Stochastic Modeling of Inducible Logical Gates in *Escherichia coli*

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Bennett Joseph Swiniarski

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Master of Science

December, 2010

Acknowledgements

There are many people that have earned my gratitude for their contribution to my time in graduate school. I would like to thank my advisor, Yiannis Kaznessis, for all the support he has given me during my graduate career. I would also like to thank my group mates, especially Kostas Biliouris, Ling Wu and Katherine Volzing, for all the discussions and help they have provided. Finally, I want to thank Emily Frandson and my parents for all the love they have shown me during my time in Minnesota.

Abstract

In this work, a systematic approach to the modeling of synthetic biological systems, with a focus on gene regulatory networks, is presented. We extend a previous study of a logical AND gate system by observing the effect of individual components of this system on its functionality using experimental techniques and computational models. The previous AND gate was synthesized using operator sequences from the well characterized prokaryotic Tetracycline and Lactose operons. This synthetic construct is responsive to the exogenous chemical inducers IPTG and aTc and produces green fluorescent protein (GFP) as an output signal. Experimentally, we broke down the AND gate promoter into its individual components by replacing the tet and/or lac operator sites by a non-binding *E. coli* DNA sequence. The individual components of the promoter were then characterized by studying the output GFP signal when varying inducer concentrations were added to the system. In addition, we have characterized systems with mutant tet operator sites to further study the regulatory circuit. Along with experimentally characterizing these systems, we have modeled the systems in detail. The models are stochastic in nature and include all the biomolecular interactions involved in transcription, translation, regulation and induction in order to quantify the influence of each individual interaction on the behavior of the system. Thus, we can quantitatively analyze the behavior of the systems, and better understand why the systems behave as they do. This approach of designing and tailoring logical regulation of gene expression can be potentially extended beyond transcriptional regulation to include other modular genetic elements, search for more complex network behaviors and assist in forward engineering of other biological circuits.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The interaction of molecules within the complex environment of the cell forms the basis of how life works. Understanding these interaction networks is of great importance, and the study of these systems is a rapidly developing field. This thesis seeks to contribute to this field by describing and analyzing biological systems using computational methods. In particular, we will be using detailed reaction networks to model the central dogma of biology in gene regulatory networks. This central dogma of biology, which is the fact that DNA is transcribed to RNA, which is translated into protein, plays a key role in gene regulatory networks. [2, 3]

Gene regulatory networks are absolutely essential to all prokaryotes and eukaryotes. It is imperative that all biological systems regulate what proteins and mRNA are made, and when those molecules are created. Without high-fidelity controls on these molecules, cells would be unable to respond to environmental stimuli, such as the availability of food, or changes in temperature. In this type of network, proteins or mRNA either activate or repress the production of other protein and mRNA molecules. Even with only these two mechanisms, systems of astonishing complexity occur in nature.

Gene regulatory networks also play a key role in the developing field of synthetic biology. Synthetic biology aims to create new functions in biological systems by engineering novel DNA sequences within the cell. In order to attain these functions, regulation of protein and mRNA is required. The field of synthetic biology has led to many exciting and innovative discoveries, such as the creation of a synthetic genome [4] and the engineering of a bacterium to produce an anti-malaria drug [5]. However, the promise of synthetic

biology is tempered by a great challenge: finding the right DNA sequence to generate new functions is not easy to do. There is no obvious correlation between any sequence and its function in biological systems.

This challenge occurs because of the very complex environment that these synthetic biological systems inhabit. Inside of a cell, which typically has a volume of 10^{-9} to 10^{-15} liters, there are literally thousands of different proteins, along with other biological molecules, such as sugars and fats. All of these different molecules have the potential to affect our engineered DNA sequence. We can figure out the major interactions between molecules inside the cell which affect our system, and write these interactions as a set of biomolecular reactions. The details of how we write these models is the topic of Chapter 4 in this work.

We now come to the next phase of our problem. We must figure out how to efficiently simulate these systems. We are faced with two major issues when attempting to simulate these models. First, we must deal with the fact that these biomolecular reactions occur on widely varying time scales ranging from microseconds for DNA binding interactions, to hours for the degradation of some proteins. These differences make our reaction network quite stiff and difficult to find solutions to. Secondly, some species in these reaction networks occur in very small quantities. In many reaction networks, all species are present in amounts great enough to approximate the number of molecules as a continuous variable. This allows us to ignore the discrete nature of the amounts of species and model the system as a set of ordinary differential equations(ODE's). However, some species in biological systems, such as operator sites, may have as few as one or two molecules present inside of the cell. We should consider the discrete nature of this species, and the fact that random fluctuations influences the system. This noise is the result of the randomness of the interactions between the small number of molecules. Our algorithm for simulating these systems must be able to deal with stiff, stochastic networks. We have been able to deal with these issues by using an algorithm developed in our group called Hybrid Stochastic Simulator for Supercomputers (Hy3S). This algorithm deals with the problems stated above using stochastic mathematics and clever partitioning of the system into fast and slow reactions. This work will describe exactly how this algorithm works mathematically in Chapter 3 of this work.

The gene regulatory network we will specifically study in this work is based on a logical

construct known as an AND gate. AND gates are an integral part of electronic systems. Their function is to take two inputs and only produce an output when both of these inputs are present. If either of the inputs is not present in the system, no output will be produced. The goal of synthetic biology is to produce desired functions from DNA sequences, and an AND gate construct will be pivotal to creating many biological functions in the future. Thus, we would like to create and analyze a simple biological AND gate for use in more complex systems.

Previously, we have created a novel AND gate construct which is contained in a small DNA sequence using small, well studied inducer molecules as inputs to the system and fluorescence as the output to the system. This system was experimentally constructed and modeled, and we found a several different novel DNA sequences which behaved as a high-fidelity AND gate. However, there is more work to be done to fully understand this system and all the factors that lead to a robust AND gate. This work focuses on breaking this system into even simpler constructs, in order to understand each part of the DNA sequence affects the entire AND gate. A complete description of how the AND gate is constructed and how we break the system down are the topics discussed in Chapter 2 of this work.

Once we constructed the system experimentally, and modeled it using a detailed modeling formalism, we were able to confirm that this modeling method allows to match the experimental results. All of the results from these experiments and simulations are the topic of Chapter 5 of this work.

To summarize,

- Chapter 1 introduces the background of synthetic biology and molecular modeling pursued in this thesis.
- Chapter 2 presents the specific biological AND gate system analyzed in this thesis.
- In Chapter 3 the modeling formalism used to analyze this system is presented.
- Chapter 4 presents the history and science behind the stochastic simulations used in this thesis.
- Chapter 5 shows and discusses the pertinent results obtained in this work.

Chapter 2

Synthetic Biology

2.1 History of Synthetic Biology

As stated before, synthetic biology is a field that implements new functions in biological systems by designing novel DNA sequences. An important class of these new functions is gene regulatory networks. Many of these different network archetypes have been studied in the literature, including oscillators, bistable switches, amplifiers, and logic gates. [1, 6–17] The common thread behind these different paradigms is that they are commonly used in the field of electronics and electrical engineering. Bistable switches are used in computer memory, and logic gates are used in nearly all electronic circuits. All electronic devices need to take a set of inputs and integrate these inputs to produce an certain output, and this is what logic gates accomplish.

When compared to other fields in engineering and biology, synthetic biology is a relatively new discipline. Several of the seminal papers in this field were published in the year 2000. These papers formed the foundation of this field, which is still developing. A very noteworthy paper, and one of the first synthetic gene networks, was published by Elowitz and Leibler. [1] In this paper, a synthetic oscillator was constructed, which allowed a reporter protein to oscillate with a defined period.

The way this synthetic network works is by using three synthetic genes, each of which represses one of the other genes to form a complete loop. With this novel system, the repression of the proteins is turned on and off at a regular interval by tuning the promoter strengths controlling each gene. This type of repression forms the basis for

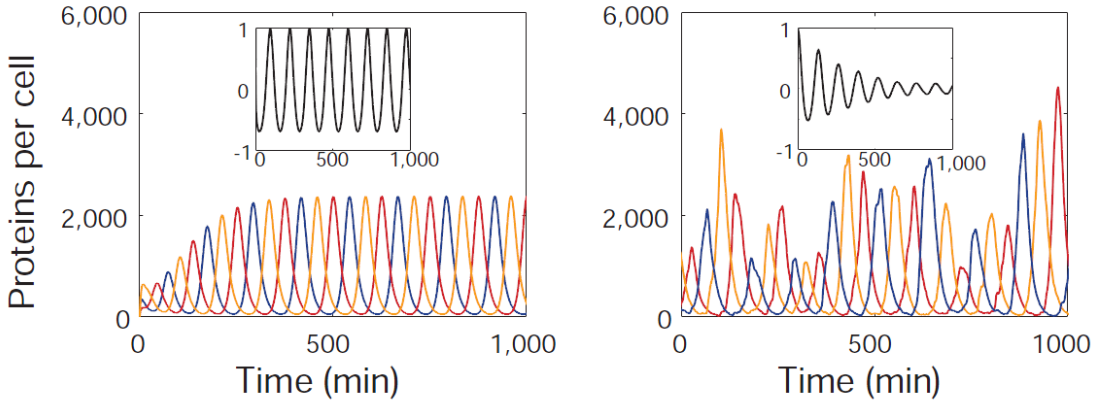

Figure 2.1: This figure was adapted from Elowitz and Leibler. [1] It shows the oscillation of three different proteins, each with a regular period, in *E. coli*. These proteins were synthetically inserted into this bacterium, and each protein represses another synthetic protein, forming a loop. The whole construct is called the repressilator.

the modeling done in this work.

It should be noted that each of these proteins was located on its own transcriptional unit, which is called monocistronic. The systems in this work follow this paradigm, but it should be noted that work has been done with polycistronic systems. [18]. The most important works in this category are the construction by Gardner and coworkers of a bistable switch in *E. coli*, and Atkinson's creation of a gene circuit which can produce both bistable behavior and oscillations. [6, 19] Also in 2000, several other important papers were published, such as a paper describing how negative regulation of a promoter by its own gene product reduces the effect of stochastic noise in the system. [7] This is of importance to this work, since we deal with the stochastic noise inherent in biological networks. This phenomenon has been studied in more depth recently as well. [20]

There has also been progress in this field when synthesizing entire biological pathways in organisms. This is an important aspect of synthetic biology, as it allows for the production of useful molecules very efficiently. In 2000, Farmer and Liao described the creation of an entire enzyme synthesis pathway. [21] This was the first example of a biological pathway synthesis. An important recent application of this is the synthesis of artemisinin, an anti-malarial drug, undertaken by Ro and coworkers in 2006. [5]

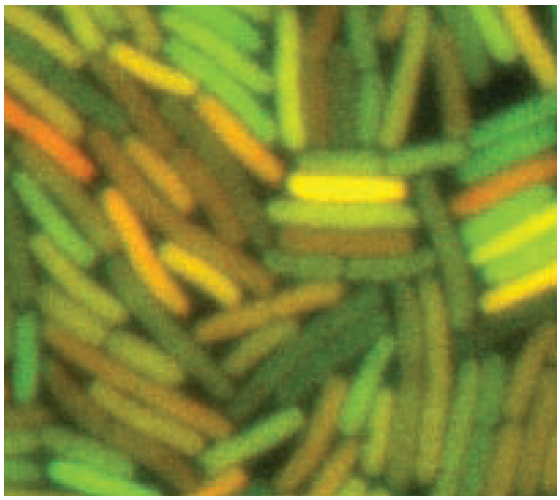

Figure 2.2: This picture shows the stochastic behavior in *Escherichia coli*. Cells with fluorescent reporter proteins were studied under identical conditions, but they show different amounts of reporter proteins. This shows the importance of stochasticity in biological systems.

In this work, the entire pathway to produce this drug is transplanted into a cell and produced. Previously, artemisinin could only be produced as the result of a costly chemical extraction process from the wormwood tree. This breakthrough allowed the drug to be produced for one-tenth of the previous cost.

The specific aspect of synthetic biology which will be studied in this work is logic gates. The first paper describing this type of system was published by Lutz and Bujard in 1997. [22] In this paper, a synthetic promoter is described which contains operator sites that bind different repressor proteins. These repressor proteins are inducible. Inducible proteins bind molecules called inducers, which change the behavior of the protein in the system. The response of the system can be turned on or off based on this induction. In Lutz and Bujard's work, the synthetic construct is not considered to be an AND gate, but in reality it is. It allowed the transcription of a gene only when the promoter is freed of repressor proteins by inducing the system. This paper served as the inspiration for the systems created in this work. A more detailed description of how this system operates will be discussed in a later section of this work.

Other logic gates have been created, such as the network of genes created by *Guet et*

al. [9] This is different from the work discussed before, because it is not created by a single promoter. Cox and Elowitz have taken a different approach to creating logic gates, using a combinatorial method to generate single promoters which showed many different logical gate phenotypes. [23]

Our work also examines the stochasticity in biological networks. Many mathematically based arguments have been made about why stochasticity should exist in dilute systems, and this was experimentally verified in 2002, when *Elowitz et. al.* used fluorescent reporter proteins in single cells to show that stochasticity is indeed present in biological systems. [24] In this work, single cells were created with identical numbers of RNApol, and ribosomes, and analyzed after the same amount of time. Even with identical starting conditions, there was still variance in the production of fluorescent protein. Figure 2.2 shows a picture of this phenomenon, with many cells having different colors. Unlike most of the systems discussed above, this work does not describe a gene regulatory network. It is still an important paper because it validates the stochastic approach we use in models discussed later in this work.

Chapter 3

Understanding the Logic AND Gate System

3.1 Importance of And Gates

Gene regulatory networks have great importance in biological systems. These networks are the key to allowing the cell to respond correctly to different stimuli in the environment. The gene regulatory network that is the focus of this work is an AND gate. An AND gate utilizes two input signals to produce an output. The table below is called a truth table, and it shows the response of the system when two different inputs are present or not.

As one can see from the truth table, the output of an AND gate is only ON when all of the inputs are present. If any of the inputs are OFF/absent then the output

Table 3.1: This is the truth table for the AND gate system. The system only shows a response when both inputs are present.

A	**B**	**Output**
0	0	0
0	1	0
1	0	0
1	1	1

is OFF. This logical operation is widely used in electronic computing, and almost all electronic programs use this logical construct. We would like to create genetic programs in the cell. Because of this, AND gates are a critical device when creating different genetic programs. In the future, an understanding of how an AND gate operates in the complex environment will be a necessary component when creating more complex genetic programs. The field of synthetic biology is looking to solve many types of different problems, some of which will require complex genetic programs to solve. The foundation of these programs will be smaller systems, such as the AND gate, which are well-understood.

Well-studied systems are ideal to use for our modeling approach. The extensive amount of literature on the components in our AND gate allows us to find all the kinetic constants and growth rates necessary for the detailed model described later on in this document. As stated before, logical AND responses are particularly common in biological networks, and studies have identified various mechanisms by which cells implement AND logic [25, 26]. As part of the effort to classify these natural genetic logic gates and to generate novel gates for use in artificial genetic circuits, a variety of AND gates have been constructed using unique regulatory mechanisms including chemical complementation [11], posttranscriptional regulation [9], and allosteric control [27]. While the complexities of these systems vary greatly, it has also been shown that AND logic can be obtained in a minimal system consisting of an individual promoter regulated by two regulatory factors [28]. A benefit of this simplified architecture is that it should allow for the easy construction of novel AND gates by the incorporation of additional cis-regulatory regions into target promoters.

3.2 Previous Work with AND Gate Systems

Our group has already demonstrated the functionality of an AND gate constructed in *Escherichia coli (E. coli)* [29]. The particular AND gate studied has two inputs: isopropyl β-D-1-thiogalactopyranoside (IPTG) and anhydrotetracycline (aTc). The output is green fluorescent protein (GFP) expression. This device works by controlling the rate of GFP expression in the absence of inducer. The construct consists of a synthetic hybrid promoter upstream of the GFP coding sequence. This hybrid promoter was

designed according to the work of Lutz and Bujard [22]. The operator sequences are from three unrelated natural regulatory elements: the tetracycline (T), lactose (L) and λ-phage operons arranged logically within a single transcriptional unit. The hybrid promoter was placed upstream of the reporter GFP gene and cloned in a high copy number plasmid [30].

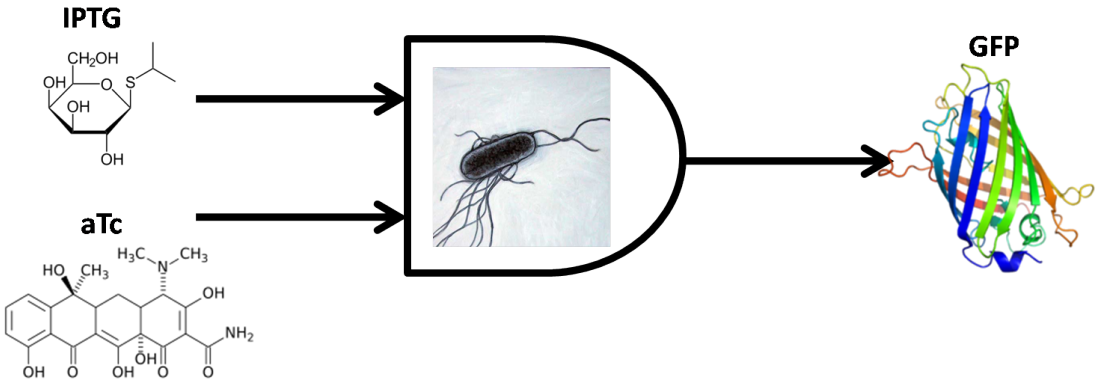

Figure 3.1: This schematic shows the inputs and outputs of our AND gate system in *E. coli*.

The AND gate system takes advantage of the fact that RNA polymerase binds to two sites, one of which is 35 bases upstream of the transcriptional start site (-35), and the second 10 bases upstream of the start site (-10). Our system changes the DNA sequence upstream of -35, in between the two sites, and downstream of the -10 site. We insert either the tetO1 (T) or LacO2 (L) site in these positions, causing TetR or LacI respectively to bind the DNA and sterically inhibit RNA polymerase when these species are present on the DNA. Thus the promoter is under the control of two different repressor proteins, TetR and LacI, both of which are constitutively produced in the E. coli DH5αPro cell strain that we used [31]. The expression of GFP was measured by fluorescence activated cell sorting in these strains. The output fluorescence depends on the input of two small molecule inducers, aTc and IPTG. Inducers interact with repressor molecules LacI and TetR and change their conformation such that repressors no longer bind to their respective operator sites. Figure 3.3 shows where this molecule bonds to the protein. These operator sites are now available for RNAp to bind and initiate transcription. The simplicity of these designs is highlighted by the minimal

number of regulatory components required to achieve high-fidelity, robust AND gate functionality.

Figure 3.2: This illustration shows the placement of the operator sites relative to the RBS and gfp in the TTL AND gate. Note that a replacement of any of these operator sites with a different operator site or "junk" DNA produces a different logic gate.

The simplicity of the AND gate, and the ability to make many systems by swapping operator sites allows us to easily create many parts which can be studied. The AND gate was broken up into smaller systems and several of the possible operator placements were experimentally created and modeled. The system was designed such that GFP expression can now be driven by just one or two components of the previously designed AND gate promoter.

An example of one such AND gate is denoted as a TTL AND gate. In this gate, there are T binding sites upstream of the -35 position, and in between the -35 and -10 positions. There is also an L binding site downstream of the -10 position. A schematic of this construct is shown in Figure 3.2 above. This construct allows control of GFP expression by both aTc and IPTG. When no inducer is present TetR and LacI preferentially bind their promoter sequences, sterically hindering the RNApol from binding and transcribing gfp. [32, 33] However, when aTc and IPTG are added to the system they cause a conformational change in TetR and LacI respectively. [34] These conformational changes cause a decrease in their binding affinity for their promoter sequences, and the promoter is then uninhibited by these repressors. [35] RNApol can then bind the free promoter and transcribe the downstream mRNA, which leads an increase in GFP protein. Therefore, this construct will produce high GFP protein levels only when both aTc and IPTG are present and very low GFP protein levels otherwise.

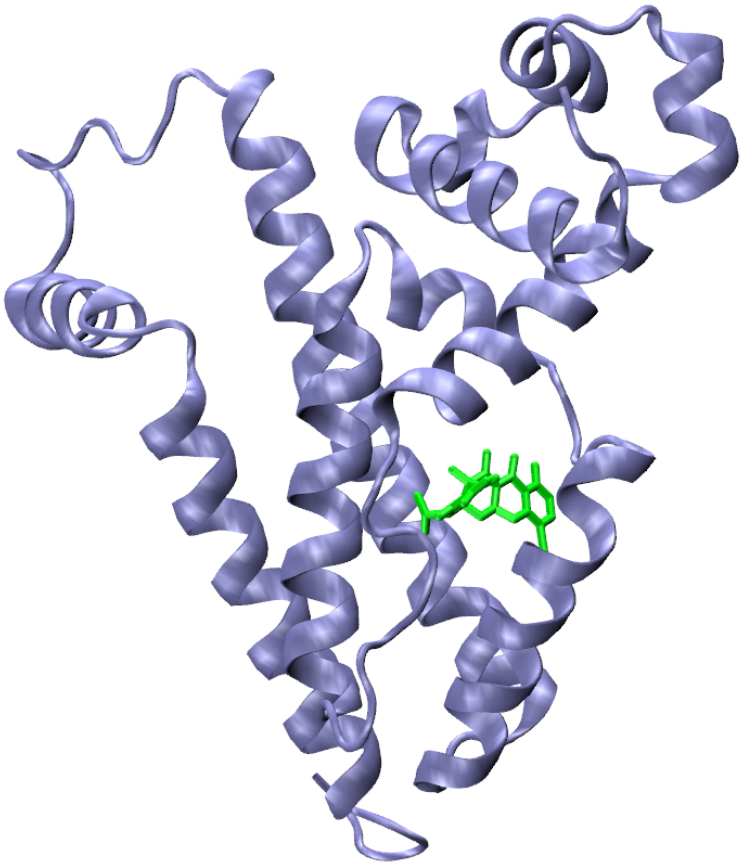

Figure 3.3: This diagram shows aTc binding the TetR monomer. This interaction causes a conformational change in the protein, reducing its affinity for the operator site. Inside the cell, this protein dimerizes, meaning it takes two aTc molecules to fully induce the conformational change. The DNA binding domain is located at the top of the molecule.

3.3 Breaking Down the AND Gate

We have broken up AND gates talked about in the previous section (Section 3.2). The complete logic gates described above consist of three different operator sites which work together to give the correct response. However, each one of these operator sites contribute differently to the make-up of construct. Using the TTL AND gate as the model system, we have created a schematic, Figure 3.4 that shows the breakdown of this system into these components. TNN, NTN, and NNL constructs have been created, where the N denotes that part of the sequence has been replaced with "junk" DNA to which neither TetR nor LacI bind. These new systems are no longer AND gates, as they will only respond to induction by either aTc or IPTG but not both.

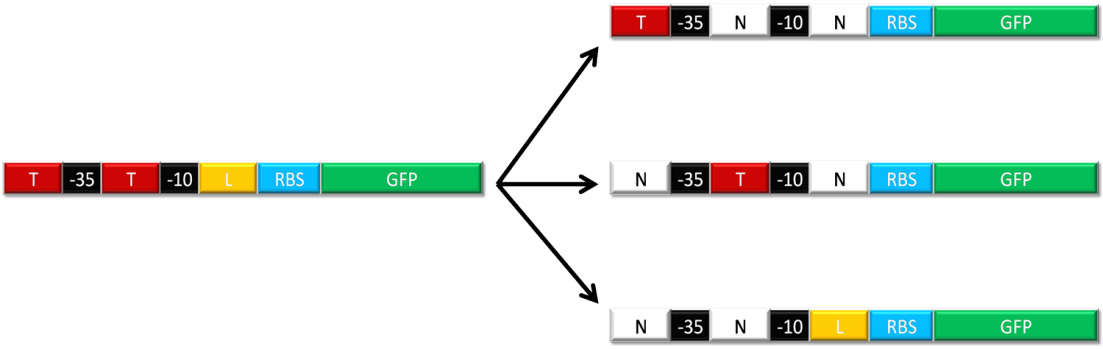

Figure 3.4: This illustration shows a TTL AND gate which is broken down into three smaller systems, the TNN, NTN, and NNL systems. The N sequence is an intronic sequence from a fungus, and is not bound by any protein in *E. coli*

When we start this process, it leads to a variety of potential systems. For the purpose of this work, we are looking at systems that have only T or only L operators. This simplifies our systems by eliminating any potential AND logical behavior, and also lowers the number of possible systems. Even with this restriction, 12 different configurations of operator sites are possible. We can consider situations where there is only a single operator site present in our systems, or situations where 2 operator sites are cloned into the promoter sequence.

In this way, we have 3 "single T" systems, TNN, NTN, and NNT. These systems all have one *tet* operator site in the promoter sequence. The N sites replace a T or L

operator, and consist of a sequence that has the correct number of DNA base pairs that will not bind any protein. It is taken from an intronic fungus sequence. [36] Since it is from another very different bacteria, we can be assured that it does not efficiently bind any proteins, especially TetR and LacI. Figure 3.5 shows these single T systems in a diagram.

Figure 3.5: This diagram shows a representation of the three possible "single T" systems created from breaking down the full AND gate to study the individual operator sites.

Similarly, we can also create single L systems. (LNN, NLN, NNL) These systems allow us to understand the impact of an L operator in each position without worrying about other effects. We can also create "double T" and "double L" systems, which contain two *tet* operators or two *lac* operators, respectively. Combining all these possibilities, we have studied twelve wild-type systems. The study of these systems will show us how each part of the AND gate affects the whole. Chapter 6 will describe how we created these systems experimentally, and how we accurately modeled them using stochastic methods. Overall, AND gates represent an important logical operation in the cell. Understanding how the create a high-fidelity AND gate in *E. coli* provides a foundation for other more complex constructs, and provides us with one more link in the path between molecules and life.

Chapter 4

Stochastic Simulations

4.1 Stochasticity in Gene Networks

Traditionally, systems involving a chemical reaction network have been represented by a system of ordinary differential equations(ODE's). This approach works well when the system in question has species with large quantities of molecules. Although these species do have a discrete number of molecules, in practice this matters little when there are so many molecules involved in the system. These systems are considered to be at the thermodynamic limit and so deterministic methods can indeed accurately represent the system. For example, let us suppose we have the reaction with rate k,

$$A + B \rightarrow C + D \qquad \text{k} = 1 \tag{4.1}$$

We can write differential equations for each species to describe their behavior through time. For example,

$$\frac{dA}{dt} = -k[A][B] \tag{4.2}$$

$$\frac{dC}{dt} = k[A][B] \tag{4.3}$$

where the rate of reaction is represented by $r = k[A][B]$. Similar equations can be written for species B and D. Once these equations are written, we can assign an initial condition to the system at t = 0, and propagate the system through time by solving

the differential equations. This type of formulation is known as mass action kinetics, because the rate of change of the species is based solely on the mass of the species in the system. These equations are fairly easy to integrate, although difficulties can occur when there are widely varying rates of reaction in the system. If we are dealing with biological systems, such a situation is likely. With widely varying constants, the system is considered to be stiff, although solvers are available to deal with this issue.

However, some of the assumptions behind this ODE formulation begin to break down for biological systems. First, biological systems contain a small numbers of molecules, meaning we can no longer approximate the population of the species as a continuous variable. We are away from the thermodynamic limit, meaning we have a small reacting volume or a very small concentration of molecules. Instead of talking about rates of reactions, we instead shift our discussion to the probability of a reaction occurring in time. The reason for this can be shown by the following example. Suppose we have only one molecule of a species in the system, and this species is degraded. At any point in time, the species is either intact or degraded. It does not make sense to talk about a rate, since the species will degrade at some instant in time. Instead, it is better to consider a probability that the species would be degraded at a specific time. There an underlying probability distribution that describes this occurrence.

4.2 Chemical Master Equation

A way to describe this underlying probability distribution was first described by McQuarrie. [37] In this work, a Chemical Master Equation was developed which describes the time evolution of all the species in the system throughout time. This equation is a differential equation which is difficult or impossible to solve for most systems. Before we delve deeper into this equation, we can define some quantities present in all reaction networks. First, we must define a vector that tells us the state of the system. In other words, this is the number of molecules present of each species at time t.

$$\begin{aligned} X_i(t) = &\text{ number of molecules of the } i^{th} \text{ unique chemical species} \\ &\text{present in the system at time t, where i = 1...N} \\ &\text{(where N is the total number of chemical species)} \end{aligned} \tag{4.4}$$

Changes to this vector will occur only when we a reaction occurs in the system. In a biological system, these reactions could be things such binding of repressor molecules to DNA, transcriptional elongation, or degradation of protein molecules. For each system we are interested in, we can define a set of M chemical reactions which describe all the possible ways species can transition. In order to easily keep track of how each reaction will change the system, we can define an M by N dimensional matrix whose i^{th}, j^{th} component will denote the change in species i once reaction j is executed.

$$\nu_{ij} = \text{the M x N reaction matrix describes the change in the number of species i once reaction j occurs.} \tag{4.5}$$

We must also define a vector which contains the possibility for a reaction to occur in the system. This is most easily described by a quantity called the reaction propensity, $a_j(t)$. This quantity is defined as the rate constant of the reaction multiplied by the number of molecules of each reactant in the system.

$$c_j = j^{th}\text{component of M-dimensional vector of rate constants} \tag{4.6}$$

$$h_j = \text{number of distinct molecular reactant combinations for reaction j, which are present at time t} \tag{4.7}$$

$$a_j dt = c_j * h_j * dt = \text{the probability that the} j^{th}\text{reaction will occur in the system in dt, given the current concentrations of reactants} \tag{4.8}$$

This propensity vector is the key variable when figuring out which reaction will occur next in the system and is also a key component of the Chemical Master Equation. The form of the equation is:

$$P(\underline{X}, t + dt|\underline{X_0}, t_0) = P(\underline{X}, t|\underline{X_0}, t_0)(1 - \sum_{j=1}^{M} a_j dt) + \sum_{j=1}^{M} P(\underline{X} - \underline{\nu}_j, t|\underline{X}_0, t_0)a_j dt \tag{4.9}$$

Simplifying this equation to a differential equation yields:

$$\frac{dP(\underline{X},t|\underline{X_0},t_0)}{dt} = \sum_{j=1}^{M}\{a_j(\underline{X}-\underline{\nu}_j,t)P(\underline{X}-\underline{\nu}_j,t|\underline{X}_0,t_0)-a_j(\underline{X},t)P(\underline{X},t|\underline{X_0},t_0)\} \quad (4.10)$$

Essentially, this equation says that the change in probability of a state of with time is the sum of the possibility that the system is at the state already and remains there and the possibility that the system is one step away from the state and changes to that state. We will define the form of this equation later, once all the pertinent variables have been defined. This equation is found by treating the system kinetics as a Markov chain, that is to say that the system has only memory of the previous state. [38] The current state is based on only the last previous state of the system. This equation, once solved, gives the probability of the system to be in any state at any time. Unfortunately, this equation becomes intractable even for very small systems. Once we see the form of this equation, it becomes clear why the solution to this equation become intractable for even modestly sized systems. As the system grows larger, the possible states of the system grows combinatorially, meaning that for even 4 or 5 reaction systems we can have nearly infinite number of possible states. As stated before, a way to avoid solving this system is to sample the underlying probability distribution. The way we can do that is to simulate an ensemble of individual trajectories for the evolution of the system through time. If we generate enough trajectories, we can sample the underlying probability distribution.

4.3 Stochastic Simulation Algorithm (SSA)

The stochastic simulation algorithm developed in the 1970's by Daniel Gillespie provides a way to generate trajectories for stochastic gene networks. [39, 40] It has been studied and improved by several groups such as Gibson and Bruck, or Li and Cao. [41, 42]. Gibson and Bruck's "Next Reaction" variant of SSA forms a component of the more complex algorithms discussed later. The overall idea behind this method is to simulate every reaction event which occurs in the system. The size of the system being simulated also has no effect on how the algorithm runs, although there will be more events to simulate for larger systems, and the algorithm will slow down. After each of these

reaction events is simulated, the system is incremented forward in time, and the reaction propensities, a_j, are used to calculate when the next reaction is predicted to occur for each reaction in the system. These times include a random component, because of the stochastic nature of the systems. Once all these next reaction times are calculated, the one with the shortest predicted time is executed, and the process is repeated until the time is incremented past the end time of the simulations. Even with optimizations, this method still must simulate each individual reaction event, rendering it very slow for systems with reactions that occur many times. It is necessary to look at other methods to get around this limitation of the SSA.

4.4 Stochastic Hybrid Methods

As shown by section 4.3, the SSA algorithm is slow when dealing with reactions that occur quickly or when reactant species are present in high concentrations. This algorithm becomes slower and more computationally intensive, and eventually must be abandoned. Several methods which approximate the exact algorithm have been developed. These methods reduce the computational burden of the SSA, but sacrifice some accuracy to do so. One of the most notable is a method developed by Gillespie (creator of the original SSA), the τ-leap method. [43] Other methods use Chemical Langevin equations, which are appropriate when the system contains fast reactions, but is still not near the thermodynamic limit. [44]

More interesting work has been done for systems where there is a clear delineation between the fast and slow reactions in the system. In this case, if the fast reaction goes to a quasi-steady state quickly, Salis has developed a method called the probabilistic steady state approximation (PSSA). [45]. Another approach to use when the system can be divided into fast and slow reactions are "hybrid" methods. These methods use one approach to simulate the fast reactions, and another to simulate the slow reactions. Salis and Kaznessis have developed this type of algorithm. [46] It is a hybrid method called which dynamically partitions the system into slow and fast reactions. The partitioning of the network is based on two parameters λ and ϵ, which are set by the user.

For a reaction to be considered "fast", the two following conditions must be true:

$$a_j(t) > \lambda \gg 1 \tag{4.11}$$

$$x_i(t) > \epsilon |v_{ji}| \tag{4.12}$$

The first condition says that the propensity of the reaction must be large. Because of the definition of the propensity, this means that either the rate constant for the reaction is large, or there are a lot of reactant molecules in the system. The second condition states that the effect of each reaction event must be small when compared to the total number of reactant species in the system. With these two conditions met, the reaction event can be approximated as a continuous variable. Typical values for ϵ and λ are 100 and 10 respectively. Of course, the state of the system and reaction propensities change over time, so these conditions must be dynamically checked throughout the simulation to reclassify reactions as "fast" or "slow". This allows us to partition the Master equation into multiple subsets.

Once we partition the system, we simulate the slow reactions quickly using Gillespie's stochastic simulation algorithm, and the fast reactions using SDE's of the form shown in equation 4.13. These equations are called Chemical Langevin equations, which approximate the occurrence of the reactions as a Gaussian distribution.

$$dX_i(t) = \sum_{j=1}^{M} \nu_{ji} a_j(\underline{X}(t))dt + \sum_{j=1}^{M} v_{ji} \sqrt{a_j(\underline{X}(t))} dW_j \tag{4.13}$$

Simulating the different subsets of the system using different methods greatly reduces the computational burden for large networks. These methods are used in the next sections of this work to accurately simulate the reaction networks for the biological systems we have created.

4.5 An Example System

In order to show that stochastic simulations are indeed needed when simulating biological systems, a small example system will be presented. We will look at a cell with a

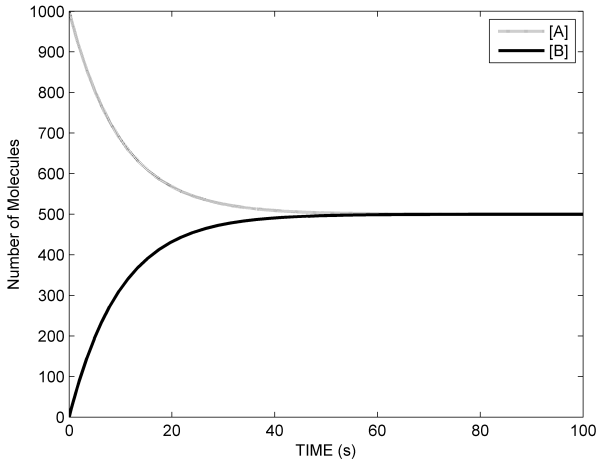

Figure 4.1: Changes in the number of molecules with respect to time when the system is solved with ODE's.

volume of 10^{-15} L, typical of a bacterial cell, with only two species inside of it. We will call these species A and B for simplicity. Only two reactions can occur in the cell:

$$A \rightarrow B \qquad \text{k} = 0.5\ s^{-1} \tag{4.14}$$

$$B \rightarrow A \qquad \text{k} = 0.5\ s^{-1} \tag{4.15}$$

Just by looking at this reaction network, we can surmise that the steady state of this reaction network will have an equal number of A and B molecules, since the forward and reverse reactions have equal rates. We will show how different initial conditions of A affect the dynamics of the system. Figure 4.1 show the deterministic solution to this system, solved in MATLAB (ode45 solver). As we expect the system exponentially approaches a steady state where half the molecules are species A and half are species B.

The first system we will stochastically simulate has an initial condition of $A_0 = 100,000$ and $B_0 = 0$. With this many molecules in the system, fluctuations and noise become less important, and the stochastic solution will approach the deterministic solution. When we graph both solutions, there are virtually indistinguishable. We can assume we are at the thermodynamic limit, and the ODE solution does a great job of describing the

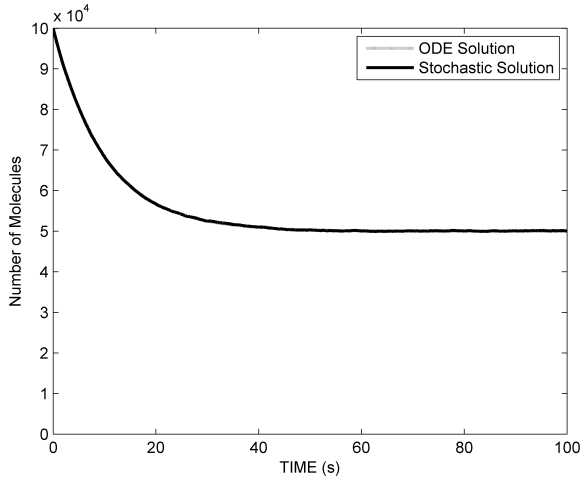

Figure 4.2: This figure shows a comparison of the stochastic and deterministic solutions when species A is initialized with 100000 molecules. Note that the solutions are so similar that it is difficult to determine the difference between them.

dynamics of the system.

Once we initialize the solution so that there are only $A_0 = 100$ molecules in the system, we no longer see the great similarity between the solutions. Instead, the noise and fluctuations start to have an impact, and the two solutions deviate. At this number of molecules, the solutions still follow the same general trend, exponentially decreasing towards the same steady state.

We can extend this study to the point where we start the system with only 10 molecules of A. At this point, we are completely within the stochastic regime, and noise completely dominates. This is apparent from Figure 4.4, which shows no resemblance between the stochastic and deterministic solutions. In this graph, the discrete nature of the stochastic simulations is also apparent. The stochastic solution can only take integer values, as opposed to the deterministic solution, which is a continuous variable.

It is important to note that each of the graphs above represent only one possible trajectory of the system. Another simulation with this algorithm will give another possible trajectory. This is different from the ODE solution, which will always give the same results. In order to sample the entire solution space, we must simulate an ensemble of trajectories. This will give us a good approximation of what the solution to the

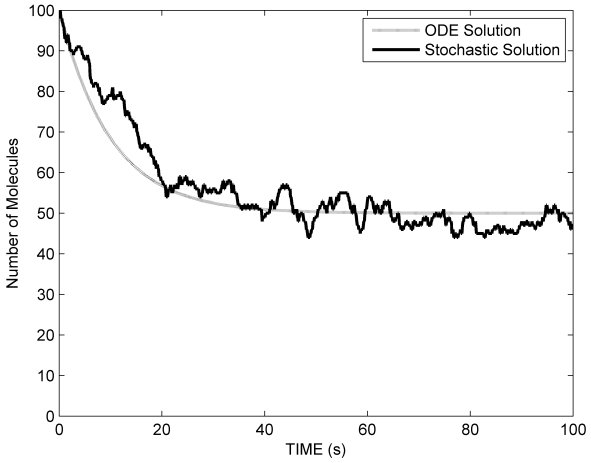

Figure 4.3: This figure shows a comparison of the stochastic and deterministic solutions when species A is initialized with 100 molecules. With this number of molecules, we can now see that the stochastic solution is significantly different from the deterministic solution.

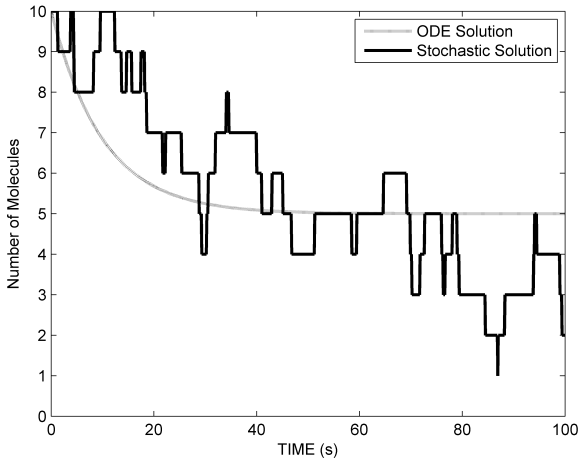

Figure 4.4: This figure shows a comparison of the stochastic and deterministic solutions when species A is initialized with 10 molecules. With this number of molecules, the discrete nature of the stochastic system becomes apparent, and the solution is substantially different than the deterministic case.

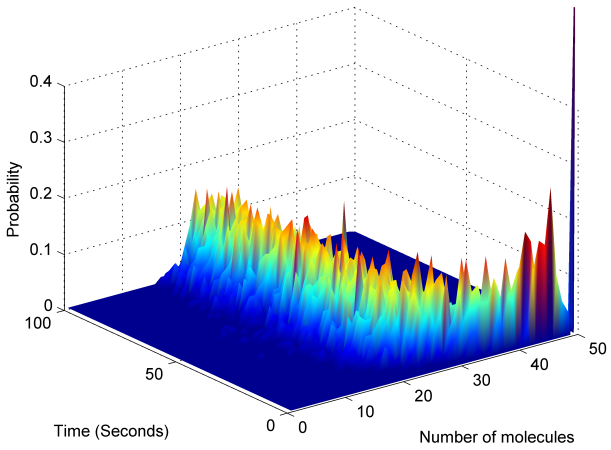

Figure 4.5: Change in the probability distribution with respect time, when the system is solved stochastically with an initial condition of 50 molecules of A. The most probable states are represented in red, and the least probable represented by blue.

Chemical Master equation would look like.

Figure 4.5 shows the solution throughout time when 1000 trajectories are simulated, starting with 50 molecules of A in the system. If one looks at the most probable distributions, highlighted by the red color, one can see that the system is most likely to follow an exponential decrease to the steady state where half the molecules are A and the other half are B. However, there are other possible trajectories, denoted by the fact that the probability distribution is not one for this exponential decrease. The system is not confined to a single state at any moment in time, but it is most probable that it would be in the mean state.

If we were to create the same graph for $A_0 = 100000$ molecules, the graph would be very sharply peaked along this ODE solution. There would be almost no probability that the state would stray far from the mean. Thus, when we have systems with larger numbers of molecules it is not necessary to simulate a large number of trajectories since they will all have essentially the same result. It would give the same result as the ODE solution.

For this example system, we have shown how the effect of stochasticity when simulating a system throughout time. With large number of molecules, these simulations

approximate the ODE solution and the effect of stochasticity is very small. As we get further from the thermodynamic limit, the effect of stochasticity becomes greater and the stochastic solution no longer identical to the deterministic solution. Since biological systems have small number of molecules, we must use the stochastic algorithms to accurately model these systems. In later sections, we will discuss these algorithms, and the results of these simulations.

Chapter 5

Systematic Modeling of Gene Regulatory Networks

5.1 Modeling Biological Interactions as Chemical Reactions

Gene regulatory networks, such as the AND gate system studied in this work, are complex systems which occur in a very crowded and complicated environment. In order to better understand what is occurring in these systems, this work will follow a systematic modeling approach. This allows us to understand how the different parts of the system interact, and we can focus on only the important molecules in the biomolecular environment. Our models describe the biology occurring inside the cells as a set of biomolecular interactions, each of which has a unique stoichiometry and rate. This reaction network can be changed for each unique system to accurately describe the differences in experimental responses. It is important to note that the modeling formalism used in this work is a detailed one. Building this type of model required us to capture the details of several different biological processes, such as transcription, translation and degradation. This section will show how we constructed these detailed models, by going through each biological process in detail, along with presenting previous modeling work.

5.2 Previous Work Modeling Biological Systems

In all of the scientific disciplines, mathematical models and simulations of experimentally observed phenomena are used. These models allow to test how well we understand what is happening in the system, and can also be quite useful when trying to create new experimental systems. In the field of synthetic biology, detailed models are a fairly recent occurrence, due to the relatively short lifespan of this field. The development of these models has also been hampered by the lack of quantitative data in biological systems. Biological systems are quite complex, and this has been a source of difficulty when finding these information. Despite these challenges, these models are rapidly being developed and show great promise for the future.

The term modeling can be applied to many different types of constructs, some qualitative and others quantitative. The simplest type of biological model consists of solely a set of interactions that describe what is occurring in the system. This model need not have any mathematical data attached to it. In almost every molecular biology textbook, this type of model is used solely for understanding what molecules interact in order for a biological system to function. These models say nothing about the rate that these interactions occur, and also say nothing about how many molecules of any species are in the system. This information is a vital component of our models, and is necessary when accurately capturing behavior of a system.

The simplest type of quantitative model is those of the type pioneered by Cox et al., which involves only a single algebraic equation. This equation describes promoter activity as a function of the amount of inducer in the system. [23] This model only uses a few kinetic parameters, and does not tie these constants to a specific biological interaction. Gardner and Collins have developed a slightly more complex model for their toggle switch system in *E. coli*. [19] This model utilizes differential equations to describe the production of 2 repressor proteins as a function of the repressor already present in the system. The kinetic constants are still not tied to a specific biological interaction, but the use of differential equations allow the system to change with respect to time. Without a correlation between the actual physical interactions in the system and the model constants, all of the parameters must be experimentally determined and consequently cannot be used in any other systems. Any models constructed in this way lose

their predictive capabilities, which makes it a time consuming process to create new models for each system studied. Because of these reasons, this work will focus on a modeling formalism pioneered by Arkin and coworkers [47] Instead of only using a couple equations, all of the biological relevant interactions were modeled. Each interaction was considered to follow an elementary rate law, and was tied to a specific interaction. This led to models of around 30 reactions to describe the behavior of a λ-phage virus. These models require more equations, but they also gain predictive characteristics for other systems. This is because each constant in the model is related to an biochemical reaction which can be measured. These constants can then be used in any system, since they are experimentally derived. Another important aspect of this work is the stochastic nature of the networks. The stochasticity inherent in the system is accounted for, and Gillespie's stochastic simulation algorithm (described in section 4.3) was used to determine the dynamics of the system. The following chapters will describe the systems analyzed, and the models used to simulate them. The modeling of synthetic biology is a promising area of research, and we will show our contribution to this exciting area in the rest of this work.

5.3 Transcription and Translation

Transcription was the first process considered, as this was the source of control in the AND gate and is an essential part of the central molecular biology dogma [2, 3]. This dogma states that DNA sequences in the cell are transcribed by RNA polymerase to form a single-stranded molecule name messenger RNA (mRNA). This mRNA molecule is then translated by a ribosome into a protein molecule. This protein molecule then folds into its active form. Proteins perform many functions in the cell, from repression of DNA to the transport of nutrients around the cell, and many others. In our system, protein's major function is to control the expression of our target DNA sequence. This is the reason why our AND gate system is considered to be a gene regulatory network. The modeling formalism used in this work takes advantage of the fact that these types of systems are well-studied in literature. [22, 26, 29] We know the mechanisms and the steps of this process, and so we can model each step individually. Each step is simple on its own, but overall, they lead to the complexity inherent in these regulatory networks.

We will model each of these steps as a 1^{st},2^{nd}, or 3^{rd} order reaction. Then we can create a step-by-step kinetic model, and evaluate the whole network based on the combination of these steps.

Modeling the transcription and translation process starts with RNA polymerase (RNApol) binding to the free promoter. We denote the promoter as species P in this description. This promoter must be free in the system in order for this reaction to occur. The promoter can also contain operator sites, which specifically bind repressor proteins. We will call these operators O1 and O2 when describing the system, although there may be other numbers of operator sites present. When bound, these proteins can sterically hinder the RNA polymerase from coming into contact with the promoter and binding. It is important to note that in *Escherichia coli*, which is the bacterial species used for our system, RNApol binds two specific DNA sequences, one 35 base pairs upstream from the transcriptional start site, and the other 10 base pairs upstream of the same site. The operator sites are located upstream, downstream or in between these sites, which is why proteins bound there do such an effective job of preventing transcription. We can write the binding of RNApol to the free promoter as follows.

$$RNApol + P + O1 + O2 \rightarrow RNApol : P : O1 : O2 \tag{5.1}$$

We also need to consider the possibility of the RNApol falling off the promoter, recreating free promoter. This reaction is the reverse of equation 5.1.

$$RNApol : P : O1 : O2 \rightarrow RNApol + P + O1 + O2 \tag{5.2}$$

We must represent the operator sites differently than the promoter because of the different interactions they participate in, although they are contained on the same stretch of DNA. This reaction is also tricky to analyze when it comes to the order of the reaction. It looks like it is a 4^{th} order reaction, but the operator sites and the promoter will always be in the same place. So essentially this reaction is describing two species coming together, the RNApol and the DNA sequence, making it a second order reaction. Writing out the rate law for this reaction will make it more apparent what exactly is going on.

$$r = k[RNApol][P][O1][O2] \tag{5.3}$$

The brackets in this equation symbolize concentration of these species. We can see that the last three species can take only 2 values, either 0 if they are bound by repressors, or 1 if they are free. Thus, this reaction can only occur when these species are all one, otherwise $r = 0$. Equation 5.3 thus reduces to equation 5.4.

$$r = k[RNApol] \tag{5.4}$$

In order to accurately parameterize this rate constant, k, we can turn to literature to see what value a typical wild type promoter constant takes. DeHaseth et. al. have given a rate constant of $10^7 M^{-1} s^{-1}$ for a system constructed *in vivo*. [48] However, this value is completely dependent on the DNA sequence of the promoter. In this work, we are constantly changing this sequence, and so this kinetic constant will be a fitted value, instead of a value found in literature. We must also determine an initial condition for all the molecules found in this reaction. For the promoter and two operators, we obviously must set the initial condition to 1, otherwise the above reactions can never happen. However, we have more flexibility when it comes to the initial condition of RNA polymerase. We can use a value also found by deHaseth et al. of a concentration of 30 nM. Doing the math for an *E. coli* cell with a volume of approximately 10^{-15} L, we get approximately 18 free RNApol molecules.

The binding of RNApol to the promoter starts the cascade of reactions which lead to mRNA production. The next step in this cascade is formation of the open complex, which is when the RNApol prepares to actually start moving down the DNA, producing mRNA. This step is much slower than the previous binding. Literature values give us a kinetic constant of $.1\ s^{-1}$. [49] No other molecules are involved in this reaction, so it is a 1^{st} order reaction of the form:

$$RNApol : P : O1 : O2 \rightarrow RNApo1 : P : O1 : O2^* \tag{5.5}$$

After the open complex is formed, the polymerase begins to move down the DNA sequence, producing mRNA. This transcriptional elongation step is shown by figure 5.1 below. In this figure, you can see the protein bound to the coding DNA, with the newly produced mRNA leaving the protein and being released into the cell.

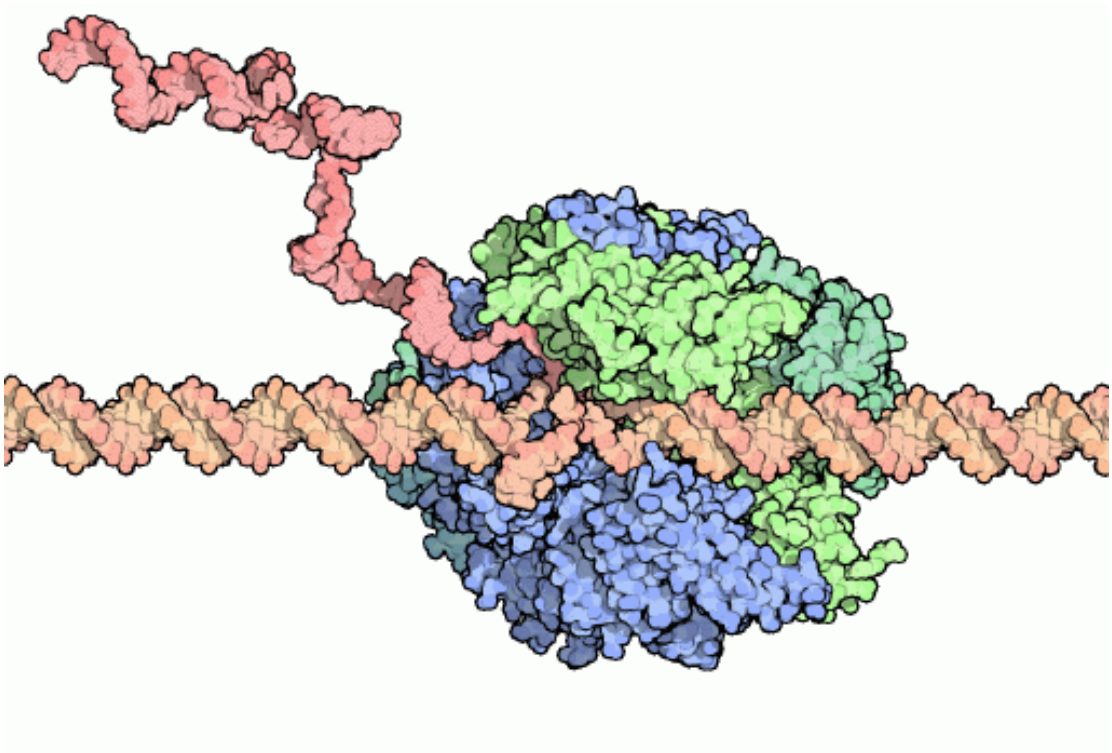

Figure 5.1: RNA Polymerase Transcribing DNA, producing mRNA [50]

This transcriptional elongation step is represented in the model by the following reaction.

$$RNApol : P^* \rightarrow P + O1 + O2 + RNApol : DNA \tag{5.6}$$

This equation seems like it produces the promoter and the operator sites. These species are not really produced, instead they are freed up by the RNApol moving further down the DNA. They are then available to participate in other reactions. The above reaction is actually a compilation of many smaller reactions which occur. Each time the RNApol moves to a new DNA base pair, it reacts and adds a base pair to the mRNA. This continues for the each base pair in the coding sequence. So the rate for the production of mRNA is directly dependent on the number of molecules in the DNA sequence. In order to account for this, the model considers this reaction to be gamma-distributed reaction(a sum of exponential reactions), with an n equal to the length of the DNA sequence. From literature, the rate of the polymerase moving along the DNA strand is about 30 base pairs per second. [47] This means the rate constant to transcribe a single nucleotide is 30 s^{-1}, which is used as the rate constant for this gamma-distributed reaction.

The final step of producing mRNA is shown by equation 5.7.

$$RNApol : DNA \rightarrow RNApol + mRNA \tag{5.7}$$

This equation frees up the molecule of RNApol to participate in other reactions and produces mRNA. Prokaryotes, such as *E. coli*, do not do any post-transcriptional modification, and so the mRNA can directly interact with the ribosome to produce protein. As before, we follow the systemic method of forming a complex, and then producing the product, which is protein in this case.

$$mRNA + rib \rightarrow mRNA : rib \tag{5.8}$$

$$mRNA : rib \rightarrow mRNA + mRNA : rib^* \tag{5.9}$$

$$mRNA : rib^* \rightarrow rib + protein \tag{5.10}$$

Equation 5.8 describes the attachment of the ribosome to the first few base pairs of the mRNA sequence. This part of the mRNA strand is intuitively called the ribosome

binding site (RBS). Similar to the promoter sequence, this binding can vary a lot in strength depending on the particular sequence. Our system keeps a constant sequence over all the systems studied, so we can fit this parameter once, and then keep it constant. Equation 5.9 deals with the possibility that more than one ribosome can be translating the same mRNA strand at the same time. It models the ribosome moving onto the coding sequence of the mRNA, freeing the RBS to be bound again. This movement forms a different complex, $mRNA : rib^*$. This reaction can again be modeled as a gamma-distributed reaction with a rate from literature, and a rate constant of 33 s^{-1} [47]. The final reaction is the translational elongation step. For the same reasons as the previous reaction, this reaction is also considered to be gamma-distributed, with the same rate. However, there is one key difference. Instead of the one to one correspondence between DNA base pairs and RNA base pairs, protein amino acids are added to the molecule once for every three mRNA base pairs. Thus, the shape parameter for the gamma-distribution is equal to one-third of the coding sequence, instead of equaling the entire length of the sequence. Once the final reaction is completed in this cascade, the model has produced protein, starting from polymerase binding onto the free promoter.

5.4 Regulatory Elements

The reaction cascade shown above is absolutely crucial to the cell's survival, but it also underscores the need for another very important element, regulation. Without regulation, the reactions above would continually produce proteins, and metabolically burdening the cell, and eventually causing its death. Our gene regulatory network includes regulation by two different repressor proteins, TetR and LacI, each of which binds to a specific operator. As stated before, this binding sterically inhibits transcription of the coding DNA. The equations shown below show how this is incorporated into the model.

$$R + O1 \rightarrow R : O1 \tag{5.11}$$

$$R : O1 \rightarrow R + O1 \tag{5.12}$$

In the equations above, we have only considered the first operator site. The other

operator, O2, has identical interactions. Species R represents the repressor protein. This protein will change based on which operator is included in the system. Equation 5.11 shows the binding of the repressor to the operator, while equation 5.12 shows the unbinding of this protein.

In order to exert precise control over the production of protein, it is important that we are able to relieve this repression as well. Our systems employ inducer molecules for precisely this purpose. These inducer molecules are small sugars which bind to the repressor molecules, and cause them to undergo a conformational change. This conformational change reduces their affinity for the DNA sequence, making them less likely to bind. We will call the inducer as species I in the model.

$$R + I \rightarrow R : I \tag{5.13}$$

$$R : I \rightarrow R + I \tag{5.14}$$

$$R : I + O1 \rightarrow R : O1 : I \tag{5.15}$$

$$R : O1 : I \rightarrow R : I + O1 \tag{5.16}$$

$$R : O1 + I \rightarrow R : O1 : I \tag{5.17}$$

$$R : O1 : I \rightarrow R : O1 + I \tag{5.18}$$

These six reactions deal with all the possible combinations between the repressor, inducer and the operator site. The kinetic constants for all the interactions are solely dependent on the choice of operator, repressor and inducer system. In our networks, we have two of these systems. First, we model the interaction of the *lac* operon, LacI protein, and isopropyl β-D-1-thiogalactopyranoside (IPTG). The other system we look at is the *tet* operon, TetR protein, and anhydro-tetracycline (aTc). Literature values are only available for a few of these systems, and we have made sure to pick from these well-studied systems. [51] For other systems, we have to fit the constants, or use thermodynamic or equilibrium data to approximate the kinetic constants in the model. A final thing we have considered regarding these inducer molecules is the fact that it takes more than one of these molecules to fully induce a conformational change in most systems. In our systems, it takes 4 IPTG molecules to fully induce LacI,a nd 2 aTc molecules to induce TetR. This adds more complexity into the model by causing to consider more possible combinations of molecules (R:I, R:2I, etc.).

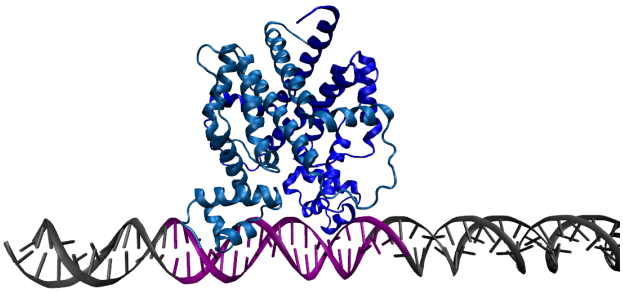

Figure 5.2: TetR binding to DNA strand, without inducer

5.5 Non-specific DNA binding

When we look at the interaction between the operator sites and the repressor molecules, we are truly looking at a protein finding and binding to a few base pairs out the 4.6 million that make up the *E. coli* genome. [52] The large size of this genome makes it less likely that the repressor regulate the transcription of our protein. Thus, we include non-specfic DNA (nsDNA) binding into the model.

$$R + nsDNA \rightarrow R : nsDNA \tag{5.19}$$

$$R : nsDNA \rightarrow R + nsDNA \tag{5.20}$$

$$R : I + nsDNA \rightarrow R : I : nsDNA \tag{5.21}$$

$$R : I : nsDNA \rightarrow R : I + nsDNA \tag{5.22}$$

We must include the possibility that an repressor protein can be bound by inducer molecules, and still have some affinity for DNA, although this affinity will be lower due to the conformation of the protein. Overall, the affinity of the repressor molecule for nsDNA is much lower than for the operator sites. If there was not so much DNA in the cell, this affinity could be neglected. However, with 4.6 million base pairs in the cell,

this small affinity becomes very important. The kinetic constants for these reactions are available in literature for many systems. [53]

5.5.1 Introduction, Degradation and Dilution of Molecules

The final part of the model deals with the process of introducing and degrading molecules. These processes are inherent cell functions that greatly affect the accuracy of the model. When deciding how to represent these reactions, we look to the experimental work done for these systems in order to accurately handle these processes. The introduction of molecules, such as inducer or proteins is handled in one of two ways. First, if the cell produces the molecule internally, we model the process as a 0^{th} reaction.

$$\rightarrow protein \tag{5.23}$$

We can approximate rate constants for these reactions based on literature values, or by measuring the amount of protein inside the cell after a known time from induction. This type of mechanism is used to model the introduction of repressor proteins into the system.

If the molecule is produced outside of the cell, we model it introduction as a first order reaction with a rate equal to the rate of transport into the cell. This is the process followed for the inducer molecules, which experimentally are introduced outside of the cell and must be transported inside.

$$I^{ext} \rightarrow I \tag{5.24}$$

Along with introduction of molecules, we must deal with the loss of molecules due to two mechanisms, degradation and dilution. Degradation is a very simple process in our model. Essentially, all of the species that can be degraded are modeled as a 1^{st} order reaction, where the product is nothing. mRNA and protein are degraded, but DNA is not. This is consistent with the biological system.

$$protein \rightarrow \emptyset \tag{5.25}$$

$$mRNA \rightarrow \emptyset \tag{5.26}$$

Dilution of molecules due to cell division is the second way molecules are lost in the model. We take into account cell growth in the model. Over the period of a cell cycle, the cell's volume doubles in size, decreasing the concentration of all the species in the cell. The algorithm used in this work models cell growth as an exponential increasing volume with a rate constant such that the cell doubles over the cell growth time. This cell growth time is a normally distributed variable with a mean and variance which we choose. Once the cell doubling occurs, all species inside the cell are halved to simulate the splitting of the original cell into two daughter cells. Several species inside the cell are not affected by this split, such as DNA, because a copy of these species was produced before the split, giving us the original amount of molecules after the split. When we combine all of the elementary reactions shown above, we can accurately describe a gene regulatory network. Many types of gene regulatory networks can be modeled using this approach, as long as the correct precautions are made with regards to the number of operator sites and how these sites are repressed. Accurate kinetic constants for the system in question must also be found. The systematic modeling approach described here was used extensively to obtain the results found in this work, and proved to be an accurate and valid approach to the modeling of biological systems.

Chapter 6

Experimental Results for the breakdown of a Biological AND Gate

6.1 Construction of Hybrid Promoters

As stated in previous sections, we experimentally constructed many different synthetic systems in *E. coli*. This experimental work was completed by Dr. Poonam Srivastava, a postdoctoral fellow in Prof. Kaznessis's group. It is vital to understand how these systems were created and the materials and methods used. The first section will be devoted to this topic. Once this has been discussed, we will analyze the results obtained.

6.1.1 Synthesis of Hybrid Promoters

The promoter was synthesized artificially using polymerase chain reaction (PCR). Primers were designed to have an overlap of 20 bp (sequence available in supplementary material). The forward primer for all the constructs was kept the same. Promoters were generated in two steps. First, SOEing PCR was done [54]. The template was generated with two overlapping synthetic oligonucleotides ($\sim$111 bp each) corresponding to the sequence of each synthetic-hybrid promoter. Each oligomer was mixed in equal amounts and incubated with Taq DNA polymerse. The reaction was carried out at 72 °C for

65 minutes to assemble the two oligo-nucleotides. The assembled DNA was purified using Qiagen PCR purification kit. Secondly, nested PCR was completed. Fragments thus obtained were used as template to further amplify the synthetic DNA. This was done with external primers corresponding to the terminal 20 bp of each larger oligonucleotide. A 203 bp DNA fragment was obtained and ran on agarose gel along with the DNA markers to check for the correct size and to trap the DNA for gel elution. After gel purification, each promoter variant was introduced in the pGLOW-TOPO plasmid (Invitrogen, Carlsbad, California) upstream of Cycle 3 GFP. Recombinant plasmids were transformed in Top 10 cells (lacI-, tetR-). Positive clones were screened by visual verification of constitutive GFP expression in Top 10 cells. Selected clones were further inoculated in 10 ml LB medium having 100 g/ml ampicillin. Plasmid isolation was done from these overnight grown cultures and integrity of the promoter sequences was confirmed by DNA sequencing. Plasmids having desired sequences were transformed into DH5αPro (lacI+, tetR+) cells to assess promoter function in the presence of repressors.

6.1.2 Plasmid and Strains

pGLOW-TOPO is a linearized vector with single 3'-thymidine (T) overhang featuring a TOPO cloning site, pUC origin of replication and ampicillin resistance gene. The reporter vector pGlow-TOPO facilitates direct insertion of promoter sequences upstream of Cycle 3 GFP (FACS optimized mutant of GFP) for an in vivo assay of promoter function based on fluorescence output [55]. Initial cloning and sequencing of the hybrid promoters was performed using Top10 cells (Invitrogen). Top 10 cells lack TetR and LacI, thus providing a platform to screen for the positive clones. Analysis of regulatory phenotype was done in DH5αPro cells. These cells constitutively express TetR and LacI from a single copy chromosomal insertion.

6.1.3 Culture Growth Conditions

E. coli strain DH5αPro was used for all in vivo promoter activity assays in the presence of IPTG and aTc inducers. A single colony of freshly transformed recombinant DH5αPro was inoculated in LB medium containing 100μg/ml ampicillin and 50 g/ml spectinomycin. Culture was grown overnight at 37°C with shaking (250 rpm). Next

day these cultures were re-inoculated 1:100 into fresh LB medium having antibiotics and varying inducer concentrations. Inducer concentrations varied from 0-1mM of IPTG and 0-200ng/ml aTc, resulting in a matrix of 18 different inducer pair combinations, each of which was monitored for GFP output over a 24-hour period. At 3, 6 and 9 hour time point samples were taken and the cells diluted 1:10 to restrict growth to logarithmic phase and retain constant cellular parameters (i.e, 70 levels). Samples were monitored for growth state by OD600. Cell growth and division contribute substantially to variability in expression levels. Thus to ensure that the cultures remained in logarithmic growth phase during the experiment, 1:10 dilutions into fresh media were performed at three hour intervals to maintain OD600 $\sim$ 0.2-0.3.

6.1.4 GFP Quantification Using Flow Cytometry

In vivo GFP fluorescence was measured using a Becton Dickinson FACS Calibur flow cytometer equipped with a 488 nm argon laser and a 515-545 nm emission filter (FL1) at low flow rate. Samples were fixed with paraformaldehyde (PFA) to halt GFP production and degradation after harvesting at each time point. One milliliter of cell culture was centrifuged to collect the cells, washed with phosphate buffered saline (PBS) and fixed for 15-30 min in 4% paraformaldehyde (PFA). The fixed cells were re-suspended in 1 ml (PBS). For each sample, 100,000 events were collected and analyzed using FlowJo software (BD Biosceinces). GFP fluorescence detected by the FL1 channel was represented as the mean fluorescence versus the normalized population distribution after subtraction of background fluorescence. Background fluorescence was determined using two sets of controls. First, from non-induced cells harboring functional pGLOW plasmids and second, from a control where all the operator sites are non-functional. Each synthetic-hybrid promoter variants was characterized under identical experimental conditions for comparison of promoter activity and AND logic gate behavior.

6.1.5 Inclusion Body Isolation

Cell pellets were dissolved in Tris Buffer (50 mM Tris-Cl, 5 mM EDTA, 1mM PMSF, pH-8.4) and sonication was done for 12-15 cycles (30 seconds burst/30 seconds cooling, 200-300W). Lysate was centrifuged at 12000RPM, 4 degrees for 30 minutes. Supernatant

was run on SDSPAGE (to check the soluble fraction). Pellets (essentially having the inclusion bodies) were resuspended in Tris buffer having 1% deoxycholate (DOC) or 1% urea. They were then incubated at 37°C for one hour with constant shaking and centrifuged again to get rid of the cell debris. Washing was done again with Tris buffer and washed pellets were resuspended and loaded on SDSPAGE for visualization.

6.1.6 Western Blot

Anti-GFP antibody was used to quantify GFP in the different systems. Inclusion bodies were run on SDSPAGE and transferred on PVDF membrane. The membrane was fixed and saturated with bovine serum albumin and washed with TBST (tris buffer saline, 0.5% tween 20 at pH 7.5). The blot was incubated overnight with anti GFP antibody (abcam cat#291), washed and then incubated with secondary anti mouse antibody. The colored reaction was developed by using alkaline phosphatase. The bands were quantified using gel documentation systems (Biorad).

6.1.7 Real Time PCR

PCR primer pairs were designed to amplify ∼150 bp fragment from GFP gene. Each real-time PCR reaction contained 50 nM primers, ∼1 ng DNA and 1X ABI SYBR PCR reaction mix. A fluorescence value proportional to the initial quantity of target DNA was calculated by a log-linear regression analysis for each triplicate amplification curve [56]. We normalized this value to an input sample, recA, which does not change with change in inducer concentration, to calculate a relative enrichment value of GFP transcripts at different growth stage.

6.2 Results

6.2.1 Construction of Hybrid Promoters

Since natural systems are more context dependent and hence more difficult to model, we constructed a synthetic AND gate using common regulatory elements, broken it down to its individual components and studied the effect of each and every component in the logical behavior of an AND gate system. The methodology behind this has been

previously discussed. The quantitative behavior of the promoter phenotype was studied both in numero and in vivo as a function of its topology. We followed the same design philosophy as in a previous study [29], and continued synthesizing promoters having either one or two functional operator sites which were used in constructing the AND gate. All of the experimental data represents the experimental results obtained after gating the data using forward and side scatter measurements to eliminate experimental artifacts. The gating for all the 100,000 events collected was done using a flow cytometry analysis software program FlowJo. 85-95% of the events were retained and were averaged for all the time points and inducer concentrations. We used this averaged data as the experimental result to inform the fitting of our models. For the T systems, we used two different concentrations of IPTG with 6 different aTc concentrations. For these systems IPTG does not have any impact on the GFP expression, and so we used them as biological replicates. A similar procedure was followed for the L systems, using IPTG as the inducer, and two different aTc concentrations as the replicates. Below we show the experimental results obtained using these methods.

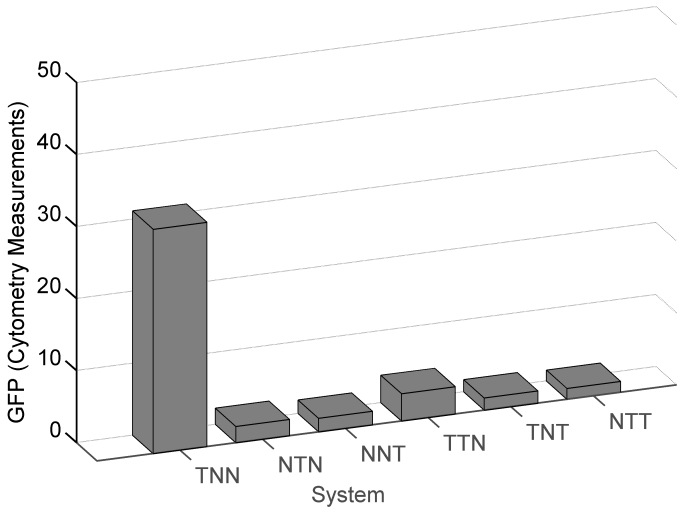

Figure 6.1: Leakiness of Different T Systems. This is the response when the system has no inducer present. For high-fidelity systems, this response should be low. All the systems show low leakiness, except the TNN system.

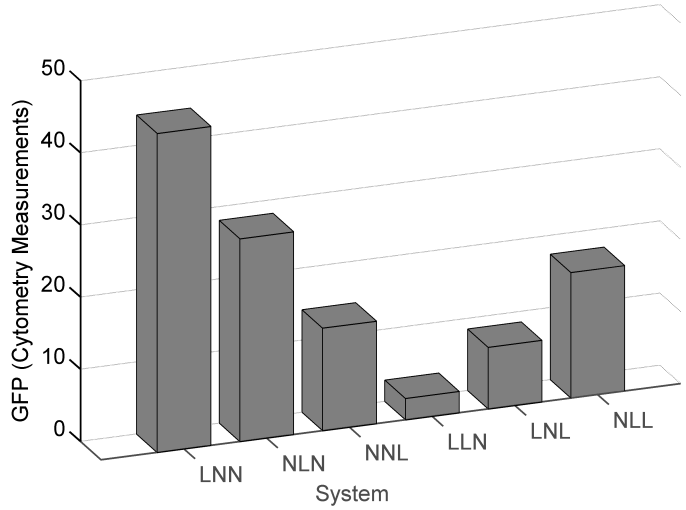

Figure 6.2: Leakiness of Different L Systems when no inducer is present. Overall, the L systems are much leakier than the T systems.

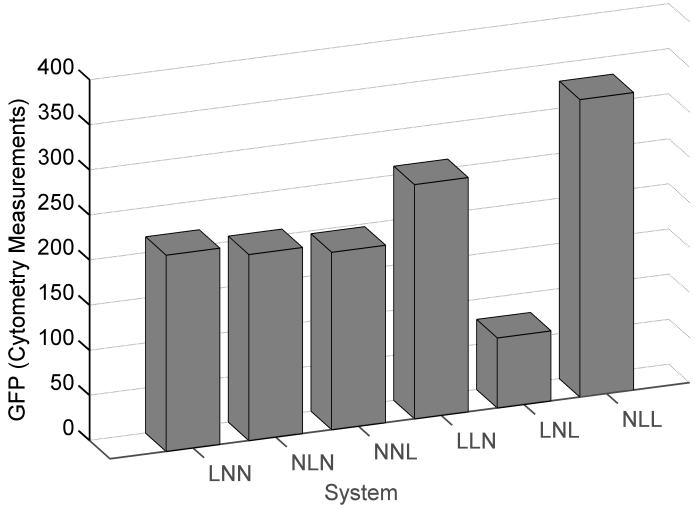

Figure 6.3: Maximum Response of L Systems. This occurred when the maximum concentration of inducer was present in the system.

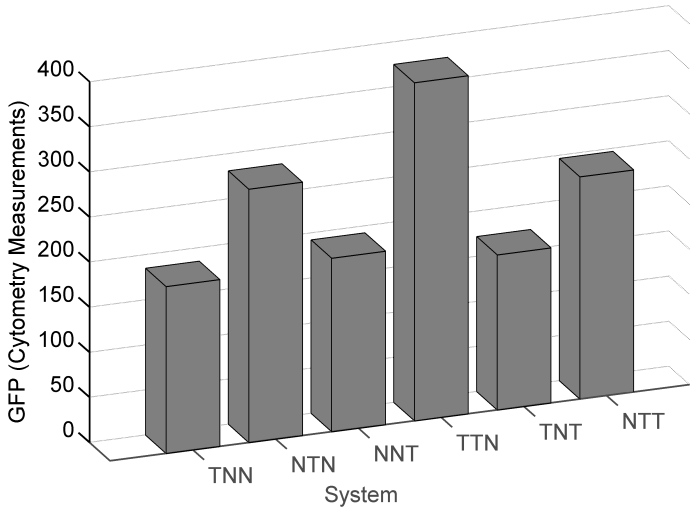

Figure 6.4: Maximum Response of T Systems. This occurred when the maximum concentration of inducer was present in the system.

6.2.2 Importance of Operator Position

In the previous study, it was found that despite the imperfect AND gate phenotype, the double tetO2 systems exhibit varying degrees of AND gate functionality. The systems TTL, TLT and LTT showed GFP signals even at zero inducer concentrations. Figure 6.1 compares the GFP expression at zero inducer concentration for all the single and double T systems. We can see that only the TNN and TTN systems show more than a baseline level of leakiness. This is due to the positioning of the T operator(s) further upstream from the transcriptional start site. Figure 6.2 shows the leakiness of the L systems, and the single L systems show a similar trend to the T systems. This observation contradicts the findings by Bujard where lac placed between the -10 and -35 positions shows maximum repression [57]. We instead see that a single L site close to the transcriptional start site is the most effective repressor. Interestingly, double L systems show the opposite trend, suggesting some cooperativity between L sites. Figure 6.4 shows the GFP expression at highest inducer concentration at 6 hours, where we see that TTN distinctively have the maximum expression compare to other T constructs. Our results agree with previous work, showing the underlying reasons why the TTL

AND gate was found to be the most robust. The double T in the first two positions give high expression and an L in the last position gives low leakiness. Overall, positioning of tet or lac played an crucial role in maintaining the fidelity of AND gate system TTL.

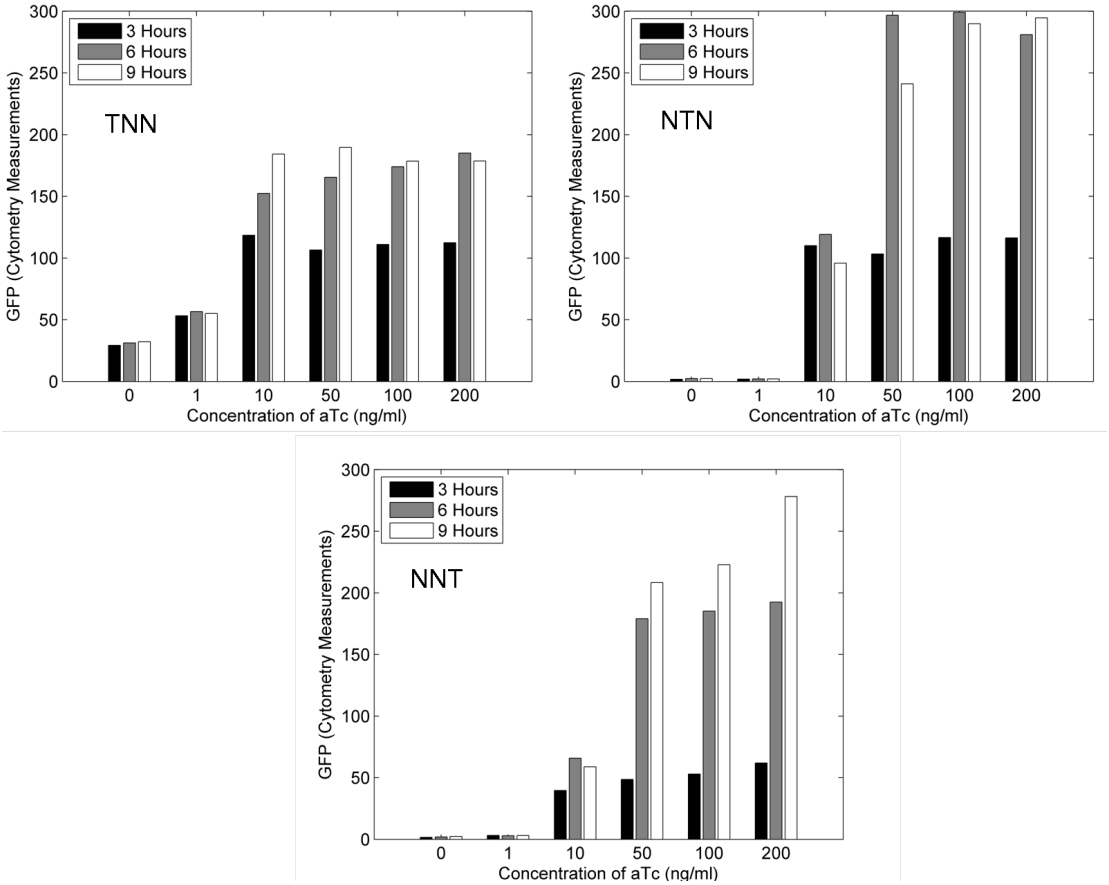

Figure 6.5: Response of Single T Systems. Each bar represents the flow cytometry measurements at a specific time point for a specific inducer concentration. These measurements quantify the amount of reporter protein in the system under these conditions.

6.2.3 Single T systems

Figure 6.5 shows the experimental response of the single T systems. Lower inducer concentrations are sufficient to induce the expression and there is negligible change in expression from 10 to 200 ng/ml of aTc. There is not much increase in the expression

after 3 hours of induction. Figure 6.6 shows that almost 40 percent of the cells are in OFF condition and does not show any GFP signal at 9 hours. Recombinant cells always try to get rid of the metabolic burden caused by the over expression of the recombinant protein by various ways [58, 59]. Among these plasmid instability and inclusion body formation are the most common reasons [60,61]. Therefore first we checked the plasmid instability by dilution plating on the antibiotic plates. Cell culture after 9 hours of induction was plated and viable colonies were counted, plasmid was isolated from these cells to check for any loss. Restriction mapping of these plasmids further confirmed that plasmid instability was not a problem here. We next focused on isolating the inclusion bodies from the induced cells. Cells were lysed by sonication and inclusion bodies were isolated as described in materials and methods and were loaded on SDSPAGE. It was found that system does form protein aggregates which were unable to show any fluorescence signal due to misfolding.

This result explains the lower GFP signal of the TNN variant compare to NTN and NNT (Figure 6.5). NTN system does not show any GFP signal at 0 and 1ng/ml of aTc, but expression improves at higher inducer concentrations though remains same for 50-200 ng/ml. Both of these trends are accurately modeled by our detailed reaction network. The systems reaches a steady state in 6 hours. Compare to TNN and NTN, NNT system showed strongest repression as GFP build up is slow and did not reach steady state till 9 hours. Expression increases with time and at higher inducer concentration. The threshold of induction is 100 ng/ml compare to 10 and 50ng/ml for TNN and NTN systems respectively. This shows that the repressor binding at position downstream to -10 hexamer shows tightest regulation which again is an interesting deviation from Lutz and Bujard [22].

6.2.4 Double T system

TTN, TNT and NTT were studied under the same conditions as the single T systems. It was observed that none of the double T systems show leaky expression (Figure 6.7). This is in contrast to the single T systems, where the TNN system did show this trait. To represent this in the models, we set the reaction rate constant for the leakiness reaction (Reaction 2) in all of these systems to zero. The fact that there is no leakiness of any of these systems is because of the unlikeliness of both operator sites being free at the same

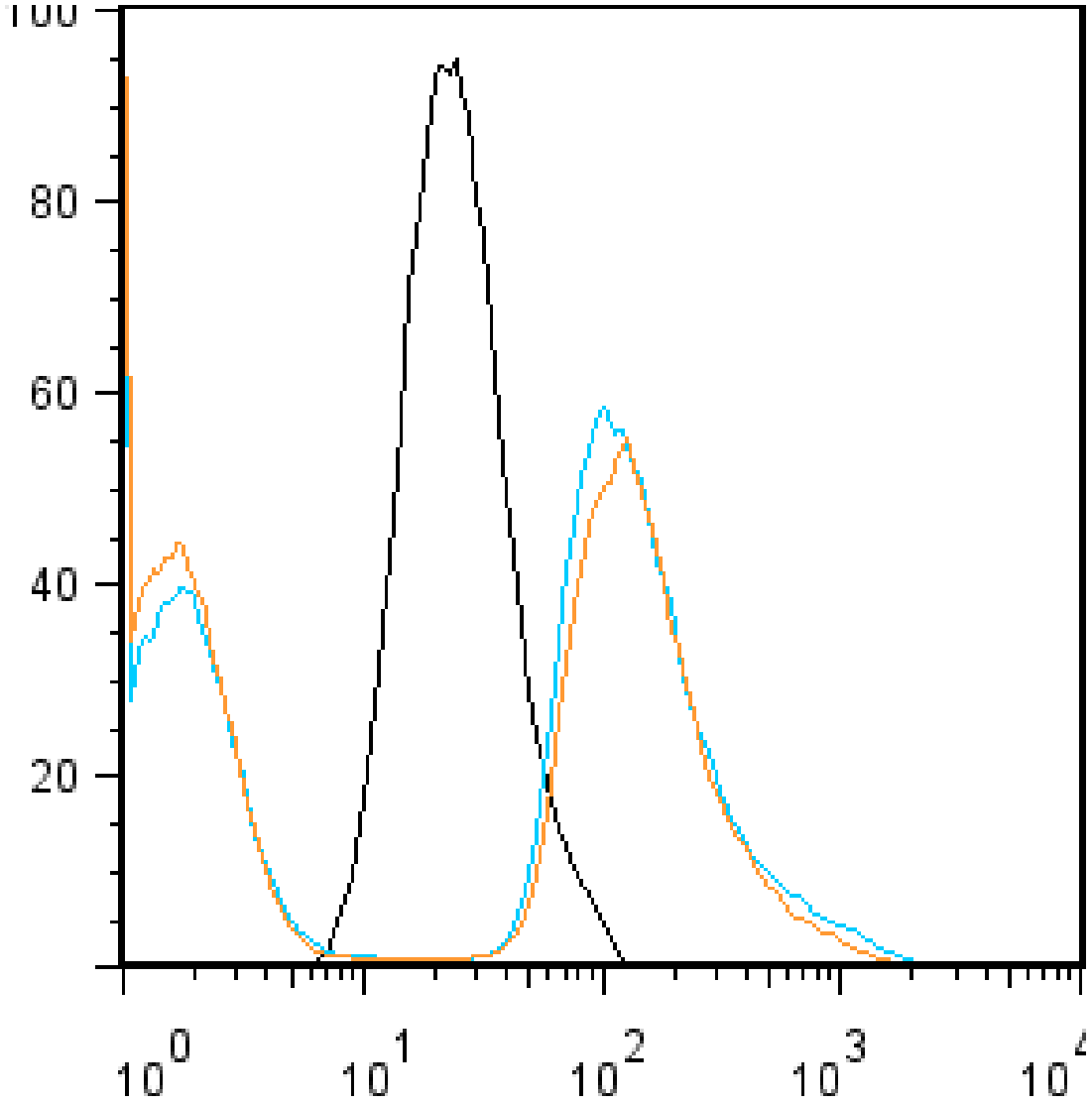

Figure 6.6: This graph shows the data from FlowJo which shows inclusion bodies. Note that there are two populations of cells, one which shows no fluorescence due to inclusion bodies, and one which shows high expression because of the lack of inclusion bodies in these cells.

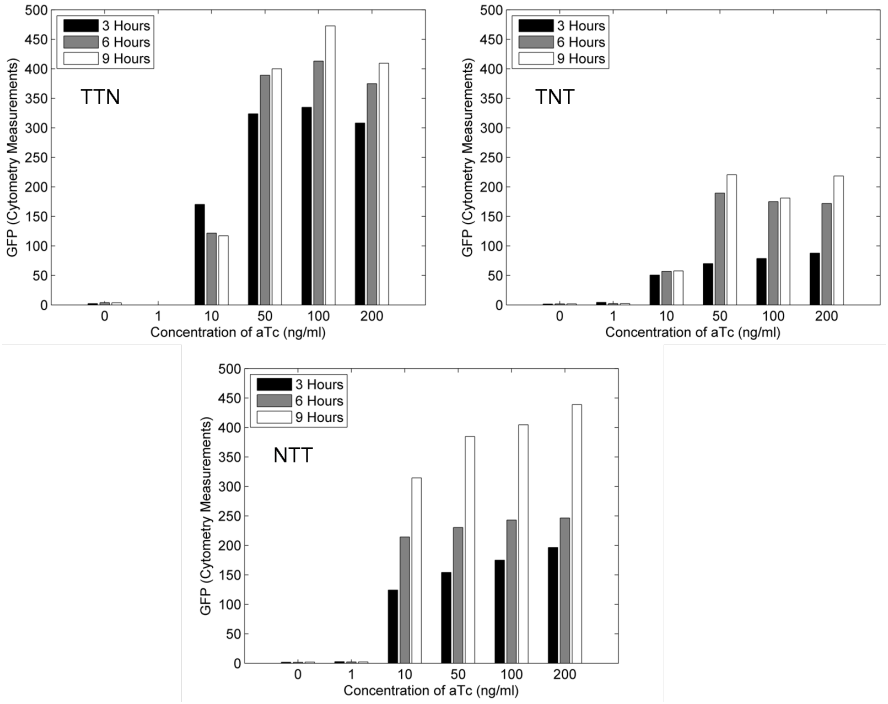

Figure 6.7: Experimental Results for Double T Systems. Each bar represents the flow cytometry measurements at a specific time point for a specific inducer concentration.

time. In a single T system, only one operator site must be unbound for transcription to occur, which is more probable than both sites being free at the same time. For TTN variant, a very strong induction was shown at the highest aTc concentrations. In fact, this system showed the highest flourescence of any of the T constructs. Compared to TTN and NTT, the GFP expression level for TNT remained low even for the high inducer concentration. This was backed by the models, which showed lower promoter strength in this system compared to the other systems. For NTT system, we observed a consistent increase in GFP fluorescence with time and inducer concentration.

6.2.5 q-RTPCR

To make sure that indeed it is the change at translational level which is affecting the overall fluorescence we performed quantitative real time PCR for the constructs showing varying level of fluorescence under same conditions. This experiment measured the mRNA molecules present in the six T systems, and shows the systems with the highest response, such as the TTN and NTT, have higher levels of mRNA (Figure 6.8). This shows that promoter strength is indeed the primary factor in determining the GFP response of the system. Besides this one difference, all other steps in the molecular biology dogma are equivalent in the systems.

According to Haseth et al. [48], the nucleotide sequence and length of the sequence present in between -35 and -10 region is decisive for the RNA polymerase binding to the promoter. In all the hybrid AND gate promoter systems the length of the operator sequence is 17 bp, but the nucleotide sequence changes according to the design. Therefore we conclude that the positioning of non coding sequence (N) in between -35 and -10 decreases the RNA polymerase activity leading to lower GFP fluorescence. This was further confirmed by quantitative real time PCR. The recombinant plasmids were transformed in DH5αPro cells (-tetR,-lacI) and RNA isolation was done for the overnight grown cultures. cDNA was prepared as described in materials and methods. Cycle threshold for GFP was compared with cycle threshold of 16 rRNA (internal control) and it was found that single T systems (NNT and TNN) and TNT show $\sim$1.8 fold less transcription compare to TTN and NTT.

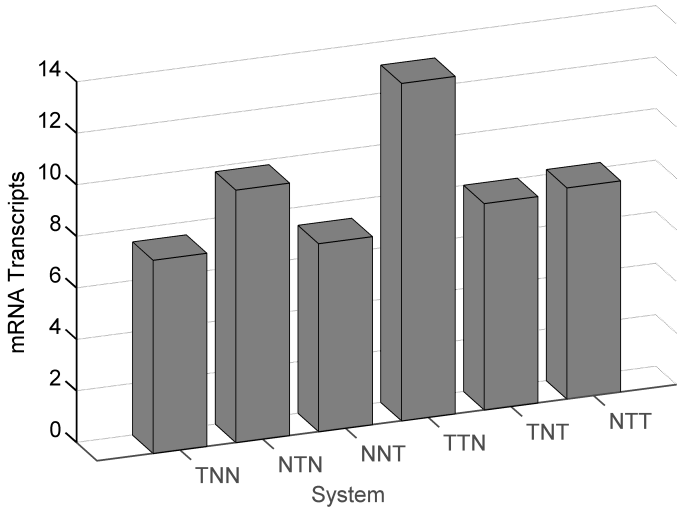

Figure 6.8: qRT-PCR Results for the six systems.

6.2.6 Single L systems

Along with characterizing and modeling systems that only responded to aTc, we also analyzed systems which responded only to IPTG. Analogous to the T systems, we studied three single L and three double L systems. Figure 6.9 shows the experimental response of the single L systems. From this figure, we can see that all of the constructs LNN, NLN and NNL had leaky expression. The most leaky construct is LNN and least leaky is NNL. This again confirms that position of operator sequence has a major role in regulation, confirming what we found in the T systems. DH5αPro cells have extra copies of lacR gene that does not allow expression of lacY gene (permease) which is present in single copy; hence there is slow diffusion of IPTG. Therefore IPTG diffuses slowly inside the cells providing slower induction when compared to aTc. Unlike TNN, all the single L systems despite being leaky did not show any inclusion body formation. A possible reason is the slow build up of protein because of the slow induction.

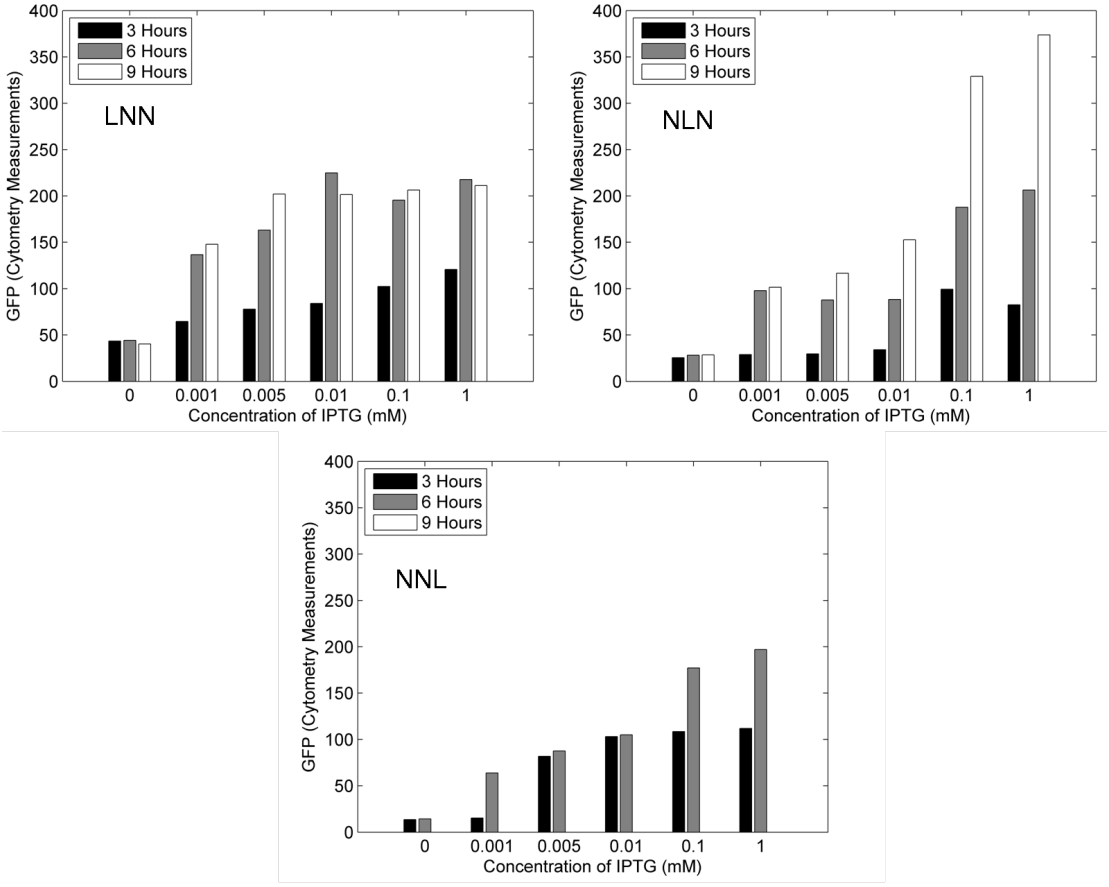

Figure 6.9: Experimental Response of Single L Systems. Each bar represents the flow cytometry measurements at a specific time point for a specific inducer concentration.

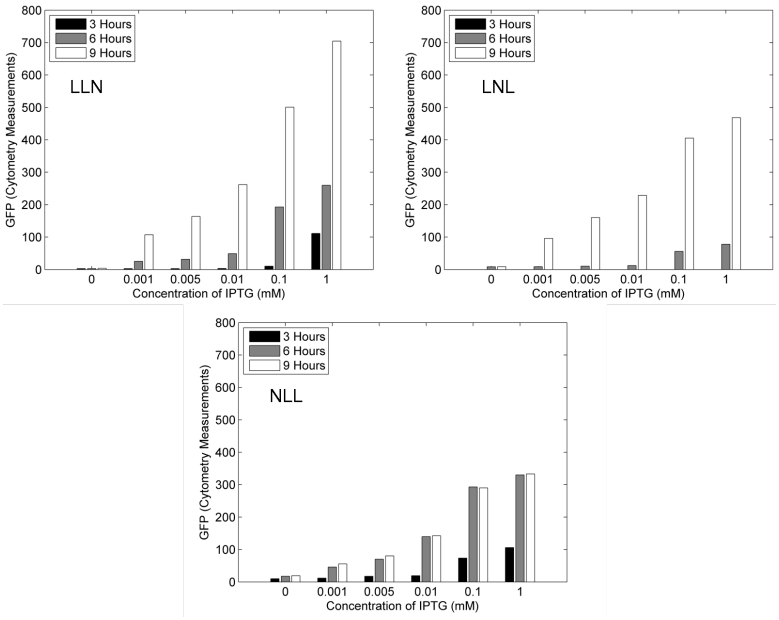

Figure 6.10: Experimental Response of Double L Systems. Each bar represents the flow cytometry measurements at a specific time point for a specific inducer concentration. The LNL is a slow-responding system, shown by the low bars at the first two time points.

6.2.7 Double L systems

Figure 6.10 shows GFP fluorescence plotted against six IPTG concentrations for the double L systems. As is show by the figure, all the constructs have almost negligible GFP fluorescence at zero inducer concentration. This is similar to the double T systems which also show very little leakiness. With two sites for repressor molecules to bind, it increases the chance that the promoter will be bound by at least one repressor, inhibiting transcription. It is also important to note that in all the constructs GFP build up was very slow. According to Dunaway et al. [51], in the presence of saturating amount of operator DNA sequences, Lac repressor shows 20 fold lower affinity for inducer. In our system, synthetic promoters for all the constructs are cloned in high copy number plasmid, providing saturating amount of operator sequence and hence forcing Lac repressor to have lower binding affinity for inducer. In the previous study with LLT, LTL and TLL, no GFP expression was observed for 6 hours; this could be because of the very slow induction of the system. Overall, we can see that LLN is the most active promoter. In 9 hours the fluorescence peaks at $\sim$700 fluorescence units, for maximum inducer concentration (1mM IPTG). The increase in inducer concentration with time definitely increases the GFP expression. For the LNL and NLL constructs, fluorescence peaks $\sim$400. This could be because of the positioning of the operator site in the promoter, where system needs more time to get induced and to reach a steady state.

6.2.8 T System Mutants

In order to explore the robustness of the AND gate system, we studied our systems with mutated repressor or operator sites. We used two of the previous systems, but with mutant operator sites in place of the wild type operators. Due to experimental limitations, such as the availability of mutant tetR expressing strains, it is much easier to engineer and use mutated operator sequences instead of mutant repressor molecules. It is also difficult to assess the changes in structural and functional properties of the repressor protein due to amino acid sequence change. We decided to use mutant operator sequences characterized by Sizemore et al [62]. The reference describes the mutations in tetO2 operator sites and the in vivo binding of the mutants to wild type TetR in the

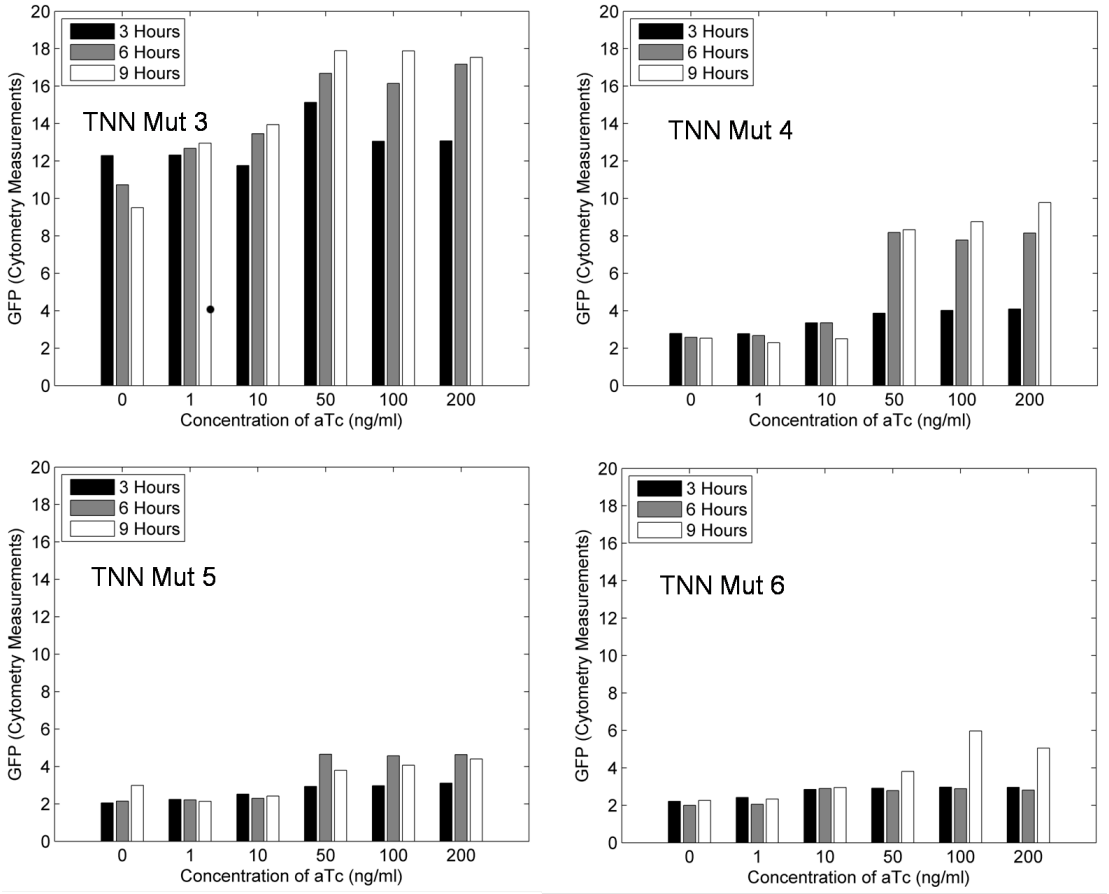

Figure 6.11: Experimental Response of Mutants of the TNN System. The response of these systems was an order of magnitude lower, due to poor binding of RNA polymerase on the mutant promoter.

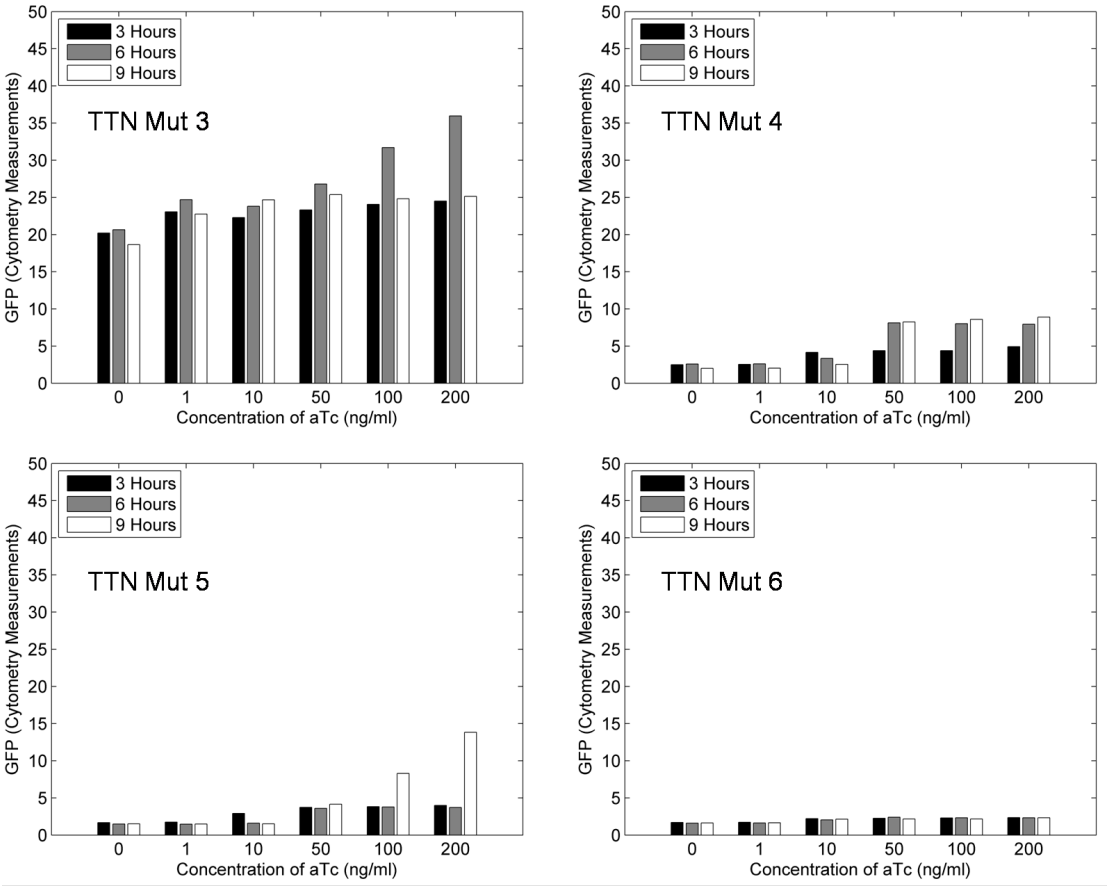

Figure 6.12: Experimental Response of Mutants of the TTN System. Similar to the single T systems, the response of these systems was an order of magnitude lower than the wild type due to poor binding of RNA polymerase on the mutant promoter.

Table 6.1: Results from a Previous Study by Sizemore et al., showing the activities of mutants we used in our study. Note that all of these mutants bind TetR well, showing at least 70% the repression of the wild-type sequence.

β-galactosidase Activity			
	-tetR	**+tetR**	**Repression(%)**
pWH1012(WT)	100±4	2±4	98%
pWH1012(3T)	88±6	52±4	41%
pWH1012(4A)	16±0.9	2±0.3	81%
pWH1012(5T)	8±0.7	1±0.3	87%
pWH1012(6C)	72±0.1	20±1	73%

presence and absence of repression. Table 6.1 summarizes the results obtained in their experiment and these results are the basis we used for our analysis. We have used mutants which have transverse point mutations in 3,4, 5 and 6th position on palindromic sequence. These mutant operator sites were used to construct 8 additional hybrid promoters. TNN and TTN constructs were created for all the 4 mutants. The promoter is constructed by artificial gene synthesis as described before, using mutated sequences. The positive colonies showing maximum GFP fluorescence were first screened in TOP10 cells. Recombinant plasmid was isolated from these cells, sequenced for further confirmation and then transformed into DH5αPro cells for study by flow cytometry. Culture was induced using 6 different concentrations of aTc and samples were collected for 3, 6, 9 and 12 hour time points post induction.

Figures 6.11 and 6.12 describes the flow cytometry results obtained after gating the rough data in FlowJo. As it is clear from the figure, all the mutants show low fluorescence when compared to all the wild type constructs. Reporter activity is highest in mutant 3 in both the presence and absence of repression. Constructs TTN and TNN of mutant 3 showed very leaky expressions. Fluorescence can be seen at zero inducer concentration for all the time points, meaning these constructs are leaky. Increases in expression are almost negligible with time and with higher inducer concentrations. Overall, we see that the TTN construct is comparatively more active than TNN. In table 6.1, Sizemore et. al report that mutants show less reporter activity compare to wild type operator [62]. Instead of using GFP, β-galactosidase was used as a reporter gene. This assay

is done by using a colorimetric assay where substrate, upon enzymatic degradation, releases colored products. The enzyme is obtained from the whole cell lysate [63]. A lactose analog, o-nitrophenyl-β-galactopyranoside (ONPG), is cleaved by β-gal to o-nitrophenyl (ONP). This molecule has a different absorbance than ONPG and can be quantified by using spectral photometer. This reporter assay is very different from the fluorescence measurements in our study. Enzyme kinetics for β-galactosidase depends on factors which are very different from fluorescence measurements. Hence the percentage of repression for the mutants could vary when GFP is used as reporter. Our system's behavior follows the description by Sizemore et. al. somewhat and the low fluorescence for mutants establishes that weaker operator sites would change the behavior of an AND gate. Since the fluorescence obtained was so low for all the constructs that it was not clearly established to what extent these mutants would affect the robustness of an AND gate.

Chapter 7

Modeling Results for the breakdown of a Biological AND Gate

7.1 Modeling Overview

In Chapter 5, we discussed the systematic method followed when constructing models of biological systems. This section will give the specific sets of reactions and rate constants used, along with results found when comparing experiments and models. We will discuss all of the different systems studied, starting with the 6 T systems, and then looking at the 6 L systems. All of the following comparisons were made by using average values for both experiments and the models. We found this value for the experiments by using the gated populations of 100,000 cells, and then finding the mean value of the fluorescence. For the models, 1000 trials were run for each experimental condition, and then the mean of these trajectories was used to compare to the experiments. All of the simulations run were stochastic in nature, and used the Hy3S algorithm developed by Salis and Kaznessis. [46]

As stated in previous sections, we chose synthetic systems in order to isolate differences in the DNA sequence of the promoter for the GFP gene. Therefore, any changes in the system's response are due solely to the position of the operator site within the

overall promoter. It is important that our models have consistent rate constants for all reactions except for two reactions. These reactions are the leakiness reaction and the promoter strength reaction.

The leakiness quantifies the promoter's response at low inducer concentrations. If the system was digital, we would expect the system to show no response, due to the repressor proteins fully occupying the promoter sites. RNA polymerase cannot bind in this instance, and no GFP protein would be produced. However, biological systems are not as robust as digital systems, and some fluorescence is observed. We will discuss the reasons behind this in the model description.

The promoter strength reaction represents the expression of the system at high inducer concentration. When the concentration of inducer is high, all of the repressor protein in the cell is bound, and it has a very low affinity for DNA. This means that the operator sites are free to bind to RNApol at any time. The binding strength of RNApol to the particular DNA sequence is the only interaction which determines the fluorescence of the system. This is called the promoter strength.

Aside from the two reactions mentioned above, all the other kinetic rate constants were determined from literature and remained constant across all the 12 systems. We can change only biologically relevant reactions in our model, and accurately capture the experimental behavior. The accuracy of our approach will be shown in the following sections.

7.2 Single T Systems

The first systems we will discuss are the single T systems. These systems all had a single T operator site, and there were three possible locations for the operator (upstream of the -35 box, in between -35 and -10 boxes, and downstream of the -10 box). This promoter design was discussed extensively in Chapter 3. This gives us the three single T systems: TNN, NTN, and NNT. A schematic of these systems is shown in Figure 7.1.

In Chapter 5, we discussed the systematic approach that was used to model these biological systems. Table 7.1 shows the exact reactions used, along with the rate constants and the sources of these kinetic rates. In several instances, only equilibrium values were available, and fitting was used to find the exact forward and reverse reaction rates used

Figure 7.1: Schematic of the single T systems. These systems were all modeled and compared to experimental results.

in the model.

7.2.1 Description of Model

Modeling the single T systems required 36 different reactions and 23 species to be included in our models. The reactions and species are shown in Table 7.1. As stated before, care was taken to include all steps of the molecular biology dogma, from transcription to translation, along with including repression and degradation. We will explain the network for the single T systems in detail. The models for the other systems are extensions of this model.

Transcription was the first process considered, as this was the source of control in the AND gate. This component of the model is denoted by reactions one to eight. Starting from RNApol binding to the free promoter, these reactions describe the initiation of transcription, the transcription elongation, and the binding of the mRNA to the ribosome to begin translation, along with the appropriate unbinding reactions. Unbinding can still occur during this process because of stochastic fluctuations in the cell. Reaction nine denotes the translation of the mRNA to the GFP protein. Reaction 2 is the leakiness reaction discussed above.

The next set of reactions, 10 - 23, addresses TetR protein binding to the operator site and its inducer, aTc. Each TetR molecule can bind up to two aTc molecules at a time, and requires thirteen reactions to describe. TetR can bind the tetO operator when it is bound to no aTc, one aTc, or two aTc. All of these possibilities must be accounted for, along with the appropriate unbinding reactions. Without all of these reactions, we could not accurately capture the regulatory behavior of TetR and aTc. Note that the unbinding constants of reactions 15, 17, 19 represent the ability of the aTc molecules to lower the affinity of TetR for tetO. These unbinding constants increase with the addition of aTc, as TetR is more likely to dissociate from the DNA with bound aTc.

Table 7.1: This table shows all the reactions involved in the single T system modeling.
* References gave equilibrium values, so forward and reverse rates are chosen.
** Degradation rate matches the dilution rate with a 60 minute cell division time.
‡ The forward and reverse rate of gfp mRNA are adjusted to give 20 polypeptides per mRNA transcript.
§ This value gives a 6 hour half life for gfp.
ϕ These reactions are gamma-distributed, since it is the addition of consecutive Poisson distributed reactions.

Reaction Number	**Reaction**	**Kinetic Constant**	**Source**
1	$RNAp + tetP + tetO1 \rightarrow RNAp : tetP$		Fitted
2	$RNAp + tetP + tetR2 : tetO1 \rightarrow RNAp : tetP + tetR2$		Fitted
3	$RNAp : tetP \rightarrow RNAp : tetP*$	10^{-2}	[49]
4	$RNAp : tetP \rightarrow RNAp + tetP + tetO1$	1	[48]
5	$RNAp : tetP* \rightarrow tetP + tetO1 + RNAp : DNAgfp$	30	[47]
6	$RNAp : DNAgfp \rightarrow RNAp + gfp_m RNA$	30	[47]ϕ
7	$gfp_m RNA + rib \rightarrow rib : gfp_m RNA$	10^6	‡
8	$rib : gfp_m RNA \rightarrow rib : gfp_m RNA_1 + gfp_m RNA$	33	[47]
9	$rib : gfp_m RNA_1 \rightarrow rib + gfp$	33	[47]ϕ
10	$tetR2 + aTc \rightarrow tetR2 : aTc$	10^8	[64]*
11	$tetR2 : aTc \rightarrow tetR2 + aTc$	10^{-3}	[64]*
12	$tetR2 : aTc + aTc \rightarrow tetR2 : aTc2$	10^8	[64]*
13	$tetR2 : aTc2 \rightarrow tetR2 : aTc + aTc$	10^{-3}	[64]*
14	$tetR2 + tetO1 \rightarrow tetR2 : tetO1$	10^8	[65]*
15	$tetR2 : tetO1 \rightarrow tetR2 + tetO1$	10^{-3}	[65]*
16	$tetR2 : aTc + tetO1 \rightarrow tetR2 : tetO1 : aTc$	10^8	[64]*
17	$te1R2 : tetO1 : aTc \rightarrow tetR2 : aTc + tetO1$	1	[64]*
18	$tetR2 : aTc2 + tetO1 \rightarrow tetR2 : tetO1 : aTc2$	10^8	[64]*
19	$tetR2 : tetO1 : aTc2 \rightarrow tetR2 : aTc2 + tetO1$	10^6	[64]*
20	$tetR2 : tetO1 + aTc \rightarrow tetR2 : tetO1 : aTc$	10^8	[64]*
21	$tetR2 : tetO1 : aTc \rightarrow tetR2 : tetO1 + aTc$	10^{-3}	[64]*
22	$tetR2 : tetO1 : aTc + aTc \rightarrow tetR2 : tetO1 : aTc2$	10^8	[64]*
23	$tetR2 : tetO1 : aTc2 \rightarrow tetR2 : tetO1 : aTc + aTc$	10^{-3}	[64]*
24	$tetR2 + nsDNA \rightarrow tetR2 : nsDNA$	10^3	[53]*
25	$tetR2 : nsDNA \rightarrow tetR2 + nsDNA$	3.2409	[53]*
26	$tetR2 : aTc + nsDNA \rightarrow tetR2 : aTc : nsDNA$	10^3	[53]*
27	$tetR2 : aTc : nsDNA \rightarrow tetR2 : aTc + nsDNA$	3.2409	[53]*
28	$aTc_e xt \rightarrow aTc$	$3.3 * 10^{-4}$	[66]
29	$\rightarrow tetR2$	10^{-11}	[29]
30	$tetR2 \rightarrow$	$2.89 * 10^{-4}$	[29]
31	$tetR2 : aTc \rightarrow aTc$	$2.89 * 10^{-4}$	[29]
32	$tetR2 : aTc2 \rightarrow 2aTc$	$2.89 * 10^{-4}$	[29]
33	$gfp_m RNA \rightarrow$	$1.16 * 10^{-3}$	‡
34	$gfp \rightarrow$	$3.21 * 10^{-5}$	§
35	$tetR2 : aTc : nsDNA \rightarrow nsDNA + aTc$	$1.93 * 10^{-4}$	**
36	$tetR2 : nsDNA \rightarrow nsDNA$	$1.93 * 10^{-4}$	**

The model must also reflect that the operator sequence of interest is one section of over 4.6 million DNA base pairs in the E. coli cell [52]. Therefore, most of the time the TetR protein will be bound to a different section of DNA despite its high affinity to tetO. This aspect of DNA binding is represented reactions 24 - 27, which give non-specific DNA binding and unbinding constants for TetR and TetR:aTc.

Reactions 28 and 29 are the two reactions corresponding to the entry of molecules into the system. In order to best mimic what is occurring in the experiment, aTc is initially considered to be external to the cell. Over time, it is transported into the cell, using a rate determined by the literature [66] TetR is constitutively produced by the strain of E. coli being used, and reaction 29 represents this by making the protein appear in the system at the correct rate. All of the other components in the system are produced by the other reactions, or their concentrations are considered to be constants. If the concentrations of the species are considered to be constants, such as ribosomes, then an initial condition is set before the simulations are run.

The final part of the model, reactions 29 - 36 deals with dilution of species by cell division and degradation which naturally occurs in the cell. In our model, these two processes are considered to be complementary, with both leading to molecules disappearing from the system. We have calculated kinetic constants that cause proteins to have realistic half lives, such as the 6 hour half life for GFP. Other molecules degradation constants are taken from literature, or fitted to give a realistic steady state number of molecules in the system.

7.2.2 Protein-Dependent Cell Growth

When we model these systems, we input the pertinent reactions along with their respective rate constants. However, we also must account for the cell volume and cell division time to accurately model the concentrations of all the species throughout time. When doing the experiments, precautions are taken to maintain the *E. coli* cells in exponential growth at all times. This allows us to give the system an initial cell volume ($10^{-}15$ liters), and increase this volume exponentially with a certain rate. This rate is based on the fact that just before the cell divides again, its volume will have doubled. Thus, it is necessary that we input an accurate cell growth time to have an correct concentrations in the system.

It is important to note that cell's growth rate is based on several factors, such as nutrients available and the metabolic burden placed on the cell. If the cell has few nutrients, or is producing a lot of protein, there are fewer resources to devote to growth, and it will divide slower. In our systems, we keep the cell's environment constant, and growth is not inhibited by the lack of nutrients. However, we are certainly burdening our cells with the production of GFP protein. This slows the cell division time from 30 minutes to about 70 minutes when the fluorescence is highest.

$$\text{CellGrowthTime} = \text{Original Growth Time} + \text{Cell Growth Constant} * \text{GFP Amount} \tag{7.1}$$

We account for this change in the cell division time using a linear approximation, shown in equation 7.1. After each cell division, a new growth time is calculated using this equation. We use an original cell growth time of 30 minutes, and cell growth constant of .057. This reflects what is occurring biologically, and makes our models more accurate.

7.2.3 Analysis of Single T Models

We will now look at the results obtained when using the models described above. Figure 7.2 shows the comparison between the modeling and experimental result. Overall, we can see very good agreement for all the systems across the range of aTc concentrations. The NTN system shows the highest free promoter response, and our models promoter strength reaction (reaction 1 above) had the largest rate constant. This is what we would expect for the system with the highest promoter strength. This trend continues for the other two systems, NNT and TNN, which shows similar maximal responses along with similar promoter strengths in our models. We also see that the TNN system is the only system which shows a significant amount of leakiness, due to the operator site position far from the transcriptional start site. The rate of the leakiness reaction is much higher for this system than the other in the models. This match between the experiments and models is a proof of concept that our detailed models can capture the biological behavior in these systems.

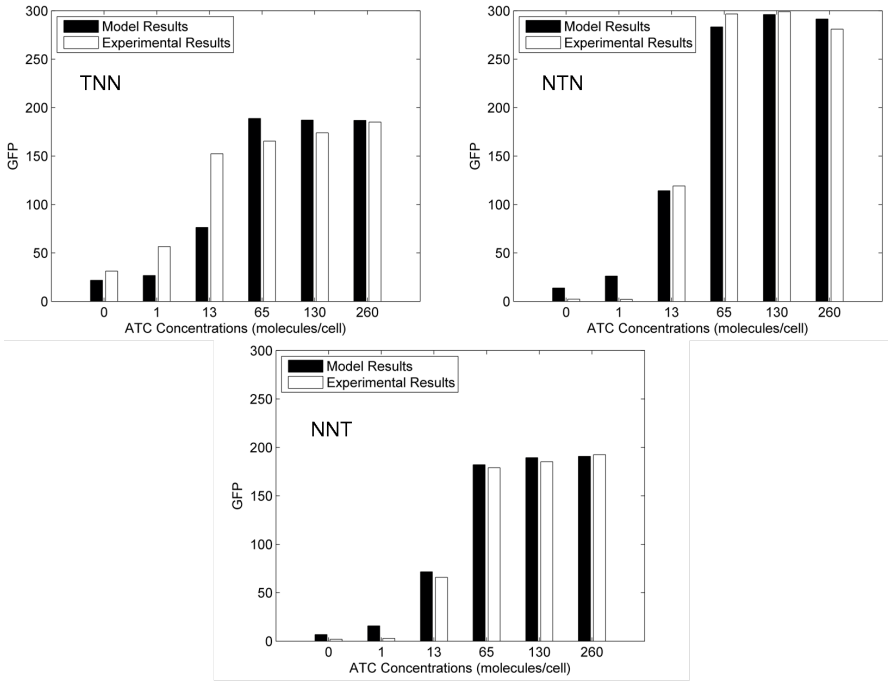

Figure 7.2: Modeling Results for Single T Systems. The models are able to closely match the experiments across a range of inducer concentrations.

Table 7.2: Characteristics of T Systems. Note that the units are different because of the different orders of reactions in the single and double T systems.

System	**Promoter Strength (Reaction 1)**	**Leakiness (Reaction 2)**
TNN	$.0036\frac{L^2}{mol^2*s}$	9*10-5
NTN	$.0058\frac{L^2}{mol^2*s}$	0
NNT	$.00365\frac{L^2}{mol^2*s}$	0
TTN	$.0085\frac{L^3}{mol^3*s}$	0
TNT	$.0035\frac{L^3}{mol^3*s}$	0
NTT	$.0045\frac{L^3}{mol^3*s}$	0

7.3 Double T Systems

We will now consider the other 3 T systems, which are the systems with two T operator sites. We refer the these as double T systems. TTN, TNT, and NTT were the systems considered. The double T models contain all the components present in the single T systems, along with additional reactions to deal with the binding of the repressor to the additional Tet operon. These reactions are quite similar to reactions 10 -23 in Table 1. Adding these reactions to the single T models, these networks have 49 reactions with 26 different species. When compared to the single T models, these models have the same kinetic constants for many of the reactions, as these constructs were created in the same cell strain, with the same initial conditions. The only reactions that have been changed between the different systems is the promoter strength and the leakiness of the system, although none of the double T systems show much leakiness.

We are able to match the experimental results quite accurately across all of the aTc concentrations for all the double T systems, similar to the single T systems (Figure 7.3). If we compare the promoter strength which best fits the experiments, shown in table 7.2, we can see that the system with the highest expression, TTN, also shows the highest promoter strength. This is expected because the strength of transcriptional initiation is the only difference between the different T systems.In fact, this system showed the most expression when compared with the other 11 systems studied. Thus, T operators in this position would be desirable components of a full AND gate.

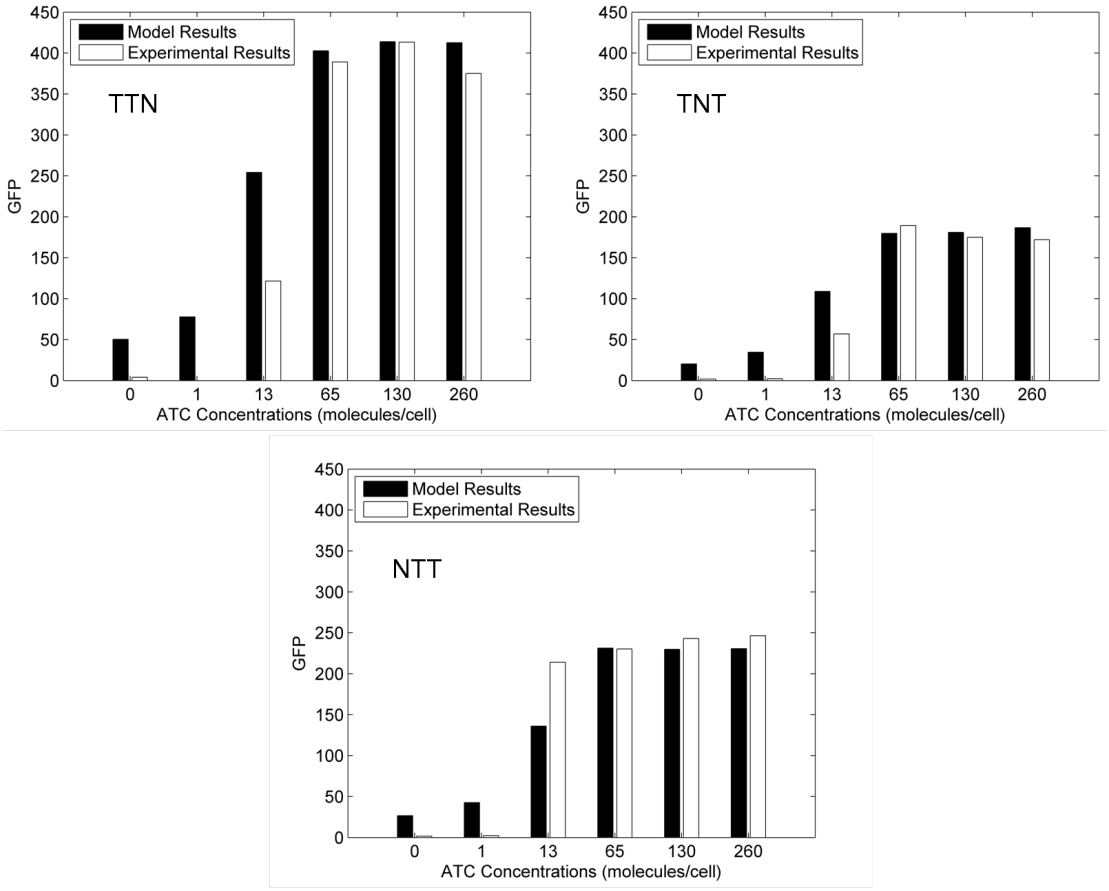

Figure 7.3: Modeling Results for Double T Systems. We are able to match the system accurately at high and low aTc concentrations.

7.4 Single L Systems

We also modeled the L systems, using the same modeling formalism as the T systems. These systems are similar to the T systems, although they use different DNA sequences, and different repressor and inducer molecules. The DNA sequences are preferentially bound by the LacI protein, instead of TetR protein. IPTG is the inducer in these systems. When this small molecule binds to LacI, it causes a conformational change in the protein, lowering its affinity for DNA and relieving repression in the system.

Because of the different molecules in the system, the species in the reaction network had to be changed substantially. However, this change is mainly cosmetic, as the system operates identically to the T system, albeit with different molecules.

Similar to the single T systems, the single L systems have only a single operator site, and this operator site can be in three possible positions. These three possible positions give us the LNN, NLN, and NNL systems. These systems were all constructed in the same cell strain, and subjected to the same conditions. For the single L systems, the reaction networks have 30 reactions, and 20 species.

For the single L systems, the models were able to capture the experimental response at the highest and lowest inducer concentrations, although they were not as accurate for the concentrations in between. This is due to some of the experimental results showing higher flourescence at a lower inducer concentration, which is a an uncommon result in the systems we studied. Experimental error and could also play a role in these results. In order to reduce this error, more biological replicates of the L systems should be done. Even so, we are still able to capture the leakiness at zero inducer concentration, and the maximum response at the highest inducer concentration, which are the two reactions we change in the models.

Although we have taken great care to eliminate any differences between the systems, it is possible that the L operators or the introduction of IPTG into the system cause unforeseen changes in the cell. This would cause the rates of the majority of the reactions, which remain constant in our models, to be inaccurate. This potential discrepancy, along with experimental error, are the major things which cause our models to be less accurate for these single L systems.

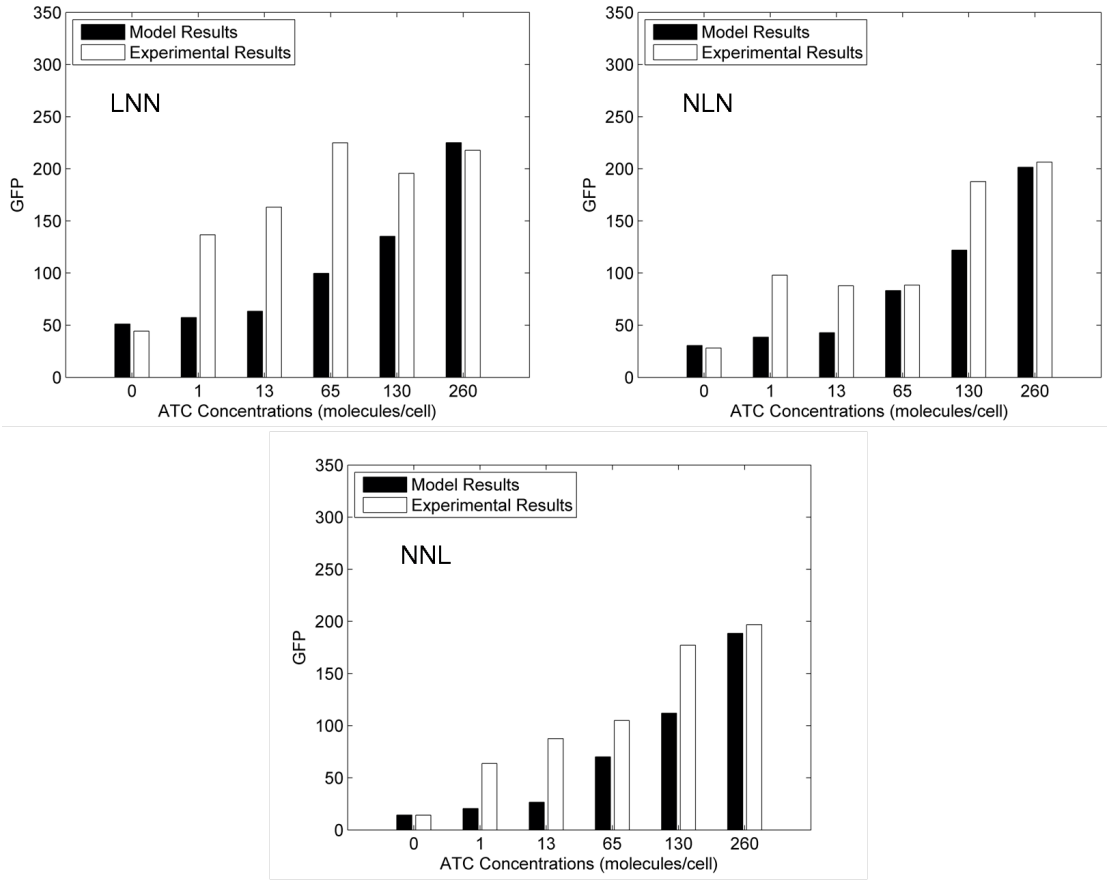

Figure 7.4: Comparison of Models and Experiments for the Single L Systems. For the L systems, there was much more variability in the experimental response, and the match between models and experiments was not as good.

7.5 Double L Systems

The final set of systems we studied were the double L systems. The three systems which make up this group are the LLN, LNL, and NLL systems. These systems are analogous to the double T systems, but with L operators in place of the T operators. The network of reactions for these systems is based on the single L networks, but contains additional reactions to account for the interactions of the second operator site. With these reactions, the system has 38 reactions and 28 species.

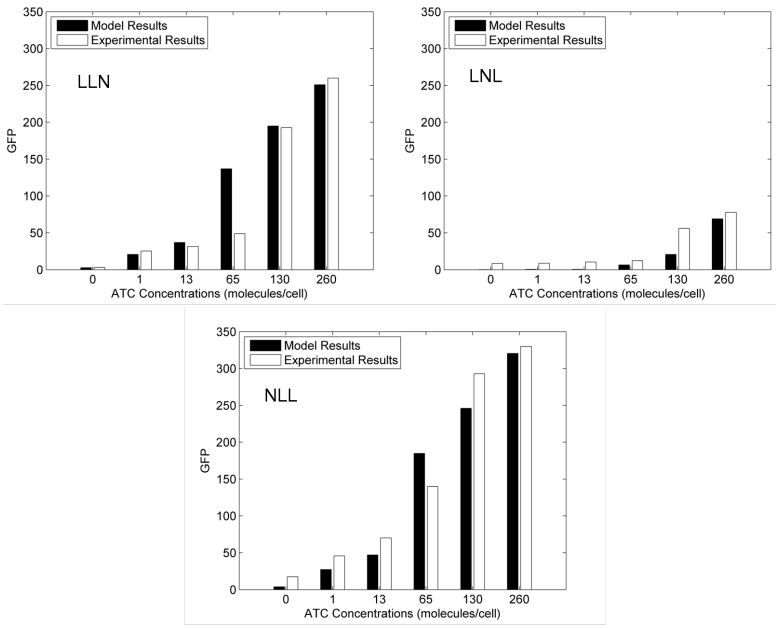

Figure 7.5: Comparison of Models and Experiments for Double L Systems. For these systems, the match between models and experiments was good across a range of aTc concentrations.

When we model these three systems, the first thing that needs to be accounted for is large disparity in maximum responses of these systems. From the experiments, we can see that the LNL system shows a much lower response than the other two systems, due to slow induction. We were able to match this behavior by lowering the promoter strength in the model, thus reducing the maximum response. The models are able to capture the experimental behavior over a wide range of responses. Overall, the double L systems

behavior is captured accurately by our models across the range of aTc concentrations. We are able to match experimental behavior changing only two biologically relevant reactions for 12 different systems. Moreover, these systems are placed in a wide range of inducer concentrations, and our models capture the behavior for these varying conditions. Thus, the modeling formalism used in this work has shown to be a useful way to quantify biological networks. Modeling these systems gives a quantitative view of the interactions in these systems, which is necessary knowledge when engineering synthetic DNA sequences. The use of this knowledge holds great promise for the construction of novel synthetic systems in the future.

Chapter 8

Conclusion and Discussion

Creating constructs based on the AND gate has given great insights into how each part of this system contributes to the overall construct, and has also provided us with the ability to study systems with large variations in promoter strength. Previous work had found that the LTT AND gate shows the highest fluorescent response with large amounts of leakiness, and that the TTL AND gate showed the lowest amount of leakiness while still maintaining a good fluorescent response. By breaking down these systems, we can see what components contribute to these responses. For the TTL system, we can see that its low leakiness is due to the position of the L operator. The L systems show a definitive trend of increased leakiness as the L operator is moved further from the transcriptional start site. This trend matches with the previously found results. It is also interesting to note that the double T systems show very little leakiness, and likely do not contribute to this aspect of the full AND gate. With regards to the maximum response of the full AND gate, our results do not match with the previous work. Our experiments show the greatest fluorescence response when the T operator are the furthest positions from the transcriptional start site, which means we would expect the TTL system to show a higher maximal response than the LTT system.

Experimentally analyzing 20 different systems within the same cell strain and with the same external conditions also allowed us to show the large effects changing the operator sequence has, and we were able to quantify these changes using models. The highest response system, TTN, showed 2 orders of magnitude greater response than the lowest system, TNN mutant 6. With such a variance in response, our library of promoters

can be used as a set of modules in a variety of larger synthetic constructs with varying promoter strength requirements.

When we modeled this library of promoters, it was found that a detailed model accurately represents the system. This model included all the steps of the molecular biology dogma also with other pertinent reactions, such as degradation and repression. As stated before, all of the systems we created were based on previous work on an AND gate, and were all constructed in the same cell strain with the same initial conditions, and measured at the same time intervals. The point of all these similarities was to isolate the difference between the systems to the DNA sequence of the promoter of GFP protein. Similarly, our detailed models only differed by two reactions, one to measure the promoter strength of the system and one to measure the leakiness of the system. By changing only these two reactions, we were able to match the experimental fluorescence of all the single and double T systems at all different inducer concentrations.

The fact that we can match behavior in these systems by only changing two reactions shows that we can model behavior of small systems within the complex environment that is the cell. The complexity of the environment does not make the system indeterminate, but instead has a constant effect across all the systems. Even though our knowledge of all the interactions and molecules inside the cell is incomplete, we can still accurately model the systems in detail. This conclusion has much broader implications than simply being able to model these systems accurately. Our work provides evidence that we can indeed deal with and understand the complexity inherent to biological systems. In our case, this means we do not have to resort to high level models based solely on empirical results. Instead, we can use our knowledge of the molecular biology dogma, combined with experimental results to get more detailed modeling.

Overall, we can conclude that creating these many different systems has allowed us to more fully understand the AND gate system and our detailed modeling scheme can accurately quantify the response of these systems by changing only biologically relevant reactions. This type of modeling scheme holds promise for quantifying other systems, and shows that insights can be gained from this approach, even in the complex environment of the cell.

References

[1] M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000. 0028-0836.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the the Cell*. Garland Science, Abdington, UK, 2008.

[3] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. *Biochemistry*. W. H. Freeman and Company, New York City, 2007.

[4] Daniel G. Gibson, Gwynedd A. Benders, Cynthia Andrews-Pfannkoch, Evgeniya A. Denisova, Holly Baden-Tillson, Jayshree Zaveri, Timothy B. Stockwell, Anushka Brownley, David W. Thomas, Mikkel A. Algire, Chuck Merryman, Lei Young, Vladimir N. Noskov, John I. Glass, J. Craig Venter, III Hutchison, Clyde A., and Hamilton O. Smith. Complete chemical synthesis, assembly, and cloning of a mycoplasma genitalium genome. *Science*, 319(5867):1215–1220, 2008.

[5] Dae-Kyun Ro, Eric M. Paradise, Mario Ouellet, Karl J. Fisher, Karyn L. Newman, John M. Ndungu, Kimberly A. Ho, Rachel A. Eachus, Timothy S. Ham, James Kirby, Michelle C. Y. Chang, Sydnor T. Withers, Yoichiro Shiba, Richmond Sarpong, and Jay D. Keasling. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086):940–943, 2006. 10.1038/nature04640.

[6] Mariette R. Atkinson, Michael A. Savageau, Jesse T. Myers, and Alexander J. Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory

behavior in escherichia coli. *Cell*, 113(5):597–607, 2003. doi: DOI: 10.1016/S0092-8674(03)00346-5.

[7] Attila Becskei and Luis Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–593, 2000. 10.1038/35014651.

[8] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7702–7707, 2003.

[9] C Guet, abreve, lin C., Michael B. Elowitz, Weihong Hsing, and Stanislas Leibler. Combinatorial synthesis of genetic networks. *Science*, 296(5572):1466–1470, 2002.

[10] Avraham E. Mayo, Yaakov Setty, Seagull Shavit, Alon Zaslaver, and Uri Alon. Plasticity of the cis-regulatory input function of a gene. *PLoS Biol*, 4(4):e45, 2006.

[11] Beat P. Kramer, Cornelius Fischer, and Martin Fussenegger. Biologic gates enable logical transcription control in mammalian cells. *Biotechnology and Bioengineering*, 87(4):478–484, 2004.

[12] Yohei Yokobayashi, Ron Weiss, and Frances H. Arnold. Directed evolution of a genetic circuit. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):16587–16591, 2002.

[13] Hideki Kobayashi, Mads Krn, Michihiro Araki, Kristy Chung, Timothy S. Gardner, Charles R. Cantor, and James J. Collins. Programmable cells: Interfacing natural and engineered gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8414–8419, 2004.

[14] Oliver Rackham and Jason W. Chin. Cellular logic with orthogonal ribosomes. *Journal of the American Chemical Society*, 127(50):17584–17585, 2005. doi: 10.1021/ja055338d.

[15] David Karig and Ron Weiss. Signal-amplifying genetic circuit enables in vivo observation of weak promoter activation in the rhl quorum sensing system. *Biotechnology and Bioengineering*, 89(6):709–718, 2005.

[16] S. Nagaraj and S. W. Davies. Design of a genetic differential amplifier. In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pages 747–750, 2005.

[17] S. Nagaraj and S. W. Davies. Inverting amplifier genetic circuit performance. In *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, pages 3142–3145, 2006.

[18] Vincent J. J. Martin, Douglas J. Pitera, Sydnor T. Withers, Jack D. Newman, and Jay D. Keasling. Engineering a mevalonate pathway in escherichia coli for production of terpenoids. *Nat Biotech*, 21(7):796–802, 2003. 10.1038/nbt833.

[19] Timothy S. Gardner, Charles R. Cantor, and James J. Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000. 10.1038/35002131.

[20] Yann Dublanche, Konstantinos Michalodimitrakis, Nico Kummerer, Mathilde Foglierini, and Luis Serrano. Noise in transcription negative feedback loops: simulation and experimental analysis. *Mol Syst Biol*, 2, 2006. 10.1038/msb4100081.

[21] William R. Farmer and James C. Liao. Improving lycopene production in escherichia coli by engineering metabolic control. *Nat Biotech*, 18(5):533–537, 2000. 10.1038/75398.

[22] R. Lutz 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. *Nucl. Acids Res.*, 25(6):1203–1210, 1997.

[23] Robert Sidney Cox, Michael G. Surette, and Michael B. Elowitz. Programming gene expression with combinatorial promoters. *Mol Syst Biol*, 3, 2007. 10.1038/msb4100187.

[24] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[25] J. Christopher Anderson, Christopher A. Voigt, and Adam P. Arkin. Environmental signal integration by a modular and gate. *Mol Syst Biol*, 3, 2007. 10.1038/msb4100173.

[26] Daniel J. Sayut, Yan Niu, and Lianhong Sun. Construction and enhancement of a minimal genetic and logic gate. *Appl. Environ. Microbiol.*, 75(3):637–642, 2009.

[27] John E. Dueber, Brian J. Yeh, Kayam Chak, and Wendell A. Lim. Reprogramming control of an allosteric signaling switch through modular recombination. *Science*, 301(5641):1904–1908, 2003.

[28] Tom Ellis, Xiao Wang, and James J. Collins. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat Biotech*, 27(5):465–471, 2009. 10.1038/nbt.1536.

[29] Kavita Iyer Ramalingam, Jonathan R. Tomshine, Jennifer A. Maynard, and Yiannis N. Kaznessis. Forward engineering of synthetic bio-logical and gates. *Biochemical Engineering Journal*, 47(1-3):38–47, 2009. doi: DOI: 10.1016/j.bej.2009.06.014.

[30] *Plasmids: Current Research and Future Trends*. Caister Academic Press, 2008.

[31] V. Helbl and W. Hillen. Stepwise selection of tetr variants recognizing tet operator 4c with high affinity and specificity. *J Mol Biol*, 276(2):313–8, 1998. Helbl, V Hillen, W Research Support, Non-U.S. Gov't England Journal of molecular biology J Mol Biol. 1998 Feb 20;276(2):313-8.

[32] Benno Muller-Hill. Lac repressor and lac operator. *Progress in Biophysics and Molecular Biology*, 30:227–252, 1976. doi: DOI: 10.1016/0079-6107(76)90011-0.

[33] M. Gossen and H. Bujard. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 89(12):5547–5551, 1992.

[34] Manfred Gossen and Hermann Bujard. Anhydrotetracycline, a novel effector for tetracycline controlled gene expression systems in eukaryotic cells. *Nucl. Acids Res.*, 21(18):4411–4412, 1993.

[35] Arthur D. Riggs, Suzanne Bourgeois, Ronald F. Newby, and Melvin Cohn. Dna binding of the lac repressor. *Journal of Molecular Biology*, 34(2):365–368, 1968. doi: DOI: 10.1016/0022-2836(68)90261-1.

[36] Purnananda Guptasarma. Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of escherichia coli. *BioEssays*, 17(11):987–997, 1995.

[37] Donald A. McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4(3):413–478, 1967.

[38] Rick Durrett. *Essentials of Stochastic Processes*. Springer-Verlag, New York, 1999.

[39] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976. doi: DOI: 10.1016/0021-9991(76)90041-3.

[40] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[41] Yang Cao, Hong Li, and Linda Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The Journal of Chemical Physics*, 121(9):4059–4067, 2004.

[42] Michael A. Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry A*, 104(9):1876–1889, 2000.

[43] Daniel T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[44] Daniel T. Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[45] H. Salis and Y. Kaznessis. An equation-free probabilistic steady state approximation: Dynamic application to stochastic simulation of biochemical networks. *J Chem Phys*, 123:214106, 2005.

[46] Howard Salis, Vassilios Sotiropoulos, and Yiannis Kaznessis. Multiscale hy3s: Hybrid stochastic simulation for supercomputers. *BMC Bioinformatics*, 7(1):93, 2006.

[47] Adam Arkin, John Ross, and Harley H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.

[48] Pieter L. deHaseth, Margaret L. Zupancic, and Jr. Record, M. Thomas. Rna polymerase-promoter interactions: the comings and goings of rna polymerase. *J. Bacteriol.*, 180(12):3019–3025, 1998.

[49] Marni Raffaelle, Elenita I. Kanin, Jennifer Vogt, Richard R. Burgess, and Aseem Z. Ansari. Holoenzyme switching and stochastic release of sigma factors from rna polymerase in vivo. *Molecular Cell*, 20(3):357–366, 2005. doi: DOI: 10.1016/j.molcel.2005.10.011.

[50] Averell L. Gnatt, Patrick Cramer, Jianhua Fu, David A. Bushnell, and Roger D. Kornberg. Structural basis of transcription: An rna polymerase ii elongation complex at 3.3 resolution. *Science*, 292(5523):1876–1882, 2001.

[51] M. Dunaway, J. S. Olson, J. M. Rosenberg, O. B. Kallai, R. E. Dickerson, and K. S. Matthews. Kinetic studies of inducer binding to lac repressor.operator complex. *J. Biol. Chem.*, 255(21):10115–10119, 1980.

[52] Frederick R. Blattner, III Plunkett, Guy, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, and Ying Shao. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–1462, 1997.

[53] Arnold Revzin and Peter H. Von Hippel. Direct measurement of association constants for the binding of escherichia coli lac repressor to non-operator dna. *Biochemistry*, 16(22):4769–4776, 2002. doi: 10.1021/bi00641a002.

[54] R. M. Horton, Z. L. Cai, S. N. Ho, and L. R. Pease. Gene splicing by overlap extension: tailor-made genes using the polymerase chain reaction. *Biotechniques*,

8(5):528–35, 1990. Horton, R M Cai, Z L Ho, S N Pease, L R Research Support, Non-U.S. Gov't United states BioTechniques Biotechniques. 1990 May;8(5):528-35.

[55] Alon Zaslaver, Anat Bren, Michal Ronen, Shalev Itzkovitz, Ilya Kikoin, Seagull Shavit, Wolfram Liebermeister, Michael G. Surette, and Uri Alon. A comprehensive library of fluorescent transcriptional reporters for escherichia coli. *Nat Meth*, 3(8):623–628, 2006. 10.1038/nmeth895.

[56] Tania Nolan, Rebecca E. Hands, and Stephen A. Bustin. Quantification of mrna using real-time rt-pcr. *Nat. Protocols*, 1(3):1559–1582, 2006. 10.1038/nprot.2006.236.

[57] M. Lanzer and H. Bujard. Promoters largely determine the efficiency of repressor action. *Proceedings of the National Academy of Sciences of the United States of America*, 85(23):8973–8977, 1988.

[58] Frank Hoffmann and Ursula Rinas. Stress induced by recombinant protein production in escherichia coli. pages 73–92. 2004.

[59] Poonam Srivastava, Palash Bhattacharaya, Gaurav Pandey, and K. J. Mukherjee. Overexpression and purification of recombinant human interferon alpha2b in escherichia coli. *Protein Expression and Purification*, 41(2):313–322, 2005.

[60] Finbarr Hayes. Toxins-antitoxins: Plasmid maintenance, programmed cell death, and cell cycle arrest. *Science*, 301(5639):1496–1499, 2003.

[61] Hauke Lilie, Elisabeth Schwarz, and Rainer Rudolph. Advances in refolding of proteins produced in e. coli. *Current Opinion in Biotechnology*, 9(5):497–501, 1998.

[62] C. Sizemore, A. Wissmann, U. Gulland, and W. Hillen. Quantitative analysis of tn10 tet repressor binding to a complete set of tet operator mutants. *Nucleic Acids Res*, 18(10):2875–80, 1990. Sizemore, C Wissmann, A Gulland, U Hillen, W Research Support, Non-U.S. Gov't England Nucleic acids research Nucleic Acids Res. 1990 May 25;18(10):2875-80.

[63] J H Miller. *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1972.

[64] Thomas Lederer, Martin Kintrup, Masayuki Takahashi, Phaik-Eng Sum, George A. Ellestad, and Wolfgang Hillen. Tetracycline analogs affecting binding to tn10-encoded tet repressor trigger the same mechanism of induction of tet repressor by a quantitative methylation protection assay. *Biochemistry*, 35(23):7439–7446, 1996.

[65] Thomas Lederer, Masayuki Takahashi, and Wolfgang Hillen. Thermodynamic analysis of tetracycline-mediated induction of tet repressor by a quantitative methylation protection assay. *Analytical Biochemistry*, 232(2):190–196, 1995. doi: DOI: 10.1006/abio.1995.0006.

[66] Sigler Albrecht, Schubert Peter, Hillen Wolfgang, and Niederweis Michael. Permeation of tetracyclines through membranes of liposomes and escherichia coli. *European Journal of Biochemistry*, 267(2):527–534, 2000. 10.1046/j.1432-1327.2000.01026.x.