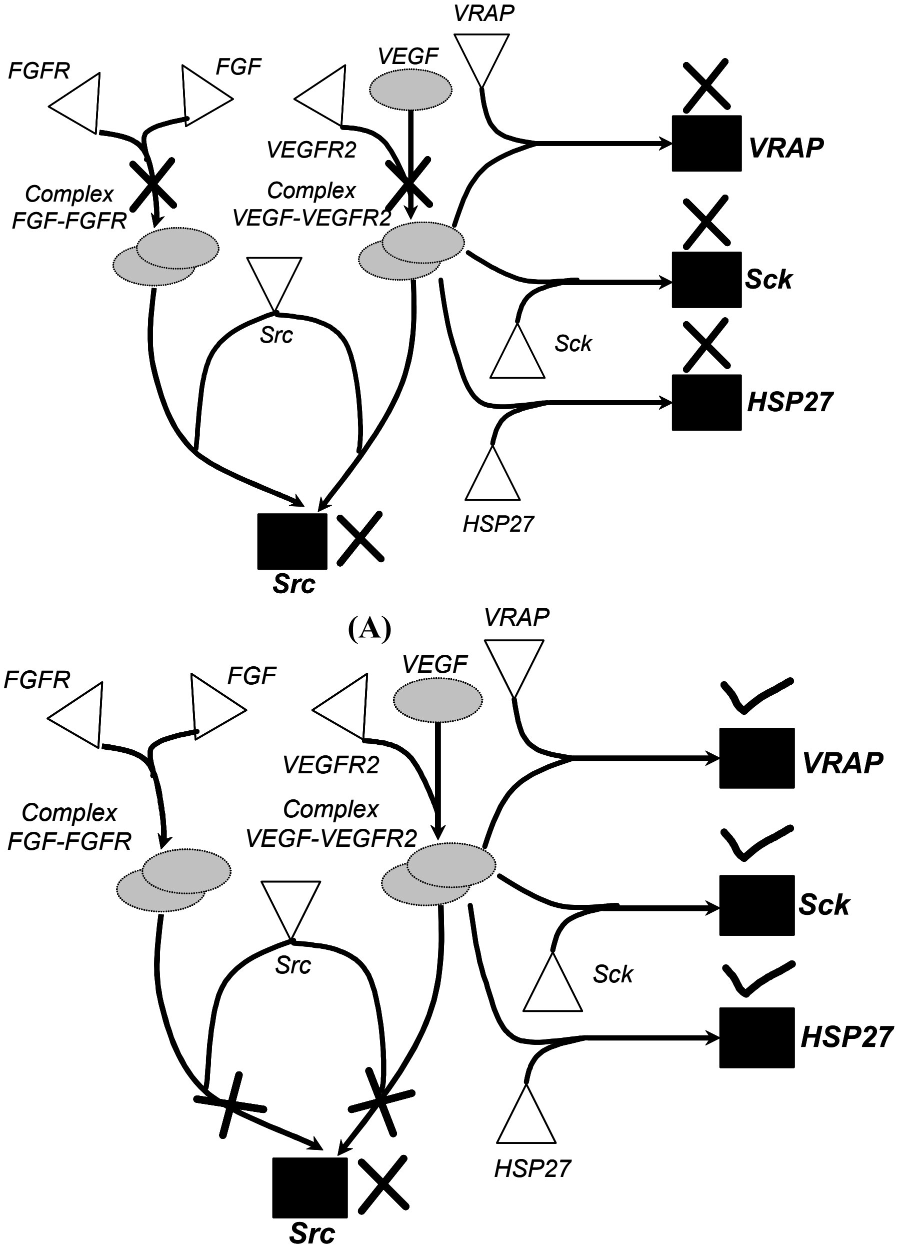
Figure 3.14: Alternative Interference strategies identified to block the formation of complex *cJun-cfos* while preserving the formation of the formation of *ETS* transcription factors and *cPLA*$_2$. The Min-Interference framework finds only single interference strategies. As shown in the figure the alternative strategies target transformations *R*6, *R7, R8* and *R*9 respectively. Note that all the disruption strategies target the terminal transformations located downstream of the MAP kinase cacades.



(A)

(B)

Figure 3.15: Interference strategies to block *eNOS* activation. Disrupting Ca^{2+} transport from ER eliminates both *cPLA$_2$* and *eNOS* activation as shown in **(A)**. In contrast disrupting *Erk* mediated activation of *eNOS* preserves *cPLA$_2$* activation as shown in **(B)**.



(A)

(B)

Figure 3.16: Interference strategies to block the activation of Src. Disrupting both VEGF\and FGF ligand-receptor binding eliminates VRAP, Sck and Hsp27 activity along with Src production as shown in **(A)**. In contrast, disruption of both VEGF-VEGFR2 and FGF-FGFR complex mediated Src activation blocks Src activation alone as shown in **(B)**.

CHAPTER 4

Mechanistic Simulation Based Approaches

4.1 Background

Gene expression is the primary method through which a living organism processes the information stored in its DNA to form all functional cellular components. Elucidation of regulation mechanisms has been an important challenge for understanding the fundamental organization and functioning of biological systems. To date, many data-driven approaches have been developed that use DNA microarray data to unravel the underlying network of genetic interactions. These broadly include clustering approaches [38, 136, 137], Boolean networks [41, 42], differential equations [44, 46, 55, 138], Bayesian networks [47] and neural networks [48]. We refer to these class of methods as "top-to-bottom" approaches as they attempt to elucidate the complex web of DNA, protein and metabolite interactions by using "snap-shot" data (top layer) to infer the inner workings (bottom layer). Alternatively "bottom-to-top" approaches rely on detailed mechanistic descriptions of the underlying molecular processes to construct a predictive model of interaction parameterized to comply with experimental observations. In this chaper, we introduce such a "bottom-to-top" simulation platform that accounts for the mechanistic detail of various processes underlying gene expression and regulation.

The fundamental processes that govern the flow of information from the DNA to a working component (proteins, ribosomes etc.) in a cell are transcription and translation. These processes, coupled with decay mechanisms and various regulatory interactions, largely control the level of gene expression in a cell. Many researchers have attempted to model the gene regulation process by abstracting these underlying processes using ordinary differential equations. Specifically, Agger and Nielsen modeled the regulation dynamics of a genetic system using equilibrium kinetics [139], Cheng *et al.* developed a model to describe the inhibition of *lac* operon by triplex forming oligos [140, 141]. Shea and Ackers developed a model for the O_R control system of bacteriophage λ [142]. Other differential equation models include efforts by Goutsias and Kim [143] and Hatzimanikatis and Lee [144].

The research cited above utilizes differential equations to represent systems that are essentially discrete in nature. Dynamics of gene expression and regulation in many cases involve interactions between relatively small numbers of molecules. For example, the number of available RNA polymerase molecules is estimated to be approximately 35 in *Escherichia coli*, while the number of available ribosomes is estimated to be approximately 350 [145]. In such discrete systems, rates of reaction are no longer deterministic; the reactions occur in a stochastic and discontinuous fashion, rendering the differential equation representation only a coarse approximation [146]. Under these conditions, stochastic fluctuations become important resulting in significant variability in the number of molecules of the species around their average value. Many experimentally verified instances of stochastic variability of genetic systems have been reported in literature. For example, the expression of plasmids containing *araBAD* promoter at subsaturating levels of inducer revealed the existence of both induced and uninduced cells in the population [147]. Elowitz and Leibler have reported that the expression of a synthetically constructed oscillating network exhibits noisy behavior [148]. On the theoretical/computational front, Monte Carlo based simulation methods have been employed by a number of researchers for studying the stochastic evolution of genetic systems [149-153]. These methods largely employ the stochastic simulation algorithm developed by Gillespie [154-156]. Alternatively, Carrier and Keasling proposed a Monte Carlo based algorithm to study the expression of prokaryotic systems [146, 157].

A systems engineering view reveals that gene expression dynamics are governed by processes that are essentially event driven, i.e., many events have to take place in a predetermined order with uncertain start and execution times to accomplish a certain task. Figure 4.1 highlights the many parallels between gene expression and manufacturing systems. In analogy to a manufacturing facility which produces a certain amount of finished product at a particular time with a certain probability, the transcription process produces mRNA transcripts with probability determined by the cellular environment and availability of required components. Similarly, accumulating mRNA and protein levels in the cell are akin to product inventory held in warehouses in a manufacturing system. Motivated by the numerous parallels between these two seemingly different settings, we propose the use of discrete event simulation, which is a powerful tool employed to model and simulate supply chains and manufacturing systems, to model and simulate gene expression systems.

To this end, in this chapter we describe the discrete event based mechanistic simulation platform DEMSIM that we have developed for testing and hypothesizing putative regulatory interactions. The key feature of the DEMSIM platform is the event-based modeling and integration of the fundamental processes underlying gene expression (such as transcription, translation and species decay) with system-specific regulatory circuitry. In the next subsection, we outline the level of mechanistic detail that is accounted for in the various biological processes followed by a description of the computational and algorithmic issues that arise while implementing the simulation framework. Subsequently, the scope of the simulation framework to answer biologically relevant questions is investigated through three examples. The extensively studied *lac* operon system is simulated for verifying that the developed tool can indeed be trained to generate the experimentally obtained biological response of a genetic system. Then, the predictive capabilities of DEMSIM are probed by applying it to simulate the SOS response in *E. coli*. Finally, the sensitivity of the proposed approach to discriminate between alternative regulatory hypotheses is examined using the *araBAD* system of *E. coli* as a benchmark.

4.2 Methods

Effective simulation of gene expression and regulation dynamics entails the detailed modeling and integration of the underlying biochemical processes with the regulatory machinery. To this end, we have modeled each of the underlying transcription, translation and decay processes as stand-alone *modules*. Each module is characterized by *physical* and *model* parameters. Physical parameters correspond to parameters which are known *a priori* from literature sources and are fixed within the simulation framework (e.g. length of gene, transcription rate, etc.). In contrast, model parameters are regression parameters that are fitted using the available experimental data.

Subsequently, the simulation is driven by communication between these modules in accordance with the specifics of the regulatory circuitry of the biological system being investigated. Furthermore, the mechanistic detail of the underlying processes is represented as a sequence of discrete events within the modules. The sequence of events that govern a given module and the associated parameters are described below.

4.2.1 Description of Discrete Event Modules

4.2.1.1 Transcription Module

The mechanism of transcription is fairly well understood compared to other biological processes [158, 159]. The physical parameters required for this module include the length of the open reading frame (ORF) L_{ORF}^{i} [nucleotides] for each gene i, the foot print size of the RNAP enzyme L_{RNAP} [nucleotides] and the rate of transcriptional elongation α_{Tp} [nucleotides/second]. The foot print size L_{RNAP} is the number of nucleotides that the RNAP has to transcribe before it clears the promoter for the subsequent transcription process. The model parameter associated with this module is the gene specific RNAP binding parameter (K_{RNAP}^{bi}) which quantifies the probability of the RNAP successfully binding to the promoter site. The discrete events constituting the transcription module are schematically shown in Figure 4.2(A). The transcription module begins with the transcription initiation event. A Monte Carlo based description is used to account for the inherent randomness associated with *all* stochastic events, including the binding events. Specifically, a uniformly distributed random number between 0 and 1 is generated and compared to the binding parameter associated with the event. If the magnitude of the generated random number is less than the binding parameter, then successful binding is assumed to have taken place otherwise the binding is assumed to have failed. If binding is successful, then the elongation phase is initiated, otherwise, promoter binding is reattempted as shown in Figure 4.2(A). The elongation phase consists of sequential elongation events whereby the mRNA transcript is produced one nucleotide at a time. Once the RNAP has transcribed L_{RNAP} nucleotides, the promoter is declared to be cleared and made available for additional transcription initiation events. This allows for the possibility of multiple RNAP molecules simultaneously transcribing a gene. We also account for the concurrent translation of an incomplete transcript, which is a well known characteristic of prokaryotic systems, by checking for the formation of the nascent ribosome binding site (RBS). This is achieved by comparing the length of the elongating mRNA to the ribosome footprint size L_{Rib}[nucleotides]. If the length of elongating mRNA is equal to L_{Rib}, then the newly formed RBS is made available for either initiation of translation or mRNA decay.

4.2.1.2 mRNA Decay Module

The complete mechanism of mRNA decay is still unresolved and many theories have been put forward to explain it [160]. However, it is largely accepted that mRNA decay is initiated when the enzyme RNase E endonuclease (RNase E) binds to the transcript [146]. In view of this relatively well established hypothesis, we have modeled the decay process as a competitive binding event where the RNase E and the ribosomal assembly both compete for the free RBS on the elongating or complete mRNA transcript (see Figure 4.2(B)). The gene specific RNase E binding parameter (K_{RNase}^{bi}) quantifies the probability of successful binding of the RNase E to an mRNA transcript. If RNase E binds to the RBS, then the mRNA transcript is cleaved, otherwise the ribosomal assembly binds to the RBS and translation is initiated.

4.2.1.3 Translation Module

Upon successful initiation by ribosome binding, a series of elongation events is executed through which the protein polypeptide chain is formed through the discrete addition of amino acid molecules (Figure 4.2(B)) at the rate determined by the translation elongation rate parameter α_{Tr} [codons/second]. RBS clearance is checked after each elongation event by comparing the number of nucleotides translated by the ribosome to L_{Rib}. If the ribosome has cleared the RBS, then the RBS is made available for the competitive binding of the RNase E and ribosomal assembly.

4.2.1.4 Protein Decay Module

Protein decay is modeled by the binding of the proteasomal assembly to the fully translated protein molecule [158] as shown in Figure 4.2©. The gene specific proteasome binding parameter $K_{\text{Proteasome}}^{bi}$ determines the frequency with which the proteasomal assembly binds to a protein molecule and cleaves it into its constituent amino acids. Table 4.1 summarizes all the modules described above along with the associated physical and model parameters.

4.2.2 Modeling of Gene Regulation

Regulation of gene expression occurs at varying degrees at all steps of the transcription through translation cascade. In our model, we assume that transcriptional initiation is the key step in gene regulation. This hypothesis has been put forth by a number of other researchers and supported by both experimental [137] and computational investigations [161]. The regulatory *logic* thus directly or indirectly alters the binding interactions of the RNAP with the promoter region of the DNA. In the context of our modeling framework, this is captured as the effect of the regulatory machinery on the probability of successful RNAP binding to the promoter region. Note that here the term regulatory logic is employed to describe a wide range of regulatory mechanisms which can be readily accounted for in our simulation framework. For example, a regulatory protein might regulate a target gene only if the concentration of the regulatory protein is beyond a threshold. In that case, the implementation of the regulatory criterion would entail checking if the concentration of the regulatory protein is above the specified threshold and subsequently making K_{RNAP}^{bi} dependent on the output of the regulatory logic. Separate RNAP binding parameters are assigned to binding events that represent alternative outcomes of the regulatory logic. The relative magnitude of these parameters quantifies the nature and strength of regulation (upregulation/down regulation). The regulatory logic employed for the test systems considered in this study are discussed in the results section.

4.2.3 Implementation of Simulation Framework

The DEMSIM software implementation consists of the following three key components: (*i*) an *event list* that contains all the events that need to be executed along with their respective execution times, (*ii)* a *global simulation clock* that records the progress of simulation time as events are sequentially executed, and (*iii*) a set of *state variables* that characterize the system and which are updated every time an event is executed. At every time step, events corresponding to all active (non–terminated) modules in the system are included in the event list. Subsequently, the event list is sorted and the event having the smallest execution time is executed. The simulation clock is advanced and the execution time of all other events is updated. Such a sequential procedure prevents the occurrence of "causality errors" by ensuring that an event with a later time stamps is not executed before an event with an earlier time stamp [162]. Furthermore, since the execution of certain events leads to the creation of new modules and the termination of existing ones, the number of active modules in the system is updated and new events are included in the event list. This procedure is then repeated for the duration of the simulation horizon and state variables such as number of mRNA and protein molecules are recorded.

We use a fixed-time step of 0.10 seconds for stepping forward in time. This time interval corresponds to the duration between two translation elongation events (since α_{Tr} = 10 codons/second) and five transcription elongation events (since α_{Tp} = 50 nt/second). Table 4.2 lists the events associated with each module and the associated execution times. This time step, which results in the lumping of 5 transcription elongation events into a compound “pseudo” transcription event, is chosen to balance computational accuracy and CPU time requirements. Other assumptions include: (*i*) transcription and translation machinery are present in excess so that dilution by cell growth and gene expression can be neglected [146]; and (*ii*) post transcriptional and post translational modifications take place instantaneously [143]. The DEMSIM framework is implemented using the C programming language on a 16 node linux cluster with dual Intel 3.4 Ghz Xeon processors.

4.3 Results

To highlight and probe its capabilities, the DEMSIM framework is applied to three different test systems. Given the stochastic nature of the underlying processes, multiple simulation runs are needed to glean a statistically complete picture of the temporal evolution of the system. The simulation runs are averaged out to extract the mean trajectory and the standard deviation is estimated at each time point. The results of the simulations are presented by plotting the mean trajectory and the ±1σ regions, where σ denotes the standard deviation.

4.3.1 Example I - *lac* operon system of *E. coli*

The *lac* operon of *E. coli* has been extensively studied as a model system for understanding prokaryotic gene regulation [53, 163, 164]. We use this relatively simple genetic system to verify that the various model parameters embedded within DEMSIM can indeed be tuned using experimental data. In particular, we focus our attention on the expression of *lacZ* gene, the first within the operon which also includes genes *lacY* and *lacA*. Transcription from the *lac* operon is inhibited by the product of *lacI* gene located upstream of the operon. However, in the presence of lactose, the gene product of *lacI* combines with lactose to form an inactive product, thus turning the operon ON. This enables transcription of the *lacY* gene which encodes the protein responsible for transport of lactose into the cell.

In addition to the basic modules described earlier, the simulation of the *lac* operon system requires a model for transport of lactose into the cell. To this end, the kinetic model developed by Wong *et al.* [53] is used. This model relates the rate of change of intracellular lactose to the amount of extracellular lactose and the amount of *lacY* protein. The mathematical form of the model is described in the Appendix. All simulation runs begin with no lactose present inside the cell and the copy number/cell of mRNA and protein of all the genes is assumed to be zero (*i.e.*, cold start). The regulatory logic is modeled by making the RNAP binding parameter for the *lac* operon conditionally dependent on the relative amounts of the inducer (lactose) and repressor (*lacI* protein) in the cell. This is achieved by utilizing the following rule based representation within the simulation framework.

$$K_{RNAP}^{b} = \begin{cases} \alpha & if \quad [lacI] \leq [Lactose] \\ \beta & if \quad [lacI] > [Lactose] \end{cases}$$

where [*lacI*] and [*Lactose*] are the number of *lacI* protein and lactose molecules respectively and $\alpha > \beta$ in accordance with the inducer/repressor role of lactose/*lacI*. In addition to these RNAP binding parameters, two other model parameters that need to be tuned are the RNase E and proteasome binding parameter for *lacZ*. These parameters are estimated by applying DEMSIM within a predictive-corrective loop whereby the parameters are tuned such that the simulation results match experimentally reported data. Specifically, we use the following experimental data for fitting [164]: *lacZ* mRNA half-life (1.3 minutes); average rate of production of *lacZ* protein (20 molecules/second); steady-state number of *lacZ* mRNA transcripts (62 molecules/cell). The values of the fitted model parameters are listed in Table 4.3. Figure 4.4(A) and (B) show the simulated profiles for the number of *lacZ* mRNA and protein molecules respectively. The simulated values for the three quantities used for fitting are: *lacZ* mRNA half-life (1.5 minutes); average rate of production of *lacZ* protein (29±2 molecules/second); steady-state number of *lacZ* mRNA transcripts (60±6 molecules/cell). These results for the *lac* operon system clearly suggest that the DEMSIM framework is able to reproduce the dynamics of gene expression using appropriately tuned model parameters.

4.3.2 Example II – SOS response system of *E. coli*

In this example, we expand both the scale of the system under consideration, in terms of the number of genes whose expression is simulated, as well as the scope of issues addressed using DEMSIM. We explore the capabilities of DEMSIM to not only reproduce experimental data with which it was trained but also its ability to predict the *de novo* response of the system to an externally imposed perturbation. To this end, the specific system that we investigate is the SOS response of *E. coli*. Irradiation of cells with UV light produces DNA lesions that transiently block the process of replication. It is now known that cells respond to this stress by upregulating the expression of several genes that function to repair the DNA lesions [165-167]. This response is termed as the SOS response (see Figure 4.5). Many of the genes involved in the repair of DNA damage are negatively regulated by the *lexA* repressor protein, which binds to a consensus sequence located upstream of the promoter. Upregulation of these genes occurs when the *recA* protein binds to the single stranded DNA created at replication forks. This introduces a conformational change in the *recA* protein, turning it into a coprotease that cleaves the *lexA* repressor. As soon as the cellular concentration of *lexA* diminishes, the genes suppressed by *lexA* are more frequently transcribed. Following repair of DNA damage, the coproteolytic activity of *recA* diminishes leading to an increase in the *lexA* concentration and thus returning the cell to its original state as shown in Figure 4.5 [168-171]. From the larger set of about 30 genes which are known to be regulated by the *lexA* repressor, we selected a subset of six genes to simulate [172, 173]. In addition to *lexA*, the genes that we considered are: *polB* (production of DNA polymerase II); *uvrA, uvrB* (nucleotide excision repair); *ruvA* (recombination process); and *dinI* (inhibitor of *umuD*).

4.3.2.1 Modeling of Gene Regulation

The regulatory logic for this system is formulated as follows. The probability of successful binding of the *lexA* protein to the protein binding region of a gene is postulated to be given by

$$K_{lexA}^{bi} = 1 - \frac{2 \cdot e^{(-\Phi(i)\cdot[lexA])}}{1 + e^{(-\Phi(i)\cdot[lexA])}} \quad \text{for } i = lexA, dinI, polB, uvrA, uvrB, ruvA$$

Here, $\Phi(i)$ is a gene specific regulatory constant and $[lexA]$ is the number of molecules of *lexA* protein. Figure 4.6 shows the dependence of K_{lexA}^{bi} on $[lexA]$ for different values of $\Phi(i)$. Parameter $\Phi(i)$ quantifies the relative binding strength of the *lexA* repressor to a particular gene *i* with a higher value of $\Phi(i)$ implying a higher magnitude for K_{lexA}^{bi} (and hence higher probability of repression). Note that the above formulation ensures that the probability of repression given by K_{lexA}^{bi} is between 0 and 1 for all values of $\Phi(i)$ and $[lexA]$ with $K_{lexA}^{bi} \to 0$ as $[lexA] \to 0$ and $K_{lexA}^{bi} \to 1$ as $[lexA] \to \infty$.

Figure 4.7 pictorially depicts the regulatory logic for the SOS response system. The *lexA* repressor binds to the operator region of the genes with a probability given by K_{lexA}^{bi}. If the repressor binds, then the gene is repressed otherwise the gene is unrepressed. The magnitude of K_{RNAP}^{bi} is made contingent on the outcome of the regulatory logic as illustrated in Figure 7 with the relative magnitudes of $\left(K_{RNAP}^{bi}\right)_{\text{Repressed}}$ and $\left(K_{RNAP}^{bi}\right)_{\text{Unrepressed}}$ quantifying the strength of repression for each of the genes. Enhanced *lexA* cleavage under *irradiated* conditions is simulated by increasing $K_{\text{Proteasome}}^{bi}$ for *lexA* gene by a factor of $\mathrm{X}_{lexA}\,(>1)$:

$$\left(K_{\text{Proteasome}}^{b\ lexA}\right)^{\text{Irradiated}} = X_{lexA} \cdot \left(K_{\text{Proteasome}}^{b\ lexA}\right)^{\text{Unirradiated}}$$

As a result of enhanced cleavage, the number of *lexA* molecules in the cell decrease reducing the magnitude of K_{lexA}^{bi}. This decreases the probability of repression of the genes in the system and the genes are more frequently transcribed. After the repair time ($\mathrm{T}_{\text{Repair}}$) has elapsed, the value of $K_{\text{Proteasome}}^{bi}$ for *lexA* is restored to its initial value thus gradually returning the cell to its original state.

4.3.2.2 Parameter Estimation

The gene specific mRNA decay parameter K_{RNase}^{bi} is estimated by matching the simulated decay of mRNA level in the absence of transcription to the experimentally observed mRNA half-life. For a given value of the decay parameter, the simulations are run by "arresting" the processes of transcription. In the context of the simulation framework, this is accomplished by setting the value of K_{RNAP}^{bi} to zero. The simulated value of half-life corresponding to the assumed decay parameter is then estimated by measuring the time needed for the initial mRNA level to drop to half. $K_{\mathrm{Proteasome}}^{bi}$ is fitted similarly by "arresting" both the transcription and translation processes. Figure 8 shows the average values of the simulated mRNA (Figure 4.8(A)) and protein (Figure 4.8(B)) half-lives as a function of the K_{RNase}^{bi} and $K_{\mathrm{Proteasome}}^{bi}$ respectively. Subsequently, the factor X_{LexA}, which accounts for the enhanced *lexA* cleavage post irradiation, is similarly fitted by adjusting its value to reproduce the experimentally observed post-irradiation half-life of approximately 1-2 minutes [168]. The time required to repair the damage to DNA is set at 45 minutes based on the observations of Courcelle *et al.* [172].

The remaining parameters are gene specific RNAP binding parameter under repressed state $(K_{RNAP}^{bi})_{\mathrm{Repressed}}$, RNAP binding parameter under unrepressed state $(K_{RNAP}^{bi})_{\mathrm{Unrepressed}}$ and the gene specific regulatory constant $\Phi(i)$. Since these parameters account for the generation of the mRNA transcripts and protein molecules in the cell, we refer to this set of parameters as generation parameters. The generation parameters are fitted by simultaneously adjusting their values to match the experimentally observed mRNA fold changes in both *irradiated* and *unirradiated* cells and the protein levels in the *unirradiated* cells. This procedure relies on the assumption that a direct correspondence exists between the mRNA transcript level and the fluorescence intensity measured in the microarray experiments. Figure 4.9 highlights the procedure employed to estimate the generation parameters. Beginning with an initial guess for the values of the generation parameters, simulations are run using the previously estimated values for the decay parameters. After the simulation equilibrates (simulation warm-up time), the mRNA and protein levels in the cells are recorded for 5,000 seconds. These measurements correspond to the mRNA and protein levels under *unirradiated* conditions. Subsequently, the cleavage of *lexA* repressor is enhanced for a duration of $\mathrm{T}_{\mathrm{Repair}}$ seconds, by multiplying $K_{\mathrm{Proteasome}}^{bLexA}$ with the previously estimated factor X_{LexA}, and the mRNA levels are recorded for a period of 5,000 seconds as shown in Figure 4.9. These measurements correspond to the mRNA levels under *irradiated* conditions. The *unirradiated* and *irradiated* mRNA levels are compared to experimentally observed mRNA fold changes reported by Courcelle *et al.* [172]. Also, the recorded protein levels are compared to experimentally reported protein levels in *unirradiated* cell cultures [165]. The generation parameters are adjusted until the simulated measurements are in reasonable agreement with experimental observations. Table 4.4 summarizes the parameters for the SOS response system and the experimental data used to estimate the parameters. The values for the estimated parameter values are provided in Table 4.5. The simulated mRNA fold changes under *unirradiated* conditions are plotted in Figure 4.10. The experimentally reported values are also plotted for comparison. Figure 4.11 shows similar comparisons for the *irradiated* conditions. Similarly, the simulated and the experimentally estimated values for the protein levels in *unirradiated* cell cultures are listed in Table 4.6. In line with the observations for the *lac* system, the fitted parameter values are able to accurately reproduce the experimental data used to train the model.

4.3.2.3 Model Validation

Next, the trained model is validated by comparing its predictions of protein levels in *irradiated* cultures to experimentally reported values. Table 4.7 lists the simulated peak protein levels estimated from the average of 120 simulation runs and the corresponding experimentally obtained values. While good agreement with experimental estimates is observed for *uvrB, polB* and *uvrA* genes, some deviation is observed for *dinI* and *ruvA* genes. One possible reason for these deviations could be that the simulation framework might not account for all regulatory interactions involving these genes. In addition to the peak protein levels, the dynamics of the temporal response of *lexA* protein on induction of SOS response are also found to be in good agreement with experimental observations of Sassanfar and Roberts [168] as shown in Figure 4.12. These results highlight how, given adequate experimental data, the DEMSIM framework can first be trained and then be used as a predictive tool for generating responses of genetic systems.

4.3.3 Example III –Induction dynamics of *araBAD* operon of *E. coli*

The ability of the simulation framework to discriminate between alternative regulatory hypotheses is probed by applying it to the *araBAD* system in *E. coli.* The *araBAD* operon has been extensively studied as it serves as an excellent model for the feed forward loop motif [26, 174-176]. The *crp* gene activates both the *araBAD* operon and the *araC* gene in presence of inducer cAMP. The *araC* gene product transcriptionally activates the *araBAD* operon in presence of inducer L-arabinose resulting in a feed forward loop motif (see Figure 4.13(A)). In addition, since the nature of regulation (*i.e.*, activation) by the *crp* gene is the same for both the operon and the *araC* gene, the motif is termed as a *coherent* feed forward loop (FFL). Theoretical studies [161] have suggested that this system acts as a sign sensitive delay element. This implies that while the motif delays the cells response to an ON step in the stimulus, no delay in response is observed in the case of the complementary OFF step. In addition, Mangan *et al.* have investigated the responses of the *araBAD* FFL motif to cAMP ON and cAMP OFF steps [177]. By comparing the response of the motif to that of *lac* promoter, which is a model for the simple AND gate motif, the authors have concluded that the *araBAD* system exhibits sign sensitive delay kinetics.

We used the DEMSIM framework to simulate two different regulatory mechanisms which both support the experimentally observed enhanced expression of the *araBAD* operon and the *araC* gene on addition of cAMP to a system saturated with L-arabinose. The first motif corresponds to a FFL (Figure 4.13(A)) and the second motif represents a parallel motif (Figure 4.13(B)). A simple AND gate motif is also considered where both *crp* and *araC* enhance the expression of the *araBAD* operon as shown in Figure 4.13(C). Identical values are assigned to gene specific decay parameters for all the three mechanisms so that the decay dynamics exhibited by the motifs are the same. Furthermore, the gene specific generation parameters are fitted such that all three motifs exhibit similar *araBAD* expression in systems which are saturated and starved of the inducer cAMP. Subsequently, the response of the motifs to cAMP ON and OFF steps is generated and the responses of the FFL and parallel motif are compared to the response of the simple AND gate response. Simulation results shown in Figure 4.14 indicate that the parallel motif model for gene regulation fails to capture the sign sensitive delay nature of the operon. In contrast, the FFL motif correctly exhibits a delayed response to a cAMP ON step (Figure 4.14(A)) while no delay is observed in response to cAMP OFF step (Figure 4.14(B)), suggesting that FFL is indeed the most plausible regulatory mechanism. These results highlight the ability of the DEMSIM framework to effectively discriminate between alternative regulatory mechanisms.

4.4 Summary and Discussion

In this chapter, we introduced a discrete event based mechanistic simulation platform (DEMSIM) and used it for testing and hypothesizing putative regulatory interactions. The key feature of the developed simulation framework was the modeling of underlying biological processes, such as transcription, translation and decay, using stand-alone modules. Each module was characterized by a sequence of discrete events in accordance with the level of mechanistic detail considered. A rule based Monte Carlo procedure was employed for capturing the randomness inherent to the molecular binding events. Subsequently, communication within the modules was driven by taking into account system specific regulatory information. A distinction was made between physical and model parameters, with the former determined either from literature or online databases and the latter determined by fitting simulation results to experimental data.

The developed tool was benchmarked by applying it to three biological systems with different levels of complexity. The relatively simple *lac* operon was used to verify that parameters embedded in DEMSIM can indeed be trained using experimental data. Subsequently, the more complex SOS response system was used to probe the predictive capabilities of the developed framework. Simulation results indicated that the tool was able to make fairly accurate predictions regarding data that was *not* used for training the model parameters. Finally, the *araBAD* system was used to highlight the developed tool's sensitivity to discriminate between relatively "close" regulatory hypotheses.

The versatility of the DEMSIM framework allows us to conduct numerous *in silico* experiments. For example, the framework employed for SOS response system can be used to make predictions regarding the gene expression dynamics in a *lexAdef* genetic context, where the genes are expressed constitutively [178]. If the model predictions are correct, then the developed model can be used to ask more complex questions regarding the biological system. For example, one could investigate the timing of induction of SOS response or the effect of single stranded DNA (ssDNA). If the model predictions are incorrect, then the experimental data can be used to refine the current model to prepare a more accurate representation of the underlying physical interactions. This exercise can provide valuable insights into the workings of the gene expression and regulatory interactions at a molecular level.

Many "top-to-bottom" computational frameworks employ high-throughput biological data to infer plausible regulatory hypotheses. For example, the GRAM algorithm proposed by Joseph *et al.* [179], utilizes gene expression data and genome-wide location analysis for DNA-binding regulators, to predict putative regulatory interactions. In contrast the DEMSIM framework takes into account the underlying mechanistic detail of the gene expression and regulation processes to construct a predictive model. Furthermore, the simulation results demonstrate the ability of the framework to verify and also discriminate between relatively “close” regulatory hypotheses. These observations suggest that DEMSIM, which adopts a “bottom-to-top” approach, can be employed in tandem with “top-to-bottom” computational frameworks such as GRAM to verify and complete the candidate regulatory hypotheses generated by the latter approaches. However, unlike “top-to-bottom” approaches, extending the simulation framework to simulate large scale gene networks entails enormous computational resources. One possible way of addressing this problem is to exploit the modular structure of large scale regulatory networks. Recent studies have indicated that the regulatory networks can be decomposed into clusters of motifs [161, 180]. Hence, the regulatory hypotheses generated by the “top-to-bottom” approaches can be investigated for their modularity and the generated sub-networks/motifs can be simulated using the proposed framework. Comparison of simulation predictions with experimental data would then serve to verify, correct and complete the inferred hypotheses.

Due to the underlying stochastic nature of the simulation framework, extending the framework to model systems with larger copy numbers of species involved is difficult as the number of events increases by many folds. In such systems we envision a hybrid simulation framework that uses both differential equation based and stochastic methods in tandem [181]. While differential equations can be used to model species with high copy number, DEMSIM can be used selectively for only low copy number species.

Table 4.1: Modules and associated parameters

Module	Value
Transcription	
• *Physical*	
L_{RNAP} (nt)	60 nt [145]
L_{Rib} (nt)	33 nt [146]
L^{i}_{ORF} (nt)	KEGG Database
α_{Tp} (nt s^{-1})	50 nt s^{-1}[182]
• *Model*	
K^{bi}_{RNAP}	Fitted
mRNA decay	
• *Model*	
K^{bi}_{RNase}	Fitted
Translation	
• *Physical*	
L_{Rib} (nt)	33 nt [146]
α_{Tr}(codons s^{-1})	10 codons s^{-1}[182]
Protein decay	
• *Model*	
$K^{bi}_{\mathrm{Proteasome}}$	Fitted

Table 4.2: Events and execution times

Module	**Execution time (s)**	**Value**
Transcription		
Initiation Event	t_{bind}	0.1 s
Elongation Event	$1/\alpha_{Tp}$	0.02 s
mRNA decay		
Initiation Event	t_{bind}	0.1 s
Translation		
Elongation Event	$1/\alpha_{Tr}$	0.1 s
Protein decay		
Initiation Event	t_{bind}	0.1 s

Table 4.3: Fitted parameter values for *lacZ* gene

Parameter	Condition	Value
K_{RNAP}^{bi}	$\left[lacI\right] \geq \left[Lactose\right]$	1.0×10^{-3}
K_{RNAP}^{bi}	$\left[lacI\right] \leq \left[Lactose\right]$	7.125×10^{-1}
K_{RNase}^{bi}	-	8.0×10^{-3}
$K_{\text{Proteasome}}^{bi}$	-	9.0×10^{-5}

Table 4.4: Parameters for SOS response system

Parameter	Reference
$(K^{bi}_{RNAP})_{\text{Unrepressed}}, (K^{bi}_{RNAP})_{\text{Repressed}}, \Phi(i)$	Adjusted to match (i) mRNA fold changes in *unirradiated* cells [172] (ii) mRNA fold changes in *irradiated* cells [172] (iii) Protein levels in u*nirradiated* cells [165]
K^{bi}_{RNase}	Selected to reproduce experimentally observed mRNA half-life [183]
$K^{bi}_{\text{Proteasome}}$	Selected to reproduce experimentally observed protein half-life; 60 minutes for *lexA* [168] 10-30 mins for other genes (Typical Value)
X_{LexA}	Selected to reproduce the *lexA* protein half-life of about 1-2 minutes post irradiation [168]
T_{Repair}	Set at 45 minutes [172]

Table 4.5: Fitted parameter values for SOS response system

Gene	$(K^{bi}_{RNAP})_{\text{Repressed}}$	$(K^{bi}_{RNAP})_{\text{Unrepressed}}$	$\Phi(i)$	K^{bi}_{RNase}	$K^{bi}_{\text{Pr}\,oteasome}$	X_{LexA}
lexA	9.5×10^{-5}	4.75×10^{-4}	3.7×10^{-3}	3.5×10^{-3}	2.4×10^{-5}	30.0
uvrA	9.5×10^{-5}	2.85×10^{-4}	2.7×10^{-3}	3.5×10^{-3}	5.0×10^{-4}	
dinI	9.5×10^{-5}	4.75×10^{-4}	5.7×10^{-3}	3.8×10^{-3}	6.0×10^{-5}	
polB	9.5×10^{-5}	4.75×10^{-4}	1.5×10^{-3}	3.8×10^{-3}	7.0×10^{-4}	
uvrB	9.5×10^{-5}	9.50×10^{-4}	4.7×10^{-3}	4.8×10^{-3}	7.0×10^{-5}	
ruvA	9.5×10^{-5}	9.50×10^{-4}	1.7×10^{-3}	3.2×10^{-3}	1.0×10^{-4}	

Table 4.6: Comparison between experimental and the fitted values of the protein levels under unirradiated conditions. The protein numbers are represented as number of copies of the protein per cell. [a] Based on Kuzminov *et al.* [165]

	No of copies/cell	
Gene	**Fitted**	**Experimental**[a]
lexA	1306	1300
uvrA	49	20
dinI	384	500
polB	72	40
uvrB	243	250
ruvA	669	700

Table 4.7: Comparison between experimental and the simulation predictions for the protein levels under irradiated conditions. The protein numbers are represented as number of copies of the protein per cell. a Based on Kuzminov et al. [165]

	No of Copies/cell	
Gene	**Predicted**	**Experimental**[a]
lexA	143	130
uvrA	112.5	250
dinI	1120	2300
polB	175	300
uvrB	1421	1200
ruvA	2158	5600

4.5 Appendix

In the kinetic model for transport of inducer (lactose) developed by [53] the rate of transport, V of inducer into the cell is given by

$$\mathrm{V} = (\mathrm{k}_{\mathrm{in}} \frac{[\mathrm{Lactose}]_{\mathrm{ext}}}{[\mathrm{Lactose}]_{\mathrm{ext}} + K_T} - \mathrm{k}_{\mathrm{out}} \frac{[\mathrm{Lactose}]_{\mathrm{in}}}{[\mathrm{Lactose}]_{\mathrm{in}} + K_T})[lacY]$$

Here, $[lacY]$ is the available amount of protein generated by the $lacY$ gene (permease); k_{in} is the specific rate constant for transport of lactose into the cell and has a value of 35.8 mol lactose/mol permease/s; $\mathrm{k}_{\mathrm{out}}$ is the specific rate constant for transport of lactose out of the cell and has a value of 1.19 mol lactose/mol permease/s; K_T is the saturation constant for lactose transport and has a value of 2.6×10^{-4} M; and $[\mathrm{Lactose}]_{\mathrm{ext}}$ is the external lactose concentration set at 0.001 M.

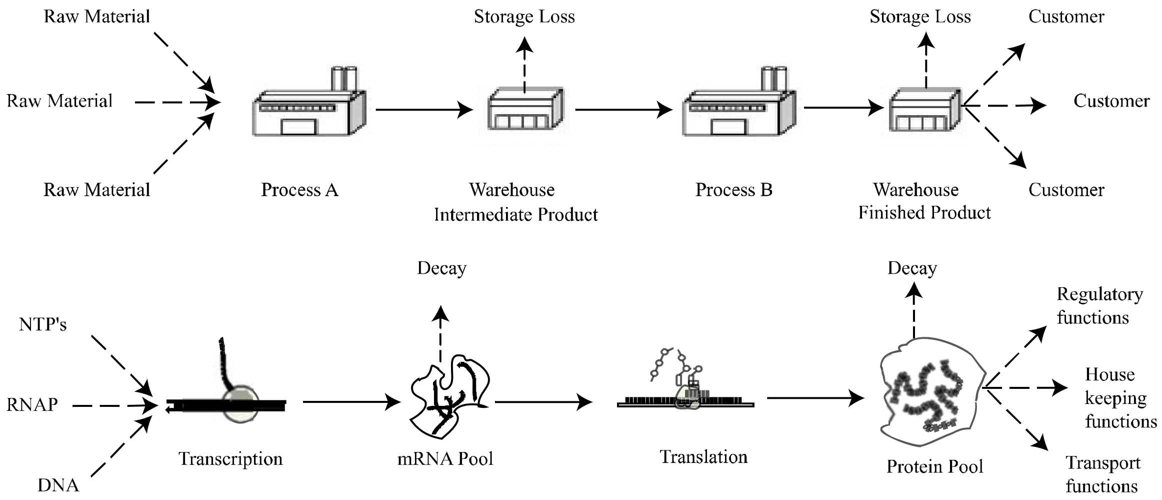


Figure 4.1: As in manufacturing processes, gene expression is also event driven implying that many events have to take place in a predetermined order to accomplish a certain task.

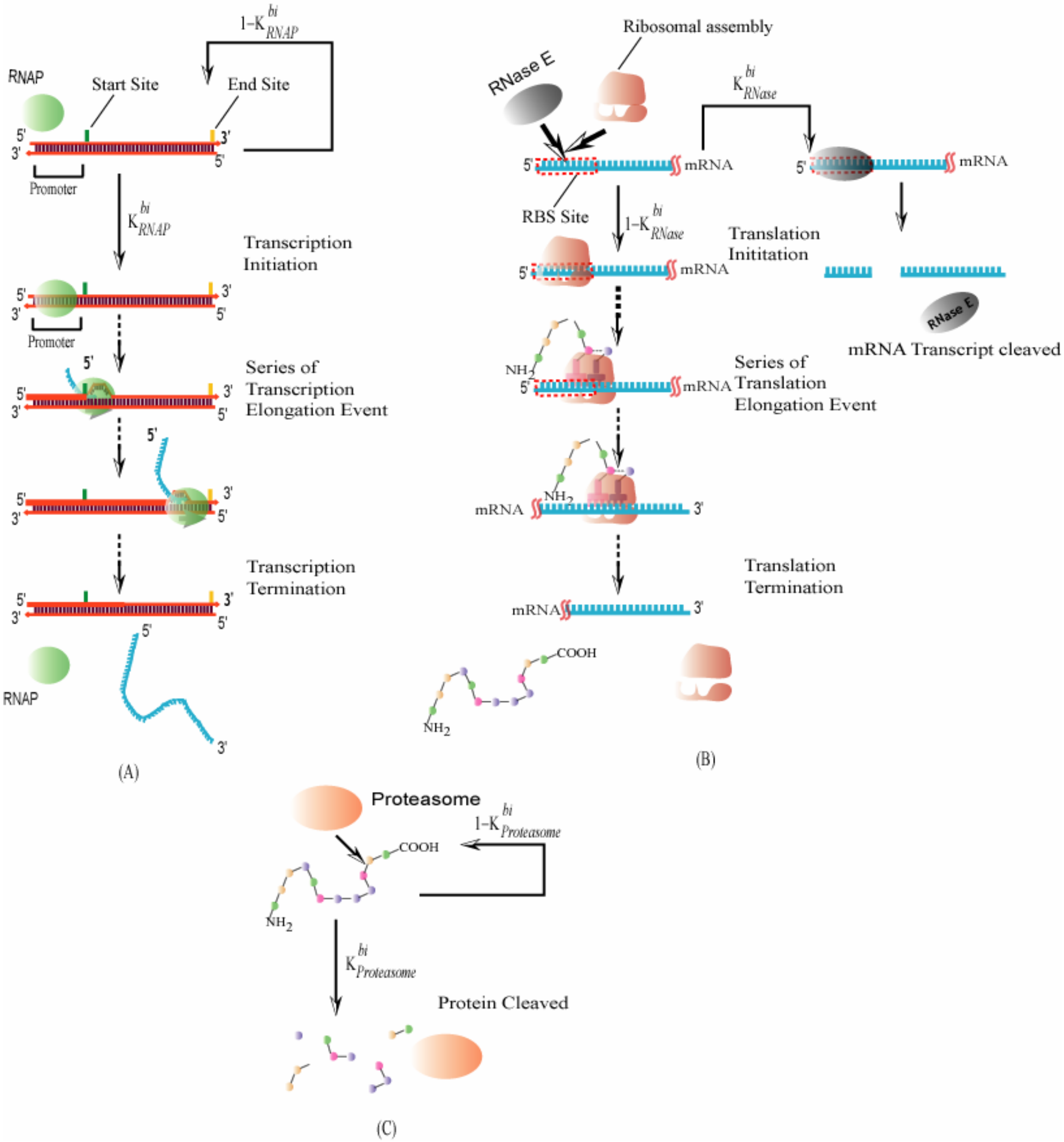


Figure 4.2: Sequence of events governing **(A)** Transcription module **(B)** mRNA decay and Translation modules **(C)** Protein decay module

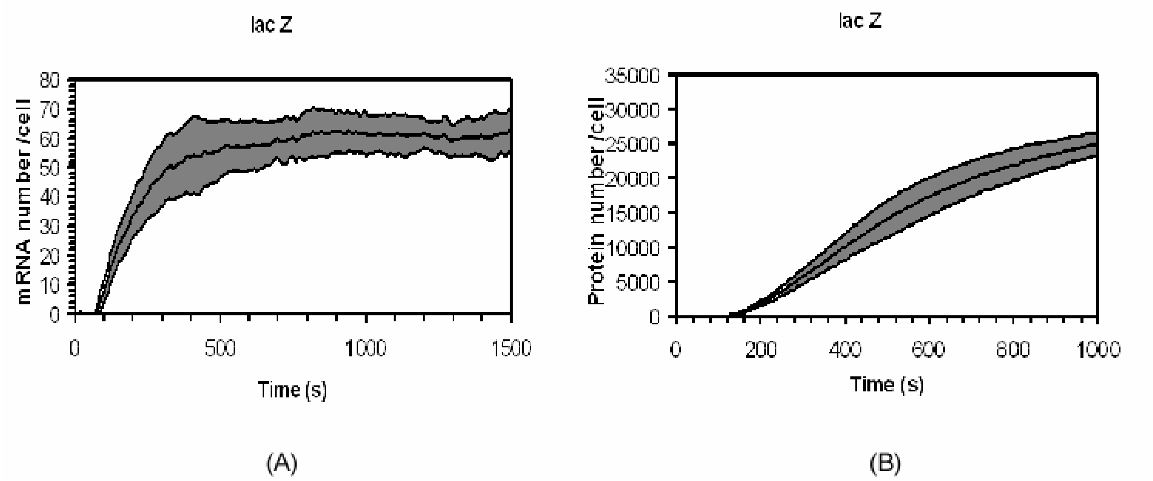


Figure 4.3: (A) The fitted profile for the *lacZ* mRNA copy number. **(B)** The fitted profile for *lacZ* protein copy number. The center solid line shows the mean profile after 50 simulation runs and the shaded region represents the ±1σ regions.

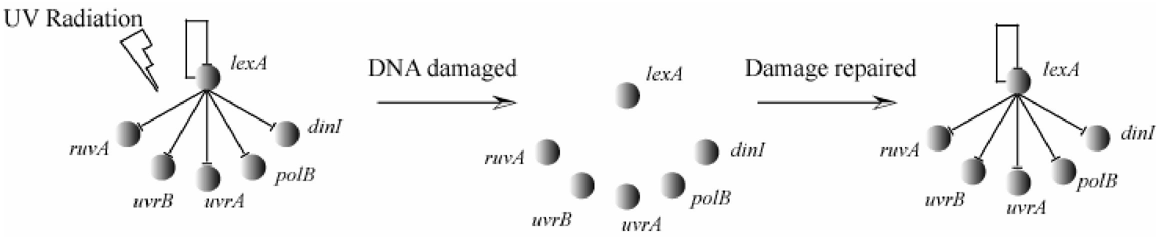


Figure 4.4: UV radiation damages the DNA duplex. The damage to DNA acts as a signal to de-repress the genes normally repressed by the lexA repressor. Consequently, these genes are more frequently expressed. After the damage to DNA has been repaired, the repressor activity of *lexA* is reestablished thus returning the cell to its original state.

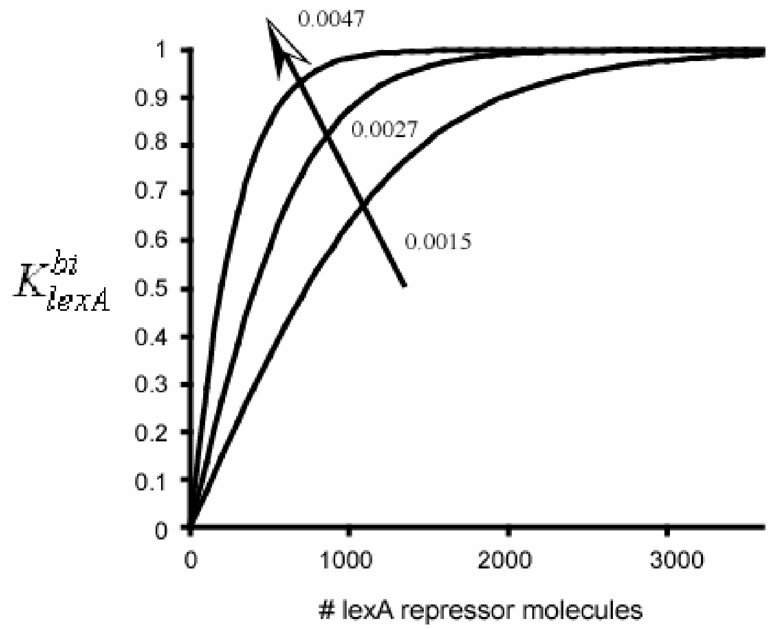


Figure 4.5: Typical profiles for the probability of repression as a function of the amount of *lexA* repressor (copy number). The probability of repression is a monotonically increasing function of the repressor level with diminishing returns. The larger the value of $\Phi(i)$, the higher the probability of repression as indicated by the black arrow.

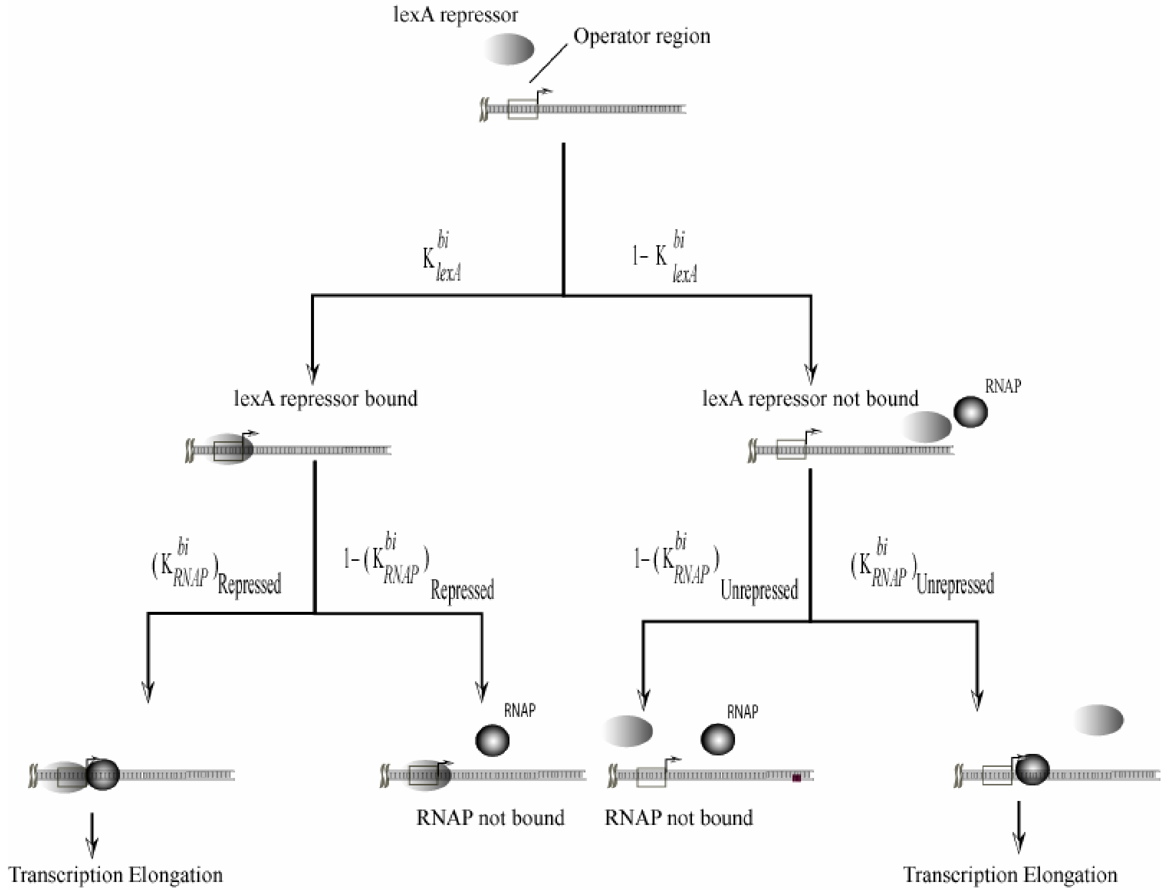


Figure 4.6: The regulatory logic employed to simulate the SOS response system.

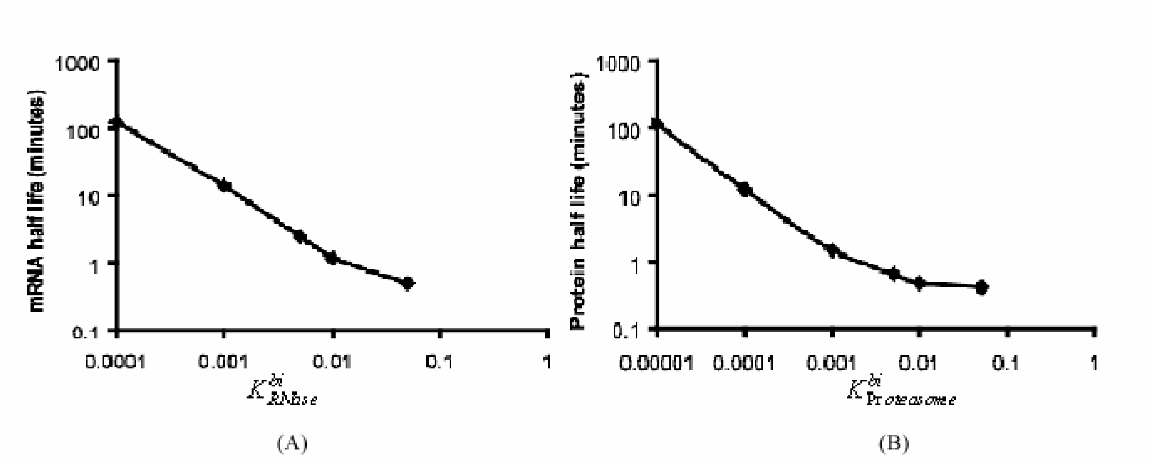


Figure 4.7: The average half life as a function of the governing decay parameter. **(A)** mRNA half life as a function of K_{RNase}^{bi}. **(B)** Protein half life as a function of $K_{\text{Proteasome}}^{bi}$

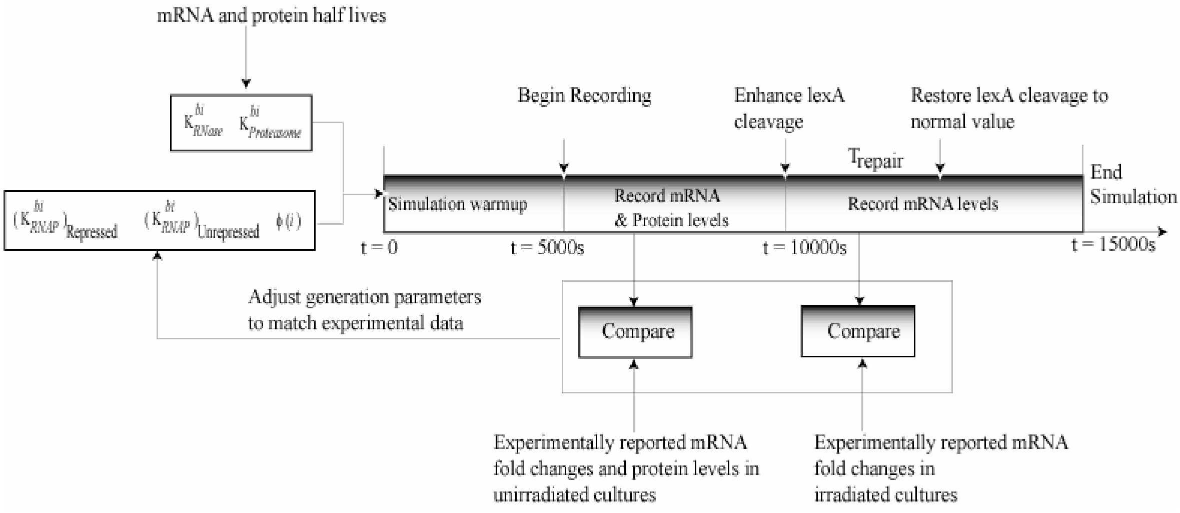

mRNA and protein half lives
K^{bi}_{RNase} $K^{bi}_{Proteasome}$
$(K^{bi}_{RNAP})_{Repressed}$ $(K^{bi}_{RNAP})_{Unrepressed}$ $\phi(i)$
Begin Recording
Enhance lexA cleavage
Restore lexA cleavage to normal value
T_{repair}
Simulation warmup
Record mRNA & Protein levels
Record mRNA levels
End Simulation
t = 0
t = 5000s
t = 10000s
t = 15000s
Adjust generation parameters to match experimental data
Compare
Compare
Experimentally reported mRNA fold changes and protein levels in unirradiated cultures
Experimentally reported mRNA fold changes in irradiated cultures

Figure 4.8: This figure illustrates how the parameters of the SOS response system are fitted to reproduce experimental data. The gene specific mRNA and protein decay parameters are estimated from the experimental mRNA and protein half lives respectively. Subsequently, the generation parameters are adjusted until the simulation results match experimental data for mRNA fold changes in *unirradiated* and *irradiated* cultures and the protein levels in *unirradiated cultures*

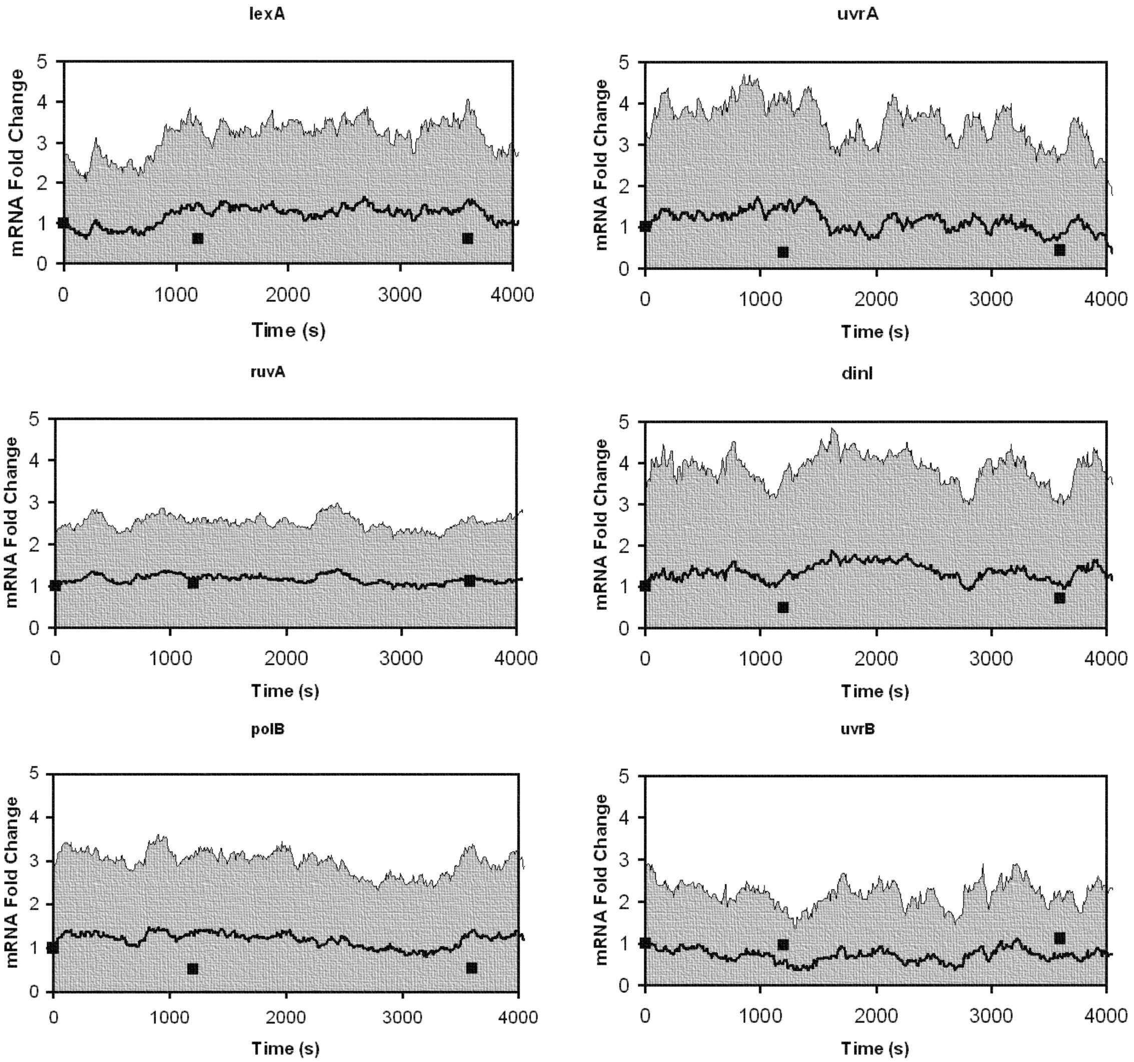


Figure 4.9: Fitted mRNA profiles under *unirradiated* conditions: The simulation results are the average of 120 simulation realizations. Both the mean trajectory and the ±1σ regions are plotted, where σ denotes the standard deviation. The balck squares are experimentally reported values [172].

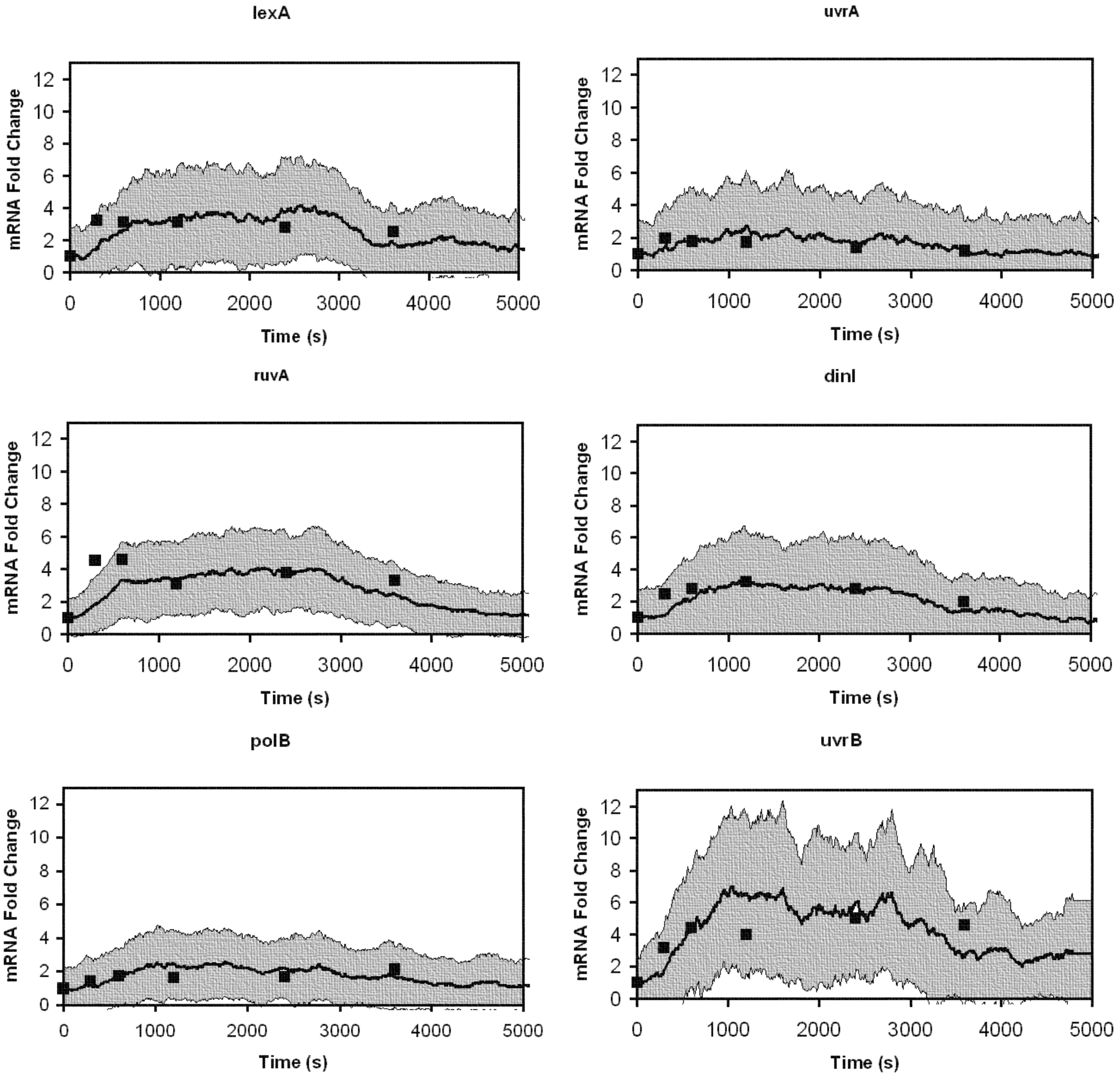


Figure 4.10: Fitted mRNA profiles *irradiated* conditions: The simulation results are the average of 120 simulation realizations. Both the mean trajectory and the ±1σ regions are plotted, where σ denotes the standard deviation. The black squares are experimentally reported values

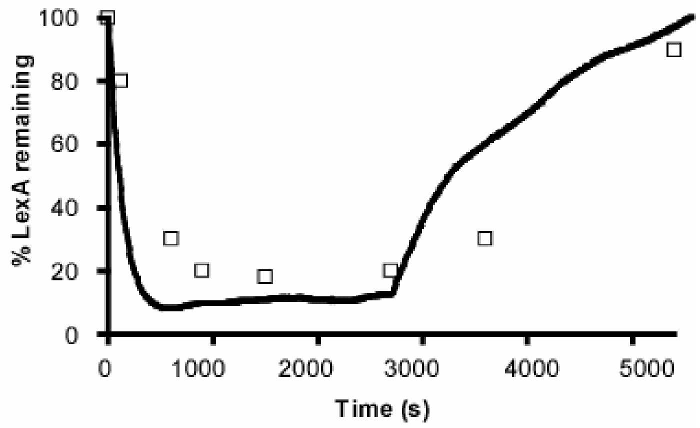


Figure 4.11: The temporal response of *lexA* gene in terms of change in the protein level following the induction of SOS response. The simulation profile is the average of 120 simulation runs. The red points are experimental measured values [168].

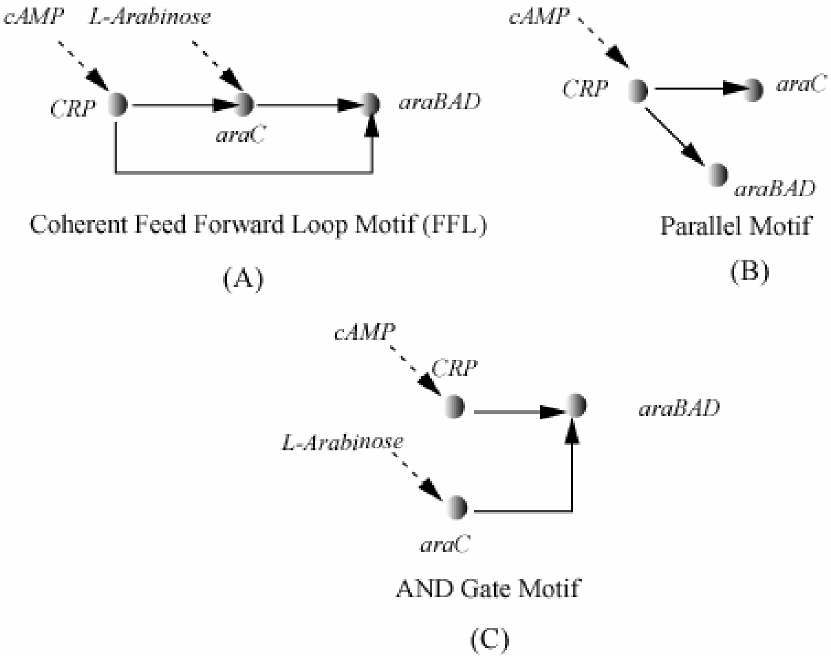


Figure 4.12: Alternative regulatory mechanisms considered in this study, **(A)** FFL motif **(B)** Parallel motif **(C)** Simple AND gate motif.

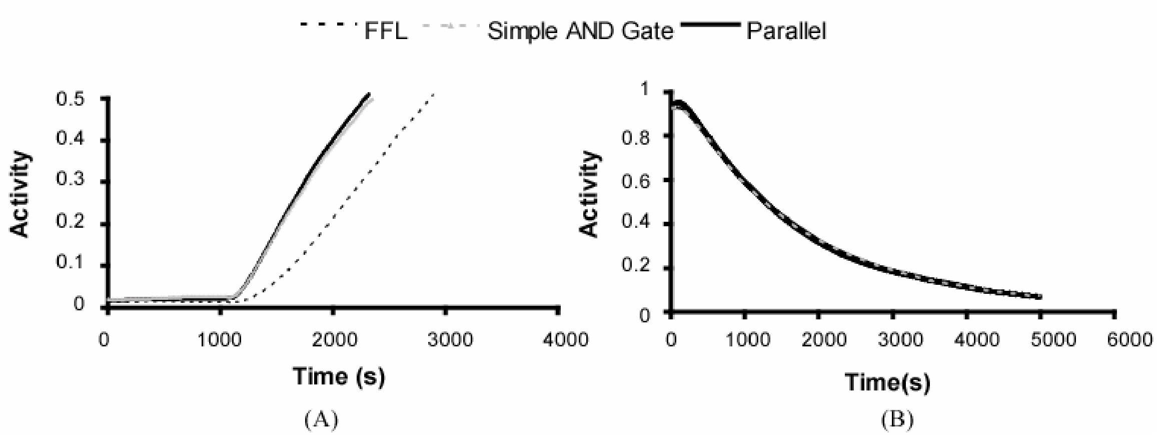


Figure 4.13: The temporal responses of the alternative regulatory motifs compared to a simple AND gate motif. **(A)** cAMP ON step **(B)** cAMP OFF step.

CHAPTER 5

Optimization Based Approach For Genetic Circuit Integration

5.1 Background

Recent years has witnessed an increasing number of studies on constructing simple synthetic genetic circuits that exhibit desired properties such as oscillatory behavior, inducer specific activation/repression, etc. The hope is that these simple circuits are the vanguards of more complex ones with far ranging implications to biotechnology and medicine bringing to fruition the promise of synthetic biology. It has been widely acknowledged that that task of building circuits to meet multiple inducer-specific requirements is a challenging one [180, 184-187]. This is because of the incomplete description of component interactions compounded by the fact that the number of ways in which one can chose and interconnect components, increases exponentially with the number of components. To meet these emerging challenges, in this chapter we introduce an optimization based framework that, given an underlying quantitative description, automatically identifies the circuit components from a list and connectivity that brings about the desired functionality.

To date, several small synthetic gene networks that accomplish a specific functionality have been constructed. For example, Gardner and co-workers have developed a genetic toggle switch- a synthetic gene regulatory network that exhibits bistability [188]. Similarly, Elowitz and Leibler have constructed a synthetic circuit termed as repressilator that was designed to produce an oscillatory response [148].

Subsequently, researchers have extended the repressilator circuit design to induce synchronous oscillations [189], design of a synthetic gene-metabolic oscillator [190] and many others [191-195]. Several researchers have employed synthetic circuits to investigate the dynamics and inner workings of more complex natural genetic networks. For example, Hooshangi et al. have constructed synthetic transcriptional cascades to investigate the ultrasensitivity and noise propagation in genetic networks [196]. Mangan et al. have investigated the structure and dynamics of the widely occurring feed forward loop motif [197, 198]. Similarly, Becskei and Serrano designed simple gene circuits to examine the effects of autoregulation in gene networks [199].

In addition to uncovering the design principles of natural genetic networks, synthetic genetic networks are now increasingly find roles in applications ranging from biotechnology, medicine and bio-sensing. For example, Martin et al. have successfully expressed enzymes from plants, yeast and *Escherichia coli* to produce amorphadine, a precursor to an anti-malarial drug artemisinin [200] and Anderson et al. have engineered the interaction between bacteria and cancer cells to depend on heterologous environmental signals [201]. Similarly, Levskaya et al. have devised a synthetic circuit that switches between different states in response to red light [202]. These ever expanding applications have spurred the interest for the development of efficient experimental, database and computational techniques to support these efforts [203] .

In response to these developments the research community has been rapidly moving towards standardization by creating the Registry of Standard Biological parts (http://parts.mit.edu/). This registry provides a comprehensive compilation of well-defined elements of a genetic circuit such as promoters, ribosome binding sites, transcripts, inducer molecules, terminator sites and plasmids among others. The impetus is that these spare parts registries will help usher the development of more rational engineering approaches for designing such circuits. The potential of using modeling and computations to better understand the function of these circuits has already been recognized and mathematical models have been proposed to describe the interactions between genetic elements [32-35].

The recent availability of well-defined spare parts lists and their interactions brings at the forefront the need to develop procedures to design and optimize genetic circuits that exhibit a desired functionality. Previous efforts in this direction include electrical circuit inspired designs proposed by Basu and Weiss [31, 204]. By constructing a library of cellular gates the authors have implemented simple logical functions such as OR, NOT and AND. Similarly, Mason et al. have investigated the behavior of an electronic model of a gene circuit to produce oscillatory behavior [205]. Other efforts include the combinatorial synthesis approach employed by Guet et al. [206]. In this work the authors varied the connectivity of genes and their corresponding promoters thus generating an ensemble of responses from the resulting genetic circuits. This approach, however, becomes intractable for circuits involving a large number of components [185]. Another important consideration associated with the design and fabrication of genetic circuits is the proper matching of kinetic rates of individual elements of the circuits. Several studies have reported that failure to generate the correct response is often due to improper assembly of the basic elements. For example, simulations conducted by Tuttle et al. have confirmed that repressillator circuits constructed by using wild-type promoters do not result in oscillations [30]. Similarly, studies conducted by Hoosangni et al. have revealed that the behavior of a transcriptional cascade depends on the promoter leakiness and expression levels at the previous stage [196]. Several researchers have stressed the need for optimizing the kinetic parameters to ensure functionality and both experimental [207] and computational approaches [185, 208, 209] have been proposed to this end.

To address these questions, in this work we introduce OptCircuit (see Figure 5.1), an optimization based framework that (i) automatically identifies the circuit components from a list and connectivity that brings about the desired functionality; (ii) Rectify or redesign an existing (non-functional) biological circuit and restore functionality by modifying an existing component (e.g., through changes in kinetic parameters) and/or identifying additional components to append to the circuit; Multiple literature sources are used to compile a set of kinetic descriptions of promoter-protein, protein-protein and protein-inducer pairs. The dynamics that govern the interactions between the elements of the genetic circuit are currently modeled using deterministic rate equations but the framework is general enough to accommodate stochastic simulations. The desired circuit response is abstracted as the maximization/minimization of an appropriately constructed objective function. Subsequently, an iterative procedure is implemented within our framework to identify an ensemble of circuits that exhibit the desired response. OptCircuit has been applied on a variety of applications ranging from the design of circuits that discriminate between inducer molecules; circuits that detect the combination of inducer molecules (i.e., 2 to 4 genetic decoder) and finally circuits whose responses are dependent on the concentration of the external inducer (concentration band detector).

5.2 Methods

5.2.1 Modeling framework

The basic elements constituting a genetic circuit include promoter elements, protein/transcript molecules and inducers. Briefly, promoters are regions of DNA where RNA polymerases bind to initiate transcription. Transcripts referred to ORFs which upon transcription and translation produce proteins which in synthetic circuits act as transcriptional regulators repressing or activating a promoter's strength. Finally, inducers are small molecules (e.g., aTC, IPTG) which by directly interacting with transcription factors can block, enable or simply modulate a transcriptional regulation event. The quantitative description of the mechanistic detail underlying the interactions embedded in the genetic circuitry requires the definition of the following sets and variables.

$Sets:$

$$I = \{i\} = \text{ set of promoters}$$
$$J = \{j\} = \text{set of transcripts}$$
$$K = \{k\} = \text{set of inducers}$$
$$T = \{\mathrm{t}\} = \text{time}$$

$Variables:$

$$P_j(t) = \text{protein level of transcript } j \text{ at time } t$$
$$Y_{ij} = \begin{cases} 1 \text{ if transcript } j \text{ is expressed from promoter } i \\ 0 \text{ otherwise} \end{cases}$$

The set *I* represents all the promoter elements investigated in this study. *J* represents the set of transcripts and finally *K* is the set of all inducer molecules. Model variables encode the structure of the synthetic circuit and quantify the protein levels. Specifically, the binary variable Y_{ij} determines which transcript *j* is expressed from a promoter *i* and $P_j\ (t)$ quantifies the level of protein *j* at a given time *t*.

5.2.2 Kinetic description of interactions

Genetic circuits are characterized by a number of interactions including protein-promoter and protein-inducer and protein-protein interactions. For example, protein *lacI,* in its tetramer form functions as a repressor for P_{lac} promoter while inducer molecule aTc suppresses the activity of protein *tetR*. In genetic circuits, unlike digital or binary logic based circuits, the presence/absence of a particular set of interactions alone is insufficient to accurately predict correctly all possible responses. In fact, several studies have reported that in addition to interactions, the kinetic rates of individual elements have to accurately match in order to ensure function. To this end, the kinetic description of each element of a genetic circuit is embedded into the OptCircuit framework.

Specifically, for every transcript *j*, the set of ordinary differential equations (ODE's) that govern the time evolution of the protein is given by Eq 5.1.

$$\frac{dP_j}{dt} = \sum_i Y_{ij}\ [\text{Rate of Production of j from i}] - K_{decay}^{j} P_j\,(t) \forall j \qquad (5.1)$$

The first term in Eq 5.1, accounts for the cumulative rate of production of a particular protein *j* from the promoter elements and the second term represents the first order decay of the protein. Also observe that the production of a protein *j* from a promoter *i* is turned ON if and only if the corresponding binary variable Y_{ij} is equal to one.

OptCircuit accounts for the activating and repressing effects on every promoter *i* within the framework by using the modeling formulation proposed by Hasty et al. [33]. Briefly, all biochemical reactions characterizing the interactions affecting a particular promoter are listed and divided into fast and slow steps. The fast reaction set typically includes protein dimerization and protein promoter binding while transcription and degradation steps compose the slow reaction set. The dynamics governing the promoter kinetics are derived using mass action kinetics with fast reactions that have rate constants in the order of seconds, assumed to be in equilibrium [33]. The modeling environment in OptCircuit is versatile enough to incorporate finer levels of mechanistic detail whenever available (e.g., modeling of mRNA [148]).

5.2.3 Objective Function Modeling

The reliance on an optimization framework for designing synthetic circuits with a desired response implies that the objective function must be carefully chosen so as its maximization or minimization is a good surrogate of the desired response(s). The type of desired responses is partitioned into inducer-free and inducer-dependent ones. Inducer-free responses translate into the design of circuits whose response is consistent with a targeted time-course. This response may be oscillatory, constant or ramping up/down. For all these case the objective function, Z minimizes the sum of the squared departures from the targeted responses at all time points:

$$Minimize\, \mathrm{Z} = \sum_{\mathrm{t}} (P_{j^*}(t) - P_{j^*}^{\exp}(t))^2 \qquad (5.2)$$

In Eq. 5.2, $P_{j^*}^{\exp}(t)$ denotes the experimentally observed profile. Inducer-dependent responses require a clear distinction between states corresponding to presence/absence of multiple inducers. Specifically, to accomplish this, an objective function is constructed as follows that maximizes the scaled separation between the inducer-present/absent responses:

$$Maximize\;\; \mathrm{Z} = \sum_{\mathrm{j}\in\mathrm{R}} \sum_{k\in K} \frac{P_j(T)\,|_k - \sum\limits_{j' \neq j \in R} P_{j'}(T)\,|_k}{P_j(T)\,|_k} \qquad (5.3)$$

In Eq. 5.3, K represents the set of inducer molecules present in the system, R represents the set of reporter proteins (e.g. GFP, YFP etc) and $P_j(T)\,|_k$ represents the steady state levels of transcript j in presence of inducer k. Alternatively, if the circuit response must be inducer concentration dependent then the objective function can be formulated again as a the minimization of a least squares sum by considering multiple inducer concentrations.

$$Minimize\, Z = \sum_{r} \left(P_j\;\; |_{k,r} - P_j^{\exp}\;\; |_{k,r}\right)^2 \qquad (5.4)$$

In Eq 5.4, r represents the discretizations levels for the inducer concentration. $P_j\big|_{k,r}$ and $P_j^{\exp}\big|_{k,r}$ represent the simulated and the desired steady-state levels of reporter transcript j at inducer discretization level r. These are only some examples of desired circuit responses. Using this optimization-based framework even more complicated responses can be modeled limited only by the imagination of the circuit designer.

5.2.4 Optimization model

Using the notation listed above, the problem of designing a genetic circuit that exhibits a desired response is formulated as the following mixed integer dynamic optimization problem (MIDO) [210-214].

$$\begin{aligned}&\text{Min/Max Z} = f(P_j(t))\\&s.t.\end{aligned} \tag{5.5}$$

$$\frac{dP_j}{dt} = \sum_i Y_{ij}\ \text{[Rate of Production if j from i] - } K^j_{decay} P_j(t) \forall j \tag{5.6}$$

$$\sum_j Y_{ij} \leq P^{\max} \quad \forall\, i \tag{5.7}$$

$$\sum_i Y_{ij} \leq T^{\max} \quad \forall\, j \tag{5.8}$$

$$\sum_i \sum_j Y_{ij} \leq M_{Max} \tag{5.9}$$

The objective function in Eq. 5.5 models the circuit response imposed by the circuit designer. Eq. 5.6, describes the time evolution of protein levels as a set of ordinary differential equations as described in the previous section. Eq. 5.7 imposes an upper limit on the number of transcripts a particular promoter i can express. Similarly, Eq. 5.8

imposes a limit on the number of times a particular transcript *j* can be expressed from different promoters. Finally, Eq. 5.9 imposes a limit on the total number of promoter-transcript pairs in the designed genetic circuit.

The boolean constraints (Eq. 5.7-5..9) offer the flexibility to incorporate the design of an existing biological circuit and probe its behavior. This can be accomplished by incorporating constraints of the form

$$Y_{ij} = 1 \, \forall \, (\{\mathrm{i}, \mathrm{j}\} \in EX) \tag{5.10}$$

where, the set *EX* contains the connectivity information of the circuit. This feature confers upon us the ability to readily extend the framework to rectify or redesign an existing (non functional) biological circuit by identifying additional components to append to the circuit to ensure its functionality.

The solution procedure for the MIDO class of optimization problems is difficult [214] due to the simultaneous presence of binary variables Y_{ij} and constraints in form of ODE's. Reliable solution methodologies that guarantee a global optimal solution for this class of problems are still in infancy [215]. Therefore, in this research we rely on a decomposition procedure to bracket an optimal solution. The basic idea of proposed approach is to generate a converging sequence of upper and lower bounds to the original problem. The solution procedure is listed in a step-wise manner below.

Step 1 : Initialize iteration counter, *iter* = 1; SET upper bound UB = ∞; SET lower bound LB = -∞; Generate an initial guess for a feasible circuit design $Y_{ij} = Y_{ij}^{iter}$

Step 2 : Integrate the system of ordinary differential equations (1.1) for fixed values of the design variables $Y_{ij} = Y_{ij}^{iter}$ to obtain the objective function value *Z*; SET UB = min(UB,Z). Store the solution corresponding to the best upper bound.

Step 3: Construct the master (lower bounding) problem as follows.

$$\min imize\ \mu$$
$$s.t.$$

$$\mu \geq Z^{iter} + \sum_{i}\sum_{i}(Y_{ij} - Y_{ij}^{k})\left(\frac{\partial Z}{\partial Y_{ij}}\right)_{Y_{ij}=Y_{ij}^{k}} \quad \forall\ k = 1,2,....iter \quad (5.11)$$

$$\sum_{j} Y_{ij} \leq P^{\max} \quad \forall i \quad (5.7)$$

$$\sum_{i} Y_{ij} \leq T^{\max} \quad \forall j \quad (5.8)$$

$$\sum_{i}\sum_{j} Y_{ij} \leq M_{Max} \quad (5.9)$$

The partial derivates are computed finite difference methods.

Solve the master problem to obtain the objective function value μ^{*} and integer solution, Y_{ij}^{*}.

SET LB = max (LB, μ^{*}).

Step 4 : If LB $\geq$UB, then STOP (crossover). Otherwise, increase iteration counter

iter → *iter* +1. $Y_{ij}^{iter} = Y_{ij}^{*}$. Return to Step 2.

In addition to identifying the optimal configuration of design variables (Y_{ij}^{*}), OptCircuit can also be employed to optimize kinetic parameters of specific elements within the genetic circuit. For example, given a genetic circuit, the task of determining the optimal promoter strength of a particular promoter i^{*} can be achieved by replacing Eq. 5.11 with

$$\mu \geq Z^{iter} + (\alpha_i - \alpha_i^k)\left(\frac{\partial Z}{\partial \alpha_i}\right)_{\alpha_i = \alpha_j^k} \quad \forall \ k = 1,2,....iter \qquad (5.12)$$

Note that given the nonlinear nature of the problem under investigation the above procedure is carried out multiple times starting for several starting initial guesses and the local optimum solution identified at each iteration is stored along with a sorted list of the best circuit configurations.

5.3 Results

In this section we highlight the capabilities of the OptCiruit framework to design circuits of varying stimulus and complexity. We first examine the test the design of simple circuit(s) against known architectures that discriminate between inducer molecules. Next, we dial up the complexity of the desired circuit response by seeking circuit configurations that can detect which combination of inducer molecules are present/absent. Finally, we test the ability of the framework to identify circuits whose responses are not only dependent on the presence/absence but also on the level of external inducers.

5.3.1 Inferring circuits with inducer-specific responses

Here we test OptCircuit by generating circuit designs whose responses are contingent on the presence/absence of different inducer molecules and compare the results with known designs [188]. Specifically, in the presence of anhydrotetracyclin (aTc) the desired circuit must express only protein *lacI* while in response to inducer IPTG the circuit must express only protein *tetR*.

The desired circuit response is imposed by maximizing the scaled difference between the expression of the desired minus the undesired florescent protein in response to the two different inducers in line with the description provided in the methods section.

$$Maximize\ \ Z = \left(\left(\frac{P_{lacI}^{aTC} - P_{lacI}^{IPTG}}{P_{lacI}^{aTC}} \right) + \left(\frac{P_{tetR}^{IPTG} - P_{tetR}^{aTc}}{P_{tetR}^{IPTG}} \right) \right) / 2 \qquad (5.13)$$

In Eq 5.13, $P_{lacI}^{aTC}, P_{tetR}^{aTC}$ represent the levels of transcripts *lacI* and *tetR* in presence of inducer *aTc* and similarly, $P_{lacI}^{IPTG}, P_{tetR}^{IPTG}$ represent the levels of *lacI* and *tetR* in presence of inducer *IPTG* respectively.

Using OptCircuit we identify multiple circuits with up to two promoter transcript pairs. The circuit configuration for the best solution is shown in Figure 5.2A. Interestingly, the configuration is reminiscent of the architecture of the well-studied genetic toggle switch [188]. The effect of dialing up the complexity of the designed circuits by allowing for as many as three and four promoter transcript elements is shown in Figure 5.2(B,C,D). Results indicate that in addition to relatively simple circuit designs akin to known ones, OptCircuit suggests non-intuitive designs with added complexity affording more opportunities for kinetic parameter tuning.

5.3.2 Design of genetic decoder

In this section, we use OptCircuit to design for more complex responses by constructed a genetic circuit equivalent of a 2-4 bit decoder. A digital decoder is a multiple-input, multiple-output logic circuit that converts coded inputs into coded outputs. Figure 5.3(A) illustrates the block diagram of a digital decoder and the corresponding truth table is shown in Figure 5. 3(C). In the context of genetic circuits, we seek the design of a circuit architecture that produces four different responses dependent on the presence and/or absence of the sugars *glucose* and *L-arabinose* respectively. Specifically, we would like the circuit to express (i) *YFP* in response to the presence of *L-arabinose* and absence of *glucose*, (ii) *RFP* in response to the absence of both *glucose* and *L-arabinose*, (iii) *BFP* when both *L-arabinose* and *glucose* are present, (iv) *GFP* when *L-arabinose* is absent but *glucose* is present (see Figure 5.3(B)).

Given *N* different promoter elements and *M* transcripts, the total number of design configurations with upto *K* promoter-transcript pairs is given by $(NM)^K$. This implies that the search space characterizing all circuit configurations is enormous even for relatively modest values for *N* and *M* thus preventing its exhaustive navigation. To alleviate this problem, we implemented the OptCircuit framework in a sequential fashion where successive elements are appended to the genetic circuit to meet, one at a time, the four desired responses. At each step, the objective function values of the ten best circuit architectures are recorded and the circuit producing the best objective value is retained for the next step. The first step shown in Figure 5.4, involves the expression of *YFP* under the $(-/+)$ condition. To this end, we borrowed the circuit configuration from the well studied, feed-forward loop architecture [197, 198]. *CRP* and *AraC* are expressed from the constitutive promoters, P_{cons}^{1} and P_{cons}^{2} respectively and *YFP* is placed under the control of the P_{BAD} promoter.

Using the circuit described in the previous paragraph as the seed, the OptCircuit framework is employed to sequentially identify additional components by following the step-wise procedure shown in Figure 5.4. After the second step (i.e., (-/-) response), our framework identifies the expression of *lacI* from the P_{BAD} promoter and the expression of *RFP* from the P_{lac} promoter. In the third step (i.e., (+,+) response) the best objective value was realized the expression of protein *tetR* from P_{BAD} and P_{lac} promoters and expression of *BFP* from the P_{tet} promoter which is repressible by protein *tetR* (Figure 5.4, step 3). Finally after the last step ((+,-) response), by allowing for expression of *GFP,* the additional elements appended to design the decoder include, the expression of protein *cI* from the P_{ara} and P_{lac} promoters and the expression of proteins *tetR* and *GFP* from the P_{λ} promoter.

The identified circuit design (Figure 5.4) happens to be consistent with a purely binary logic viewpoint of regulation. This is not the case with all identified designs. For example, one such circuit configuration involves the expression of protein *lacI* from P_{ara} promoter instead of the expression of protein *tetR* from the P_{BAD} promoter (see Figure 5.5) leading to a behavior that it is inconsistent with Boolean-only regulation. To illustrate this, consider the truth table of the design shown in Figure 5.5. When both the sugars are present, then *YFP* is expressed and *RFP* and *GFP* are shut-off. However, unlike the circuit described in previous paragraph, expression of *YFP* is not accompanied by expression of *tetR* and hence the P_{tet} promoter is free to express the fluorescent protein *BFP*. Nevertheless, OptCircuit identified this circuit configuration as an optimal architecture for a genetic decoder because the employed kinetic description accounts for *not only the presence but also the level* of each participating molecule needed to activate transcription. Figure 5.6, provides a comparison of the steady-state levels of proteins *tetR* and *BFP* for the circuits described in Figure 5.4 (step 4) and Figure 5.6. In circuit (A), the level of *tetR* is relatively high (~ 60 nm) which in turn strongly represses the expression from the P_{tet} promoter. This is expected since, in circuit A, expression of *YFP* is accompanied by expression of *tetR*. In contrast, in circuit B, even though the level of protein *tetR* is relatively low (~10 nm), examination of the level of protein *BFP* suggests that even low levels of protein *tetR* are able to effectively repress the expression of *BFP* from the P_{tet} promoter. The low level of *tetR* is a manifestation of the leaky repression exerted on the P_{lac} promoter by the *lacI* protein. This observation is further substantiated by expression of protein *RFP*, albeit at low levels. These results indicate that by taking into account the underlying kinetic description of the interactions, the OptCircuit framework is able to expand upon possible circuit designs by drawing from architectures that may not be valid based on digital logic viewpoint though adequately meet the imposed requirement due to the careful matching of kinetic parameters as often observed in nature.

This study also sheds light onto the design principles for the construction of a genetic decoder. Figure 5.7, illustrates the binary logic schematics for the genetic circuits characterizing each step shown in Figure 5.5(B). Observe that at each step, the OptCircuit framework allows for the addition of components that are activated only when the corresponding inducer conditions are met. In addition, by expressing appropriate repressor molecules, the OptCircuit framework ensures the repression of all the other promoters expressing fluorescent proteins. For example, after Step 1 the AND gate expressing *YFP* is active only when both the inducers are present (see Figure 5.7). Subsequently, a NOT gate logic is introduced after the second step to turn OFF the P_{lac} promoter when *YFP* is expressed. After the third step, an OR gate with two inputs followed by an NOT gate is introduced. The OR gate combines the indirect repressive effect that turns OFF the production of *GFP* if either *YFP* or *RFP* are expressed. Finally, after the last step an OR gate with three inputs followed by a NOT gate is introduced. This ensures that if either one of *GFP, RFP* or YFP are expressed then *BFP* is turned OFF and conversely *BFP* is expressed only if none of the three are expressed.

5.3.3 Design of Concentration Band Detector

With this example, we explore whether OptCircuit can pinpoint design configurations whose responses are dependent not only on the presence/absence of external inducers but also on their concentrations. We use the concentration band detector example [216] to demonstrate the OptCircuit application. Briefly, this circuit expresses high levels of a reporter protein only when the concentration of the external inducer (i.e.

L-arabinose) is within a specific range [216] (i.e., neither too high or too low) as shown in Figure 5.8.

In line with the design proposed by Basu and coworkers, OptCircuit first places the reporter protein under the control of a repressible promoter (i.e., P_{tet} promoter) which is repressed by protein *tetR* (dotted line in Figure 5.8(A)). Subsequently, we use OptCircuit to design two circuits, a low threshold detector (LTD) and a high threshold detector (HTD). The LTD circuit expresses high levels of *tetR* at low levels of *L-arabinose* and low levels of *tetR* at high levels of *L-arabinose* (see Figure 5.8(B)). In contrast, the HTD circuit is designed to express low levels of *tetR* at low levels of inducer and high levels of *tetR* at high levels of *L-arabinose* (see Figure 5.8(C)). Finally, the LTD and HTD circuits are fused together to obtain an inverted bell shaped response for protein *tetR*.

The best circuit configurations proposed by OptCircuit are shown in Figure 5.9. The only difference between the LTD and HTD is that while *tetR* is expressed from the P_{lac} promoter in the LTD, it is expressed from P_{BAD} promoter in the HTD. Examination of the circuit behavior reveals that at low levels of *L-arabinose*, the P_{BAD} promoter is not sufficiently activated ensuring low levels of protein *lacI*. This in turn implies that the P_{lac} promoter is free to express *tetR* from the LTD circuit (see Figure 5.9 and Figure 5.10(A)). As the amount of *L-arabinose* accumulates in the system, the transcriptional expression from the P_{BAD} promoter is enhanced leading to expression of *lacI* from LTD and *tetR* from HTD (see Figure 5.9 and 5.10 (B)). Finally, expression of *lacI* from HTD turns off expression of *tetR* from HTD. The final OptCircuit design enables the expression of protein *tetR* from P_{lac} and P_{BAD} promoters, *lacI* from P_{BAD} promoter and reporter protein *GFP* from P_{tet} promoter. The level of protein *tetR* as a function of level of *L-arabinose* is shown in Figure 5.10(C). As shown in Figure 5.10(C), we find that the circuit response deviates significantly from the desired response implying that by simply reshuffling existing components the desired response is not attainable.

To address this remaining challenge we next explore whether modifying any existing component in the circuit will shift the circuit response closer to the desired response. Specifically, the circuit described in the previous paragraph is ``fixed'' and subsequently starting from the current parameter values as an initial guess we optimize the kinetic parameter values using Eq 5.12. Results indicate that a considerable improvement in circuit response is obtained when the transcriptional efficiency of the constitutive promoter expressing protein *CRP* is decreased 13 fold. This resulted in a 18.69% decrease in the objective value (8.1071→6.5976). The effect of this parameter modifications are quantified in Figure 5.10(D) demonstrating that OptCircuit can be used to pinpoint kinetic parameter modifications improving its functionality.

5.4 Summary and Discussion

In this chapter, we introduced an optimization-based approach termed OptCircuit that (i) automatically identifies the circuit components from a list and connectivity that brings about the desired functionality; (ii) Rectify or redesign an existing (non-functional) biological circuit and restore functionality by modifying an existing component and/or identifying additional components to append to the circuit. The dynamics that govern the interactions between the elements of the genetic circuit were modeled using deterministic rate equations and the desired circuit response is abstracted as the maximization/minimization of an appropriately constructed objective function. Subsequently, an iterative procedure was implemented within our framework to identify an ensemble of circuits that exhibit the desired response. The capabilities of the developed tool were investigated by synthesizing circuits that exhibit a wide array of responses. The genetic toggle switch example demonstrated the ability of the framework to suggest simple or more complex circuit configurations capable of discriminating between inducer molecules. The 2 to 4 genetic decoder example led to complex circuit designs consisting of as many as 13 promoter-transcript pairs that may or may not be identifiable through a digital logic based design procedure. Finally, the concentration band detector example illustrated how OptCircuit can be used to design not the architecture of the synthetic circuit but also suggest modifications on its kinetic parameters for optimized performance. OptCircuit can also be employed in tandem with existing computational methods for fine-tuning circuit performance by providing initial configurations. For example, Feng and co-workers developed a global sensitivity analysis based approach to identify the optimal parameter configuration [185]. In their approach, they start from a representative circuit configuration and then proceed to identify the optimal parameter set by estimating the sensitivity of the parameter variation on the circuit response.

It is important to emphasize that all the kinetic parameter modifications suggested by OptCircuit can be realized using a host of experimental strategies. For example in the construction of repressilator circuit, the authors control the rate of protein degradation by *ssrA* tagging whereby an amino acid sequence is introduced into the proteins which makes them a target for all proteases [217]. Similarly Yokobayashi and co-workers have used directed evolution to restore the performance of an unoptimized circuit [207]. Specifically, by focusing on the *cI* gene and its corresponding ribosome binding site, the authors report mutations that potentially reduce the translational efficiency or reduce the ribosome binding affinity. Other promising strategies include the approaches developed by Lutz and Bujard [218] to control the promoter activity and repression for the P_{tet} and P_{lac} promoters.

In recent years, researchers have deposited several standard and interchangeable biological parts in the registry of standard biological parts (e.g. composite parts such as Isoamyl alcohol generating device (BBa_J45400), Elowitz repressilator (BBa_I5610)). Currently efforts are underway to specify the functionality of these parts interms of parameter estimates and behavior. Our results and those proposed by other researchers conclusively demonstrate that proper parameter compatibility is essential to ensure funtionality. As the characterization of these parts is moving at a fast pace, the OptCircuit framework could serve as a design platform to aid in the construction and finetuning of integrated biological circuits.

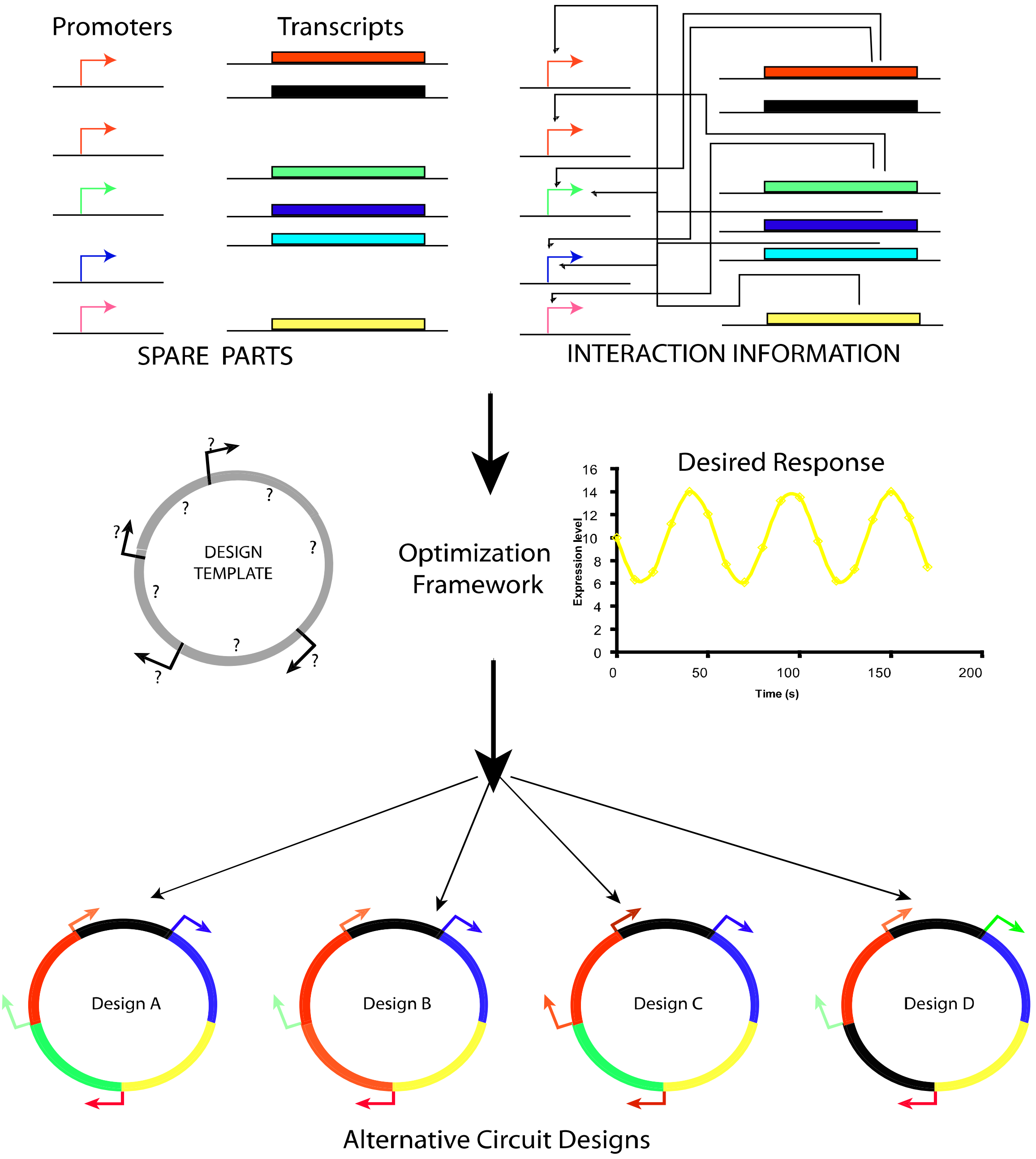


Figure 5.1: A pictorial illustration of the OptCircuit framework. The three key components of the framework are the basic genetic elements (promoters, transcripts, inducers); the underlying kinetic mechanisms that drive the circuit response and finally the desired behavior of the circuit under construction. These three components are integrated by OptCircuit using an optimization based formulation.

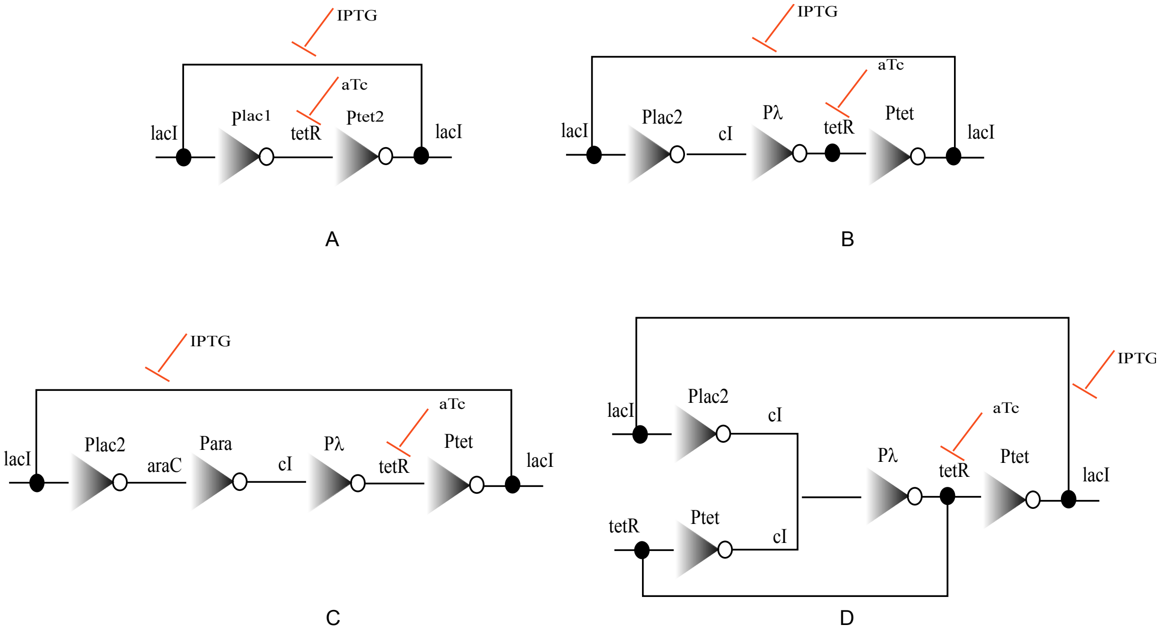
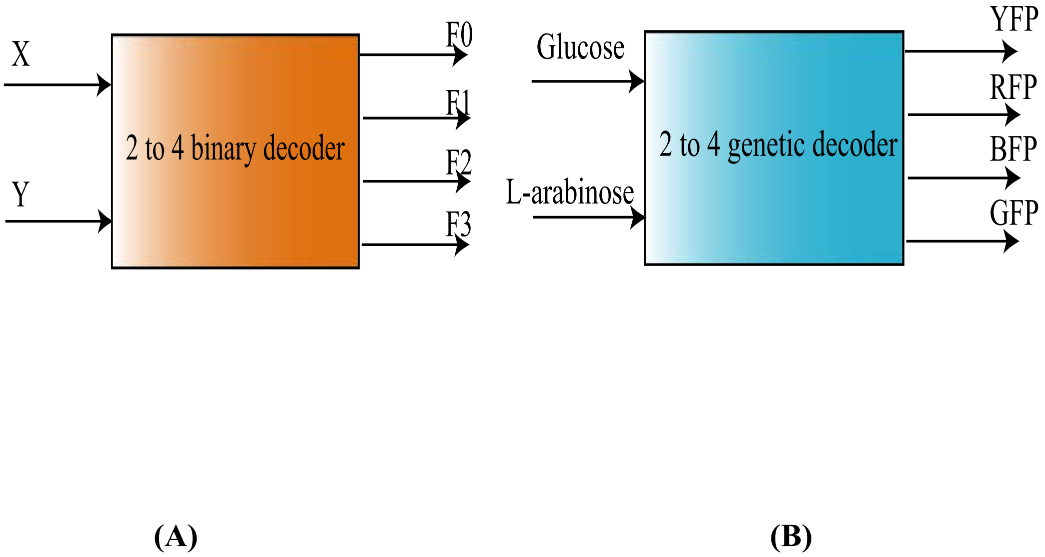


Figure 5.2: Alternative circuit configurations proposed by OptCircuit for the first example. OptCircuit is able to identify more complex architectures to realize a particular outcome. For example, in (A), in presence of aTc, protein *tetR* is suppressed enabling expression of *lacI* from P_{tet} promoter. On the other hand, in presence IPTG, protein *lacI* is suppressed enabling expression of *tetR* from P_{lac1} promoter.



X	Y	F0	F1	F2	F3
0	0	1	0	0	0
1	0	0	1	0	0
0	1	0	0	1	0
1	1	0	0	0	1

(C)

Figure 5.3: A pictorial illustration of the 2-4 bit decoder. In **(A)**, a block diagram for a digital decoder is shown. In **(C)**, the corresponding genetic decoder is illustrated. **(C)**, provided the truth table associated with the decoders shown in **(A)** and **(B)**.

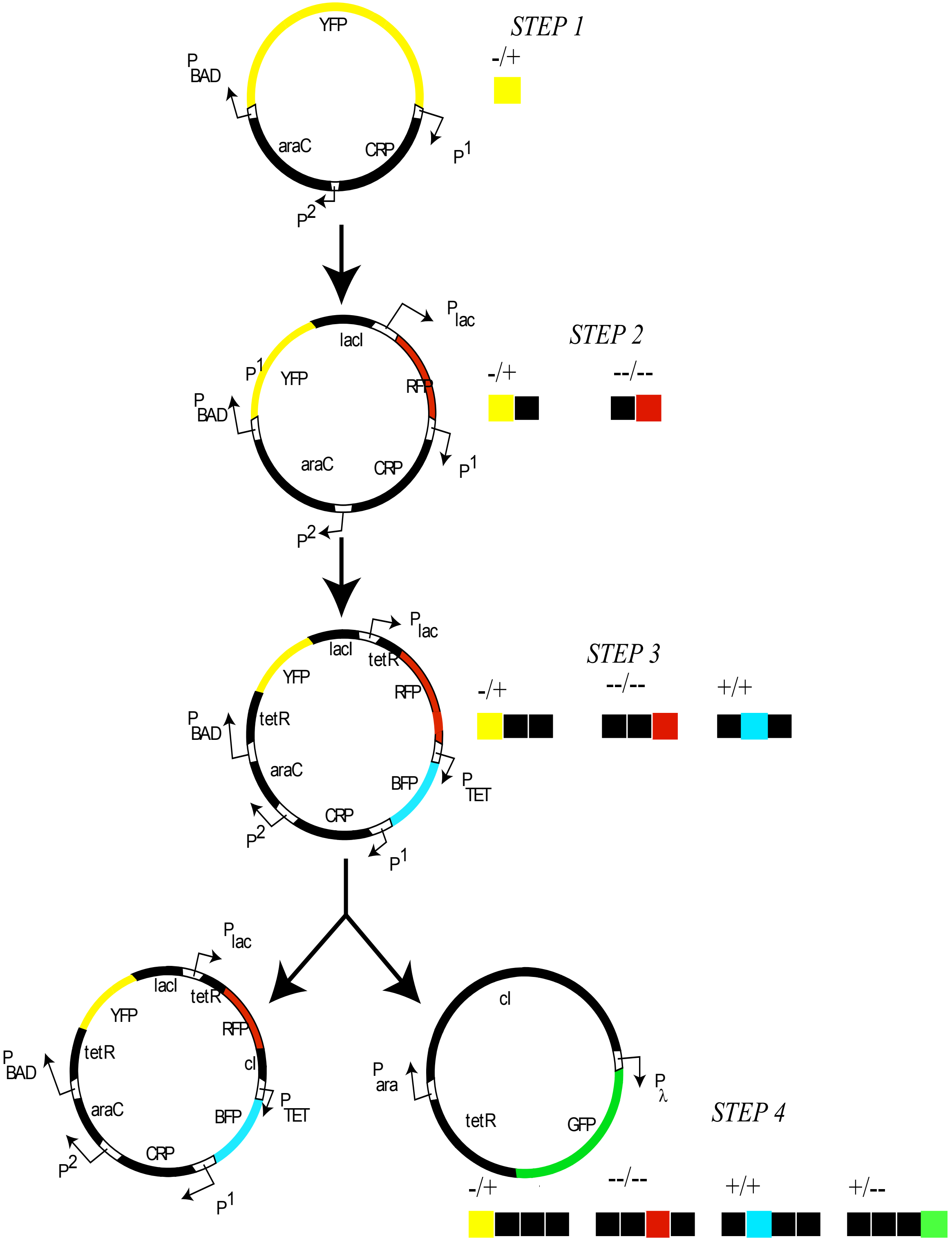


Figure 5.4: The circuit configurations predicted by OptCircuit The final step (step 4) is shown as a combination of two architectures.

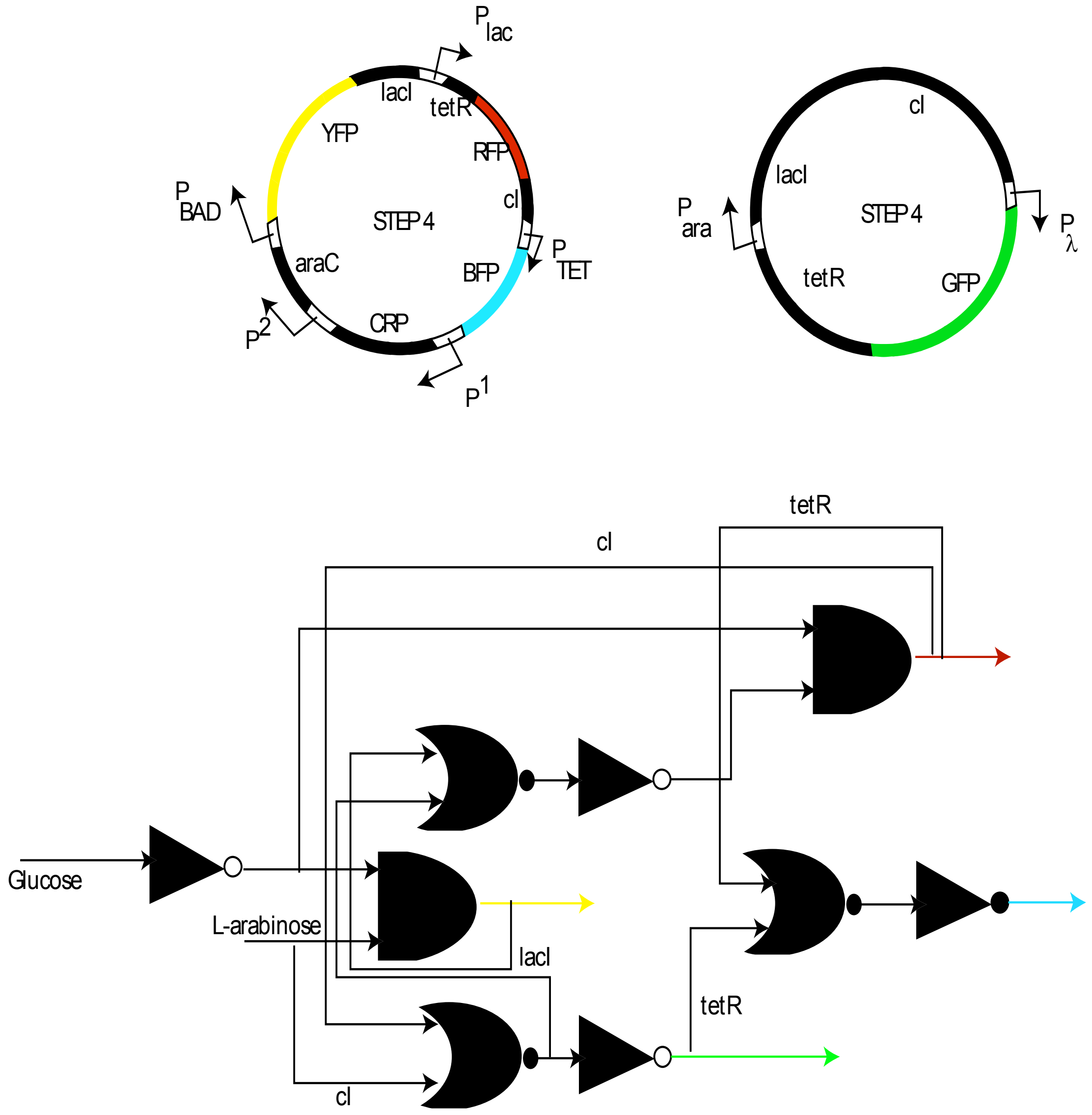


Figure 5.5: Alternative circuit design predicted by the OptCircuit framework. The corresponding binary logic diagram is also provided for comparison. The colored arrows indicate the expression of corresponding colored florescent proteins. The logic gates are represented using established conventions. The gates include, AND, the triangles represent NOT gates or inverters and the crescent shaped gates are the OR gates.

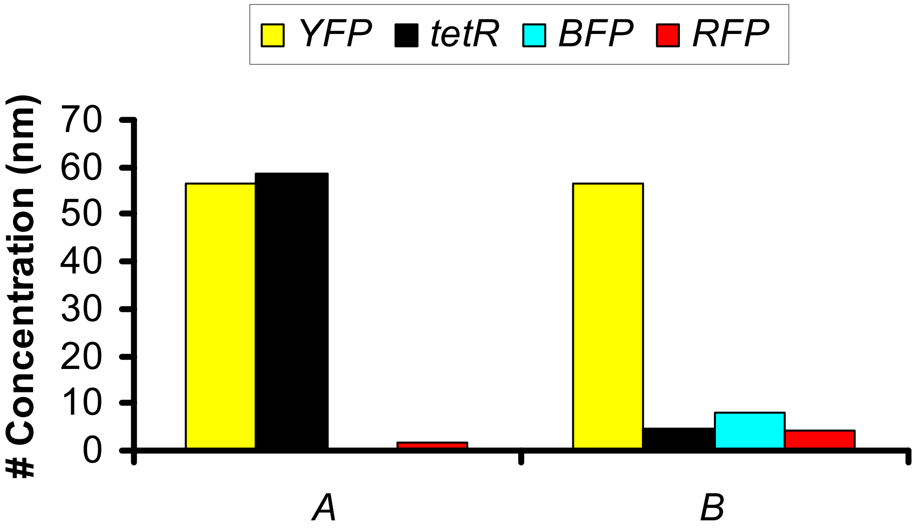


Figure 5.6: The steady-state levels of proteins for the circuits shown in Figure 5.4(step 4) and Figure 5.5. Circuit (A), represents the circuit shown in final step of Figure 5.4. (B) represents the circuit in Figure 5.5.

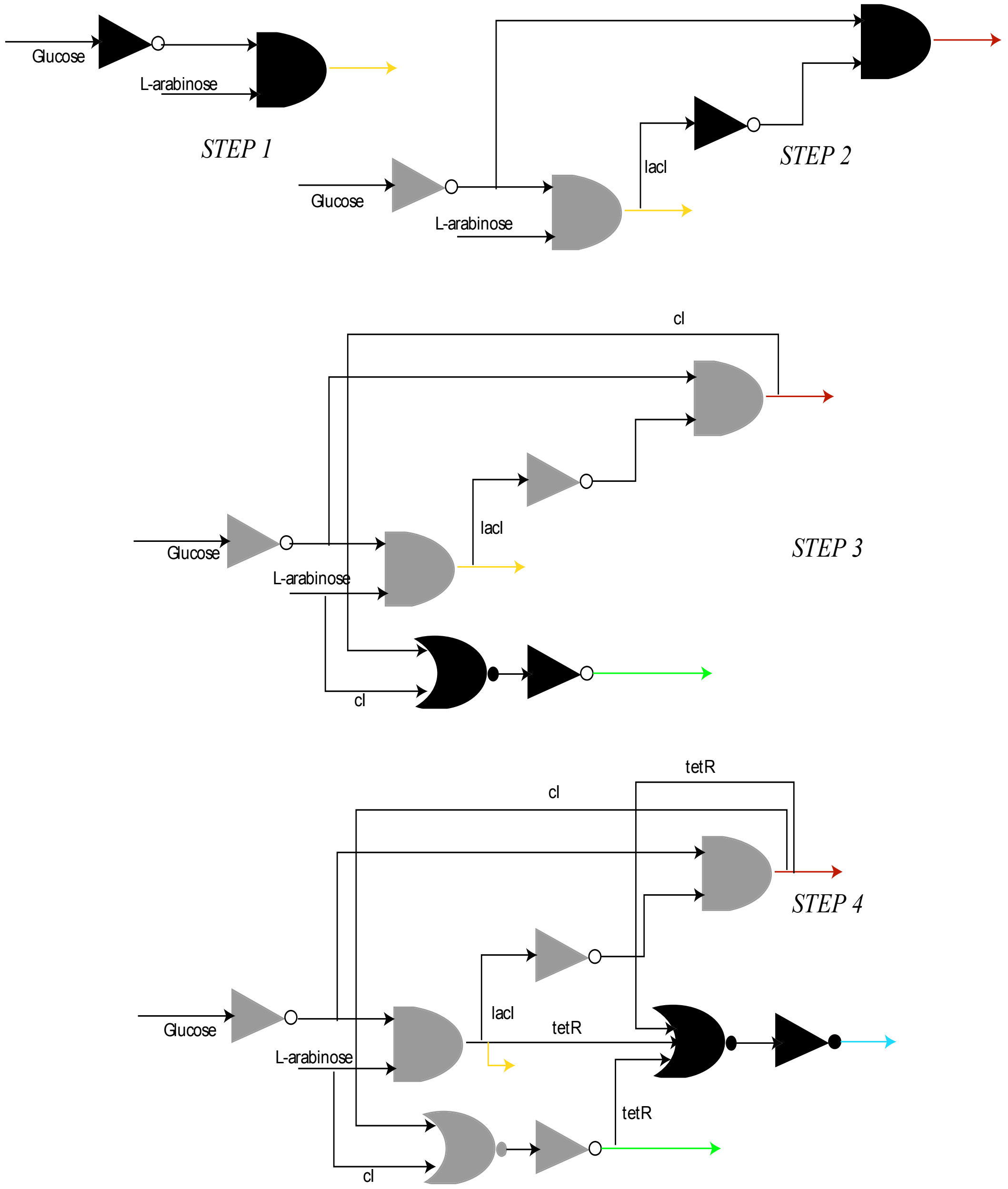


Figure 5.7: Binary logic representation of the circuits obtained at in Figure 5.5. The additional elements appended at each step are indicated using black colored logic gates. Observe that at each step, the OptCircuit framework appends an OR gate followed by a NOT gate to generate the required response. The OR gate serves to integrate the indirect effect of the expression under corresponding inducer conditions.

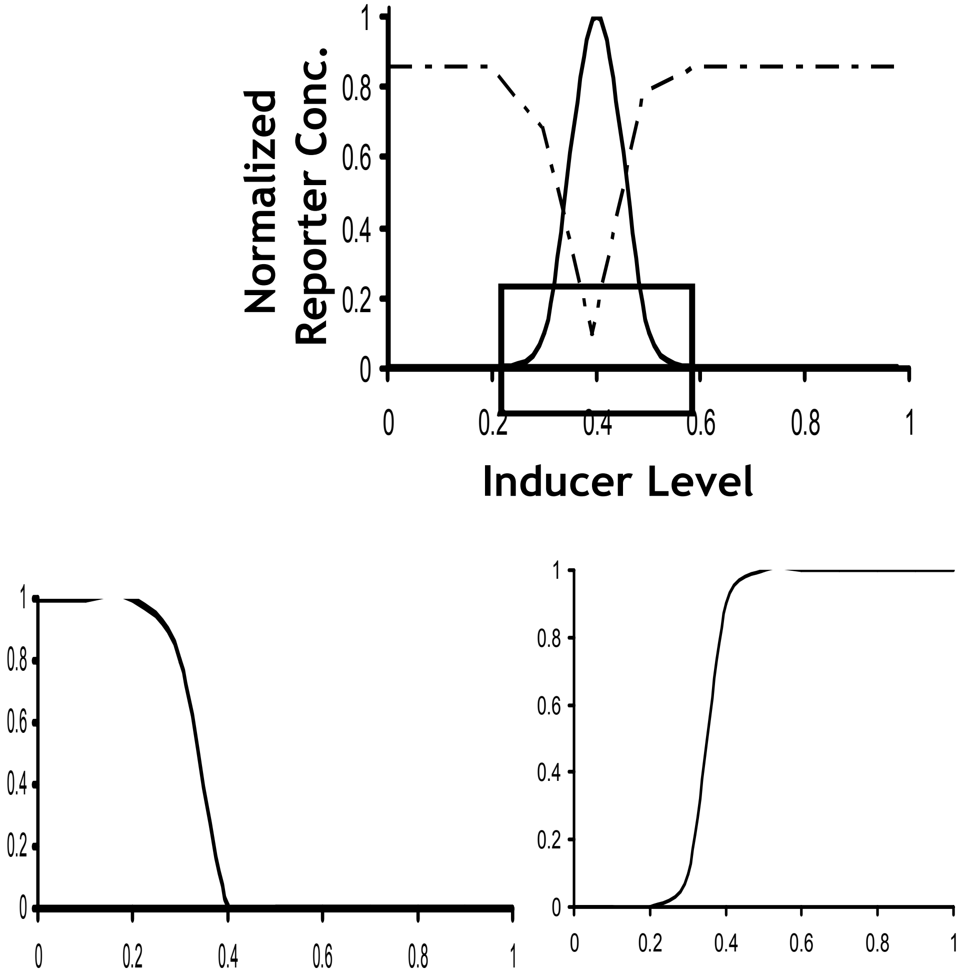


Figure 5.8: The desired response from the concentration band detector example. The solid line represents the desired response of the inducer molecule (*GFP)*. The dotted line represents the desired response of the protein *tetR*. In (B), the desired response of protein *tetR* in the LTD circuit is shown, and in (C), the corresponding response form the HTD circuit is illustrated.

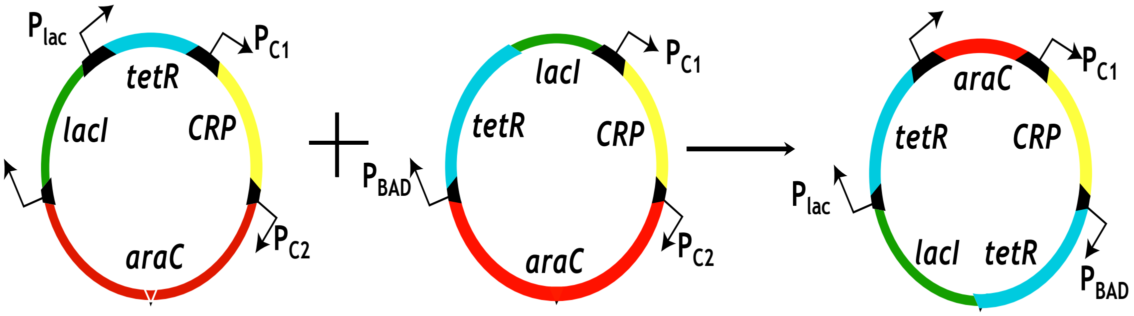


Figure 5.9: The best circuit configurations proposed by OptCircuit. In (A), the LTD circuit is shown; In (B), HTD circuit is illustrated and finally the combination of both the circuit designs is shown in (C)

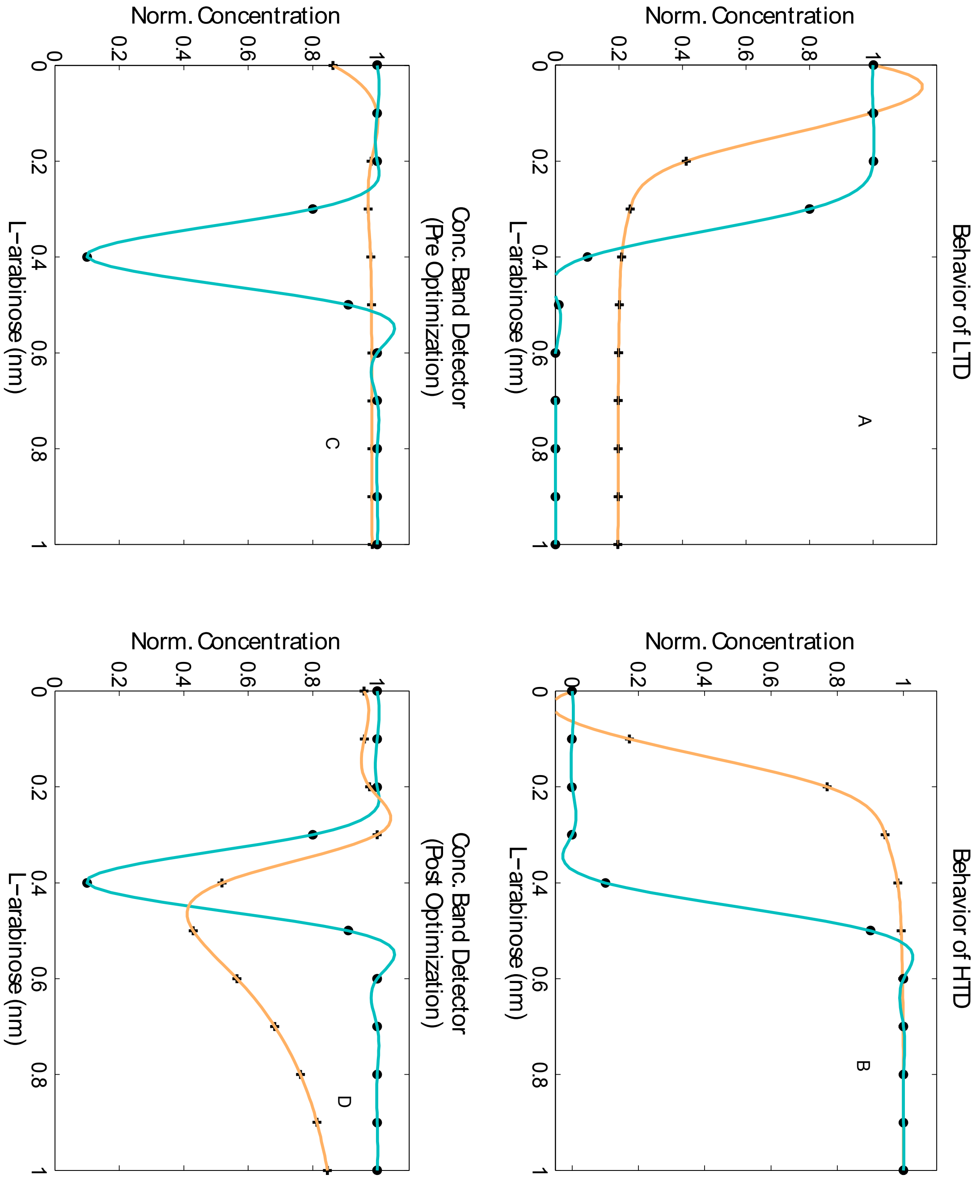


Figure 5.10: In (A), the response of the LTD circuit is shown. In (B), the response obtained from HTD is illustrated. In (C), the response obtained prior to optimization of kinetic parameters is shown. Note that this response deviates significantly from the desired response. In (D), the response obtained after optimizing kinetic parameters is shown.

CHAPTER 6
Synopsis

6.1 Summary

Recent advances in generation of high-throughput biological data have engendered the need to develop efficient computational platforms to enable the integration of huge tracts of diverse biological data and make predictions regarding the behavior of biological system. In response to this challenge, in this work we developed a suite of computational frameworks which can serve as powerful tools to aid various stages of biological research and discovery.

In Chapter 2, an optimization based modeling and solution framework that enabled the inference of gene regulatory networks from gene expression data while accounting for time delay was proposed. By analyzing the amount of variance in the data explained by the model, it was shown that the proposed methodology explained more variance in real data as compared to randomized data implying it is accurately able to capture the underlying network structure. Subsequently, the performance of the proposed framework was benchmarked by applying it to both *in numero* and real microarray datasets. Numerous regulatory relationships with time delay were uncovered suggesting that time delay is ubiquitous in gene regulation. We demonstrated that by neglecting key system properties such as time-delay results in a significant increase in the number of parameters that are required to explain the system dynamics. Finally, the magnitude of recovered weights is found to be low indicating that the predicted networks are less sensitive to random fluctuations in gene expression.

Next, a computational framework was developed to analyze the topological properties (Min-Input) of cell signaling networks. Computational results for large-scale cell signaling networks governing the progression of prostate cancer were presented. It was shown that the Min-Input framework uncovers a number of interesting organizational principles of signaling networks. Specifically, by identifying all cellular stimuli that can elicit the formation of a particular response we demonstrated that outputs of signaling networks can be classified into two distinct sets, highly degenerate or highly specific depending on whether they can be elicited by many different input combinations or a few dedicated ones. The presence of alternative strategies to realize an outcome can be rationalized as an evolutionary adaptation to protect against failure, thus improving response robustness. These classifications have important ramifications for guiding the development of therapeutic strategies. For example, blocking the formation of a highly degenerate outcome (for e.g. *cyclinD*) is hard to accomplish because it requires the disruption of multiple steps. Overall the results for the Min-Input problem indicate that cellular outputs can be stimulated by several different signaling molecules hinting at the enormous complexity associated with disrupting signal transduction. The Min-Interference framework was then introduced to address this challenge. Given a set of input signaling molecules, the Min-Interference framework suggests targeted disruptions in these architectures to negate an undesirable cellular outcome while preserving the desirable ones. Computational results indicated that the framework was able to suggest a host of disruption strategies to negate an outcome. While these include intuitive strategies that involved the disruption of terminal transformations, it was shown that the framework is able to suggest non-intuitive strategies involving upstream transformations whose disruption is propagated downstream to accomplish the required objective. Further, we also underline the biologically relevance of the identified strategies as several drug molecules exist to carry out the proposed disruptions. Furthermore, by proactively preserving desirable outputs disruption strategies were identified that were less likely to involve side-effects by contrasting them against the action and reported side-effects of existing drug molecules. This computational base hence provides a novel and versatile tool to guide the development of systematic design and analysis of therapeutic interventions.

In Chapter 4, we turn our attention to the development of mechanistic simulation platforms to test and validate alternative regulatory hypothesis. Specifically the DEMSIM simulation platform was introduced to support this endeavor. The key feature of the DEMSIM platform was the event-based modeling of the basic processes of transcription, translation, species decay and integration of the fundamental processes with system-specific regulatory circuitry. The stochasticity inherent to gene expression and regulation was captured using a Monte Carlo based sampling algorithm. Subsequently, the abilities of the proposed frameworks were tested by performing several computational studies. The *lac* promoter study highlighted that the parameters embedded in the framework can indeed be tuned to reproduce experimental measurements. The results for the SOS response system of *E. coli* show that the platform can be used to make predictions regarding the behavior of the biological systems under novel conditions. Finally, in the *araBAD* study we demonstrate the sensitivity of the framework to discriminate between alternative regulatory hypotheses that were postulated to explain an experimental observation.

In Chapter 4, we have discussed the development of simulation platforms for investigating the behavior of naturally occurring systems. In Chapter 5, we take the next step and raise the question, "How can we integrate well studied biological components (promoters, transcripts etc.) to construct functional biological circuits that exhibit a targeted response?" It was recognized that the task of constructing circuits that exhibit a particular response is a challenging one and matching of kinetic rates of individual elements of the circuit is essential to ensure its performance. To address this challenge, we introduced the OptCircuit framework, an optimization based framework to design and optimize synthetic biological circuits. Multiple literature sources were used to populate a set of kinetic descriptions of promoter-protein, protein-protein and protein-inducer pairs. The dynamics that govern the interactions between the elements of the genetic circuit were simulated using deterministic rate equations in the current implementation but the framework is general enough to accommodate stochastic simulations. The desired circuit response is abstracted as the maximization/minimization of an appropriately constructed objective function. Subsequently, an iterative procedure was implemented within our framework to identify an ensemble of circuits that exhibit the desired response. Computational results for the well-studied genetic toggle switch example highlighted the ability of the framework to suggest circuit designs of varying complexity affording more opportunities for kinetic parameter tuning. In the genetic decoder example, it was shown that by accurately accounting for the underlying mechanistic description, the framework can suggest circuit designs that may or may not be compatible with boolean logic based description. Finally, in the concentration band detector study, we demonstrate that the framework can rectify or redesign an existing (non-functional) biological circuit and restore functionality by modifying an existing component (e.g., through changes in kinetic parameters). Overall, our results suggest that the OptCircuit platform can serve as an efficient tool to aid the design and fine-tuning of integrated biological systems.

6.2 Future Perspectives

In silico approaches are vital as we proactively attempt to tame the astounding complexity associated with biological systems. In this work we introduced several modeling and solution frameworks that address different levels of biological complexity and all our approaches yield testable predictions regarding the behavior of biological systems. While promising, these tools have to be augmented with several capabilities to further improve their performance and predictive power.

A promising direction of research appears to lie within the realm of synthetic biology. Multiple modeling studies have demonstrated the strong influence of noise and stochastic events on circuit performance [196, 219]. This motivates the need to design circuits that are inherently robust to noise and leakiness of specific components. Key features that confer robustness are redundancy, modularity and the ability to decouple perturbations [220]. While most of the current literature regarding biological robustness has focused on elucidating the architectural and mechanistic features of a network, much less effort has been devoted to developing quantitative and qualitative criteria for quantifying robustness. Efforts in this direction include the work of Cherry and Adler [221] who have proposed that large separation between steady states is likely render the biological switch immune to stochastic fluctuations. Subsequently, H. El-Farra et al. [222] employed these performance measures to develop optimization problems to identify parameters that confer robustness. We believe that OptCircuit can be extended to incorporate these principles. For example, performance measures such as separation between steady-states could be imposed as appropriately formulated objective functions to systematically synthesize circuits that are likely to be robust to stochastic fluctuations. Similarly, other qualitative metrics such as redundancy can be incorporated by enforcing alternative ways of realizing an outcome.

The accuracy of the *in silico* approaches relies on the completeness of the underlying biological models. This is particularly true in case of network analysis based approaches where the network of interactions provides an abstraction of the biological system. A prominent enhancement is to improve the accuracy of the qualitative models of interactions that serve as inputs to network analysis based frameworks. It is widely acknowledged that most of the genome-scale networks are inherently incomplete with several functionalities missing and hence there is a need to develop approaches to reconcile these inconsistencies. The first task entails the identification of these “gaps” in network reconstructions obtained using automated model generation tools such as SimPheny® and Pathway tools [223]. For example, in case of metabolic networks, these gaps manifest as metabolites that are neither produced or consumed and hence are disconnected from the rest of the network. Efforts are already underway to identify such pathologies in metabolic networks and these approaches could be potentially extended to signal transduction pathways. The exercise of bridging gaps involves generating a list of candidate genes that could potentially support the missing functionality and some approaches have already been proposed to this end [224-226]. While promising, the efficacy of these current approaches may be improved by reconciling the predicted models with experimental data regarding gene essentiality, growth/no growth under different uptake conditions. Inconsistencies with experiments will then offer avenues for model refinement. Obviously, stoichiometric models of interactions represent only an approximation of the biological system and conclusions drawn based on these models maybe specific to the chosen problem setting and the underlying assumptions made. To address this concern a natural direction of research would be to extend our current network analysis models to investigate kinetic models of biological interactions to make more realistic predictions regarding their behavior. For example, the Min-Input and the Min-Interference frameworks can be readily extended to account for signaling network dynamics. However, these modifications enhance the complexity of the problem significantly on several fronts. First, the computational complexity is elevated and the resulting problems fall under the class of mixed integer dynamic optimization problems (MIDO). Hence we need to work towards the development of efficient algorithms and some promising directions include the solution frameworks proposed by Chachuat and co-workers [215]. Second and the most important issue arises with the paucity of the available kinetic information which limits the completeness of the current kinetic models. However, in recent years kinetic information of signaling networks is being revealed (the Database of Quantitative Cellular Signalling (DOQCS) provides a repository of modules of signaling pathways containing approximately one-third of all published kinetic models of signaling pathways [69]) and these promising developments will only hasten the shift to kinetic models in the future.

Finally, it is important to underling the significance of carefully designed experimental protocols to completely extract the potential of intelligent *in silico* approaches in biology. The predictive power of these approaches can be enhanced by carrying out successive rounds of model refinements by comparing experimental observations to model predictions. This iterative exercise promises to result in models that can potentially make more realistic predictions in wake of biological complexity.

Bibliography

1. Chu, S., et al., *The transcriptional program of sporulation in budding yeast.* Science, 1998. **282**: p. 699.
2. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Molecular Biology of the Cell, 1998. **9**: p. 3273.
3. Wen, X., et al., *Large-scale temporal gene expression mapping of central nervous system development.* PNAS, 1998. **95**: p. 334.
4. Cohen, B.A., et al., *Discrimination between paralogs using microarray analysis: application to the Yap1p and Yap2p transcriptional networks.* Molecular Biology of the Cell, 2002. **13**: p. 1608.
5. Lee, T.I., N.J. Rinaldi, and F. Robert, *Transcriptional regulatory network in Saccharomyces cerevisiae.* Science, 2002. **298**: p. 799.
6. Gardner, T.S., et al., *Inferring genetic networks and identifying compound mode of action via expression profiling.* Science, 2003. **301**(5629): p. 102-5.
7. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**: p. 680.
8. Ideker, T., et al., *Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.* Science, 2001. **292**: p. 929.
9. Herrgard, M.J., M.W. Covert, and B.O. Palsson, *Reconciling gene expression data with known genome-scale regulatory network structures.* Genome Research, 2003. **13**: p. 2423-2434.
10. Oh, M., et al., *Global expression profiling of acetate-grown Escherichia coli.* The Journal of Biological Chemistry, 2002. **277**: p. 13175-13183.
11. Stephanopoulos, G., et al., *Mapping physiological states from microarray expression measurements.* Bioinformatics, 2002. **18**: p. 1054-1063.
12. Misra, J., et al., *Interactive exploration of microarray gene expression patterns in a reduced dimensional space.* Genome Research, 2002. **12**: p. 1112-1120.
13. Gill, R.T., et al., *Genome-Wide Dynamic Transcriptional Profiling of the Light-to-Dark Transition in Synechocystis Sp. Strain PCC 6803.* J Bacteriol, 2002. **184**: p. 3671-3681.
14. Demeter, J., et al., *The Stanford Microarray Database: implementation of new analysis tools and open source release of software*

10.1093/nar/gkl1019. Nucl. Acids Res., 2007. **35**(suppl_1): p. D766-770.

15. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG*

10.1093/nar/gkj102. Nucl. Acids Res., 2006. **34**(suppl_1): p. D354-357.

16. Faith, J.J., et al., *Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles.* PLoS Biology, 2007. **5**(1): p. e8.
17. Salgado, H., et al., *RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions 10.1093/nar/gkj156.* Nucl. Acids Res., 2006. **34**(suppl_1): p. D394-397.
18. Krull, M., et al., *TRANSPATH: an integrated database on signal transduction and a tool for array analysis.* Nucleic Acids Res, 2003. **31**(1): p. 97-100.
19. Schacherer, F., et al., *The TRANSPATH signal transduction database: a knowledge base on signal transduction networks.* Bioinformatics, 2001. **17**(11): p. 1053-7.
20. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic Acids Res, 2003. **31**(1): p. 374-8.
21. Gilman, A.G., et al., *Overview of the Alliance for Cellular Signaling.* Nature, 2002. **420**(6916): p. 703-6.
22. Li, J., et al., *The Molecule Pages database.* Nature, 2002. **420**(6916): p. 716-7.
23. Albert, R., *Scale-free networks in cell biology 10.1242/jcs.02714.* J Cell Sci, 2005. **118**(21): p. 4947-4957.
24. Milo, R., et al., *Network Motifs: Simple Building Blocks of Complex Networks 10.1126/science.298.5594.824.* Science, 2002. **298**(5594): p. 824-827.
25. Burgard, A., P. Pharkya, and C. Maranas, *Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnology and Bioengineering, 2003. **84**(6): p. 647-657.
26. Schleif, R., *AraC protein: a love-hate relationship.* BioEssays, 2003. **25**: p. 274-282.
27. Grossman, A., *Genetic networks controlling the initiation of sporulation and the development of genetic competence in Bacillus subtilis.* Annu Rev Genet, 1995. **29**: p. 477-508.
28. von Dassow, G., et al., *The segment polarity network is a robust developmental module.* 2000. **406**(6792): p. 188-192.
29. Walker, G., *The SOS response of Escherichia coli.* Escherichia coli, 1996: p. 1400–1416.
30. Tuttle, L.M., et al., *Model-driven designs of an oscillating gene network.* Biophys J, 2005. **89**(6): p. 3873-83.
31. Basu, S. and R. Weiss. *The Device physics of Cellular Logic Gates*. in *First Workshop on Non-Silicon Computation*. 2002.
32. Tyson, J.J., K.C. Chen, and B. Novak, *Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell.* Curr Opin Cell Biol, 2003. **15**(2): p. 221-31.
33. Hasty, J., et al., *Designer gene networks: Towards fundamental cellular control.* Chaos, 2001. **11**(1): p. 207-220.
34. Gilman, A. and A.P. Arkin, *Genetic "code": representations and dynamical models of genetic components and networks.* Annu Rev Genomics Hum Genet, 2002. **3**: p. 341-69.

35. Glass, L., et al., *Chaotic Dynamics in an Electronic Model of a Genetic Network.* Journal of Statistical Physics, 2005. **121**(516): p. 969-994.

36. Bolouri, H. and J.M. Bower, *Computational Modeling of Genetic and Biochemical Networks*, ed. J.M. Bower. 2001, Cambridge,Massachusetts: The MIT Press.

37. Bolouri, H. and E.H. Davidson, *Modeling transcriptional regulatory networks.* BioEssays, 2002. **24**(12): p. 1118-1127.

38. Spellman, *Comphrensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microrray Hybidization.* Mol. Biol. Cell, 1998. **9**: p. 3273-3297.

39. D.Hwang, et al., *Determination of minimum sample size and discriminatory expression patterns in microarray data.* Bioinformatics, 2002. **18**: p. 1184-1193.

40. Stephanopoulous, G., et al., *Mapping physiological states from microarray expression measurements.* Bioinformatics, 2002. **18**: p. 1054-1063.

41. Akutsu, T. and S. Miyano, *Algorithms for Inferring Qualitative Models of Biological Networks.* Pac. Symp. Biocomput., 2000. **5**: p. 290-301.

42. Ideker, T.E., V. Thorsson, and R.M. Karp, *Discovery of Regulatory Interactions through Pertubations:Inference and Experimental Design.* Pac. Symp. Biocomput., 2000. **5**: p. 302-313.

43. D.Jong, H., *Modleing and Simulation of genetic regulatory systems:a literature review.* Journal of Computational Biology, 2002. **9**(1): p. 67-103.

44. Chen, T., H.G.L. He, and G.M. Church, *Modeling Gene Expression with Differential Equations.* Pac. Symp. Biocomput., 1999. **4**: p. 102-111.

45. Yeung, M.K.S., J. Tegner, and J.J. Collins, *Reverse Engineering gene networks using singular value decomposition and robust regression.* PNAS, 2002. **99**: p. 6163-6168.

46. Hoon, M.J.L., et al., *Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations.* Pac. Symp. Biocomput., 2003. **8**: p. 17-28.

47. Friedman, N., et al., *Using Bayesian Networks to Analyze Expression Data.* Journal of Computational Biology, 2000. **7**: p. 601-620.

48. Vohradsky, J., *Neural Model of the Genetic Network.* The Journal of Biological Chemistry, 2001. **276**: p. 36168-36173.

49. Jagle, U., et al., *Role of Positive and Negative Cis-regulatory Elements in the Transciptional Activation of the Lysozyme Locus in Developing Macrophages of Transgenic Mice.* The Journal of Biological Chemistry, 1997. **272**: p. 5871-5879.

50. Gill, R.T., et al., *Genome -Wide Transciptional Profiling of the Light-to-Dark Transition in Synechocystis sp . Strain PCC 6803.* Journal of Bacteriology, 2002. **184**(13): p. 3671-3681.

51. Nitzan Rosenfeld, U.A., *Response Delays and the Structure of Transcription Networks.* J.Mol.Biol, 2003. **329**(645-654).

52. Yildirim, N. and M.C. Mackey, *Feedback Regulation in the Lactose Operon: A Mathematical Modeling Study and Comparison with Experimental Data.* Biophysical Journal, 2003. **84**: p. 2841-2851.

53. Wong, P., S. Galdney, and J.D. Keasling, *Mathematical model of the lac operon: inducer exclusion,catabolite repression, and diauxic growth on glucose and lactose.* Biotechnology Progress, 1997. **13**: p. 132-143.
54. Quin, J., et al., *Beyond Synexpression Relationships:Local Clustering of Time Shifted and Inverted Gene Expression Profiles Identities New, Biologically Relavant Interactions.* J. Mol. Biol., 2001. **314**: p. 1053-1066.
55. D'haeseleer, P., L. Shoudan, and R. Somogyi, *Linear Modeling of mRNA expression Levels During CNS Development and Injury.* Pac. Symp. Biocomput., 1999. **4**: p. 41.
56. Weaver, D.C., C.T. Workman, and G.D. Stormo, *Modeling Regulatory Network with Weight Matrices.* Pac. Symp. Biocomput., 1999. **4**: p. 112-123.
57. Someren, E.P.V., L.F.A. Wessels, and M.J.T. Reinders, *Linear Modeling of Genetic Networks from experimental Data.* 2000.
58. Someren, E.P.V., et al., *Robust Genetic Network Modeling By Adding Noisy Data.* Proceedings of the 2001 IEEE - EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP01), Baltimore, Maryland, June 2001., 2001.
59. Someren, E.P.V., L.F.A. Wessels, and M.J.T. Reinders, *Multi-criterion optimization for genetic network modeling.* Signal Processing, 2003. **83**: p. 763-775.
60. Winston, W.L. and M. Venkataraman, *Introduction To Mathematical Programming*. 4 ed. Vol. 1. 2003, Pacific Grove: Brooks/Cole-Thomson Learning.
61. Brooke, A., et al., *GAMS-The Solver Manuals,GAMS Development Corporation: Washigton, DC.* 1998.
62. Brooke, A., et al., *GAMS: A user's guide. GAMS Development Corp*. 2002, Washington D.C.
63. S.Jin, M.D. Jesus-Berrios, and A.L.Sonenshein, *A Bacillus Subtilis malate dehydrogenase gene.* J Bacteriol, 1996. **178**(2): p. 560-3.
64. P.Miller, et al., *Transcriptional regulation of a promoter in the men gene cluster of Bacillus subtilis.* J Bacteriol, 1988. **170**(6): p. 2742-8.
65. Ross, S.M., *Introdution to Probability and Statistics for Engineers and Scientists*. 2 ed. 2000: Harcourt Academic Press.
66. Takahashi, K., S.N.V. Arjunan, and M. Tomita, *Space in systems biology of signaling pathways-towards intracellular molecular crowding in silico.* FEBS letters, 2005. **579**: p. 1783-1788.
67. Schlessinger, J., *Common and Distinct Elements in Cellular Signaling via EGF and FGF Receptors.* Science, 2004. **306**: p. 1506-1507.
68. Cross, M.J., et al., *VEGF-receptor signal transduction.* TRENDS in Biochemical Sciences, 2003. **28**(9): p. 488-494.
69. Sivakumaran, S., et al., *The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks.* Bioinformatics, 2003. **19**(3): p. 408-415.

70. Pappin, J.A. and B.O. Palsson, *Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk.* Journal of theoretical biology, 2004. **227**: p. 283-297.

71. Pappin, J.A., et al., *Reconstruction of cellular signalling networks and analysis of their properties.* Nature Reviews Molecular Cell Biology, 2005.

72. Aksenov, S.V., et al., *An integrated approach for inference and mechanistic modeling for advancing drug development.* FEBS letters, 2005. **579**: p. 1878-1883.

73. Apic, G., et al., *Illuminating drug discovery with biological pathways.* FEBS letters, 2005. **579**: p. 1872-1877.

74. Bader, G.D., D. Betel, and C.W. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res, 2003. **31**(1): p. 248-50.

75. Bader, G.D., et al., *BIND--The Biomolecular Interaction Network Database.* Nucleic Acids Res, 2001. **29**(1): p. 242-5.

76. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic Acids Res, 2002. **30**(1): p. 303-5.

77. Mi, H., et al., *The PANTHER database of protein families,subfamilies,functions and pathways.* Nucleic Acids Research, 2005. **33**(Database issue): p. D284-D288.

78. Scita, G., et al., *Signaling from Ras to Rac and beyond: not just a matter of GEF's.* The EMBO Journal, 2000. **19**(11): p. 2393-2398.

79. Weng, G., U.S. Bhalla, and R. Iyengar, *Complexity in biological signaling systems.* Science, 1999. **284**(5411): p. 92-6.

80. Bhalla, U.S. and R. Iyengar, *Emergent properties of networks of biological signaling pathways.* Science, 1999. **283**(5400): p. 381-7.

81. Jordon, J.D., E.M. Landau, and R. Iyengar, *Signaling Networks: The Origins of Cellular Multitasking.* Cell, 2000. **103**: p. 193-200.

82. Neves, S.R. and R. Iyengar, *Modeling of signaling networks.* BioEssays, 2002. **24**: p. 1110-1117.

83. Haugh, J.M., *A unified model for signal transduction reactions in cellular membranes.* Biophys J, 2002. **82**(2): p. 591-604.

84. Shvartsman, S.Y., et al., *Spatial range of autocrine signaling: modeling and computational analysis.* Biophys J, 2001. **81**(4): p. 1854-67.

85. Heinrich, R., B.G. Neel, and T.A. Rapoport, *Mathematical models of protein kinase signal transduction.* Mol Cell, 2002. **9**(5): p. 957-70.

86. Femenia, F.J. and G. Stephanopoulos, *Activation Ratios for Reconstruction of Signal Transduction Networks.* Mol Eng. Biol. Chem. Systems (MEBCS), 2003. **1-9**.

87. Hoffmann, A., et al., *The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation.* Science, 2002. **298**(5596): p. 1241-5.

88. Bhalla, U.S., *The chemical organization of signaling interactions.* Bioinformatics, 2002. **18**(6): p. 855-63.

89. Kremling, A., et al., *The organization of metabolic reaction networks: a signal-oriented approach to cellular models.* Metab Eng, 2000. **2**(3): p. 190-200.

90. Kremling, A. and E.D. Gilles, *The organization of metabolic reaction networks. II. Signal processing in hierarchical structured functional units.* Metab Eng, 2001. **3**(2): p. 138-50.

91. Schoeberl, B., et al., *Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors.* Nat Biotechnol, 2002. **20**(4): p. 370-5.

92. von Dassow, G., et al., *The segment polarity network is a robust developmental module.* Nature, 2000. **406**(6792): p. 188-92.

93. Von Dassow, G. and G.M. Odell, *Design and constraints of the Drosophila segment polarity module: robust spatial patterning emerges from intertwined cell state switches.* J Exp Zool, 2002. **294**(3): p. 179-215.

94. Ingolia, N.T., *Topology and robustness in the Drosophila segment polarity network.* PLoS Biol, 2004. **2**(6): p. E123.

95. Albert, R. and H.G. Othmer, *The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster.* J Theor Biol, 2003. **223**(1): p. 1-18.

96. Yi, T.M., et al., *Robust perfect adaptation in bacterial chemotaxis through integral feedback control.* Proc Natl Acad Sci U S A, 2000. **97**(9): p. 4649-53.

97. Alon, U., et al., *Robustness in bacterial chemotaxis.* Nature, 1999. **397**(6715): p. 168-71.

98. Kholodenko, B.N., et al., *Quantification of information transfer via cellular signal transduction pathways.* FEBS Lett, 1997. **414**(2): p. 430-4.

99. Ptashne, M. and A. Gann, *Signal transduction. Imposing specificity on kinases.* Science, 2003. **299**(5609): p. 1025-7.

100. Park, S.H., A. Zarrinpar, and W.A. Lim, *Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms.* Science, 2003. **299**(5609): p. 1061-4.

101. Stelling, J., et al., *Robustness of cellular functions.* Cell, 2004. **118**: p. 675-685.

102. Oikawa, T., *ETS transcription factors: Possible targets for cancer therapy.* Cancer Science, 2004. **95**(8): p. 626-633.

103. McCarty, M.F., *Targeting Multiple Signaling Pathways as a Strategy for Managing Prostate Cancer: Multifocal Signal Modulation Therapy.* Integrative Cancer Theraoies, 2004. **3**(4): p. 349-380.

104. Hucka, M., et al., *The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Netwrok Models.* Bioinformatics, 2003. **19**(4): p. 524-531.

105. Brown, M., *Perl programmers's reference*. 1999, Berkley,CA: Osborne/McGraw-Hill.

106. Price, N.D., J.L. Reed, and B.O.Palsson, *Genome-scale Models of Microbial Cells: Evaluating the consequences of constraints.* Nature Reviews Microbiology, 2004. **2**: p. 886-897.

107. Schilling, C.H., D. Letscher, and B.O. Palsson, *Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective.* J Theor Biol, 2000. **203**(3): p. 229-48.

108. Schuster, S. and C. Hilgetag, *On elementary flux modes in biochemical reaction systems at steady state.* J Biol Syst, 1994. **2**: p. 165-182.

109. Schuster, S., D.A. Fell, and T. Dandekar, *A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks.* Nat Biotechnol, 2000. **18**(3): p. 326-32.

110. Klamt, S. and J. Stelling, *Combinatorial complexity of pathway analysis in metabolic networks.* Mol Biol Rep, 2002. **29**(1-2): p. 233-6.

111. Burgard, A.P., et al., *Flux coupling analysis of genome-scale metabolic network reconstructions.* Genome Res, 2004. **14**(2): p. 301-12.

112. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng, 2003. **84**(6): p. 647-57.

113. Burgard, A.P., S. Vaidyaraman, and C.D. Maranas, *Minimal reaction sets for Escherichia coli metabolism under different growth requirements and uptake environments.* Biotechnol Prog, 2001. **17**(5): p. 791-7.

114. Burgard, A.P. and C.D. Maranas, *Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions.* Biotechnol Bioeng, 2001. **74**(5): p. 364-75.

115. Koretzky, G.A., *The role of Grb2-associated proteins in T-cell activation.* Immunology today, 1997. **18**(8): p. 401-406.

116. Burgard, A.P., P. Pharkhya, and C.D. Maranas, *Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnology Bioengineering, 2003. **74**: p. 364-375.

117. Hartl, M., A.G. Bader, and K. Bister, *Molecular targets of the oncogenic transcription factor jun.* Current Cancer Drug Targets, 2003. **3**(1): p. 41-55.

118. Hommes, D.W., M.P. Peppelenbosch, and S.J.H.v. Deventer, *Mitogen activated protein (MAP) kinase signal transduction pathways and novel anti-inflammatory targets.* Gut, 2003(52): p. 144-151.

119. Gratton, J.P., et al., *Selective inhibition of tumor microvascular permeability by cavtratin blocks tumor progression in mice.* Cancer Cell, 2003. **4**: p. 31-39.

120. Duda, D.G., D. Fukumura, and R.K. Jain, *Role of eNOS in neovascularization: NO for endothelial progenitor cells.* TRENDS in Molecular Medicine, 2004. **10**(4): p. 143-154.

121. Bonventre, J.V., et al., *Reduced fertility adn postischaemic brain injury in mice deficient in cytosolic phopholipase A_2.* Nature, 1997. **390**: p. 622-625.

122. Weis, S., et al., *Endothelial barrier disruption by VEGF-mediated Src activity potentiates tumor cell extravasation and metastasis.* Journal of Cell Biology, 2004. **167**(2): p. 223-229.

123. Richardson, P., T. Hideshima, and K. Anderson, *Thalidomide: Emerging Role in Cancer Medicine.* Annual Reviews in Medicine, 2002. **53**: p. 629-657.

124. Wu, L., et al., *VRAP is an adaptor protein that binds KDR, a receptor for Vascular Enthothelial Cell Growth Factor.* The Journal of Biological Chemistry, 2000. **275**(9): p. 6059-6062.

125. Das, D.K. and N. Maulik, *Physiological role of Heat shock protein 27*, in *Heat shock protein in Myocardinal Protection*. 2000.

126. Goldman, D.A., *Thalidomide use: past history and current implications for practice.* Oncol Nurs Forum, 2001. **28**(3): p. 471-477.

127. Chen, X., Z.L. Li, and T.Z. Chen, *TTD: Therapeutic Target Database.* Nucleic Acids Research, 2002. **30**(1): p. 412-415.

128. Nikolsky, Y., T. Nikolskaya, and A. Bugrim, *Biological networks and analysis of experimental data in drug discovery.* Drug Discovery Today, 2005. **10**(9): p. 653-662.

129. Rhodes, D.R. and A.M. Chinnaiyan, *Integrative analysis of the cancer transcriptome.* Nature Genetics, 2005. **37**: p. S31-S37.

130. Zhou, X.F., X.Q. Shen, and L. Shemshedini, *Ligand-Activated Retinoic Acid receptor inhibits AP-1 Transcactivation by disrupting c-Jun/c-FOS dimerization.* Molecular Endocrinology, 1999. **13**(2): p. 276-285.

131. Brockmann, D., et al., *Repression of the c-Jun trans-Activation function by the adenovirus type 12 E1A 52R protein correlates with the Inhibition of phosphorylation of the c-Jun Activation domain.* Journal of Biological Chemistry, 1995. **270**(18): p. 10754-10763.

132. Siegmund, B. and M. Zeitz, *Therapeutic approaches in inflammatory bowel disease based on the immunopathogenesis.* Annales Academiae Medicae Bialostocenis, 2004. **49**: p. 22-30.

133. Duncia, J.V., J.B.S. 3rd, and R.E. Olson, *"MEK inhibitors: the chemistry and biological actovity of U0126, its analogs , and cyclizatio products.* Bioorg. Med. Chem. Lett, 1998. **8**(20): p. 2839-2844.

134. Park, S., et al., *RKIP downregulates B-Raf kinase activity in melanoma cancer cells.* Oncogene, 2005. **24**: p. 3535-3540.

135. Orning, L., et al., *A cyclic pentapeptide derived from the second EGF-like domain of Factor VII is an inhibitor of tissue factor dependent coagulation and thrombus formation.* Thrombosis Haemostatis, 2002. **87**(1): p. 13-21.

136. Ang, S., et al., *Acid-Induced Gene Expression in Helicobacter pylori: Study in Genomic Scale by Microarray.* Infection and Immunity, 2001. **69**(3): p. 1679-1686.

137. Helmann, J.D., et al., *The global transcriptional response of Bacillus subtilis to peroxide stress is coordinated by three transcription factors.* Journal of Bacteriology, 2003. **185**: p. 243-253.

138. Dasika, M.S., A. Gupta, and C.D. Maranas, *A Mixed Integer Linear Programming(MILP) Framework For Inferring Time Delay in Gene Regulatory Networks.* Pac Symp Biocomput., 2004. **9**: p. 474-485.

139. Agger, T. and J. Nielsen, *Genetically Structured Modeling of Protein Production in Filamentous Fungi.* Biotechnology Bioengineering, 1999. **66**: p. 164-170.

140. Cheng, B., R.L.Fournier, and P.A.Relue, *The Inhibition of Escherichia coli lac Operon Gene Expression by Antigene Oligonucleotides-Mathematical Modeling.* Biotechnology Bioengineering, 1999. **70**(4): p. 467-472.

141. Cheng, B., et al., *An Experimental and Theoretical Study of Escherichia coli lac operon Gene Expression by Antigene Oligonucleotides.* Biotechnology Bioengineering, 2000. **74**(3): p. 220-229.

142. Shea, M.A. and G.K. Ackers, *The O_R Control System of Bacteriophage Lambda A Physical-Chemical Model for Gene Regulation.* Journal of Molecular Biology, 1985. **181**: p. 211-230.

143. Goutsias, J. and S. Kim, *A Nonlinear Discrete Dynamical Model for Transcriptional Regulation: Construction and Properties.* Biophysical Journal, 2004. **86**: p. 1922-1945.

144. Hatzimanikatis, V. and K.H. Lee, *Dynamical Analysis of Gene Networks Requires Both mRNA and Protein Expression Information.* Metabolic Engineering, 1999. **1**: p. 275-281.

145. Kierzek, A.M., J. Zaim, and P. Zielenkiewicz, *The effect of Transcription and Translation Frequencies on the Stochastic Fluctuations in Prokaryotic Gene Expression.* The Journal of Biological Chemistry, 2001. **276**(11): p. 8165-8172.

146. Carrier, T.A. and J.D. Keasling, *Mechanistic Modeling of Prokaryotic mRNA decay.* Journal of Theoretical Biology, 1997. **189**: p. 195-209.

147. Siegele, D.A. and J.C. Hu, *Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations.* Proc. Natl. Acad. Sci, USA, 1997. **94**: p. 8168-8172.

148. Elowitz, M.B. and S. Leibler, *A synthetic oscillatory network of transcriptional regulators.* Nature Genetics, 2000. **403**: p. 335-338.

149. Kepler, T.B. and T.C. Elston, *Stochasticity in Transcriptional Regulation: Origins,Consequences, and Mathematical Representations.* Biophysical Journal, 2001. **81**: p. 3116-3136.

150. McAdams, H.H. and A. Arkin, *Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci, USA*, 1997. **94**: p. 814-819.

151. Arkin, A., J.Ross, and H.H. McAdams, *Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage lambda-Infected Escherichia coli Cells.* Genetics, 1998. **149**: p. 1633-1648.

152. Kurata, H., N. Matoba, and N. Shimizu, *CADLIVE for constructing a large-scale biochemical network based on a simulation-directed notation and its application to yeast cell cycle. Nucleic Acids Research*, 2003. **31**(14): p. 4071-4084.

153. Kastner, K., J. Solomon, and S. Fraser, *Modeling a HOX Gene Network in Silico Using a Stochastic Simulation Algorithm.* Developmental Biology, 2002. **246**: p. 122-131.

154. Gillespie, D.T., *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions.* Journal of Computational Physics, 1976. **22**: p. 403-434.

155. Gillespie, D.T., *Exact Stochastic Simulation of Coupled Chemical Reactions.* The Journal of Physical Chemistry, 1977. **81**(25): p. 2340-2361.

156. Gibson, M.A. and J. Bruck, *Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels.* Journal of Physical Chemistry, 2000. **104**: p. 1876-1889.

157. Carrier, T.A. and J.D. Keasling, *Investigating Autocatalytic Gene Expression Systems through Mechanistic Modeling.* Journal of Theoretical Biology, 1999. **201**: p. 25-36.
158. Alberts, B., et al., *Molecular Biology of THE CELL*. 1994, New York & London: Garland Publishing,Inc.
159. Hardinson, R.C., *Molecular Genetics-Volume II*. Vol. II. 2002: McGraw-Hill Primis Custom Publishing.
160. G.M.Marianne, *Messenger RNA Stability and Its Role in Control of Gene Expression in Bacteria and Phages. Annual Review of Genetics*, 1999. **33**: p. 193-227.
161. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nature Genetics, 2002. **31**: p. 64-68.
162. Tropper, C., *Parallel Discrete-Event Simulation Applications.* Journal of Parallel and Distributed Computing, 2002. **62**(327-335).
163. Vilar, J.M.G., C.C. Guet, and S. Leibler, *Modeling network dynamics:the lac operon , a case study.* The Journal of Cell Biology, 2003. **161**(3): p. 471-476.
164. Kennell, D. and H. Riezman, *Transcription and Translation Initiation Frequencies of the Escherichia coli lac Operon.* Journal of Molecular Biology, 1977. **114**: p. 1-21.
165. Kuzminov, A., *Recombinatorial Repair of DNA Damage in Escherichia coli and Bacteriophage Lambda.* Microbiology and Molecular Biology Reviews, 1999. **63**(4): p. 751-813.
166. Janion, C., *Some aspects of the SOS response system - A critical survey.* Acta Biochimica Polonica, 2001. **48**(3): p. 599-610.
167. Henestrosa, A.R.F.d., et al., *Identification of additional genes belonging to the LexA regulon in Escherichia coli.* Molecular Microbiology, 2000. **35**(6): p. 1560-1572.
168. Sassanfar, M. and J.W. Roberts, *Nature of the SOS-inducing Signal in Escherichia coli The Involvement of DNA Replication.* Journal of Molecular Biology, 1990. **212**: p. 79-96.
169. Rehrauer, W.M., et al., *Interaction of Escherichia coli RecA Protein with the LexA Repressor.* The Journal of Biological Chemistry, 1996. **271**(39): p. 23865-23873.
170. Brent, R. and M. Ptashne, *Mechanism of action of the lexA gene product.* Proc. Natl. Acad. Sci, USA, 1981. **78**(7): p. 4204-4208.
171. Betrand-Burggraf, E., et al., *Promoter Properties and Negative Regulation of the uvrA Gene by the LexA Repressor and its Amino-terminal DNA Binding Domain.* Journal of Molecular Biology, 1987. **193**: p. 293-302.
172. Courcelle, J., et al., *Comparitive Gene Expression Profiles Following Exposure in Wild-Type and SOS-Deficient Escherichia coli.* Genetics, 2001. **158**: p. 41-64.
173. Khil, P.P. and P.D. Camerini-Otero, *Over 1000 genes are involved in the DNA damage response of Escherichia coli.* Molecular Microbiology, 2002. **44**(1): p. 89-105.

174. Schleif, R., *Regulation of the L-arabinose operon of Escherichia coli.* Trends in Genetics, 2000. **16**(12): p. 559-565.

175. Wu, M. and R. Schleif, *Mapping Arm-DNA-binding Domain Interactions in AraC.* Journal of Molecular Biology, 2001. **307**: p. 1001-1009.

176. Seabold, R.R. and R. Schleif, *Apo-Arac Actively Seeks to Loop.* Journal of Molecular Biology, 1998. **278**: p. 529-538.

177. Mangan, S., A. Zaslaver, and U.Alon, *The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks.* Journal of Molecular Biology, 2003. **334**: p. 197-204.

178. Quillardet, P., M.A. Rouffaud, and P. Bouige, *DNA array analysis of gene expression in response to UV radiation in Escherichia coli. Research in* Microbiology, 2003. **154**: p. 559-572.

179. Joseph, Z.B., et al., *Computational discovery of gene modules and regulatory networks.* Nature Biotechnology, 2003. **21**(11): p. 1337-1342.

180. Alon, U., *Biological Networks: The Tinkerer as an Engineer.* Science, 2003. **301**: p. 1866-1867.

181. Kiehl, T.R., R.M. Mattheysses, and M.K. Simmons, *Hybrid simulation of cellular behavior.* Bioinformatics, 2004. **20**(3): p. 316-322.

182. Hardinson, R.C., *Molecular Genetics-Volume I.* Vol. I. 2002: McGraw-Hill Primis Custom Publishing.

183. Bernstein, J.A., et al., *Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays.* Proc. Natl. Acad. Sci,USA, 2002. **99**(15): p. 9697-9702.

184. Sprinzak, D. and M.B. Elowitz, *Reconstruction of genetic circuits.* Nature, 2005. **438**(7067): p. 443-8.

185. Feng, X.J., et al., *Optimizing genetic circuits by global sensitivity analysis.* Biophys J, 2004. **87**(4): p. 2195-202.

186. Hasty, J., D. McMillen, and J.J. Collins, *Engineered gene circuits.* Nature, 2002. **420**(6912): p. 224-30.

187. Endy, D., *Foundations for engineering biology.* Nature, 2005. **438**(7067): p. 449-53.

188. Gardner, T.S., C.R. Cantor, and J.J. Collins, *Construction of a genetic toggle switch in Escherichia coli.* Nature, 2000. **403**(6767): p. 339-42.

189. McMillen, D., et al., *Synchronizing genetic relaxation oscillators by intercell signaling.* Proc Natl Acad Sci U S A, 2002. **99**(2): p. 679-84.

190. Fung, E., et al., *A synthetic gene-metabolic oscillator.* Nature, 2005. **435**(7038): p. 118-22.

191. Atkinson, M.R., et al., *Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli.* Cell, 2003. **113**(5): p. 597-607.

192. Basu, S., et al., *A synthetic multicellular system for programmed pattern formation.* Nature, 2005. **434**(7037): p. 1130-4.

193. Hasty, J., et al., *Noise-based switches and amplifiers for gene expression.* Proc Natl Acad Sci U S A, 2000. **97**(5): p. 2075-80.

194. Judd, E.M., M.T. Laub, and H.H. McAdams, *Toggles and oscillators: new genetic circuit designs.* Bioessays, 2000. **22**(6): p. 507-9.

195. Kobayashi, H., et al., *Programmable cells: interfacing natural and engineered gene networks.* Proc Natl Acad Sci U S A, 2004. **101**(22): p. 8414-9.

196. Hooshangi, S., S. Thiberge, and R. Weiss, *Ultrasensitivity and noise propagation in a synthetic transcriptional cascade.* Proc Natl Acad Sci U S A, 2005. **102**(10): p. 3581-6.

197. Mangan, S., A. Zaslaver, and U. Alon, *The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks.* J Mol Biol, 2003. **334**(2): p. 197-204.

198. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 11980-5.

199. Becskei, A. and L. Serrano, *Engineering stability in gene networks by autoregulation.* Nature, 2000. **405**(6786): p. 590-3.

200. Martin, V.J., et al., *Engineering a mevalonate pathway in Escherichia coli for production of terpenoids.* Nat Biotechnol, 2003. **21**(7): p. 796-802.

201. Anderson, J.C., et al., *Environmentally controlled invasion of cancer cells by engineered bacteria.* J Mol Biol, 2006. **355**(4): p. 619-27.

202. Levskaya, A., et al., *Synthetic biology: engineering Escherichia coli to see light.* Nature, 2005. **438**(7067): p. 441-2.

203. Tian, J., et al., *Accurate multiplex gene synthesis from programmable DNA microchips.* Nature, 2004. **432**(23): p. 1050-1054.

204. Weiss, R., G.E. Homay, and T.F. Knight. *Toward in vivo Digital Circuits*. in *DIMACS Workshop on Evolution as Computation*. 1999.

205. Mason, J., et al., *Evolving complex dynamics in electronic models of genetic networks.* Chaos, 2004. **14**(3): p. 707-15.

206. Guet, C.C., et al., *Combinatorial synthesis of genetic networks.* Science, 2002. **296**(5572): p. 1466-70.

207. Yokobayashi, Y., R. Weiss, and F.H. Arnold, *Directed evolution of a genetic circuit.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 16587-91.

208. Francois, P. and V. Hakim, *Design of genetic networks with specified functions by evolution in silico.* Proc Natl Acad Sci U S A, 2004. **101**(2): p. 580-5.

209. Battogtokh, D., et al., *An ensemble method for idenitfying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 16904-16909.

210. Mohideen, M.J., J.D. Perkins, and E.N. Pistikopoulos, *Towards an Efficient Numerical Procedure for Mixed Integer Optimal Control.* Computers and Chemical Engineering, 1997. **21**(Suppl): p. S457-S462.

211. Sirdeshpande, A.R., M.G. Ierapetritou, and I.P. Androulakis, *Design of Flexible Reduced Kinetic Mechanisms.* Process Systems Engineering, 2001. **47**(11): p. 2461-2473.

212. Tlacuahuac, A.F. and L.T. Beigler, *A Robust and Efficient Mixed-Integer Non-Linear Dynamic Optimization Approach for Simultaneous Design and Control.* 2004.

213. Bansal, V., J.D. Perkins, and E.N. Pistikopoulos, *A Case Study in Simultaneous Design and Control Using Rigorous Mixed-Integer Dynamic Optimization Models.* Ind.Eng.Chem.Res, 2002. **41**: p. 760-778.

214. Bansal, V., et al., *New algorithms for mixed-integer dynamic optimization.* Computers and Chemical Engineering, 2003. **27**: p. 647-688.

215. Chachuat, B., A.B. Singer, and P.I. Barton, *Global Mixed-Integer Dynamic optimization.* AIChe Journal, 2005. **51**(8): p. 2235-2253.

216. Basu, S., D. Karig, and R. Weiss, *Engineering signal processing in cells: Towards molecular concentration band detection.* Natural Computing, 2003. **2**(4): p. 463-478.

217. Keiler, K.C., P.R. Waller, and R.T. Sauer, *Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA.* Science, 1996. **271**(5251): p. 990-3.

218. Lutz, R. and H. Bujard, *Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements*

10.1093/nar/25.6.1203. Nucl. Acids Res., 1997. **25**(6): p. 1203-1210.

219. Hooshangi, S. and R. Weiss, *The effect of negative feedback on noise propagation in transcriptional gene networks.* Chaos, 2006. **16**(2): p. 26108-26108.

220. Kitano, H., *Biological robustness.* Nature Reviews Genetics, 2004. **5**(11): p. 826-837.

221. Cherry, J.L. and F.R. Adler, *How to make a biological switch.* J Theor Biol, 2000. **203**(2): p. 117-33.

222. El-Farra, N.H. *An Optimization-Based Method for the Design of Robust Synthetic Switches in Biological Networks*. in *AICHE*. 2005. Cincinnati.

223. Karp, P.D., S. Paley, and P. Romero, *The Pathway Tools software.* Bioinformatics, 2002. **18 Suppl 1**: p. S225-32.

224. Kharchenko, P., et al., *Identifying metabolic enzymes with multiple types of association evidence.* BMC Bioinformatics, 2006. **7**: p. 177.

225. Kharchenko, P., D. Vitkup, and G.M. Church, *Filling gaps in a metabolic network using expression information.* Bioinformatics, 2004. **20 Suppl 1**: p. I178-I185.

226. Green, M.L. and P.D. Karp, *A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.* BMC Bioinformatics, 2004. **5**: p. 76.

VITA **Madhukar S. Dasika**

Madhukar Suri Dasika was born on October 9th, 1980 in Hyderabad, India. He graduated from Little Flower Junior College, Uppal, Hyderabad in 1998. Subsequently, he proceeded to do a Bachelors degree in Chemical Engineering at the Indian Institute of Technology, Madras (Chennai), which he completed in 2002. In Fall 2002, Madhukar joined the graduate school at the Pennsylvania State University, where he pursued a Ph.D. in chemical engineering under the guidance of Dr. Costas D. Maranas. His research focused on the development of optimization tools to enable the inference analysis/redesign and validation of biological networks and resulted in the following publications.

1. Dasika, M. S., A. Gupta and C.D. Maranas (2004), "A Mixed Integer Linear Programming Framework For Inferring Time Delay in Gene Regulatory Networks," Pacific Symposium on Biocomputing, 9,474-485.
2. Dasika, M. S., A. Gupta and C.D. Maranas (2005), "DEMSIM: A discrete event based mechanistic simulation platform for gene expression and regulation dynamics," Journal of theoretical Biology, 232, 55-69.
3. Dasika, M. S., A. Burgard and C.D. Maranas (2006), "A computational framework for topological analysis and targeted disruption of signal transduction networks," Biophysical journal, 91,382-398.
4. Dasika, M.S., V. Satishkumar and C.D. Maranas (2006), "Analysis and redesign of biological networks: Metabolic and Signaling networks", Proceedings of ICCSB, Shanghai.
5. Dasika, M.S., and C.D. Maranas (2007), "OptCircuit: An optimization based method for computational design of genetic circuits," (Under Review).
6. Satishkumar, V, M.S. Dasika and C.D. Maranas (2007),"Optimization based automated reconstruction of genome-scale metabolic networks", BMC Bioinformatics, *In Press.*