MODELING REGULATION OF TRANSCRIPTION INITIATION

by

ELIANE ZERBETTO TRALDI

A dissertation submitted to the

Graduate School–New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Konstantin Mischaikow

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2011

ABSTRACT OF THE DISSERTATION

Modeling Regulation of Transcription Initiation

By ELIANE ZERBETTO TRALDI

Dissertation Director:
Konstantin Mischaikow

The concept of activation in transcriptional regulation is based on the assumption that product mRNA increases monotonically as a function of regulator concentration. We analyze the Shea-Ackers model of transcription and find this assumption to be correct only for the simplest of promoters. We define a new regulatory constant that is a nonlinear combination of association and transcription initiation constants characterizing activation and repression for more complicated promoters. Our results can guide the synthesis of new promoters and lead to a deeper understanding of the constraints guiding the natural promoters evolution.

Using a validated mathematical model based on the Shea-Ackers transcription rate function, we then show that two modes of upregulation have very different effects on the function of promoter P_{RM} in phage lambda. We predict that if CI_2 bound to O_R2 produced equal increase in RNAP-DNA binding constant (compared to wild-type increase in the closed-open transition probability), the lysogen would be significantly less stable.

We then focus on the promoter clearance process during transcription initiation. Our work builds upon an initial sequence-dependent three-pathway model proposed

by Xue *et al.* After making several modifications to this model and not being able to satisfactorily match experimental data, we introduce a new parameter to the model: the possible formation of secondary structure in the single stranded scrunched DNA accumulated before RNA polymerase is able to escape the promoter .

Acknowledgements

I would like to sincerely thank my advisor Konstantin Mischaikow for the support, guidance and encouragement he gave me throughout this process.

I would like to thank Tomáš Gedeon for the discussions and collaboration, and for being so helpful and accessible.

I would like to thank Richard Ebright for his time and patience to discuss Biology with mathematicians.

I would like to thank Kate for the collaboration and friendship.

I would like to thank Ariella and Steve for letting me have a home inside their home. It would have been very difficult to have Daniel here with me if I were in any other place.

I would like to thank Marcio for being such a good dad and for sharing with me the important, and sometimes intercontinental, task of parenting Daniel. Also for the many times he told me so confidently I could do this.

I would like to thank the friends I made here at Rutgers. They all contributed in some way to make this happen. In special Julie, with whom I shared not only many laughs but also many tears. I would not have made through the last year and a half here without her.

Dedication

This thesis is dedicated to my son Daniel, for understanding I had to be away and could not be with him all the time, for helping me get things into perspective and see what is really important. I would like also to dedicate this thesis to my parents who are always so proud of me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

All cells, whether from a single-celled organism, the complex human organism with more than 10^{13} cells, or from any other living organism, amazingly share many fundamental features.

In all species individual cells carry all the genetic information storing it as double-stranded DNAs – long complementary paired polymer chains formed from four possible different monomers, called nucleotides. The genetic information in all living organisms can be seen as a code based on a four-letter alphabet – A, T, G and C, corresponding to the four monomer types.

In order to carry out the instructions given in the DNA, all cells produce two other types of polymers: RNAs and proteins. RNA is a polymer closely related to DNA and it is also a code based on a four-letter alphabet – A, U, G and C. The process of RNA synthesis, called transcription, uses the same strategy of template polymerization used in DNA replication in which segments of the DNA sequence are used as templates for the production of RNAs. Different types of RNAs exist, but most of them are messenger RNAs (mRNAs) whose function is to instruct protein synthesis, a process called translation. Proteins, just like DNAs and RNAs, are long polymer chains, and all cells translate RNA to protein essentially in the same way. Proteins are formed from 20 different monomers called amino acids. Proteins are used in all cells to direct most of the chemical reactions. The specific function of each protein is determined by its amino acid sequence, which in turn is determined by the nucleotide sequence of the corresponding DNA segment. Each DNA segment corresponding to one protein is called a gene.

In all cells, gene expression is regulated. That is, instead of producing all possible

proteins all the time, decisions are continuously made about at which rate transcription and translation of different genes should occur, depending on what the cell needs to accomplish. While these decisions are made on a case-by-case basis using a wide range of cellular and extra-celllular signals, the execution of the orders happend through the fundamental processes of DNA, RNA and protein synthesis, and those processes are essentially the same in all living cells. See Figure 1.1.

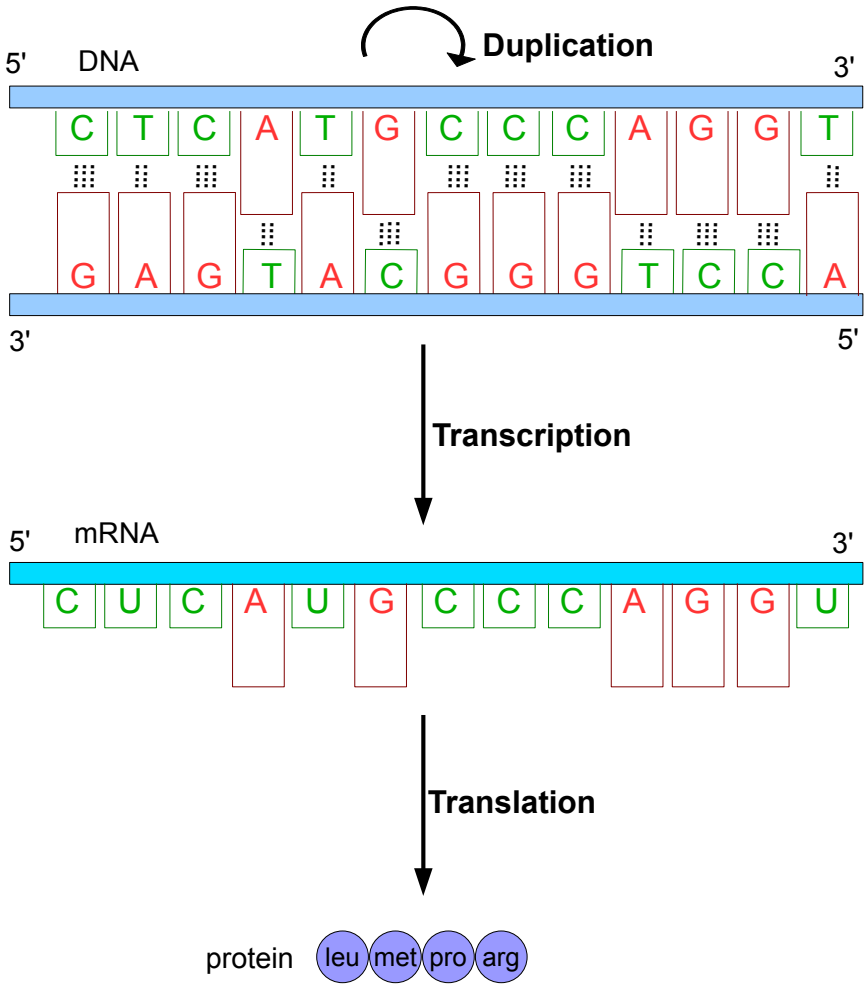


Figure 1.1: Central Dogma of Molecular Biology: genetic information flows from DNA to RNA to protein.

Most of our understanding of fundamental processes like DNA replication, transcription and translation have come to light through studies of *Escherichia coli*, or *E. coli*. Since, in essence, these mechanisms are the same in all living organisms, we learn a lot through studies in *E. coli*, or other model organisms. Bacteria and their phages (or bacteriophages, viruses that infect bacteria) have been of central importance to the development of molecular genetics.

E. coli, as a model system, has been intensively studied, and much more is known about *E. coli* than about any other living organism. *E. coli*, like other bacteria, has properties that facilitate genetic experiments, making it a good organism of choice when trying to study basic cellular processes:

- *E. coli* is a haploid organism, that is, it has only one copy of each gene, making it easier to identify cells with a particular mutation.
- *E. coli* reproduces asexually by cell division, producing offspring that are genetically identical to their parent and to each other.
- *E. coli* has a very short generation time.
- *E. coli* can be easily and inexpensively grown in laboratories and it can adapt to variable chemical conditions.

The focus of this thesis is on bacterial transcription, more specifically, on modeling the initial phase of transcription in *E. coli*.

Transcription

Transcription is the synthesis of RNA from a DNA template, and is the first step in the process leading to the highly regulated mechanism of gene expression.

Transcription is catalyzed by the enzyme RNA polymerase and, in contrast to DNA replication in which the whole DNA strand is copied, in transcription only a comparatively short molecule is produced. Therefore RNA polymerase must have the ability to recognize where along the DNA strand to start and and where to terminate transcription.

RNA polymerase is highly conserved from bacteria to humans [1, 2, 3, 4], making the simpler bacterial RNA polymerase a good model for the study of RNA polymerases in general. *E. coli*'s RNA polymerase is by far the best studied and characterized of all RNA polymerases.

Bacteria have only one type of RNA polymerase, while eukaryotic cells have three [5]. The bacterial RNA polymerase core enzyme has five subunits: two copies of the α subunit, and one copy of each β, β' and ω subunits [6]. The core enzyme alone is capable of binding the RNA with no specificity, and can initiate RNA synthesis at any point on the DNA, without recognizing the promoter. This can be observed in vitro [5]. In cells, RNA polymerase will only initiate transcription at promoters and this happens due to the addition of another subunit to the core enzyme, the initiation factor called the σ factor. Its addition decreases the affinity of RNA polymerase to non-specific DNA and increases its affinity to DNA promoters. Therefore the σ factor directs RNA polymerase to the promoters to ensure transcription will only be initiated there [7]. The core enzyme together with the σ factor is referred to as the RNA polymerase holoenzyme.

There are different types of σ factor and *E. coli* has 7 types [8]. The primary one is called σ^{70} (so called since the protein is 70 kD in size).The holoenzyme with σ^{70} transcribes most genes in a growing cell including most of the housekeeping genes.

Promoters recognized by RNA polymerase containing σ^{70} are often referred to as σ^{70}-promoters. The majority of σ^{70}-promoters share some specific sequence characteristics: two conserved sequences of six nucleotides, that are separated by a non-specific sequence of $17 - 19$ nucleotides, and are centered, respectively, at positions -10 and -35 [9]. These sequences are called -35 and -10 regions. See Figure 1.2.

Some σ^{70}-promoters have an additional DNA element, called the UP-element, that increases RNA polymerase binding through an additional specific interaction between the enzyme and the DNA [10, 11]. Other σ^{70}-promoters lack a -35 region and instead have an "extended -10 region", which is the standard -10 region with an additional short sequence element at its upstream end [5].

An additional DNA element just downstream from the −10 region, called the discriminator, has recently been found to bind RNA polymerase and influences the stability of the holoenzyme-promoter complex [12].

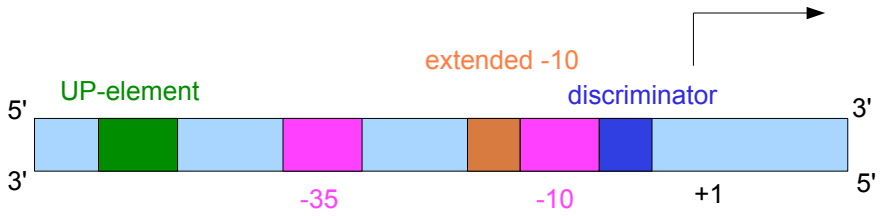


Figure 1.2: Common features of σ^{70} promoters include typical −35 and −10 regions, the UP-element, the extended −10 region and the discriminator.

Transcription can be divided in three phases: initiation, elongation and termination.

Initiation

Transcription initiation is the first and most highly regulated of the three transcription phases. Transcription initiation itself is divided in three steps.

In the first step, called *RNA polymerase binding* or just *binding*, RNA polymerase binds the DNA in the promoter region, forming the RNA polymerase-promoter closed complex, in which the DNA is in double-stranded form (Figure 1.3). The promoter determines where transcription should start and which strand should be used. As described above, this is guided by the binding specificity of the σ factor to the promoter.

In the second step, called *isomerization* or *open complex formation*, RNA polymerase melts, or unwinds, 13–14 base pairs of DNA between positions −11 and +2/+3, forming the RNAP-promoter open complex, or transcription bubble. In the open complex the bases of the coding strand are exposed, and therefore available for base-pairing of the complementary NTPs for RNA synthesis. The +1 nucleotide is positioned in the active site of RNA polymerase, where the polymerization reactions occur. See Figure 1.4. In contrast with RNA polymerase binding, isomerization is essentially an irreversible process and usually guarantees that transcription will initiate [5].

The third step in transcription initiation is called *promoter clearance* or *promoter*

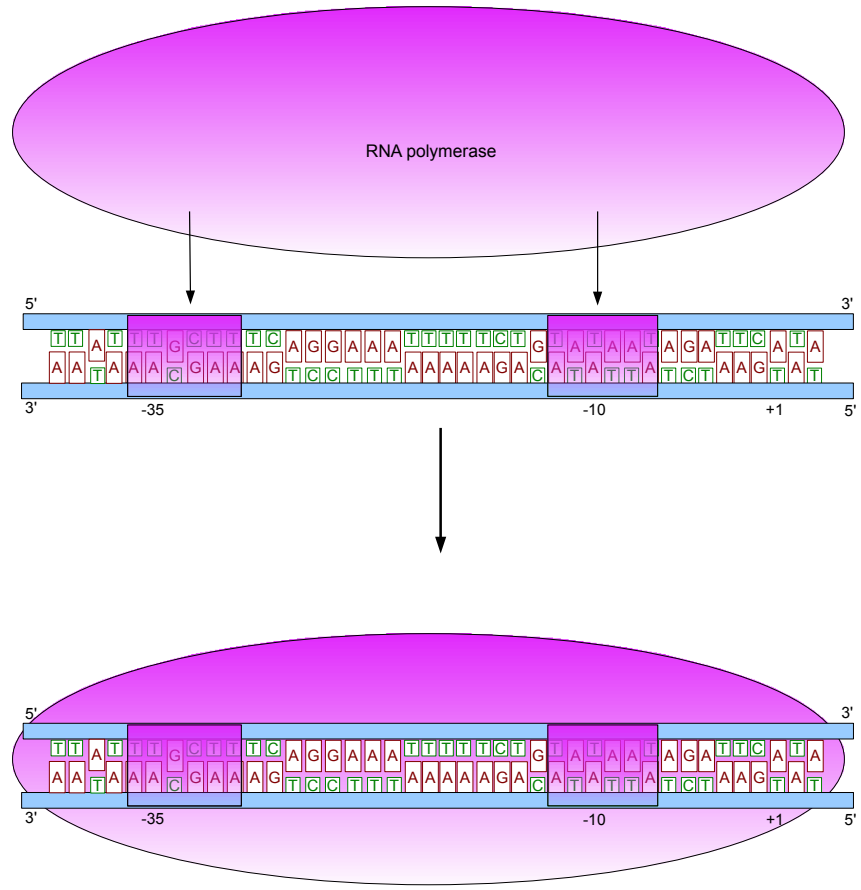


Figure 1.3: The first step in transcription initiation is RNA polymerase binding.

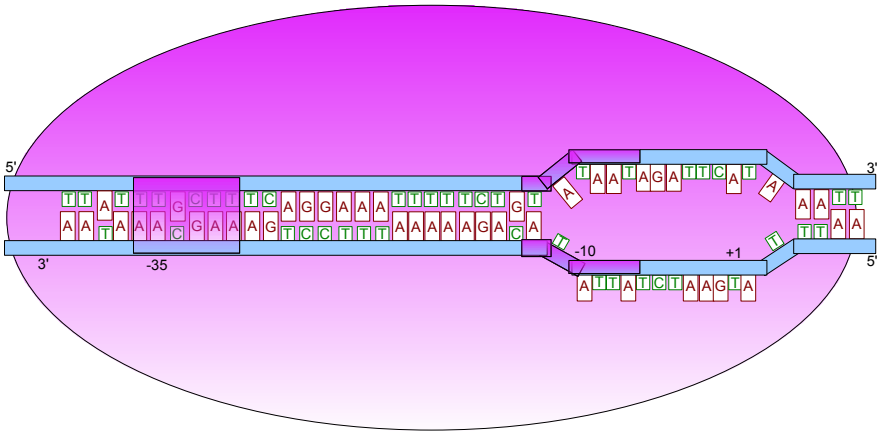


Figure 1.4: RNA polymerase-promoter open complex

escape, and it is in this phase that RNA synthesis is initiated, while RNA polymerase is still physically attached to the promoter. Since RNA polymerase synthesizes a new RNA chain on a DNA template, it does not need a primer. Instead, the NTPs must enter its active site through its secondary channel, or NTP-uptake channel. In order to initiate transcription, the first complementary NTP enters the RNA polymerase active site and is held stably on the template while the next complementary NTP arrives for polymerization to occur. After the first polymerization reaction is performed, RNA polymerase's active site must be made available again for the next reaction. See Figure 1.5. While several models have been proposed for how the enzyme's active site translocates along the DNA template during transcription initiation, recent experiments [13, 14] have shown that RNA polymerase remains stationary on the promoter while it unwinds the downstream DNA and pulls, or scrunches, that DNA into itself. The DNA accumulated within the enzyme is accomodated as single stranded bulges. See Figure 1.6.

Typically before RNA polymerase is able to break the bonds with the promoter to enter the elongation phase of transcription, it goes through a process called *abortive transcription* or *abortive initiation*. During abortive initiation RNA polymerase synthesizes and releases short RNA segments, or abortive transcripts, typically ranging in size from 2 to 15 nucleotides. Abortive transcripts were first observed in [15] in transcription reactions containing only the first two NTP substrates. Later experiments have shown existence of longer abortive transcripts. For a review on abortive initiation see [16]. Abortive transcripts have also been detected in vivo [17].

After the release of an abortive transcript, RNA polymerase goes back to the inital open complex conformation, and RNA synthesis starts again. This process is repeated until RNA polymerase is able to clear, or escape, the promoter and enter the elongation phase of transcription.

Promoter escape can therefore be seen as the transition between transcription initiation and elongation. It is a complicated process and involves large conformational changes. The transition to the elongation phase is associated with the breaking of all the interactions between RNA polymerase and the promoter (including σ-promoter

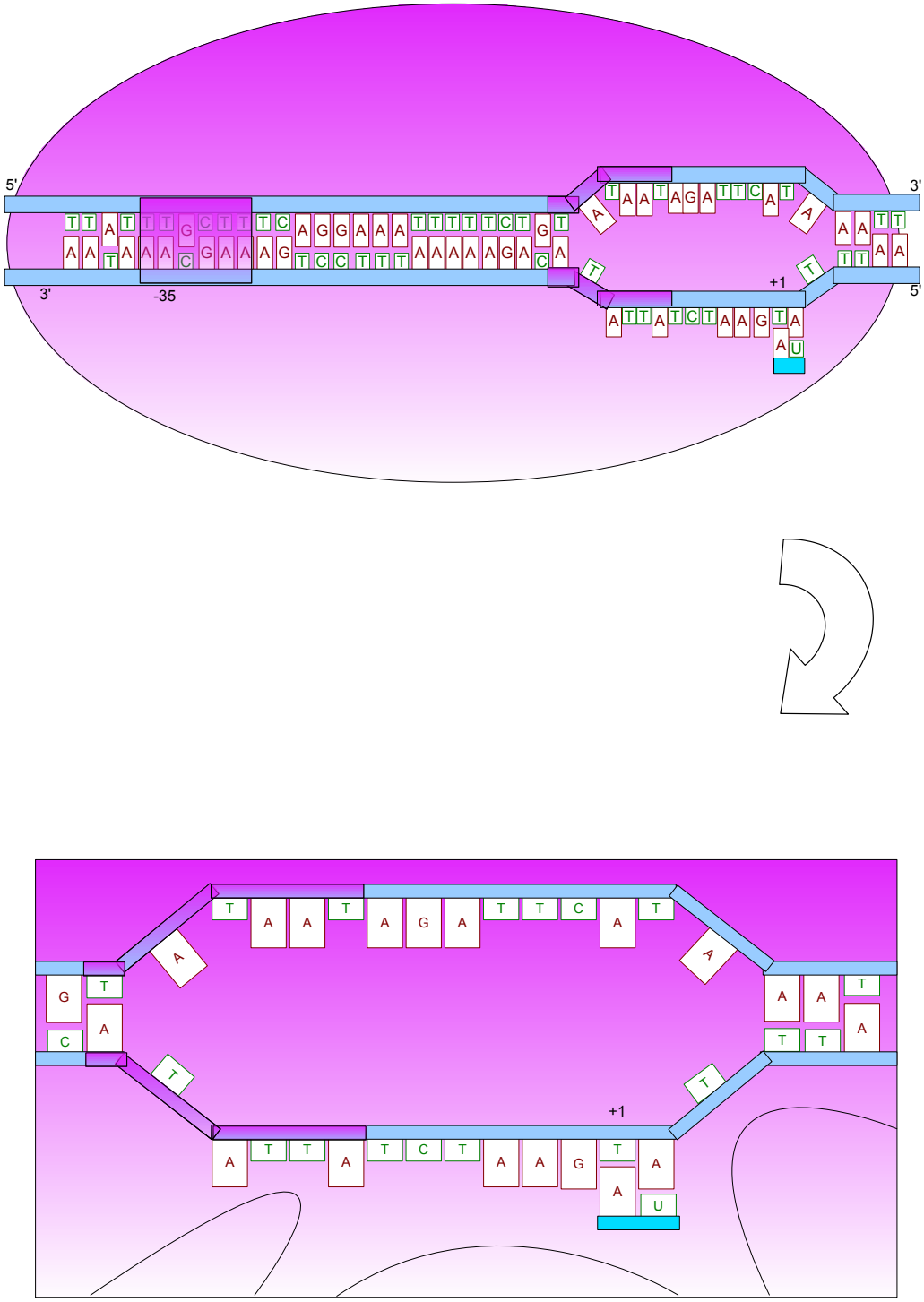


Figure 1.5: After the first polymerization reaction is performed, RNA polymerase needs to make its active site available for the next reaction, in order to elongate the nascent RNA.

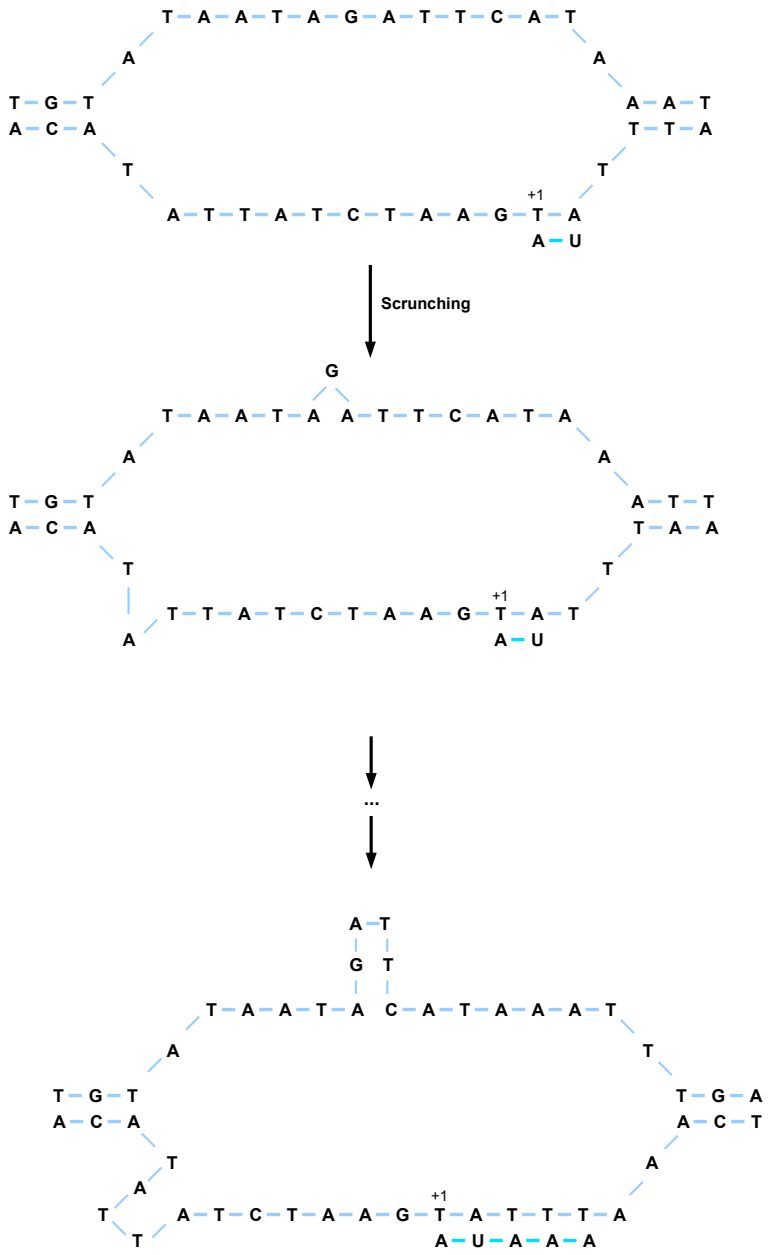


Figure 1.6: While RNA polymerase is still bound to the promoter it pulls, or scrunches, the DNA strands into itself in order to make its active site available for the next polymerization reaction.

contacts), simultaneous collapse of the transcription bubble back to a size of $12-14$ nucleotides, and forward movement of RNA polymerase. Promoter escape was thought to coincide with the release of the σ-factor [18, 19, 20]. Recent studies showed that while loss of σ-promoter occurs during promoter escape, the σ-factor may not necesseraly be released immediately [21, 22]. Likely, σ-factor release occurs at different positions for different promoters [23, 24, 25, 26].

Elongation

During elongation, RNA polymerase is no longer bound to the promoter, and therefore is free to move along the DNA template in a $5'$ to $3'$ direction, elongating the growing RNA chain one nucleotide at time using the nucleotides that enter its active site through its secondary channel, and guided by the DNA template. While doing so it performs several functions. In order to move along the DNA, at each step it unwinds one base pair of the downstream DNA. It re-anneals the upstream DNA one base pair at time, keeping the size of the transcription bubble constant throughout elongation. It dissociates the growing RNA chain from the template, directing the RNA chain thorough the exit channel, leaving only the last 8 or 9 incorporated nucleotides base-paired to the template DNA. In addition, RNA polymerase also performs proofreading functions.

Termination

Termination occurs after RNA polymerase has copied, or transcribed, a stretch of DNA corresponding to a gene (or genes in the case of an operon). At this stage, it stops elongating the RNA chain, releases the RNA product and dissociates itself from the DNA, in order to become available to perform another transcription reaction. Sequences called terminators, present at the end of genes, are the triggers for RNA polymerase dissociation and RNA release.

Motivation

Our motivation to study transcription initiation comes from the thermodynamical approach introduced by Ackers et. al in 1982 [27]. Let $\{x_1, x_2, ..., x_n\}$ be the set of regulatory proteins for a gene. Let $\mathcal{S}$ be the set of all possible binding states for the regulatory region of this gene. Given a state $s \in \mathcal{S}$, s_i denotes the number of molecules of protein x_i bound to the regulatory region; and $s_0 \in \{0, 1\}$ indicates whether RNA polymerase is bound to the promoter. The probability of a binding state $\sigma \in \mathcal{S}$ is

$$\mathbb{P}_\sigma = \frac{e^{-\frac{\Delta G_\sigma}{RT}} [RNAP]^{\sigma_0} \prod_{i=1}^{n} [x_i]^{\sigma_i}}{\sum_{s \in \mathcal{S}} e^{-\frac{\Delta G_s}{RT}} [RNAP]^{s_0} \prod_{i=1}^{n} [x_i]^{s_i}}, \tag{1.1}$$

where $[\cdot]$ denotes concentration, ΔG_s is the free energy of the the state s, R is the universal gas constant, and T is the temperature. The derivation of equation (1.1) is taken from [28] and presented in Appendix A.

The transcription rate is then defined as

$$f([RNAP], [x_1], ..., [x_n]) = \sum_{\{s | s_0 = 1\}} k(s) \mathbb{P}_s, \tag{1.2}$$

where $k(s)$ is the rate of transcription initiation. Notice that both RNA polymerase and protein binding are represented in the probability function. Both opening and clearance processes must then be incorporated in this constant $k(s)$, and therefore $k(s)$ must be fitted to the data. Since it is difficult to distinguish experimentally between rates corresponding to different states, only one rate is often used for all states with bound RNA polymerase. It should be emphasized that $k(s)$ and ΔG_s in (1.1) are both functions of the DNA sequence and it is clear that $k(s)$ and ΔG_s have a non-linear relationship. We want to explore ways to quantify this non-linear relationship, and hopefully this will bring insights on how to compute the rates $k(s)$ as functions of the DNA sequence using modeling of the opening and escape processes.

Outline

When modeling transcriptional regulation, the concept of activation is commonly based on the assumption that product mRNA increases monotonically as a function of regulator concentration. In Chapter 2 we present a mathematical analysis of the Shea-Ackers transcription rate function given by (1.2) and find this assumption to be correct only for the simplest promoters. We define a new regulatory constant that is a nonlinear combination of association and transcription initiation constants characterizing activation and repression for more complicated promoters. The material in this chapter is part of collaborative work with Konstantin Mischaikow, Kate Patterson and Tomáš Gedeon, and has been published in [29]. Reproduction of this material here is done with kind permission from Springer Science and Business Media: Bulletin of Mathematical Biology, When activators repress and repressors activate: A qualitative analysis of the Shea-Ackers model, volume 70, 2008, 1660–1683, Tomáš Gedeon, Konstantin Mischaikow, Kate Patterson, and Eliane Traldi.

In Chapter 3 we use the Shea-Ackers model to show that two different modes of upregulation have very different effects on the promoter P_{RM} function in the bacteriophage λ. More specifically, we show that in the context of proper functioning of the phage λ induction, the binding constant K_B plays a fundamentally different role from the opening and clearing constant k. The material in this chapter is part of collaborative work with Konstantin Mischaikow, Kate Patterson and Tomáš Gedeon, and has been published in [30]. Reproduction of this material here is done with kind permission from Elsevier: Biophysical Journal, Binding Cooperativity in Phage λ is Not Sufficient to Produce an Effective Switch, volume 94, 2008, 3384–3392, Tomáš Gedeon, Konstantin Mischaikow, Kathryn Patterson, and Eliane Traldi.

In Chapter 4 we look only at the last phase of transcription initiation, the promoter clearance process. We start by presenting a model by Xue, Liu and Ou-Yang [31]. While still following the main idea of their model, we introduce several modifications and improvements, without being able to satisfactory match experimental data. We introduce a new feature to the model: the formation of secondary structure in the

scrunched DNA. While there is no biological evidence that secondary structure will form in the scrunched DNA, there seems to be no biological evidence against its formation. We believe the addition of this feature results in an overall improvement to the model. The material in this chapter is part of collaborative work with Konstantin Mischaikow,Tomáš Gedeon and Richard Ebright.

Chapter 2

A Qualitative Analysis of the Shea-Ackers Model

Synthetic biology suggests the possibility of developing organisms with different functional abilities that may provide solutions to a wide variety of fundamental problems ranging from medicine to renewable energy. Producing such organisms may require a deep understanding of existing as well as novel signal transduction/gene regulatory network designs. Recent work has shown the feasibility of complete genome transplantation [32], thus, in theory, completely original networks could be employed. In practice synthetic circuits have already been constructed [33, 34]. However, for many aspects ranging from the construction of the individual components to the design of the architecture of the networks themselves, much remains to be understood.

On the network level Alon [35] provides a compelling framework for understanding the design principles of biological circuits as it relates local models for transcriptional regulation and network design to phenomenological function of the system as a whole. The local model is in accordance with the concept of regulated recruitment [36], wherein the rate of transcription of mRNA is determined by the local structure of the DNA and concentrations of regulatory proteins, often referred to as activators and repressors. As the names suggest activators enhance and repressors decrease the rate of transcription. For the most part, Hill functions are used in [35] to model the transcription rate: $f(r) = ar^n/(b + r^n)$ for activators and $f(r) = a/(b + r^n)$ for repressors with $n \geq 1$. Observe that these are monotone functions of the regulatory protein r.

The assumption of monotone regulatory interaction is widespread. The most common representation of a regulatory network is a graph with vertices corresponding to the chemical species or genes and edges corresponding to reactions. Each reaction

is usually labelled with a positive or a negative sign corresponding to up- or down-regulation. Considerable effort has been spent deducing dynamics and function from such representations of a network [35, 37]. The theory of motifs is a result of such activity.

Assuming all chemical reactions on the regulatory region involving the regulatory proteins and RNA polymerase (RNAP) equilibrate on a much faster time scale than transcription, Shea and Ackers [38] construct a nonlinear model for the rate of transcription. Since the first time scale is on the order of milliseconds (bacteria) to seconds (eukaryotes) and the other on the order of minutes [35], this is a reasonable assumption in both bacterial and eukaryotic cells. The Shea-Ackers model provides a broadly accepted quantitative framework [39] and has been experimentally validated for a variety of gene networks [27, 30, 38, 40, 41]. It should also be noted that since the Hill function is derived from the assumption of equilibrium binding of one transcription factor to the promoter, the Shea-Ackers nonlinearity is a generalization of the Hill function that naturally allows for multiple binding factors.

There are many ways in which transcription initiation is controlled and very likely more ways will be discovered in the future, but, as described before, there are three main steps in this process. The first is the binding of RNA polymerase to the DNA (characterized by the association constant K_B that is directly related to the binding energy of RNA polymerase to DNA). The second is the isomerization of the closed RNAP-DNA complex to an open complex (characterized by a rate constant k_f), and finally successful clearance of the promoter by RNA polymerase (characterized by the constant k_{clear}). In the Shea-Ackers framework the last two processes are modeled as a *transcription initiation rate* and are lumped into one constant k. Both the binding energy of RNA polymerase and the transcription initiation rate are controlled by the transcription factors. Thus to each control state s, which is a particular configuration of regulatory proteins and RNA polymerase bound to the DNA, there is an association constant $K_B(s)$ and an initiation rate constant $k(s)$.

Within the context of the model of regulated recruitment the set of control states, the association constants and the initiation rate constants are the fine levers by which

the cell controls transcription. These can be measured. Of course, what is of interest to systems biologists is the effect of particular regulators on transcription. Transcription factors are activators if an increase in their concentration leads to an increase in the rate of transcription and they are repressors if an increase in their concentration leads to a decrease in the rate of transcription. The Shea-Ackers transcription rate function is a sufficiently well established quantitative model of these interactions to justify a mathematical investigation of its behavior as a function of association constants K_B, and transcription initiation rates k.

Observe that the above definition of an activator or repressor is equivalent to an assumption of monotonicity with respect to the concentration of the regulatory protein. While this is true for Hill functions, we show here that it need not be the case for the Shea-Ackers function. While this should not be a surprise to biologists - in low concentrations the regulatory protein CI_2 in the phage lambda switch is an activator for the cI gene, but at high concentrations it becomes a repressor - the theoretical extent to which non-monotonicity may occur has not, to the best of our knowledge, been made clear.

The mathematical implications of non-monotone reaction functions can be significant. As an example, the global dynamics of cyclic feedback systems with arbitrarily many components with monotone reaction functions exhibits very simple dynamics; asymptotically one can have only equilibria or periodic orbits [42]. However, if the reaction functions are not monotone, then one can have chaotic dynamics [43].

In principle, the lack of monotonicity of the Shea-Ackers function could have an equally significant impact on the conclusions expressed in [35] concerning the design principles of biological circuits. In reality, it is quite possible that the biologically constrained parameters prevent this non-monotonicity. Understanding and design of transcriptional regulation require the ability to easily identify the appropriate constraints on the set of states, their association constants, and their initiation rate constants. With this in mind we introduce what we refer to as the *regulatory constant*, ρ, which is a nonlinear combination of various association and initiation rate constants, that

reduces the determination of regions of monotonicity to linear equations. If r is a regulatory protein with regulatory constant ρ_r, then in the absence of any other regulatory proteins $\rho_r > 1$ implies that r is an activator and $\rho_r < 1$ implies that r is a repressor.

An outline of this chapter is as follows. In Section 2.1 we review the Shea-Ackers model and illustrate it in the context of the *trp* and *lac* operons of *E. coli*. In Section 2.2 we introduce various concepts and notation. We begin our analysis of the Shea-Ackers model by showing that in this model the transcription rate is entirely controlled by the association and initiation rate constants (see Theorem 2.2.6). Section 2.3 contains the main results of this chapter. In Section 2.3.1 we examine the case of a regulatory region with a single binding site and a single regulatory protein. Though the Shea-Ackers function is more general than a Hill function, monotonicity is still preserved. We also derive a relationship between the association constant and the initiation rate constant for the regulatory protein that determines whether the protein is an activator or a repressor (see Figure 2.6).

This relation leads to the definition of the regulatory constant ρ_r. Since it is well known [44] that multiple regulatory proteins can bind at the same site, in Section 2.3.2 we consider the case of a regulatory region with a single binding site but multiple regulatory proteins. Formulas which exactly determine when proteins will be activators or repressors as a function of their regulatory constants are presented. A complete classification for the case of two regulatory proteins is given in Theorem 2.3.6 and Corollary 2.3.7. Section 2.3.3 examines the unequal impact of the association constant K_B and transcription initiation rate constant k on the Shea-Ackers function. In Section 2.3.4 we extend the results of Section 2.3.2 to a generic gene with one regulatory protein that has two possible binding sites. We define a *regulatory constant for a pair* ρ_{12} and again we are able to determine if the regulatory protein r is an activator or a repressor using values of the regulatory constants ρ_1, ρ_2, and ρ_{12}. We apply our results to a phage λ model. Finally in Section 2.3.5 we calculate the Shea-Ackers function for a gene with two regulators and two binding sites where one of the regulator binding sites overlaps the RNA polymerase binding site. Imposing further restrictions we recover a model of the *lac* operon.

2.1 The Shea-Ackers Model

The model for regulated recruitment begins with the concept of regulatory proteins binding in various configurations and at different sites in the regulatory region of a particular gene. To capture this we consider a collection of control states. The simplest state is the *empty state* which occurs when no regulatory proteins and no RNA polymerase is bound to the regulatory region. We denote this by $s_\emptyset$. The set of possible non-empty states is denoted by $\mathcal{S}$. For the purposes of this analysis a state in $\mathcal{S}$ is typically determined by the configurations of the regulatory proteins for that particular gene, $\{r_1, \ldots, r_n\}$, and the presence or absence of RNA polymerase, though in principle other control factors could be included. The simplest non-empty states consist of those for which a single regulatory protein or a single RNA polymerase is bound to the DNA. These states are called *elementary states* and denoted by $\mathcal{E} \subset \mathcal{S}$. Within the context of the model of regulated recruitment we can use the elementary states to describe the minimal information associated with any non-empty state $s \in \mathcal{S}$. This leads to the following definition.

Definition 2.1.1 A *decomposition* of the state $s \in \mathcal{S}$ is the list of elementary states $\{s_i \mid i = 1, \ldots, I\} \subset \mathcal{E}$ which indicates whether regulatory proteins and/or RNA polymerase are bound to the DNA when the state s occurs. In an abuse of notation we will often write $s = \{s_i \mid i = 1, \ldots, I\}$.

RNA polymerase plays an essential role in that without its presence transcription cannot occur. We use $[\cdot]$ to denote concentration, and RNAP represents RNA polymerase. Although the concentration of RNA polymerase is a variable, [RNAP], in order to keep our focus on the effects of regulatory proteins, it will be treated as a constant. The elementary state where only RNA polymerase is bound to the DNA is denoted by s_P. Let $\mathcal{S}_0 \subset \mathcal{S}$ be the set of states which do not have RNA polymerase bound to the promoter.

Under the assumption that the binding of RNA polymerase and proteins r_i to the DNA is sufficiently more rapid than the transcription process Ackers *et al.* [27] define

the probability of the occurrence of the control state s to be

$$\mathbb{P}_s = \mathbb{P}_s([RNAP], [r_1], \ldots, [r_m]) = \frac{K_B(s)[RNAP]^{\alpha_s}[r_1]^{\alpha_s^1}[r_2]^{\alpha_s^2}...[r_m]^{\alpha_s^m}}{Z},$$

where

$$K_B(s) := e^{-\frac{\Delta G_s}{RT}} \tag{2.1}$$

and the partition function Z is given by

$$Z([RNAP], [r_1], \ldots, [r_m]) = 1 + \sum_{s \in \mathcal{S}} K_B(s)[RNAP]^{\alpha_s}[r_1]^{\alpha_s^1}[r_2]^{\alpha_s^2}...[r_m]^{\alpha_s^m}. \tag{2.2}$$

In this formula ΔG_s denotes the energy associated to the state $s \in \mathcal{S}$ under the normalization that $\Delta G_{s_\emptyset} = 0$. The exponents α_s^i indicate the number of r_i molecules bound to the regulatory region in state s and similarly, α_s denotes the number of RNA polymerase molecules bound to the regulatory region in state s. As is standard, T is the temperature and R is the universal gas constant [28].

Let $k(s)$ be the rate of transcription initiation of the binding state s. In particular, if $\alpha_s = 0$, i.e. $s \in \mathcal{S}_0$, then it is assumed that $k(s) = 0$. Under these assumptions the Shea-Ackers transcription rate function [38] of the gene in question is

$$f([RNAP], [r_1], \ldots, [r_m]) = \sum_{s \in \mathcal{S}} k(s)\mathbb{P}_s. \tag{2.3}$$

Remark 2.1.2 From (2.3) it should be clear that to describe transcription regulation of a gene within the context of the Shea-Ackers function it is sufficient to know the set of states $\mathcal{S}$ and for each state s to know the association constant $K_B(s)$ and the transcription initiation rate $k(s)$. Because of its frequent use we define $K_P := K_B(s_P)$ and $k_P := k(s_P)$.

Example 2.1.3 The *trp* operon of *E. coli* is regulated by the TrpR repressor protein (see Figure 2.1). When tryptophan is present, it binds the TrpR repressor inducing conformational change in that protein and enabling it to bind the *trp* operator. This binding prevents transcription, since the operator overlaps with the RNA polymerase

binding site. When tryptophan is limiting, the TrpR repressor is free of its corepressor (tryptophan) and cannot bind to the operator, allowing RNA polymerase to bind the promoter and start transcription [45].

For this genetic regulatory region, there are three states, $s_\emptyset$, s_P, and the elementary state s_r in which TrpR is bound to the DNA. Let $[r]$ denote the concentration of TrpR. To simplify the notation let $K_r := K_B(s_r)$. The partition function is given by $Z([r],[RNAP]) = 1+K_r[r]+K_P[RNAP]$. Since s_P is the only regulatory state leading to transcription, the Shea-Ackers function for the *trp* operon is

$$f([r],[RNAP]) = \frac{k_P K_P[RNAP]}{1+K_r[r]+K_P[RNAP]}. \tag{2.4}$$

This function can be viewed as a generalization of the Hill function for protein binding, with the addition of RNA polymerase. When $[RNAP]$ is assumed to be constant f becomes a Hill function for a repressor with $a = k_P K_P[RNAP]$ and $b = 1+K_P[RNAP]$.

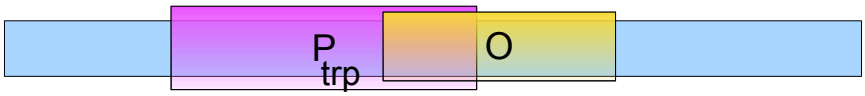


Figure 2.1: Regulatory region of *trp* operon in *E. Coli*: The *trp* operator O (binding site for TrpR) overlaps the *trp* promoter P_{trp} (RNA polymerase binding site). When TrpR is bound to the DNA transcription cannot occur.

Remark 2.1.4 The *trp* operon is also subject to transcription attenuation and feedback inhibition [45]. Within the context of this model we do not consider these types of regulation.

Example 2.1.5 The transcription of the *E. coli lac* operon is controlled by lacI and CAP-cAMP complex. The lacI binding region (operator O_1) overlaps with the RNA polymerase binding site and the CAP-cAMP complex binding site is located upstream of the promoter [44] (see Figure 2.2). In the absence of the inducer allolactose, lacI will bind to the operator and prevent binding of RNA polymerase which prevents transcription. In the presence of allolactose, the repressor can no longer bind to the operator

because of the interaction of the inducer with the repressor. Only in the absence of adequate glucose supply will the CAP-cAMP complex bind to the DNA, which is required for RNA polymerase to effectively bind the promoter. Therefore transcription will occur only in the presence of allolactose and absence of glucose, that is, only when the CAP-cAMP complex is bound to the DNA and lacI is not bound. The set of possible regulatory states are $s_\emptyset$, s_P, s_c, s_r, $s_{cP} = \{s_c, s_P\}$, and $s_{cr} = \{s_c, s_r\}$. The elementary states s_c and s_r correspond to binding of the CAP-cAMP complex to the DNA and the repressor to the DNA, respectively. To simplify notation let us denote $k(s_*)$ and $K_B(s_*)$ by k_* and K_*, respectively. The Shea-Ackers function is given by

$$f([c],[r],[RNAP]) = \frac{[RNAP]}{Z}(k_P K_P + k_{cP} K_{cP}[c]), \tag{2.5}$$

where

$$\begin{aligned} Z([c],[r],[RNAP]) &= 1 + K_c[c] + K_r[r] + K_{cr}[c][r] + K_P[RNAP] \\ &\quad + K_{cP}[c][RNAP]. \end{aligned}$$

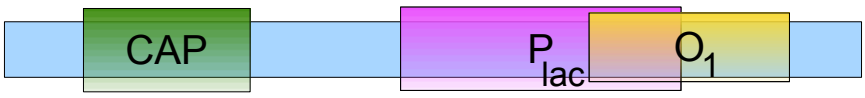


Figure 2.2: Regulatory region of *lac* operon in *E. Coli*: The *lac* operator O_1 (binding site for lacI) overlaps the *lac* promoter P_{lac}. When lacI is bound to the DNA, transcription cannot occur. The CAP binding site is upstream of the *lac* promoter and when bound by the CAP-cAMP complex it enhances the probability of RNA polymerase binding to the promoter.

2.2 Activators and Repressors

The simplest control design would involve each regulatory protein acting as either an "activator" or a "repressor." Heuristically, increasing the presence of an activator should result in a higher expression of the gene, while increasing the presence of a repressor

should lead to a lower level of gene expression. To analyze the behavior of the Shea-Ackers function, we begin our analysis with a generic regulatory region. We use this to extract a binding dependence constant and a normalized transcription initiation constant. These constants define control within the Shea-Ackers function. This is the macroscopic characterization of the activator and the repressor. With this in mind we make the following definitions.

Definition 2.2.1 A regulatory protein r is a *phenomenological activator* for a gene if the transcription rate of this gene always increases with the concentration of r, that is, $\frac{\partial f}{\partial [r]} > 0$ for all $[r] \geq 0$. Conversely, r is a *phenomenological repressor* for the gene in question if the transcription rate of this gene always decreases with the concentration of r, that is, $\frac{\partial f}{\partial [r]} < 0$ for all $[r] \geq 0$.

Example 2.2.2 Consider the *trp* operon as in Example 2.1.3. The Shea-Ackers function is given by (2.4). Differentiation gives

$$\frac{\partial f}{\partial [r]}([r], [RNAP]) = -\frac{k_P K_r K_P [RNAP]}{Z^2} < 0,$$

and hence the regulator r is a phenomenological repressor.

Example 2.2.3 Consider the *tox* gene regulation by DtxR protein. The diphtheria toxin is composed of two subunits that are synthesized from the *tox* gene. The regulatory region for the *tox* gene consists of an operator overlapping with the RNA polymerase binding site [45] (see Figure 2.3). The DtxR protein binds to the operator for the *tox* gene only in the presence of ferrous ions (Fe^{2+}) and prevents transcription. The *tox* gene is turned on when there is a low level of free iron. Since the regulatory region for the *tox* gene has the same configuration of Example 2.2.2, the Shea-Ackers function will be given by (2.4) with r representing the DtxR protein. It follows from Example 2.2.2 that the regulator DtxR is a phenomenological repressor.

An argument following the lines of Examples 2.2.2 and 2.2.3 leads to the following result.

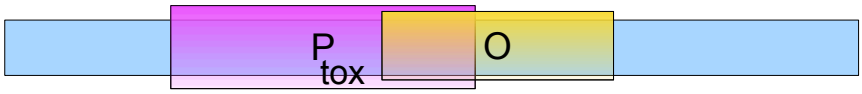


Figure 2.3: The *tox* operon regulatory region: The *tox* operator O (binding site for DtxR) overlaps with the *tox* promoter P_{tox}. When DtxR is bound to the DNA transcription cannot occur.

Proposition 2.2.4 *Consider a gene with regulators r_i all of whose binding sites overlap with the RNA polymerase binding site, and whose binding excludes binding of RNA polymerase. Then the regulators r_i are phenomenological repressors.*

As the name suggests, we use the adjective phenomenological to indicate the most directly observable relationship between the concentration of a regulatory protein and the production of the associated mRNA of the gene. However, given a set of possible states, the Shea-Ackers function has two free parameters, the association constant, $K_B(s)$, and the transcription initiation rate constant, $k(s)$. In principle these quantities can be determined by experiment. Thus, it makes sense to try to understand the phenomenological properties of regulatory proteins in terms of their association constants and transcription initiation rates. We begin with the following definition, which is justified by Theorem 2.2.6.

Definition 2.2.5 The decomposition $\{s_i \mid i = 1, \ldots, I\}$ of a state s in terms of its elementary states is *independent* if

$$K_B(s) = \prod_{i=1}^{I} K_B(s_i)$$

or equivalently

$$\Delta G_s = \sum_{i=1}^{I} \Delta G_{s_i}.$$

The following theorem shows that within the Shea-Ackers function all the regulation of gene transcription occurs because of the interactions of binding energies and transcription initiation rates. Without this, the Shea-Ackers function reduces to a Hill function describing interaction of RNA polymerase and the promoter. In particular,

the transcription rate is constant with respect to concentrations of regulatory proteins.

Theorem 2.2.6 *Consider a gene with the set of regulatory states $\mathcal{S}$ which satisfies the following two conditions:*

1. *for each state $s \in \mathcal{S}$ its decomposition into its elementary states is independent,*
2. *the rate of transcription initiation $k(s)$ does not depend on the state, i.e. $k(s) = k$ for all the states $s \notin \mathcal{S}_0$.*

Then the transcription rate is given by

$$f([RNAP], [r_1], \ldots, [r_m]) = k\frac{K_P[RNAP]}{1 + K_P[RNAP]}.$$

Proof. Consider a gene with a regulatory region that contains n distinct binding sites for m proteins, $\{r_1, \ldots, r_m\}$, and one binding site for RNA polymerase. Let k be the state independent rate of transcription initiation. To simplify notation we will denote $K_B(s_i)$ by K_{s_i}. Since each state is decomposed independently into its elementary states

$$\mathbb{P}_s = k\frac{K_{s_1}K_{s_2}\ldots K_{s_n}K_P[r_1]^{\alpha_s^1}[r_2]^{\alpha_s^2}\ldots[r_m]^{\alpha_s^m}[RNAP]}{Z},$$

and the partition function $Z([RNAP], [r_1], \ldots, [r_m])$ is given by

$$Z = 1 + \sum_{s\in\mathcal{S}}(K_{s_1}K_{s_2}\ldots K_{s_n}K_{s_{n+1}}[r_1]^{\alpha_s^1}[r_2]^{\alpha_s^2}\ldots[r_m]^{\alpha_s^m}[RNAP]^{\alpha_s}),$$

where $s_i \in \{r_1, \ldots, r_m, \emptyset\}$ for $i = 1, \ldots, n$ and either $s_{n+1}{=}s_P$ when RNA polymerase is bound, or $s_{n+1}{=}s_\emptyset$ when the RNA polymerase binding site is empty.

Therefore the transcription rate $f([RNAP], [r_1], \ldots, [r_m])$ is of the form

$$f = k\frac{A([RNAP], [r_1], [r_2], \ldots, [r_m])}{A([RNAP], [r_1], [r_2], \ldots, [r_m]) + B([r_1], [r_2], \ldots, [r_m])},$$

with $B = \sum_{s\in\mathcal{S}_0} K_{s_1}\ldots K_{s_n}[r_1]^{\alpha_s^1}\ldots[r_m]^{\alpha_s^m}$ and $A = K_P[RNAP]B$.

Then

$$f([RNAP],[r_1],\ldots,[r_m]) = k\frac{K_P[RNAP]B}{K_P[RNAP]B+B} = k\frac{K_P[RNAP]}{1+K_P[RNAP]},$$

which is a Hill function of the concentration of RNA polymerase, which is independent of the concentration $[r_i]$ of the transcription factors. □

Theorem 2.2.6 indicates that for control to occur there must be dependence of states and/or differences in the transcription initiation rate. To quantify these differences we introduce two new parameters.

Definition 2.2.7 Given a state $\{s_i \mid i = 1,\ldots,I\}$ its *binding dependence constant* is defined by

$$\beta_s := \frac{K_B(s)}{\prod_{i=1}^{I} K_B(s_i)}$$

and if $s \in \mathcal{S} \setminus \mathcal{S}_0$ its *normalized transcription initiation constant* is

$$\phi_s := \frac{k(s)}{k_P}.$$

Example 2.2.8 (Case of $\phi < 1$) Gyrase, an enzyme found in bacteria and plants, is composed of two subunits GyrA and GyrB, both of which are inhibited by Fis, a nucleoid protein [46, 47]. Fis inhibits GyrA by directly competing with RNA polymerase for the *gyrA* promoter. The control of GyrB expression is more interesting. In the presence of Fis, RNA polymerase stably binds the *gyrB* promoter, and even forms an open complex, but transcription still fails to initiate [48], see Figure 2.4. Because RNA polymerase in the presence of Fis freely and stably binds to the *gyrB* promoter, but transcription fails, this is an example of $\phi < 1$.

Example 2.2.9 (Case of $\phi > 1$) The *right operator* $\mathrm{O_R}$ in phage λ has three regions designated $\mathrm{O_R}1$, $\mathrm{O_R}2$ and $\mathrm{O_R}3$ (see Figure 2.5). The $\mathrm{O_R}$ region also contains two disjoint promoters $\mathrm{P_R}$ (*R*ight promoter) and $\mathrm{P_{RM}}$ (*R*epression *M*aintenance promoter). The promoter $\mathrm{P_R}$ completely overlaps $\mathrm{O_R}1$ and partially overlaps $\mathrm{O_R}2$; $\mathrm{P_{RM}}$ completely

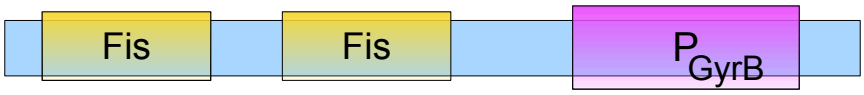


Figure 2.4: The regulatory region of *gyrB* as an example of $\phi < 1$. In the regulatory region of *gyrB* there are Fis binding sites upstream of the RNA polymerase binding site. When Fis is bound to the DNA it enhances RNA polymerase binding, but transcription fails to initiate.

overlaps O_R3 and partially overlaps O_R2. The gene cI, that codes for the repressor protein, and a gene cro, that codes for Cro protein, flank the O_R region. The binding of RNA polymerase to P_R initiates transcription of cro gene, while RNA polymerase binding to P_{RM} initiates transcription of the cI gene. When a CI_2 protein binds O_R2 it assists P_{RM} bound RNA polymerase to isomerize from a closed complex to an open complex, increasing the transcription rate [44]. This is an example of $\phi > 1$.

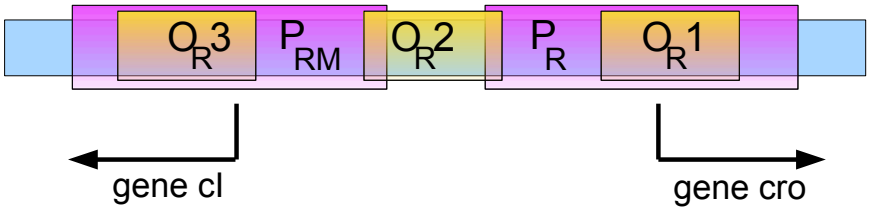


Figure 2.5: The right operator of phage λ as an example of $\phi > 1$. The cro gene is transcribed from the P_R promoter, while the cI gene is transcribed from P_{RM} promoter. The DNA regions O_R1, O_R2 and O_R3 are binding sites for either CI or Cro proteins.

2.3 Binding and Initiation Regulation

The key question we want to address is the correspondence between activation and repression on the biochemical level and on the macroscopic, or phenomenological level. We define a regulatory constant ρ which has a non-linear dependence on both the binding dependence constant and the normalized transcription initiation constant. In Proposition 2.3.3 we characterize a phenomenological regulator using ρ and show that in the simplest of settings a phenomenological activator is equivalent to $\rho > 1$ and a phenomenological repressor is equivalent to $\rho < 1$.

In Section 2.3.2 we find that the constant ρ determines whether a regulator is an activator or a repressor in a situation where multiple regulators compete for the same binding site. In section 2.3.3 we explore in more depth the unequal effect of K_B and k, two key parameters of the Shea-Ackers function, on the rate of transcription. In Section 2.3.4 and Section 2.3.5 we discuss activation and repression for the operators with two binding sites and one, or two regulators, respectively. We illustrate our results on examples of phage λ and *lac* operon.

2.3.1 The Simple Regulatory Region

The simplest nontrivial regulatory region has one binding site for the RNA polymerase and another for a single regulatory protein. We capture this in the following definition.

Definition 2.3.1 A *simple regulatory region* is defined by the set of states $\mathcal{S} = \{s_\emptyset, s_r, s_P, s_{rP}\}$ where $s_{rP} = \{s_r, s_P\}$.

The existence of the state s_{rP} implies that both RNA polymerase and the protein r can be bound to the DNA simultaneously. To simplify the notation, let

$$\beta_r := \frac{K_B(s_{rP})}{K_B(r)K_P} \quad \text{and} \quad \phi_r := \frac{k(s_{rP})}{k_P}.$$

Observe that the Shea-Ackers function is

$$\begin{aligned} f([r],[RNAP]) &= \frac{k_P K_P[RNAP] + k(s_{rP})K_B(s_{rP})[r][RNAP]}{1 + K_B(s_r)[r] + K_P[RNAP] + K_B(s_{rP})[r][RNAP]} \\ &= \frac{k_P K_P[RNAP]}{Z}\left(1 + \phi_r\beta_r K_r[r]\right). \end{aligned} \tag{2.6}$$

Whether r is an activator or repressor is determined by the sign of the derivative of f. Differentiating (2.6) gives

$$\frac{\partial f}{\partial [r]}([r],[RNAP]) = \frac{k_P K_r K_P[RNAP]}{Z^2}\left(\phi_r\beta_r(1 + K_P[RNAP]) - 1 - \beta_r K_P[RNAP]\right). \tag{2.7}$$

Failure of the regulatory protein to be an activator or a repressor at a particular concentration is equivalent to $\frac{\partial f}{\partial [r]}([r],[RNAP]) = 0$, that is,

$$\frac{\phi_r\beta_r(1+K_P[RNAP])}{1+\beta_r K_P[RNAP]} = 1.$$

This leads to the following definition.

Definition 2.3.2 Consider a regulatory region with a regulatory protein r for which the state $s_{rP} = \{s_r, s_P\}$ exists. The *regulatory constant* of r is

$$\rho_r := \frac{\phi_r\beta_r(1+K_P[RNAP])}{1+\beta_r K_P[RNAP]}. \tag{2.8}$$

As an example, consider an *E. coli* culture with growth rate $\mu \approx 0.02$ min^{-1}, which corresponds to a doubling time of 30 minutes, then there are approximately 1500 active RNA polymerase molecules per cell [49]. This corresponds to $[RNAP] \approx 3.0\ \mu M$ and hence

$$\rho_r := \frac{\phi_r\beta_r(1+3.0\cdot K_P)}{1+3.0\cdot\beta_r K_P}.$$

Rewriting (2.7) in terms of the regulatory constant we obtain

$$\frac{\partial f}{\partial [r]}([r],[RNAP]) = \frac{k_P K_r K_P[RNAP]}{Z^2}\left(1+\beta_r K_P[RNAP]\right)(\rho_r - 1). \tag{2.9}$$

Thus, as expected the sign of the derivative is determined by ρ_r. This gives the following result.

Proposition 2.3.3 *Consider a simple regulatory region with regulatory protein r. Then*

$$r \text{ is a phenomenological activator} \iff \rho_r > 1$$

and

$$r \text{ is a phenomenological repressor} \iff \rho_r < 1.$$

As immediate consequences of Proposition 2.3.3 we have the cases when either the binding constant β_r or the normalized transcription initiation constant ρ_r is one.

Corollary 2.3.4 *Consider a simple regulatory region and suppose that the rate of transcription initiation is independent of state, i.e.* $k(s_P) = k(s) = k$. *Then,*

$$r \text{ is a phenomenological activator} \iff \beta_r > 1$$

and

$$r \text{ is a phenomenological repressor} \iff \beta_r < 1.$$

Corollary 2.3.5 *Consider a simple regulatory region and assume that the decomposition of* s *into its elementary states is independent. Then,*

$$r \text{ is a phenomenological activator} \iff \phi_r > 1$$

and

$$r \text{ is a phenomenological repressor} \iff \phi_r < 1.$$

Observe that if $\rho_r = 1$, then $\frac{\partial f}{\partial [r]} \equiv 0$, and hence r has no regulatory impact. Figure 2.6 indicates the $\rho_r = 1$ isocline in the binding dependence constant and normalized transcription initiation constant plane.

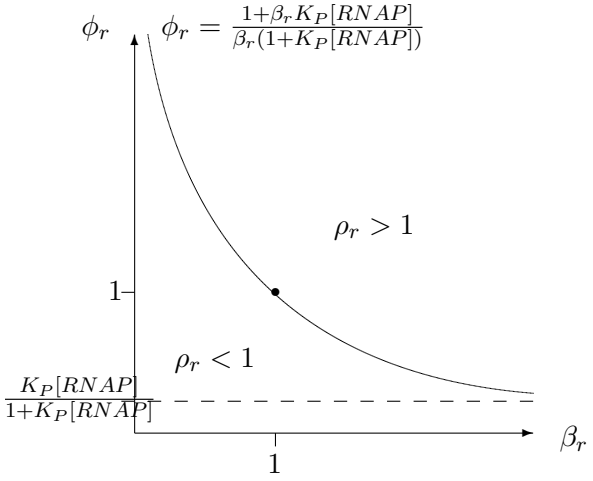


Figure 2.6: The relative importance of the dependence constant as opposed to the normalized transcription initiation constant in determining whether a single regulator is a phenomenological activator or repressor.

One of the consequences of this result is that the regulatory protein r is either a phenomenological activator or phenomenological repressor, but it cannot take on both functions. As is indicated in later subsections, the introduction of multiple regulatory proteins or multiple binding sites changes this.

2.3.2 Multiple Regulators, One Binding Site

The examples of multiple proteins competitively binding to the same site are ubiquitous and range from λ phage Cro and CI proteins [44] to eukaryotes [5].

With this in mind we turn our attention to the setting of a gene with n regulatory proteins r_i, $i = 1, \ldots, n$, one regulator binding site, and one RNA polymerase binding site. The associated collection of states is $\mathcal{S} = \{s_\emptyset, s_i, s_P, s_{iP} \mid i = 1, \ldots, n\}$, where $s_{iP} = \{s_i, s_P\}$. The initiation rate function is given by

$$f([r_1], \ldots, [r_n], [RNAP]) = \frac{k_P K_P [RNAP]}{Z} \left(1 + \sum_{i=1}^{n} \phi_i \beta_i K_i [r_i] \right) \tag{2.10}$$

where $Z = 1 + \sum_{i=1}^{n} K_i [r_i] + K_P [RNAP] + \sum_{i=1}^{n} \beta_i K_i K_P [r_i][RNAP]$.

Straightforward differentiation and substitution of (2.8) gives

$$\begin{aligned} \frac{\partial f}{\partial [r_i]} &= \frac{k_P K_i K_P [RNAP]}{Z^2} \cdot \frac{1 + \beta_i K_P [RNAP]}{1 + K_P [RNAP]} \\ &\quad \cdot \left((1 + K_P [RNAP])(\rho_i - 1) + \sum_{j=1}^{n} K_j (1 + \beta_j K_P [RNAP])[r_j](\rho_i - \rho_j) \right). \end{aligned} \tag{2.11}$$

From an experimental point of view, perhaps the easiest test for the regulatory nature of the protein r_i is to measure whether production of mRNA increases or decreases with respect to $[r_i]$ in the absence of the other regulatory proteins. Observe that

$$\left. \frac{\partial f}{\partial [r_i]} \right|_{[r_j]=0, j \neq i} = \frac{k_P K_i K_P [RNAP]}{Z^2} \cdot (1 + \beta_i K_P [RNAP])(\rho_i - 1). \tag{2.12}$$

Therefore the sign of $\frac{\partial f}{\partial [r_i]}$ for low concentrations of other proteins $[r_j], j \neq i$, is determined by the regulatory constant ρ_i.

To determine whether or not a regulator can change between an activator and

repressor requires identifying the regions where $\frac{\partial f}{\partial [r_i]} = 0$. Solving (2.11) for zero gives rise to the following hyperplane

$$\sum_{j=1}^{n} K_j(1 + \beta_j K_P[RNAP])(\rho_i - \rho_j)[r_j] = (1 + K_P[RNAP])(1 - \rho_i). \tag{2.13}$$

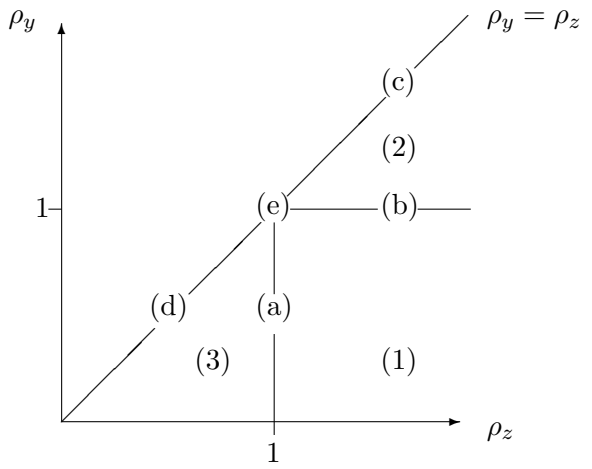


Figure 2.7: The regulatory constants ρ_y and ρ_z determine the roles of the regulatory proteins y and z as activators or repressors. In region (1) z is a phenomenological activator and y is a phenomenological repressor. For all values of ρ_z and ρ_y in region (2), there is a particular $[z]^*$ which designates whether y is an activator or repressor. Similarly, region (3) has a $[y]^*$ where z is an activator or repressor.

As a first application of this formalism we completely characterize the behavior of the regulatory proteins when $n = 2$. To make the notation more transparent let $y = r_1$ and $z = r_2$. Without loss of generality we restrict our attention to the case where $\rho_z \geq \rho_y$. (See Figure 2.7.)

Theorem 2.3.6 *Consider a gene with two regulatory proteins y and z, one regulator binding site, and one RNA polymerase binding site.*

1. *If $\rho_z > 1 > \rho_y$, then z is a phenomenological activator and y is a phenomenological repressor.*

2. *If $\rho_z > \rho_y > 1$, then z is a phenomenological activator and there exists $[z]^* > 0$ such that if $[z] > [z]^*$ then y is an activator and if $[z]^* > [z]$ then y is a repressor.*

3. *If $1 > \rho_z > \rho_y$, then y is a phenomenological repressor and there exists $[y]^* > 0$ such that if $[y] > [y]^*$ then z is an activator and if $[y]^* > [y]$ then z is a repressor.*

We also have the following special cases.

Corollary 2.3.7 *Consider a gene with two regulatory proteins y and z, one regulator binding site, and one RNA polymerase binding site.*

a. *If $1 = \rho_z > \rho_y$, then z is an activator for all $[y] > 0$ and y is a phenomenological repressor.*

b. *If $\rho_z > \rho_y = 1$, then z is a phenomenological activator and y is a repressor for all $[z] > 0$.*

c. *If $\rho_z = \rho_y > 1$, then z and y are phenomenological activators.*

d. *If $1 > \rho_z = \rho_y$, then z and y are phenomenological repressors.*

e. *If $\rho_z = \rho_y = 1$, then $\frac{\partial f}{\partial [z]} = \frac{\partial f}{\partial [y]} \equiv 0$. Thus neither z nor y are phenomenological activators or repressors.*

Proof of Theorem 2.3.6. In this simpler setting, (2.13) reduces to two equations which can be solved explicitly:

$$[y]^* = \frac{(1 + K_P[RNAP])(\rho_z - 1)}{K_y(1 + \beta_y K_P[RNAP])(\rho_y - \rho_z)} \tag{2.14}$$

and

$$[z]^* = \frac{(1 + K_P[RNAP])(\rho_y - 1)}{K_z(1 + \beta_z K_P[RNAP])(\rho_z - \rho_y)}. \tag{2.15}$$

Observe that there exists at most one positive value for $[y]^*$ and $[z]^*$.

1. Both $[y]^* < 0$ and $[z]^* < 0$, thus f is monotone in $[z]$ and $[y]$. The result follows from (2.12).

2. In this case, $[y]^* < 0$ and $[z]^* > 0$. Again, the signs of the derivatives are determined by (2.12).

Similar arguments prove *3* as well as Corollary 2.3.7 □

A presentation of the complete characterization of the behavior of more than two regulatory proteins is possible, but tedious. Instead we present a typical result in the case of three regulatory proteins.

Proposition 2.3.8 *Consider a gene with three regulatory proteins r_i, $i = 1, 2, 3$, one regulator binding site, and one RNA polymerase binding site. Assume $\rho_3 > \rho_2 > 1 > \rho_1$. Then*

1. *r_3 is a phenomenological activator.*

2. *r_1 is a phenomenological repressor.*

3. *If*

$$[r_3] > \frac{K_1(1+\beta_1 K_P[RNAP])(\rho_1-\rho_2)}{K_3(1+\beta_3 K_P[RNAP])(\rho_2-\rho_3)}[r_1] + \frac{(1+K_P[RNAP])(1-\rho_2)}{K_3(1+\beta_3 K_P[RNAP])(\rho_2-\rho_3)}$$

then r_2 is a phenomenological repressor, and if

$$[r_3] < \frac{K_1(1+\beta_1 K_P[RNAP])(\rho_1-\rho_2)}{K_3(1+\beta_3 K_P[RNAP])(\rho_2-\rho_3)}[r_1] + \frac{(1+K_P[RNAP])(1-\rho_2)}{K_3(1+\beta_3 K_P[RNAP])(\rho_2-\rho_3)}$$

then r_2 is a phenomenological activator. (See Figure 2.8.)

We note that the curve in the $[r_1], [r_3]$ plane that separates regions where r_2 is an activator and where r_2 is a repressor is a line where both the slope and the intercepts are functions of ρ_1, ρ_2 and ρ_3. This underscores the effectiveness of the regulatory constants in the characterization of activation and repression.

2.3.3 K_B- versus k-cooperativity

Consider a regulatory region with states $\mathcal{S}$. Let r be a regulatory protein. Denote the elementary state in which r is bound to the DNA by s_r. Define $\mathcal{S}^r \subset (\mathcal{S} \setminus \mathcal{S}_0)$ to be the set of states s which contain the elementary state s_r in their decomposition. The protein r exhibits *K_B-cooperativity* (*k-cooperativity*) if $\beta(s) > 1$ ($\phi(s) > 1$) for all $s \in \mathcal{S}^r$. We wish to compare the relative effect of K_B-cooperativity against k-cooperativity.

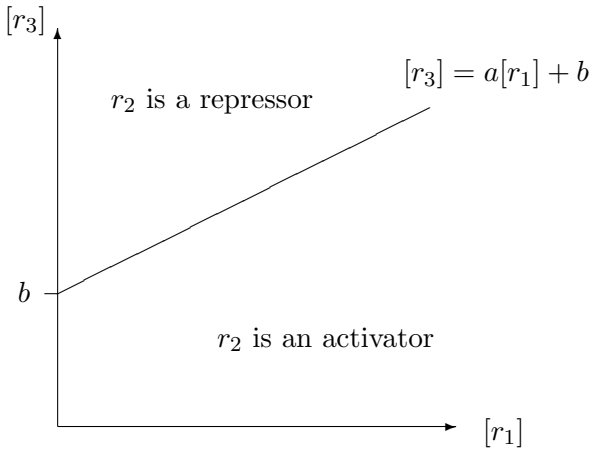


Figure 2.8: Under the assumption that $\rho_3 > \rho_2 > 1 > \rho_1$ we show that r_3 is a phenomenological activator and r_1 is a phenomenological repressor. A line in $[r_1], [r_3]$ plane separates regions where r_2 is an activator and where r_2 is a repressor, see Proposition 2.3.8. The slope $a = \frac{K_1(1+\beta_1 K_P[RNAP])(\rho_1-\rho_2)}{K_3(1+\beta_3 K_P[RNAP])(\rho_2-\rho_3)}$ and the intercept $b = \frac{(1+K_P[RNAP])(1-\rho_2)}{K_3(1+\beta_3 K_P[RNAP])(\rho_2-\rho_3)}$ are functions of ρ_1, ρ_2 and ρ_3.

The following theorem indicates that if a particular regulatory protein can produce either an "equal" amount of K_B-cooperativity or k-cooperativity, then the latter results in a greater rate of production of mRNA. Apart from the maximal production of mRNA there are very likely other evolutionary constraints imposed on the cell. A corollary of our result is that if a particular protein interacts with RNA polymerase by only K_B cooperativity, then there must be additional constraints worth the trade-off of decreased mRNA production.

Theorem 2.3.9 *Consider a regulatory region with states $\mathcal{S}$ and regulatory proteins $\{r, r_1, \ldots, r_n\}$. Let $s_{rP} = \{s_r, s_P\}$ where s_r is the elementary state where r is bound to the DNA. Assume that for all $s \in \mathcal{S}^r$, as defined above,*

$$K_B(s) = K(s_{rP})K_o \quad \textit{and} \quad k(s) = k(s_{rP}),$$

where K_o represents the association constant for the possible binding of $\{r_1, \ldots, r_n\}$. Let $f^{a,b}$ denote the initiation rate function under the assumption that $\beta_{s_r} = \frac{K(s_{rP})}{K_r K_P} = a$

and $\phi_{s_r} = \frac{k_r}{k_P} = b$. *If* $c > 0$, *then*

$$f^{1,1+c}([r],[r_1],\dots[r_n],[RNAP]) > f^{1+c,1}([r],[r_1],\dots[r_n],[RNAP]).$$

Proof. The initiation rate function $f^{a,b}$ has the form

$$\frac{ab\sum_{s\in\mathcal{S}^r} k_P K_r K_P K_o [r]^{\alpha_0}\cdots[RNAP]^{\alpha} + \sum_{s\in(\mathcal{S}\setminus(\mathcal{S}^r\cup\mathcal{S}_0))} k(s)K_B(s)[r_1]^{\alpha_1}\cdots[RNAP]^{\alpha}}{1 + a\sum_{s\in\mathcal{S}^r} K_r K_P K_o [r]^{\alpha_0}\cdots[RNAP]^{\alpha} + \sum_{s\in\mathcal{S}\setminus\mathcal{S}^r} K_B(s)[r_1]^{\alpha_1}\cdots[RNAP]^{\alpha}}$$

Thus, $f^{1,1+c}$ is given by

$$\frac{(1+c)\sum_{s\in\mathcal{S}^r} k_P K_r K_P K_o [r]^{\alpha_0}\cdots[RNAP]^{\alpha} + \sum_{s\in(\mathcal{S}\setminus(\mathcal{S}^r\cup\mathcal{S}_0))} k(s)K_B(s)[r_1]^{\alpha_1}\cdots[RNAP]^{\alpha}}{1 + \sum_{s\in\mathcal{S}^r} K_r K_P K_o [r]^{\alpha_0}\cdots[RNAP]^{\alpha} + \sum_{s\in\mathcal{S}\setminus\mathcal{S}^r} K_B(s)[r_1]^{\alpha_1}\cdots[RNAP]^{\alpha}}$$

while $f^{1+c,1}$ is given by

$$\frac{(1+c)\sum_{s\in\mathcal{S}^r} k_P K_r K_P K_o [r]^{\alpha_0}\cdots[RNAP]^{\alpha} + \sum_{s\in(\mathcal{S}\setminus(\mathcal{S}^r\cup\mathcal{S}_0))} k(s)K_B(s)[r_1]^{\alpha_1}\cdots[RNAP]^{\alpha}}{1 + (1+c)\sum_{s\in\mathcal{S}^r} K_r K_P K_o [r]^{\alpha_0}\cdots[RNAP]^{\alpha} + \sum_{s\in\mathcal{S}\setminus\mathcal{S}^r} K_B(s)[r_1]^{\alpha_1}\cdots[RNAP]^{\alpha}}$$

The numerators of the two terms are identical but the denominator of $f^{1+c,1}$ is larger since $c > 0$. □

The difference between K_B- and k-cooperativity may have consequences for the function of the organism. On the P_{RM} promoter of the phage λ, the CI_2 repressor interacts with RNA polymerase using k-cooperativity [44]. By using a detailed model of the induction process, which is based on experimental data, we predict in Chapter 3 that replacing k-cooperativity with the same amount of K_B-cooperativity yields a defective phage [30]. This mutant phage induces at a much lower level of radiation and is inherently unstable to noise.

Finally, our result can be viewed in the context of design of tightly controllable promoters in synthetic biology. Lanzer and Bujard [50] studied which factors most affect repressibility of promoters. They found that both the association constant of the RNA polymerase K_P and the promoter clearance rate k_P play key roles. In a later paper Lutz and Bujard [51] put the emphasis on the rate K_P, since stronger binding of RNA polymerase puts the repressor at a competitive disadvantage and hence a gene

with high K_P is difficult to repress. They construct tightly repressible and highly inducible synthetic operators from promoters with low to moderate K_P. Our results suggest that on highly inducible operators the cooperativity between the regulator and RNA polymerase will be characterized by a moderate binding dependence constant β and a very high normalized transcription initiation constant ϕ.

2.3.4 One Regulator, Two Binding Sites

We now consider a gene with one regulator r, two regulator binding sites, and one RNA polymerase binding site. The collection of non-empty states is

$$\mathcal{S} = \{s_1, s_2, s_{12}, s_P, s_{1P}, s_{2P}, s_{12P}\}$$

where s_i, $i = 1, 2$ denotes the elementary state of r bound in the i-th binding site and

$$s_{12} = \{s_1, s_2\}, s_{1P} = \{s_1, s_P\}, s_{2P} = \{s_2, s_P\}, s_{12P} = \{s_1, s_2, s_P\}.$$

To simplify the notation, let $K_* := K_B(s_*)$ and $\phi_* := \phi_{s_*}$.

Then the Shea-Ackers function takes the form

$$\begin{aligned} f([r],[RNAP]) &= \frac{k_P K_P [RNAP]}{Z}(1 + (\phi_{1P}\beta_{1P}K_1 + \phi_{2P}\beta_{2P}K_2)[r] \\ &\quad + \phi_{12P}\beta_{12P}K_1K_2[r]^2), \end{aligned} \tag{2.16}$$

where

$$\begin{aligned} Z([r],[RNAP]) &= 1 + K_P[RNAP] + (K_1 + K_2)[r] \\ &\quad + (\beta_{1P}K_1 + \beta_{2P}K_2)K_P[r][RNAP] + \beta_{12}K_1K_2[r]^2 \\ &\quad + \beta_{12P}K_1K_2K_P[r]^2[RNAP]. \end{aligned}$$

Recall the regulatory constants for each protein-binding site pair

$$\rho_1 := \frac{\phi_{1P}\beta_{1P}(1 + K_P[RNAP])}{1 + \beta_{1P}K_P[RNAP]} \quad \text{and} \quad \rho_2 := \frac{\phi_{2P}\beta_{2P}(1 + K_P[RNAP])}{1 + \beta_{2P}K_P[RNAP]}.$$

With two binding sites there is an additional state where both sites are occupied by the protein. We define a *regulatory constant for a pair* by

$$\rho_{12} := \frac{\phi_{12P}\beta_{12P}(1 + K_P[RNAP])}{\beta_{12} + \beta_{12P}K_P[RNAP]}. \tag{2.17}$$

Then after tedious computation the derivative of (2.16) can be written as

$$\begin{aligned} \frac{\partial f}{\partial [r]}([r],[RNAP]) &= \frac{k_P K_P[RNAP]}{Z^2}\Big(\alpha_1(\rho_1 - 1) + \alpha_2(\rho_2 - 1) \\ &\quad +2\alpha_{12}(\rho_{12} - 1)[r] \\ &\quad +[r]^2\frac{\alpha_{12}}{\alpha_P}\left(\alpha_1(\rho_{12} - \rho_1) + \alpha_2(\rho_{12} - \rho_2)\right)\Big), \end{aligned} \tag{2.18}$$

where

$$\begin{aligned} \alpha_P &:= 1 + K_P[RNAP] \\ \alpha_1 &:= K_1(1 + \beta_{1P}K_P[RNAP]) \\ \alpha_2 &:= K_2(1 + \beta_{2P}K_P[RNAP]) \\ \alpha_{12} &:= K_1K_2(\beta_{12} + \beta_{12P}K_P[RNAP]) \end{aligned}$$

are all positive constants.

Observe that $\frac{k_P K_P[RNAP]}{Z^2}$ is nonzero for all $[r]$ and the remainder of the derivative is a quadratic function

$$\frac{df}{d[r]} = \frac{k_P K_P[RNAP]}{Z^2}(A[r]^2 + B[r] + C) \tag{2.19}$$

where

$$\begin{aligned} A &:= \frac{\alpha_{12}}{\alpha_P}(\alpha_1(\rho_{12} - \rho_1) + \alpha_2(\rho_{12} - \rho_2)) \\ B &:= 2\alpha_{12}(\rho_{12} - 1) \\ C &:= \alpha_1(\rho_1 - 1) + \alpha_2(\rho_2 - 1). \end{aligned} \tag{2.20}$$

When the coefficients A, B, C have the same sign we have the following corollary.

Corollary 2.3.10 *Consider a gene with one regulator r, two regulator binding sites, and one RNA polymerase binding site. If*

$$\rho_{12} \geq \rho_1 \geq 1 \text{ and } \rho_{12} \geq \rho_2 \geq 1,$$

and at least one inequality is strict, then r is a phenomenological activator. If

$$\rho_{12} \leq \rho_1 \leq 1 \text{ and } \rho_{12} \leq \rho_2 \leq 1$$

and at least one inequality is strict, then r is a phenomenological repressor.

However, if the signs of the coefficients are not the same then r may be an activator for small $[r]$ and a repressor for large $[r]$. This indeed is the case in phage λ P_{RM} promoter where CI acts as an activator at low concentrations and as a repressor at high concentrations [44].

Example 2.3.11 The center of the regulatory processes in the phage λ is the *right operator* O_R, see Figure 2.5.

The lysogenic pathway corresponds to the state of the O_R where CI dimers are bound to both O_R2 and O_R1, blocking the P_R promoter and thus transcription of the cro gene, while RNA polymerase is free to bind P_{RM}, maintaining the transcription of the cI gene. In the lysogen O_R1 is almost always bound by a CI_2 protein and thus the production of Cro is very low. We simplify the situation by assuming that in fact O_R1 is always occupied by CI_2 and there is no production of Cro in lysogeny. Therefore we will only consider O_R2 and O_R3 binding sites and only the regulatory protein CI.

These assumptions imply that we are in the setting of a single regulatory protein with two binding sites. Let s_1 and s_2 correspond to the elementary states where CI is bound to O_R2 and O_R3, respectively. However, in the phage λ, O_R3 overlaps with P_{RM}. Mathematically this is incorporated by setting $\beta_{2P} = \beta_{12P} = 0$ which implies that $\rho_{12} = \rho_2 = 0$. This immediately implies that $A < 0$ and $B < 0$. The value of C can in principle be of both signs; if $C < 0$ then the CI would be a phenomenological repressor and if $C > 0$ then there is a unique positive value $[r]^*$ at which r switches from being

an activator to being a repressor. At the value $[r]^*$ the transcription initiation rate is at its maximum and in this respect the promoter is at its peak performance. Based on the experimental values collected in Santillan and Mackey [40] $\beta_{1P} = 1$, $\phi_{1P} = 12.26$ and thus $\rho_1 = 12.26$. Further $K_1 > K_2$ and $(1 + K_P[RNAP]) > 1$ and therefore $C > 0$ in phage λ.

The next question we address is whether it is possible to choose values of A, B, C in such a way that equation (2.19) has two positive roots. If it is possible and $A < 0$, then the regulatory protein r is an activator at low $[r]$, a repressor at intermediate $[r]$, and then an activator for large $[r]$. (For $A > 0$ the switch would be from repressor to activator and back to the repressor.)

While such A, B and C certainly exist for a general quadratic equation, as the next Proposition shows, for A, B and C as specified in (2.20) this is not possible.

Proposition 2.3.12 *Consider a gene with one regulator r, two regulator binding sites, and one RNA polymerase binding site. Then there is at most one positive value $[r]^*$ at which the derivative of the transcription rate function $f([r])$ can change the sign.*

Proof. For (2.19) to have two positive zeros either $A < 0, C < 0$ and $B > 0$ or all signs are reversed. We will show that this cannot happen. Assume $A < 0, C < 0$ and $B > 0$, the other case being analogous. The condition $B > 0$ is equivalent to

$$\rho_{12} > 1. \tag{2.21}$$

The conditions $A < 0$ and $C < 0$ are equivalent to solving

$$\begin{aligned} \alpha_1(\rho_{12} - \rho_1) + \alpha_2(\rho_{12} - \rho_2) &< 0 \\ \alpha_1(\rho_1 - 1) + \alpha_2(\rho_2 - 1) &< 0. \end{aligned} \tag{2.22}$$

Since $\alpha_i > 0, i = 1, 2$ the terms $\rho_{12} - \rho_2$ and $\rho_{12} - \rho_1$ have opposite signs. The same is true of $\rho_1 - 1$ and $\rho_2 - 1$. Assume that $\rho_1 - 1 > 0$ and $\rho_2 - 1 < 0$, the opposite case being analogous. Since $\rho_{12} > 1$ this forces $\rho_{12} - \rho_2 > 0$ and thus $\rho_{12} - \rho_1 < 0$. Then

the solution of the set of inequalities (2.22) is the region in the positive quadrant of the α_1, α_2 plane given by

$$\alpha_2 < \frac{-(\rho_{12} - \rho_1)}{\rho_{12} - \rho_2}\alpha_1, \quad \alpha_2 > \frac{\rho_1 - 1}{-(\rho_2 - 1)}\alpha_1$$

where both slopes are positive. This set has non-empty intersection in the positive quadrant if and only if

$$\frac{-(\rho_{12} - \rho_1)}{\rho_{12} - \rho_2} > \frac{\rho_1 - 1}{-(\rho_2 - 1)}$$

which is equivalent to

$$(\rho_{12} - \rho_1)(\rho_2 - 1) > (\rho_{12} - \rho_2)(\rho_1 - 1).$$

After simplification this inequality is equivalent to

$$\rho_{12}(\rho_2 - \rho_1) > \rho_2 - \rho_1.$$

Since $\rho_2 < \rho_1$ this contradicts (2.21). □

Example 2.3.13 (Optimal transcription depends on $\mathbf{K}_B$- vs. k-cooperativity.) We revisit our simplified λ phage lysogen maintenance model and discuss the dependence of the critical value $[r]^*$ on ϕ_{1P} and β_{1P} (see Section 2.3.3). The cooperativity between CI_2 and RNA polymerase is accomplished by O_R2 bound CI_2 increasing ϕ about 12-fold ($\phi_{1P} \approx 12$) without having any significant effect on binding probability of the polymerase) [52, 53] ($\beta_{1P} \approx 1$).

When $\beta_{12P} = \beta_2 = 0$ the coefficients of the quadratic equation in (2.19) are

$$A = -K_1^2 K_2 \beta_{12} \phi_{1P} \beta_{1P}, \quad B = -2K_1 K_2 \beta_{12}$$

and

$$C = K_1(\phi_{1P}\beta_{1P} - 1) - K_2) + \beta_{1P} K_1(\phi_{1P} - 1) K_P[RNAP].$$

Solving for the positive root of the quadratic equation in $[r]$ we get $[r]^* = \frac{1}{2A}(-B + \sqrt{B^2 - 4AC})$.

Now we discuss two cases. First, corresponding to the wild type phage, we let $\beta_{1P} = 1$ and $\phi_{1P} = \delta > 1$. Then

$$[r]^*_{wt} = \frac{1}{-K_1^2 K_2 \beta_{12} \delta} \left(-B + \sqrt{B^2 + 4K_1^2 K_2 \beta_{12} \delta (K_1(\delta - 1)(1 + \beta_{1P} K_P [RNAP]) - K_2)} \right). \tag{2.23}$$

The second case we analyze is in certain sense the opposite of the first one. For this fictitious mutant we set $\beta_{1P} = \delta > 1$ and $\phi_{1P} = 1$, which means that there a CI-RNAP binding cooperation, but CI does not enhance the transcription initiation. Then the critical concentration value is

$$[r]^*_{mut} = \frac{1}{-K_1^2 K_2 \beta_{12} \delta} \left(-B + \sqrt{B^2 + 4K_1^2 K_2 \beta_{12} \delta (K_1(\delta - 1) - K_2)} \right). \tag{2.24}$$

Since $1 + \beta_{1P} K_P [RNAP] > 0$ comparing (2.23) and (2.24) we see that

$$[r]^*_{wt} > [r]^*_{mut} \tag{2.25}$$

at the same δ.

Li *et. al.* [53] removed the positive control of the phage λ by an Arg to His change in the σ^{70} subunit of RNA polymerase. This corresponds to $\beta_{1P} = 1$ and $\phi_{1P} = 1$ in our model. In the same paper Li *et. al* report that when the mutant RNA polymerase was combined with the wild type CI, β_{1P} was increased, without significantly affecting ϕ_{1P}. By comparing the value of $[r^*]$ in such a mutant with the wild type value of $[r^*]_{wt}$ the prediction (2.25) can be verified experimentally.

Using k-cooperativity ($\phi_{1P} = \delta > 1$ and $\beta_{1P} = 1$) it takes a lower value of cooperation level δ to guarantee that CI_2 is an activator at low concentrations, compared to K_B-cooperativity ($\beta_{1P} = \delta > 1$ and $\phi_{1P} = 1$).

2.3.5 Two Binding Sites for Two Regulators

A canonical example for transcriptional control using two regulators is the E. coli *lac* operon. The transcription of the *lac* operon is controlled by lacI and CAP-cAMP complex (see Figure 2.1). This type of promoter has two regulators y and z, one binding site for regulator y, one binding site for regulator z and one RNA polymerase binding site. The feasible states are

$$\mathcal{S} = \{s_\emptyset, s_z, s_y, s_P, s_{zP}, s_{zy}\},$$

where $s_{zP} = \{s_z, s_P\}$ and $s_{zy} = \{s_z, s_y\}$. In particular, we assume that the states $s = \{s_z, s_y, s_P\}$ and $s_{yP} = \{s_y, s_P\}$ are not possible because of the mutual overlap between the y and P binding sites. Then the transcription rate is (compare (2.5))

$$f([z],[y],[RNAP]) = \frac{k_P K_P [RNAP]}{Z}(1 + \phi_z \beta_{zP} K_z[z]), \tag{2.26}$$

where

$$\begin{aligned} Z([z],[y],[RNAP]) &= 1 + K_z[z] + K_y[y] + \beta_{zy} K_z K_y [z][y] \\ &\quad + K_P[RNAP](1 + \beta_{zP} K_z[z]). \end{aligned}$$

Furthermore,

$$\begin{aligned} \frac{\partial f}{\partial [z]}([z],[y],[RNAP]) &= \frac{k_P K_P K_z [RNAP]}{Z^2}\Big((\rho_z - 1)(1 + \beta_{zP} K_P [RNAP]) \\ &\quad + (\phi_z \beta_{zP} - \beta_{zy}) K_y [y]\Big) \qquad (2.27) \\ \frac{\partial f}{\partial [y]}([z],[y],[RNAP]) &= -\frac{k_P K_P [RNAP]}{Z^2}(1 + \phi_z \beta_{zP} K_z [z]) \\ &\quad \cdot (K_y + \beta_{zy} K_z K_y [z]) \qquad (2.28) \end{aligned}$$

It is clear from expression (2.28) that the regulatory protein y is a phenomenological repressor. But in general we cannot label z as a phenomenological activator or repressor.

Example 2.3.14 Now we specialize further to the *lac* operon where y represents lacI and z represents the CAP-cAMP complex. By [54] it is known that $\phi_{zP} = 1$, and by [41] we know $\beta_{zy} = 1$ and $\beta_{zP} > 1$, therefore (2.27) simplifies to

$$\frac{\partial f}{\partial [z]}([z],[y],[RNAP]) \quad = \quad \frac{k_P K_P K_z [RNAP]}{Z^2}(\beta_{zP} - 1)(1 + K_y[y]) > 0.$$

It follows that z (CAP-cAMP) is a phenomenological activator. Therefore in the *lac* operon setup with the repressor blocking transcription by preventing RNA polymerase binding we recover the correspondence between biochemical and macroscopic markers of activation and repression.

An immediate consequence of having $\phi_{zP} = 1$ is that the sign of (2.27) has no dependence on RNA polymerase concentration. It is interesting to notice though that if $\phi_{zP} > 1$ instead of $\phi_{zP} = 1$, then z would still be a phenomenological activator and y a phenomenological repressor. Not only that, but z would be a more effective activator and y a more effective repressor since the derivatives would be greater in absolute value, but would keep the same signs. This situation could be achieved by moving the CAP binding region. This is the situation in the *gal* operon promoter P_1 [54]. It is not clear why this is not the regulation process adopted in the *lac* operon, since it seems that would be more effective regulation. It would be interesting to investigate whether there are other constraints that force *E. coli* to use this less than optimal regulator.

Another observation is that $\beta_{zy} = 1$ is also not a requirement for z to be a phenomenological activator and y a phenomenological repressor. Assuming $\beta_{zP} > 1$, it is sufficient to have $\beta_{zy} \leq \phi_{zP}\beta_{zP}$ to still have the same result. However, with these changes z would be less effective as an activator, and y less effective as a repressor. On the other hand, if $\beta_{zy} > \phi_{zP}\beta_{zP}$, then for high concentrations of y the regulator z would work as a repressor instead.

2.4 Discussion

Most of the conceptual models of transcriptional regulation assume monotonicity of the function relating product mRNA and the concentration of the regulator. This perhaps reflects the prevailing mode of data collection through knockout experiments where the absence of the putative regulator causes either an increase or a decrease of the mRNA production. On the modeling front this assumption leads to widespread use of Hill type response functions.

The Shea-Ackers model of transcriptional regulation was introduced more then 20 years ago. Using the chemical equilibrium assumption and using experimentally accessible parameters the resulting Shea-Ackers function relates concentrations of regulatory proteins, RNA polymerase and the geometry of the promoter to the transcription rate. The model has been matched to experimental data and the necessary parameters have been measured for at least a couple of canonical examples like *lac* operon in E. coli and phage λ switch.

The Shea-Ackers function reduces to a Hill function only in the case when there are no regulatory proteins. We show however that the Shea-Ackers function is still monotone for a promoter that contains a single binding site for a single regulatory protein in addition to a RNA polymerase binding site. If there are more binding sites, or more regulatory proteins, then non-monotonicity is common. While this non-monotonicity is used by certain organisms (CI_2 control of its own expression in phage λ), it may be tightly controlled in other cases by keeping concentrations of regulatory proteins in monotone regions. This opens up many new questions about regulatory circuit design and perhaps points to a need to revisit results that were obtained using the Hill model response function.

For all but the simplest of operators the key parameters of the Shea-Ackers model have a complicated, nonlinear effect of monotonicity of the transcription rate. We define a new regulatory constant ρ which greatly simplifies characterization of activation and repression for several complicated promoter designs. Since the constant is experimentally accessible it provides a new tool for the understanding of existing operators

as well as the design of new ones.

Chapter 3

Binding Cooperativity in Phage λ is Not Sufficient to Produce an Effective Switch

Transcriptional control plays a fundamental role in gene expression. The initiation of transcription involves a series of reactions which, as described before, can be summarized into three steps: binding, open complex formation and promoter escape. The activation and repression of transcription initiation is primarily caused by regulatory proteins and the structure of DNA. Regulated recruitment [44] provides a conceptual model for this process. Considerable progress has been made in understanding the biochemistry of the various reactions in the process [55, 14] and, in particular, it is clear that while the three steps are physically coupled there is considerable freedom for varying the respective energy profiles. To model these steps in the simplest way we will again treat opening and escape as a single chemical reaction with forward reaction rate k determined by the regulatory proteins and their interaction with the DNA. Binding will be treated as a reversible reaction with an equilibrium constant $\mathrm{K_B}$.

This simplification of the biochemistry allows one to develop thermodynamic models to quantify the rates of transcription initiation [55, 27, 39] that can be validated against experimental data [56, 41]. However, as described in Chapter 2, the combination of activators, repressors, and the above mentioned steps implies that control of transcription initiation is a highly nonlinear process, which in turn suggests that systematic mathematical analysis may lead to a deeper understanding of this regulatory mechanism. Given the goal of synthetic biology, claims based on the mathematical models must be experimentally verifiable.

More is known about the phage λ machinery than any other gene regulation mechanism [44, 57]. After infecting *E. coli*, the phage λ follows one of two pathways: *lysis*,

where it uses the bacterial molecular machinery to make many viral copies, kills the host bacterium and leaves to infect other cells; or *lysogeny*, where it integrates its DNA into the bacterial DNA and divides for generations with the bacterium. The lysogen exhibits great stability, yet it induces (switches to lysis) readily when the bacteria are irradiated with ultraviolet light.

The primary objective here is to use the above mentioned mathematical models to demonstrate that, in the context of the proper functioning of the phage λ induction, the binding constant K_B plays a fundamentally different role from the opening and clearing constant k. In particular, they are *not* interchangeable; that is, modifications in K_B cannot be directly compensated for by modifications in k and vice versa. To make this argument we begin, in Section 3.1 with a review of a simplified biological model of the phage λ switch and a precise statement of why increases in K_B are not equivalent to increases in k. In Section 3.2 we recall and explain the associated mathematical model and in Section 3.3 relate it back to the biology. We validate the model in Section 3.4 by considering several mutants, where our model recovers experimental observations of the lysogen stability. With this justification, in Section 3.5 we make several mathematical predictions concerning the unequal role played by RNA polymerase binding versus closed-open complex transition in transcription initiation process. These predictions are in principle experimentally testable.

3.1 The Phage λ Switch

The central controlling region for the lysogen maintenance is the *right operator* O_R, even though the long range cooperative binding with the O_L operator plays a crucial role in stability of the lysogen. (For a more complete description of the regulatory mechanisms, refer to [44].) O_R has three subregions designated O_R1, O_R2 and O_R3 (see Figure 2.5). The O_R region also contains two disjoint promoters P_R (*R*ight promoter) and P_{RM} (*R*epression *M*aintanence promoter). The promoter P_R completely overlaps O_R1 and partially overlaps O_R2, while P_{RM} completely overlaps O_R3 and partially overlaps O_R2. The gene cI, that codes for the repressor protein CI and the gene cro, that codes for Cro protein, flank the O_R region. Binding of either CI or Cro dimers (CI_2, Cro_2) to

O_R2 prevents binding of RNA polymerase to P_R, but it does not prevent such binding to P_{RM}. The initiation of transcription of cro occurs only if RNA polymerase binds to P_R. Similarly, the initiation of transcription of cI occurs only if RNA polymerase binds to P_{RM}.

The lytic pathway corresponds to a state where Cro_2 protein is bound to O_R3, blocking the P_{RM} promoter and thus transcription of cI. At the same time RNA polymerase is free to bind P_R, thus maintaining the transcription of cro. The lysogenic pathway corresponds to the state of O_R where CI_2 binds to both O_R2 and O_R1 blocking the P_R promoter and hence the transcription of cro. RNA polymerase is free to bind P_{RM} and thus maintain the transcription of cI. Even though these pathways are stable, the change from lysogeny to lysis, called *induction*, is experimentally well documented. When the bacterial population is subject to irradiation by UV light, the phage λ starts to lyse the bacteria and emerge in about 45 minutes. The irradiation causes RecA protein-mediated cleavage of CI which lowers its effective concentration [44, 58, 59, 60]. There are several key features which makes lysogen very stable and the induction "switch-like" [44].

1. High level of cooperativity between CI molecules: CI forms dimers CI_2 in the solution; when bound to neighboring regions O_R2 and O_R1 (or O_R2 and O_R3) it forms tetramers, and as described in [44], it forms octomers with CI_2 bound to the O_L operator, which is fairly distant, at 3.6kb, from O_R along the DNA strand.

2. Cooperative binding of CI_2 to O_R2 and O_R1: binding of CI_2 to O_R1 facilitates binding of another CI_2 molecule to O_R2.

3. Variable binding affinities of CI_2 and Cro_2 to different O_R regions: CI_2 has the highest affinity to O_R1, lower for O_R2 and lowest for O_R3, while Cro_2 has the highest affinity to O_R3, lower for O_R2 and O_R1.

4. Cooperative binding of CI_2 to O_R2 and RNA polymerase at P_{RM}: that is, O_R2 bound CI_2 increases the forward rate constant k at P_{RM} about 10-fold without having any significant effect on the binding of the RNA polymerase to the DNA [52].

We refer to the cooperativity in 4 as k-cooperativity. In an intriguing paper Li *et. al.* [53] have shown that after an Arg to His change in the σ subunit of RNA polymerase, the wild type CI activates mutant RNA polymerase by increasing K_B. We will refer to this cooperativity as K_B-cooperativity. This suggests that mutations allowing for an increase in K_B were (and are) evolutionary accessible to the phage. It is therefore likely that k-cooperativity, as opposed to an increase in K_B, has been selected for functional reasons. Further support for this hypothesis is provided by the fact that not all activators increase k. In fact in phage λ the factor CII acting on P_{RE} promoter uses both the K_B- and k-cooperativity [61] and the CAP activation of the *lac* operon in *E. coli* uses K_B-cooperativity [36].

To investigate this hypothesis we model the dynamics of the entire switch and study the effect of the K_B- and k-cooperativity on the stability of the lysogenic state. We show that the stability of the lysogen depends crucially not only on the fact that CI_2 interacts cooperatively with RNA polymerase, but also on the fact that this cooperativity increases k rather than K_B. In fact, our computations suggest that increasing K_B 100 fold while abolishing k-cooperativity yields phage with lysogen that is significantly less stable than the wild type.

3.2 The Mathematical Model

We make use of a delay differential equation model developed by Santillán and Mackey [40]:

$$\begin{aligned}
\frac{d[M_{cI}]}{dt} &= [O_R] f^c_{RM}([CI_2]_{\tau_M}, [Cro_2]_{\tau_M}) \qquad (3.1)\\
&\quad + [O_R] f_{RM}([CI_2]_{\tau_M}, [Cro_2]_{\tau_M}) - (\gamma_M + \mu)[M_{cI}]\\
\frac{d[M_{cro}]}{dt} &= [O_R] f_R([CI_2]_{\tau_M}, [Cro_2]_{\tau_M}) - (\gamma_M + \mu)[M_{cro}] \qquad (3.2)\\
\frac{d[CI]}{dt} &= \nu_{cI}[M_{cI}]_{\tau_{cI}} - (\gamma_{cI} + \mu)[CI] \qquad (3.3)\\
\frac{d[Cro]}{dt} &= \nu_{cro}[M_{cro}]_{\tau_{cro}} - (\gamma_{cro} + \mu)[Cro] \qquad (3.4)
\end{aligned}$$

which, as is explained below, tracks concentrations of cI mRNA, cro mRNA, CI protein and Cro protein. Concentrations are denoted by square brackets; that is $[CI]$ is the

total concentration of CI protein while $[M_{cro}]$ is the concentration of cro mRNA.

We will use $[Cro_2]$ and $[CI_2]$ to denote the concentration of CI and Cro dimers and $[RNAP]$ to denote concentration of the RNA polymerase. The concentration of the right operator is $[O_R]$. The subscript notation $[M_{cro}]_{\tau_{cro}}$ indicates that the concentration of cro mRNA is evaluated at time $t - \tau_{cro}$ where t is the present time. The time delays τ_{cI} and τ_{cro} are incorporated to take into account the fact that the production of the proteins from the associated mRNA and the actual process of transcription is not instantaneous.

Equations (3.3) and (3.4) are based on the assumption that the changes in protein concentrations are linear functions of the corresponding mRNA concentrations. There are two sets of positive decay constants. Since the volume of the growing bacteria increases, concentrations of all chemicals in a cell decrease. This is modeled by the decay constant δ which is the same in all equations. In addition, each chemical species experiences a specific degradation rate denoted by γ_*. Of particular interest is the constant γ_{cI}. We will model the effect of UV light, which, as observed earlier, lowers the effective concentration of CI dimers, by increasing the degradation rate γ_{cI} of the CI protein.

The ν_* are positive translation initiation constants.

The change in concentration of mRNA is described by equations (3.1) and (3.2). The nonlinear function $f_R([CI_2]_{\tau_M}, [Cro_2]_{\tau_M})$ describes the rate of transcription initiation at the promoter $\mathrm{P_R}$. For the sake of clarity the rate of transcription initiation at the promoter $\mathrm{P_{RM}}$ is expressed as the sum of two functions $f^c_{RM}([CI_2]_{\tau_M}, [Cro_2]_{\tau_M})$ and $f_{RM}([CI_2]_{\tau_M}, [Cro_2]_{\tau_M})$, where the first applies to the state of the operator in which $\mathrm{CI_2}$ is bound to $\mathrm{O_R}2$ and the second when it is not.

Santillán and Mackey's [40] construction of these functions is based on the work of Ackers *et. al.* [27] and, as described in Chapter 2, begins with expressions of the probability of binding of RNA polymerase to the promoter in the presence or absence of the regulatory proteins. The probability of a particular macroscopic state s of the

operator takes the form

$$\mathbb{P}_s([CI_2],[Cro_2]) = \frac{K_B(s)[Cro_2]^{\alpha_s}[CI_2]^{\beta_s}[RNAP]^{\gamma_s}}{\sum_i K_B(s_i)[Cro_2]^{\alpha_i}[CI_2]^{\beta_i}[RNAP]^{\gamma_i}} \tag{3.5}$$

where

$$K_B(s) = e^{\frac{-\Delta G_s}{RT}} \tag{3.6}$$

and the summation in the denominator is taken over all possible states. Since ΔG_s denotes the binding energy of the state, $K_B(s)$ determines the equilibrium constant for the biochemical reaction that results in binding of the regulatory proteins and/or RNA polymerase to the DNA in a closed form. The right (O_R), the left (O_L) operator (each of which has three subdomains) and the three promoters (P_R, P_{RM}, and P_L) are included in the model of Santillán and Mackey [40]. Therefore the state s of the operator is a description of which of the nine sites are empty or occupied by which of the three possible molecules CI_2, Cro_2, or RNA polymerase.

These probabilities need to be multiplied by an appropriate constant, $k(s)$, to incorporate the forward reaction rates of the opening and escape steps in order to obtain a rate of transcription initiation. Thus for each state, the transcription initiation rate has the form

$$f_s([CI_2],[Cro_2]) = k(s)\frac{K_B(s)[Cro_2]^{\alpha_s}[CI_2]^{\beta_s}[RNAP]^{\gamma_s}}{\sum_i K_B(s_i)[Cro_2]^{\alpha_i}[CI_2]^{\beta_i}[RNAP]^{\gamma_i}}. \tag{3.7}$$

Though clearly a simplification, we assume that the rate constants $k(s)$ take on three values: k_{cro} when RNA polymerase is bound to P_R, k_{cI}^c when RNA polymerase is bound to P_{RM} and CI_2 is bound to O_R2, and k_{cI} when RNA polymerase is bound to P_{RM} and CI_2 is not bound to O_R2.

Finally, f_R is the sum of all combinations of (3.7) with the restriction that each state s has a RNA polymerase bound to P_R, with O_R1 and O_R2 unbound. Similarly, f_{RM}^c is the sum of (3.7) for all states s which have RNA polymerase bound to P_{RM} and CI_2 bound to O_R2, and f_{RM} the sum of (3.7) for all states s which have RNA polymerase

bound to P_{RM} but CI_2 is not bound to O_R2.

To compare this model against experimental data, requires knowledge of the above mentioned constants. The experimentally determined values are taken from [40] and presented in Tables 3.1 and 3.2.

The binding energies ΔG_s are calculate using the following formula

$$\begin{aligned}\Delta G_s = &\sum_{X=R,L}\sum_{Y=CI_2,Cro_2}\sum_{\nu=1}^{3}\Delta G^Y_{O_{X^\nu}}\Gamma^Y_{O_{X^\nu}}(s)\\ &+\sum_{X=R,L}\sum_{Y=CI_2,Cro_2}\sum_{\nu=1}^{2}\Delta G^Y_{O_{X^{\nu\nu+1}}}\Gamma^Y_{O_{X^\nu}}(s)\Gamma^Y_{O_{X^{\nu+1}}}(s)\Gamma^{Cro_2}_{O_{X^{123}}}(s)\\ &+\sum_{X=R,L}\Delta G^{Cro_2}_{O_{X^{123}}}\Gamma^{Cro_2}_{O_{X^1}}(s)\Gamma^{Cro_2}_{O_{X^2}}(s)\Gamma^{Cro_2}_{O_{X^3}}(s)\\ &+\sum_{X=RM,R,L}\Delta G^{RNAP}_{P_X}\Gamma^{RNAP}_{P_X}(s)+\sum_{\nu=1}^{3}\Delta G_{RL}\Gamma^{CI_2}_{O_{R^\nu}}(s)\Gamma^{CI_2}_{O_{L^\nu}}(s)\end{aligned}$$

where

$$\Gamma^Y_X(k)=\begin{cases}1, & \text{if molecule } Y \text{ is bound to site } X;\\ 0, & \text{otherwise}\end{cases}$$

and

$$\Gamma^{Cro_2}_{O_{X^{123}}}(s)=\begin{cases}0, & \text{if } Cro_2 \text{ is bound to } O_R1,\ O_R2, \text{ and } O_R3\\ 1, & \text{otherwise}\end{cases}$$

All ΔG^*_* values in Table 3.2 are computed from [62]. The detailed explanation of how these energies have been computed can be found in [40]. The first sum includes all binding energies of transcription factors to the six binding sites on both left and right operator. The second sum includes all cooperation energies between any two adjacent factors and the third takes into account cooperativity that results from having Cro bound to all three binding sites on either O_R or O_L. It should be noted that in the measurements by Darling *et. al.* [62], the cooperative binding energies when Cro is bound to all three subdomains of O_R or O_L are not equal to the sum of the cooperative binding energies $\Delta G^{Cro_2}_{Ox12}$ and $\Delta G^{Cro_2}_{Ox23}$ (see Table 3.2). The term $\Gamma^{Cro_2}_{O_{X^{123}}}(s)$ in the second sum guarantees that when Cro occupies all three subdomains in O_R or O_L, the cooperative energies $\Delta G^{Cro_2}_{Ox12}$ and $\Delta G^{Cro_2}_{Ox23}$ are not included in this sum. The energies $\Delta G^{Cro_2}_{Ox123}$ are then added in the third sum. The fourth sum adds the RNA polymerase binding energy for the state, and the last one contributes any cross cooperation between CI_2 molecules

bound to P_R and P_L.

Table 3.1: Estimated parameter values from [40] (with the addition of ϕ) for equations (3.1)-(3.4).

$\mu \simeq 2.0 \times 10^{-2}$ min^{-1}	$k_{cro} \simeq 2.76$ min^{-1}
$k_{cI}^{c} \simeq 4.29$ min^{-1}	$k_{cI} \simeq 0.35$ min^{-1}
$\gamma_M \simeq 0.12$ min^{-1}	$\gamma_{cI} \simeq 0.0$ min^{-1}
$\gamma_{cro} \simeq 1.6 \times 10^{-2}$ min^{-1}	$\nu_{cI} \simeq 0.09$ min^{-1}
$\nu_{cro} \simeq 3.2$ min^{-1}	$\tau_{cI} \simeq 0.24$ min
$\tau_{cro} \simeq 6.6 \times 10^{-2}$ min	$\tau_M \simeq 5.1 \times 10^{-3}$ min
$K_D^{cI} \simeq 5.56 \times 10^{-3} \mu$M	$K_D^{cro} \simeq 3.26 \times 10^{-1} \mu$M
$[O_R] \simeq 5.0 \times 10^{-3} \mu$M	[RNAP] $\simeq 3.0$ μM
$\Delta G_{RL} \simeq -3.1$ kcal/mol	$\phi \simeq 4.29/.35 = 12.26$

3.3 Interpreting the Model

Based on the biochemistry of the phage λ switch, the phenomenological state of lysogeny is associated with low levels of Cro and high levels of CI. Similarly, lysis is associated with low levels of CI and high levels of Cro. With this in mind, we look for equilibria of the system (3.1)-(3.4) and declare that an equilibrium for which $0 \approx$ [Cro] $<<$ [CI] is a lysogenic equilibrium and an equilibrium for which $0 \approx$ [CI] $<<$ [Cro] is a lytic equilibrium.

The equilibria of this system are steady (time independent) states of the system and thus are not dependent on delays. Notice that since both CI and Cro proteins form dimers, the right hand side of the equations (3.1)-(3.4) depend on the concentration of dimers. We need the conversion formula for computing the concentration of dimers from total concentration of monomers.

Consider the chemical reaction

$$2a_1 \underset{k_-}{\overset{k_+}{\longleftrightarrow}} a_2$$

where a_1 is a free monomer form of the protein a and a_2 represents a dimer of protein a, k_+ and k_- are the forward and backward rate constants respectively.

In chemical equilibrium with $K_D = k_-/k_+$, we have the following relation:

$$[a_1]^2 = K_D[a_2]. \tag{3.8}$$

Table 3.2: Estimated binding energies from [40].

$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_R}1} \simeq -12.5$ kcal/mol	$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_L}1} \simeq -11.5$ kcal/mol
$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_R}2} \simeq -10.5$ kcal/mol	$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_L}2} \simeq -9.7$ kcal/mol
$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_R}3} \simeq -9.5$ kcal/mol	$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_L}3} \simeq -9.7$ kcal/mol
$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_R}12} \simeq -2.7$ kcal/mol	$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_L}12} \simeq -2.7$ kcal/mol
$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_R}23} \simeq -2.9$ kcal/mol	$\Delta G^{\mathrm{CI}_2}_{\mathrm{O_L}23} \simeq -2.9$ kcal/mol
$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_R}1} \simeq -12.0$ kcal/mol	$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_L}1} \simeq -12.0$ kcal/mol
$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_R}2} \simeq -10.8$ kcal/mol	$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_L}2} \simeq -10.8$ kcal/mol
$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_R}3} \simeq -13.4$ kcal/mol	$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_L}3} \simeq -13.4$ kcal/mol
$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_R}12} \simeq -1.0$ kcal/mol	$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_L}12} \simeq -1.0$ kcal/mol
$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_R}23} \simeq -0.6$ kcal/mol	$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_L}23} \simeq -0.6$ kcal/mol
$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_R}123} \simeq -0.9$ kcal/mol	$\Delta G^{\mathrm{Cro}_2}_{\mathrm{O_L}123} \simeq -0.9$ kcal/mol
$\Delta G^{\mathrm{RNAP}}_{\mathrm{P_R}} \simeq -12.5$ kcal/mol	$\Delta G^{\mathrm{RNAP}}_{\mathrm{P_L}} \simeq -11.3$ kcal/mol
$\Delta G^{\mathrm{RNAP}}_{\mathrm{P_{RM}}} \simeq -11.5$ kcal/mol	

K_D is the dissociation constant. In addition, if $[a]$ is the total monomer concentration,

$$[a] = [a_1] + 2[a_2]. \tag{3.9}$$

The equations (3.8)-(3.9) can be used to solve for $[a_2]$ leading to

$$[a_2] = \frac{[a]}{2} - \frac{K_D}{8}\left(\sqrt{1 + 8\frac{[a]}{K_D}} - 1\right)$$

from which follows that

$$[CI_2] = \frac{1}{2}[CI] - \frac{K_D^{cI}}{8}\left(\sqrt{1 + 8\frac{[CI]}{K_D^{cI}}} - 1\right) \tag{3.10}$$

$$[Cro_2] = \frac{1}{2}[Cro] - \frac{K_D^{cro}}{8}\left(\sqrt{1 + 8\frac{[Cro]}{K_D^{cro}}} - 1\right) \tag{3.11}$$

Let

$$\phi := \frac{k_{cI}^c}{k_{cI}}.$$

Observe that this provides a measure of the effect of O_R2 bound CI_2 on the forward reaction rate associated with opening and escape. In particular, $\phi > 1$ implies that the rate of transcription initiation with O_R2 bound CI_2 is higher than that without. We refer to this as *k-cooperativity*.

As is indicated in Section 3.2, γ_{cI} indicates the degradation rate of $[CI]$, induced for example by exposure to UV radiation. Since this is known to trigger induction of phage, we study the equilibria as a function of γ_{cI}. Observe that the equilibria satisfy the two equations

$$\Phi([CI],[Cro],\gamma_{cI}) = 0 \qquad \text{and} \qquad \Theta([CI],[Cro]) = 0$$

where

$$\begin{aligned}
\Phi([CI],[Cro],\gamma_{cI}) &= \frac{\nu_{cI}}{\gamma_M+\mu}[O_R]\left(f^c_{RM}([CI_2],[Cro_2]) + f_{RM}([CI_2],[Cro_2])\right) \\
&\quad -(\gamma_{cI}+\mu)[CI] \\
\Theta([CI],[Cro]) &= \frac{\nu_{cro}}{\gamma_M+\mu}[O_R]f_R([CI_2],[Cro_2]) - (\gamma_{cro}+\mu)[Cro].
\end{aligned}$$

The intersection of these two curves in the [CI], [Cro] plane determines two protein concentrations at a dynamical equilibrium; the remaining two concentrations $[M_{cI}]$ and $[M_{cro}]$ can be found from equations (3.3) and (3.4) with the left hand side set equal to zero.

Observe that Θ is independent of γ_{cI}. The set $\Theta([CI],[Cro]) = 0$ is given by the black curve in Figure 3.1. According to Table 3.1, for wild type phage in the absence of UV radiation, $\gamma_{cI} = 0$ min^{-1}. The set $\Phi([CI],[Cro],0) = 0$ is plotted in red dash in Figure 3.1. There is a unique equilibrium, i.e. intersection point of $\Theta([CI],[Cro]) = 0$ and $\Phi([CI],[Cro],0) = 0$, for which $[CI] = 0.528\,\mu M$ and $[Cro] = 1.04 \times 10^{-5}\mu M$. This is a *lysogenic equilibrium*.

As the parameter γ_{cI} increases the $\Phi = 0$ curve shifts its relative position relative to the $\Theta = 0$ curve. When γ_{cI} is 0.00039 min^{-1}, a pair of new intersections corresponding to new equilibria appear. Plotted in blue dots in Figure 3.1 is $\Phi([CI],[Cro],0.05) = 0$. The equilibrium with high value of $[Cro]$ and low value of $[CI]$ corresponds to lytic state and we call it a *lytic equilibrium*. Observe that there are three equilibria: a lysogenic equilibrium, a lytic equilibrium, and an unstable intermediate equilibrium. Finally, the green dash-dot curve represents $\Phi([CI],[Cro],0.35) = 0$ which intersects $\Theta = 0$ in a single point corresponding to the lytic equilibrium.

Clearly, the set of equilibria changes as a function of γ_{cI}. This is indicated in the bifurcation diagram of Figure 3.2, where the equilibrium values of $[Cro]$ are plotted on the vertical axis as a function of γ_{cI}. This graph allows us to describe the induction process. When no UV

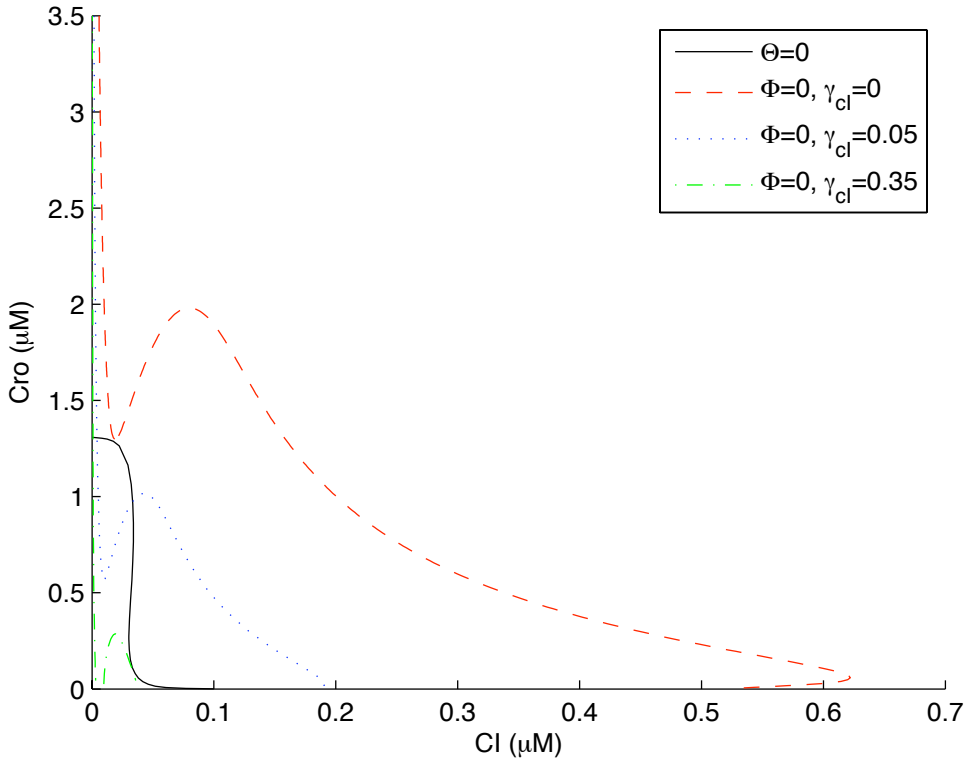


Figure 3.1: Nullclines for $\Theta = 0$ (black solid) and $\Phi = 0$ with $\gamma_{cI} = 0$ min^{-1} (red dash), $\gamma_{cI} = 0.05$ min^{-1} (blue dots) and $\gamma_{cI} = 0.35$ min^{-1} (green dash-dot).

radiation is applied to bacterial population, $\gamma_{cI} = 0$ min^{-1} and the phage occupies lysogenic equilibrium. As γ_{cI} is slowly increased, the lysogenic equilibrium moves and the phage state tracks this slowly moving equilibrium. Immediately after γ_{cI} crosses the value of 0.343 the lysogenic equilibrium disappears and the state rapidly approaches the lytic equilibrium.

Therefore we define the value $\gamma^*_{WT} := 0.343$ min^{-1} as the *wild type induction value*. The color code in Figure 3.2 shows the values of γ_{cI} that correspond to the same color curves in Figure 3.1.

In Sections 3.4 and 3.5 we make use of bifurcation diagrams such as that of Figure 3.2, thus we point out some of the important features. For the parameter values 0.00039 min^{-1} $\leq \gamma_{cI} \leq$ 0.343 min^{-1} the wild type phage λ switch is *bistable*; that is there are two stable equilibria, the lysogenic equilibrium (corresponding to the lower branch) and the lytic equilibrium (corresponding to the upper branch), and furthermore, for some initial concentrations the state of the phage will evolve toward the lysogenic equilibrium and for other initial concentrations toward the lytic equilibrium.

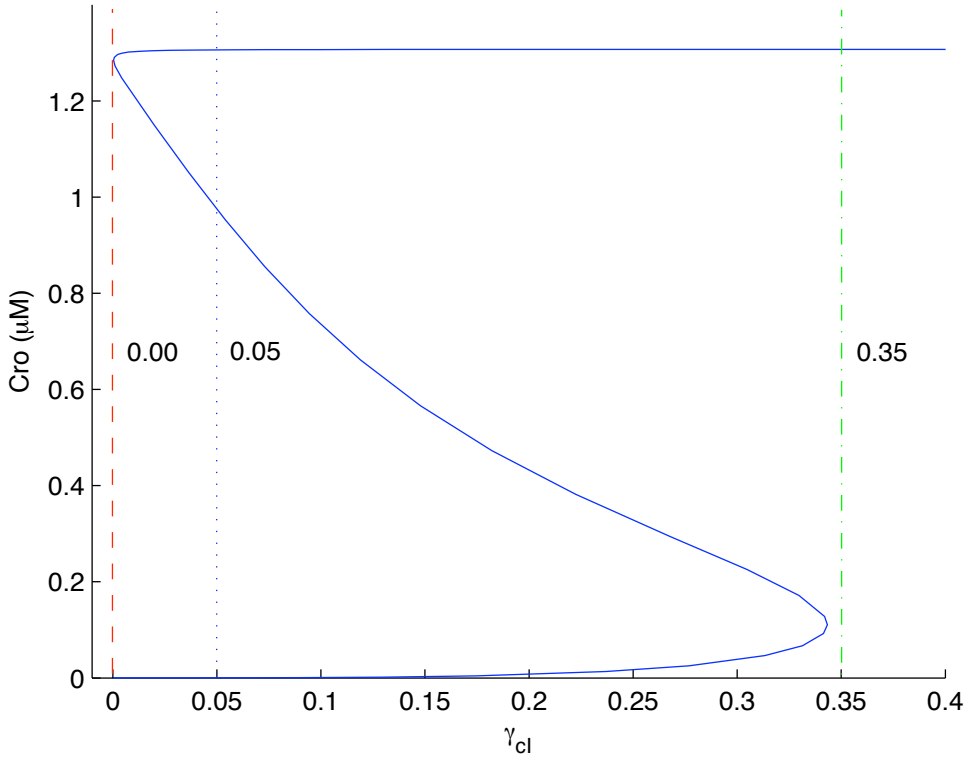


Figure 3.2: Bifurcation diagram of γ_{cI} versus [Cro].

We introduced the dimensionless parameter ϕ to have a measure of the change in the forward reaction rate associated with opening and escape. We wish to have a similar measure for the binding probabilities. When the binding of a transcription factor increases RNA polymerase residence time on the promoter, it is reflected in the Ackers model in the cooperative increase of the binding energy of the transcription factor-RNAP pair. We denote the binding energy between CI_2 and O_R2 by $\Delta G^{CI_2}_{O_R2}$ and binding energy between RNA polymerase and P_{RM} by $\Delta G^{RNAP}_{P_{RM}}$. In the absence of binding cooperation, as is the case in the wild type phage λ, the binding energy contribution from O_R2-bound CI and P_{RM}-bound RNA polymerase to any state s that contains them is

$$\Delta G_{\text{ind}}(s) := \Delta G^{CI_2}_{O_R2} + \Delta G^{RNAP}_{P_{RM}} + \Delta G_{\text{rest}}(s),$$

where subscript 'ind' stands for independent binding of the binding factors and $\Delta G_{\text{rest}}(s)$ is the binding energy of the other factors in state s.

The cooperative binding between CI_2 and RNA polymerase is reflected in additional binding energy $\Delta G^{CI_2RNAP}_{O_R2P_{RM}}$. If this energy is positive we refer to this as *K_B-cooperativity*. We express

the cooperativity in terms of the binding constant $K_B(s)$ (see (3.6))

$$K_B(s) := \beta K_B^{\text{ind}}(s)$$

where $K_B^{\text{ind}}(s) = exp(-\frac{1}{RT}(\Delta G_{\text{ind}}(s)))$ and the state s independent multiplicative factor

$$\beta := exp(-\frac{1}{RT}(\Delta G_{\text{O}_\text{R}2\text{P}_\text{RM}}^{\text{CI}_2\text{RNAP}})).$$

In this formulation $\beta > 1$ represents the cooperative binding.

In summary, the k-cooperativity is manifested by the constant $\phi > 1$ (see Section 3.2) and K_B-cooperativity by $\beta > 1$.

3.4 Model Validation

In order to validate our biological interpretation of the equilibria of equations (3.1)-(3.4) we model the induction scenarios for several different phage mutants which are described in the literature.

3.4.1 O_R323 Mutant

Little *et. al.* [63] constructed a mutant O_R323 in which the O_R1 domain was replaced by O_R3 and reported the following results:

R1 O_R323 can lysogenize;

R2 O_R323 has a threshold response, but at lower doses of UV radiation and at a higher level of free phage in the lysogen than the wild type;

R3 in the lytic state the burst size i.e. the number of phages per infected cell, of O_R323 is lower than that of wild type.

This mutation is easily incorporated into the mathematical model. To replace the O_R1 binding site by the O_R3 binding site we set the binding energy of CI_2 to O_R1 to be that of CI_2 to O_R3 (-9.5 kcal/mol). Similarly, the binding energy of Cro to O_R1 is set to that of Cro to O_R3 (-12.0 kcal/mol).

The bifurcation curves for the O_R323 mutation as compared with the wild type are presented in Figure 3.3. The graph shows the concentration of Cro as a function of γ_{cI}. The blue solid curve represents the wild type phage, while the black dashed curve represents the O_R323. The

lower branch on both curves corresponds to the lysogenic equilibrium and the upper branch to the lytic equilibrium.

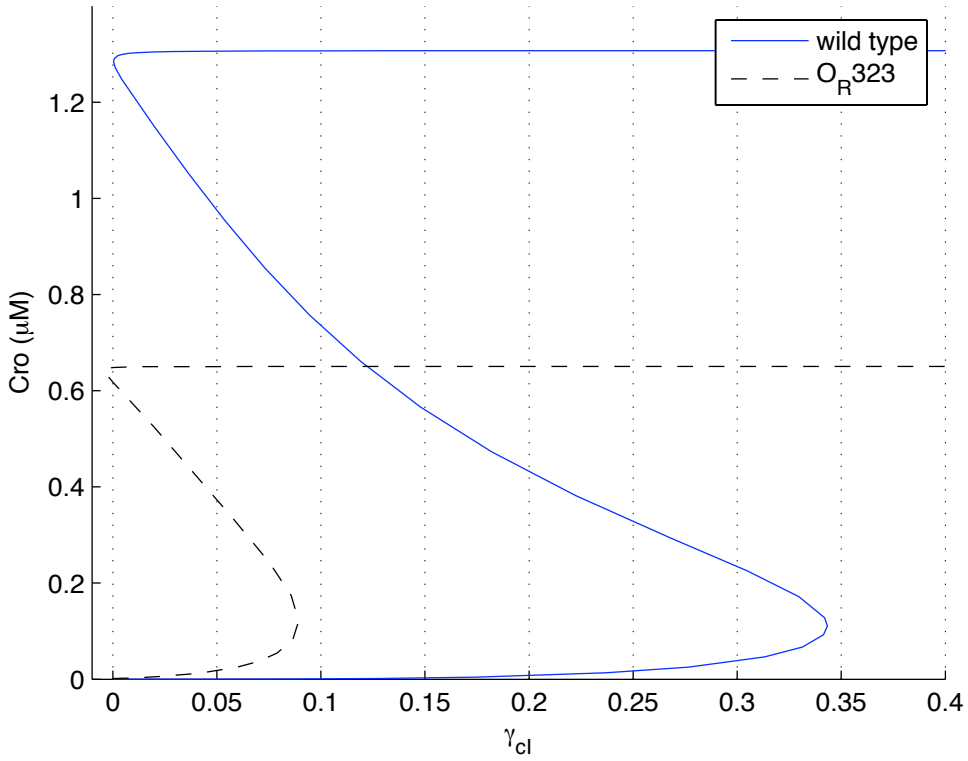


Figure 3.3: Bifurcation diagrams for wild type and O_R323. The concentration of Cro is graphed as a function of γ_{cI}. The blue solid curve represents the wild type phage, while the black dashed curve represents the mutant.

The existence of the lower branch in Figure 3.3 implies that O_R323 can lysogenize (compare **R1**). However, the induction value for the O_R323 mutant is $\gamma^*_{O_R323} = 0.09$ min^{-1} $<$ 0.34 min^{-1} $= \gamma^*_{WT}$, which suggests that a lower level of UV radiation is required to induce lytic growth (compare **R2**). Observe that when $\gamma_{cI} = 0$ min^{-1} there are three equilibria in the system describing O_R323. Thus a stable lytic equilibrium is present even in the absence of UV radiation and thus in the presence of noise some phages can spontaneously induce and switch to lytic state. This would manifest itself experimentally in increased number of free phages (compare **R2**).

Finally, it is possible that the burst size (number of phages per infected cell) is proportional to the transcription level of the lytic pathway in phage's genome, which in turn may be proportional to the level of Cro production in the lytic state. This theory is in agreement with Figure 3.3 in which the Cro production in the lytic state for O_R323 (the upper black dashed

branch) is significantly lower than in the wild-type lytic state (the upper blue solid branch) (compare **R3**). Of course, the burst size can also be determined by energetics of the cell or by available resources, and therefore the suggested relationship between Cro production and the burst size is, at best, speculative.

3.4.2 P_{RM} Mutant

Michalowski and Little [64] (see also [65]) obtained multiple mutants of phage λ by subjecting the P_{RM} binding site to mutagenesis. These were then compared to wild type by three criteria: the ability to grow lytically, the ability to establish and maintain a stable lysogenic state, and the ability to undergo prophage induction. In the experiments they were particularly careful not to affect the O_R2 and O_R3 binding sites. Of these isolates they further analyzed nine which were selected because they were comparable to or more difficult to induce than the wild type. When compared to wild type these nine strains seem to share three properties: they had an equal or increased P_{RM} binding affinity, a decreased P_R binding affinity, and an increase in the k-cooperativity between CI_2 and RNA polymerase. To model such mutant we set $P_{RM}= -12.5$ kcal/mol, $P_R= -10.5$ kcal/mol, and $\phi = 4.5/.35$, which should be compared to wild type values $P_{RM}= -11.5$ kcal/mol, $P_R= -12.5$ kcal/mol and $\phi = 4.29/.35$. The resulting bifurcation diagrams are presented in Figure 3.4 The induction parameter $\gamma^*_{P_{RM}} \simeq 0.85$ min^{-1} for the mutation is much higher then the wild type $\gamma^*_{WT} \simeq 0.35$ min^{-1} implying greater stability of the lysogen.

3.4.3 cI-pc Mutant

When a *pc* mutation is introduced to CI it eliminates the k-cooperativity between CI_2 protein bound to O_R2 and RNA polymerase [44]. This mutant forms lysogen in a wild-type bacteria, but suffers from high rate of spontaneous induction and induction at a very low levels of UV light.

To model this mutant we replace the k^c_{cI} in the function f^c_{RM} (see equation (3.2)) by k_{cI}. This implies $\phi = 1$. The associated bifurcation curves are indicated in Figure 3.5. Observe that our model predicts that the induction value is dramatically lower ($\gamma^*_{WT} = 0.34$ min^{-1} in wild type, $\gamma^*_{CIpc} = 0.01$ min^{-1} in the mutant). In the noisy environment of a cell we expect that this low stability threshold will yield a high spontaneous induction rate.

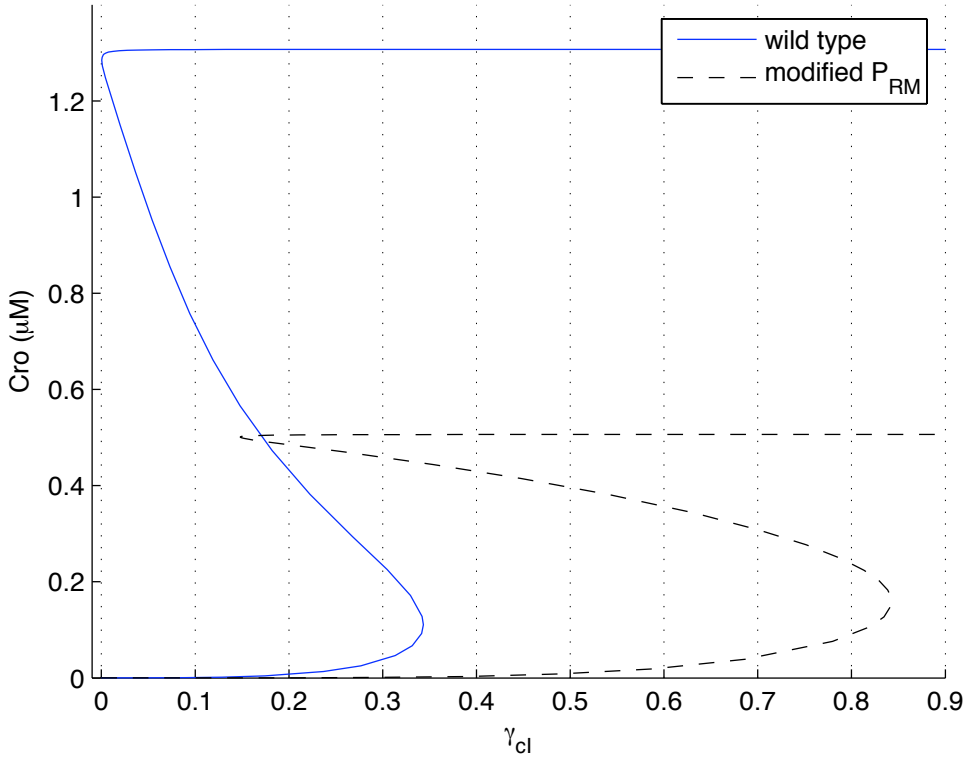


Figure 3.4: Bifurcation diagram of wild type vs. a phage with mutated P_{RM} binding site which resulted in having P_{RM}= −12.5 kcal/mol, $\phi = 4.5/.35$, and P_R= −10.5 kcal/mol. For comparison, the wild type values were P_{RM}= −11.5kcal/mol, $\phi = 4.29/.35$, and P_R= −12.5kcal/mol.

3.5 K_B- and k-cooperativity Are Not Interchangeable

Our most significant prediction is that K_B- and k-cooperativity affect the stability of the lysogen differently, and thus are not interchangeable. To demonstrate this we compare the stability of the lysogen under k-cooperativity, $\beta = 1, \phi = \alpha > 1$, against K_B-cooperativity, $\phi = 1, \beta = \alpha > 1$, for different values of α. The analysis of the stability of the cI-pc mutant in Section 3.4.3 provides the first step of this analysis. In this mutant both $\phi = 1$ and $\beta = 1$, thus all cooperation is abolished and our model predicts that the induction value is dramatically lower.

To test the ability of K_B-cooperativity to restore the lysogen stability, we fix $\phi = 1$ and solve for the equilibria at $\beta = 10$ and $\beta = 100$. The bifurcation diagrams are presented in Figure 3.6 where they can be compared against the cI-pc mutant and the wild type (recall that for the wild type $\phi \approx 12$ and $\beta = 1$). Observe that when $\beta = 10$, the induction value is $\gamma^*_{\beta=10} = 0.04$ min^{-1} which is much lower than $\gamma^*_{WT} = 0.34$ min^{-1}. We predict that this produces a very unstable lysogen. Even in the case of unrealistically strong K_B-cooperativity, $\beta = 100$, the induction

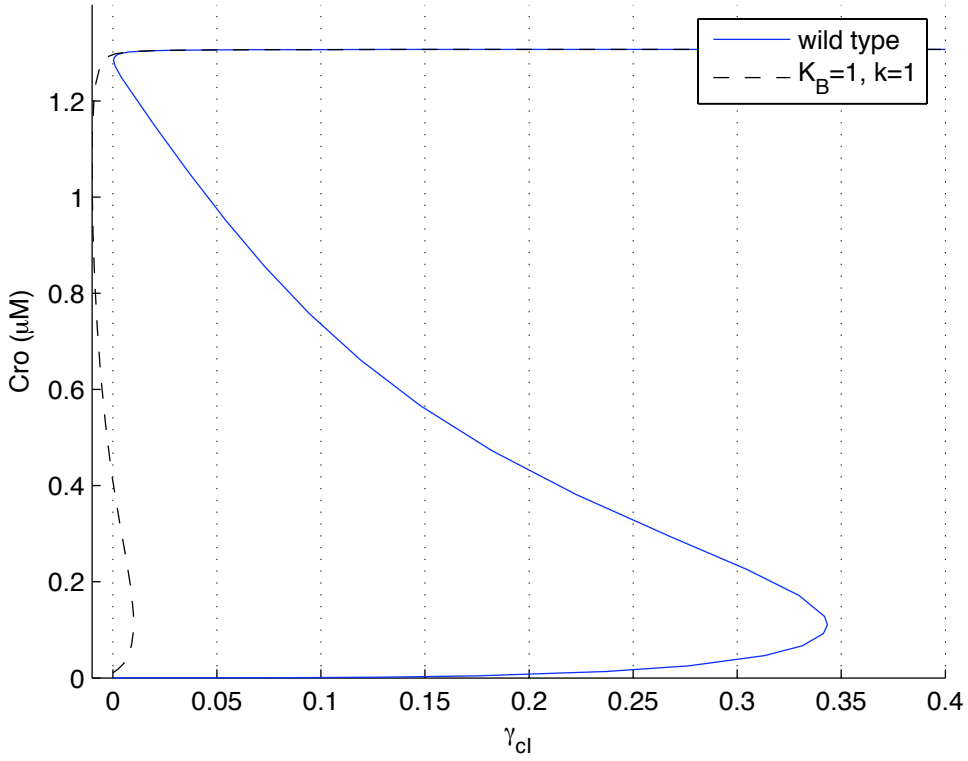


Figure 3.5: Bifurcation diagram of wild type vs. cI-pc mutant where all positive control between CI_2 and RNA polymerase has been eliminated ($\beta = 1, \phi = 1$).

value is only $\gamma^*_{\beta=100} = 0.07$ min^{-1}.

Figure 3.6 clearly indicates that K_B- and k-cooperativity are not equivalent. This difference is highlighted in Figure 3.7 where isoclines of the induction value γ_* are plotted as a function of β and ϕ. The deviation of symmetry across the diagonal $\beta = \phi$ indicates the extent to which K_B- and k-cooperativity fail to be equivalent in maintaining the stability of the lysogenic state.

While Figures 3.6 and 3.7 clearly indicate that there is a difference between K_B- and k-cooperativity, they provide no explanation for this difference. Since the interactions between the binding factors are mediated through nonlinear functions we do not expect there to be a simple, but complete quantitative description of this difference. However, there are two mathematical results that provide a partial explanation.

The first has to do with the rate of production of CI. Let

$$f^{\beta,\phi}_{RM}([CI_2],[Cro_2]) := f^{c}_{RM}([CI_2],[Cro_2]) + f_{RM}([CI_2],[Cro_2])$$

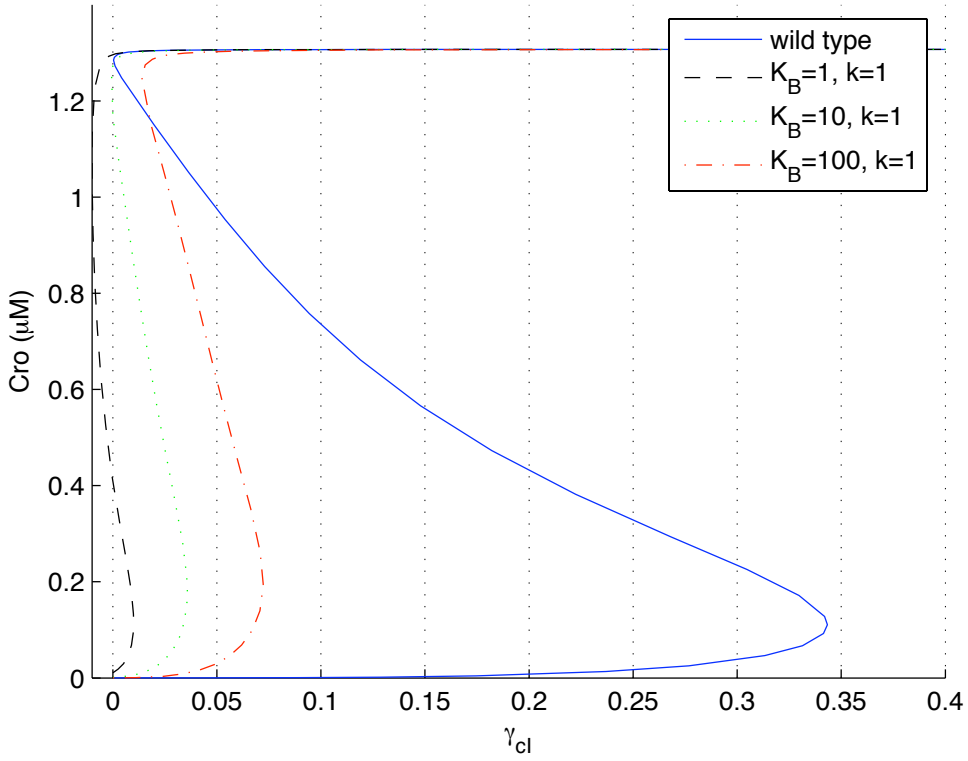


Figure 3.6: Results from eliminating positive control ($\phi = 1$) with values of $\beta = 1$ (black dashed curve), $\beta = 10$ (green dotted curve) and $\beta = 100$ (red dash-dot curve). We graph concentration of Cro as a function of γ_{cI}.

for fixed values of β and ϕ. By Theorem 2.3.9, if $\alpha > 1$, then

$$f_{RM}^{1,\alpha}([CI_2],[Cro_2]) > f_{RM}^{\alpha,1}([CI_2],[Cro_2]).$$

This means that the rate of transcription of cI mRNA is greater under k-cooperativity than under an equal amount of K_B-cooperativity.

The second has do with the biological fact that at low concentrations CI_2 up regulates its own transcription, while at high concentrations is down regulates its own transcription [44]. In the lysogen O_R1 is almost always bound by CI_2 protein and thus the production of Cro is very low. To produce a simple model that can be easily analyzed we assume CI_2 is always bound to O_R1, and thus the states of interest involve the binding of CI_2 to O_R2 and O_R3. In Example 2.3.13 it is proven that under these assumptions there exists a unique critical concentration κ, such that if $[CI_2] < \kappa$, then CI_2 is an activator and if $[CI_2] > \kappa$, then CI_2 is a repressor. This implies that the maximal production rate of CI mRNA occurs at $[CI]_2 = \kappa$.

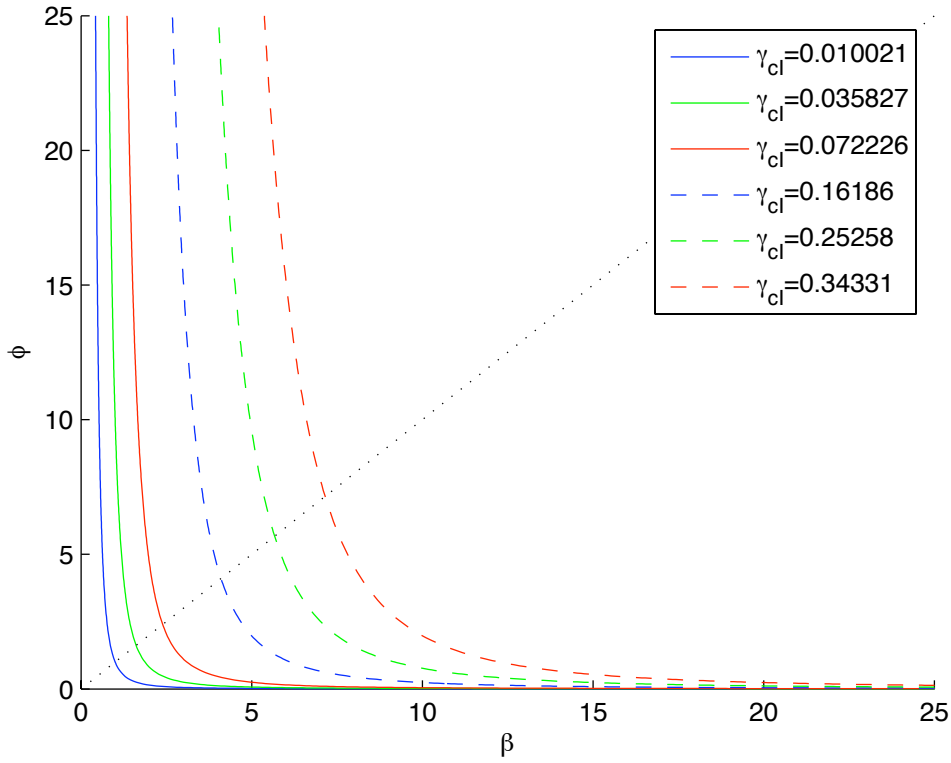


Figure 3.7: Level curves of the induction value γ_* as a function of both β and ϕ. Here $\beta > 1$ represents K_B-cooperativity and $\phi > 1$ represents k-cooperativity.

As is shown in Example 2.3.13 κ is larger under k-cooperativity than under an equal amount of K_B-cooperativity. In particular, the critical concentration for the wild type is greater than the critical concentration for the cI-pc mutant.

3.6 Discussion

One of the common features of transcriptional control in bacteria and eukaryotes is "activation by recruitment", where subtle interactions between the transcription factors and RNA polymerase control the rate of transcription. The three essential steps in this process (binding, opening and escape) coalesce in the Ackers modeling framework into two sets of constants. One set captures binding energies, while the other models the transcription initiation process which includes both opening and escape. If for some state of the operator the binding of a factor increases the binding probability of RNA polymerase we call it K_B-cooperativity. If on the other hand the factor increases the probability of transcription initiation we call it k-cooperativity.

At the first glance it may appear that these two types of activation are interchangeable.

We have shown using an experimentally validated dynamic model of phage λ that with respect to induction of the lysogenic state k- and K_B-cooperativity are not substitutable. Without k-cooperativity the lysogenic state of the phage λ switch is quite unstable and comparable to some known mutants like O_R323 [63].

Our model produced experimentally verifiable predictions and can serve to test hypothesis about induction of phage λ various mutants before they are constructed in the lab. Furthermore, the mathematical techniques and arguments used to obtain these predictions are quite general and thus in the long run we believe that this type of analysis will prove useful for bioengineers who are trying to design novel genetic control units.

Chapter 4
Modeling Promoter Clearance

Several theoretical models have been proposed for the elongation phase of transcription, and they are in reasonably good agreement with what is observed experimentally [66, 67]. However, much less is available in the literature in terms of models for the initiation phase.

Much progress has been made in understanding the biochemistry of the various reactions in transcription initiation [68, 69, 70, 13, 14] and it is clear that binding, opening and promoter clearance are physically coupled processes. Each one of these steps depends on the DNA sequence and its binding affinity to RNA polymerase and to the regulatory proteins. This opens up the possibility of providing a comprehensive model of the entire initiation process, where a DNA sequence is the input and expected rate of transcription initiation is the output.

As described before, transcription initiation can be divided in three main processes: binding, open complex formation and promoter clearance. The binding of RNA polymerase to the promoter is a well understood process, and characterized by the binding constant K_B. In Chapters 2 and 3, using the Shea-Ackers framework, we combine open complex formation and promoter clearance. This is represented by the constant k that we call transcription initiation rate. Here we present our work on modeling promoter clearance.

In Section 4.1 we present and discuss a model of promoter clearance introduced by Xue, Liu and Ou-Yang [31]. While we agree with the main idea of the model, we propose modifications, which we introduce in Section 4.2. However, this model fits experimental data poorly. In Section 4.3 we introduce a new feature to the model: formation of secondary structure in the scrunched DNA. Since the secondary structure is sequence-dependent, it will have a non-uniform impact to the model (as opposed to just varying one parameter that will have the same value for all sequences).While we are still not satisfied with the results, we believe the introduction of secondary structure in the scrunched DNA provides an overall improvement to the model.

4.1 XLO-Y Model

Based on the description of the promoter clearance process given in the Introduction, it seems natural that a model of this phase of transcription should include the scrunching, abortive and escape processes.

In [31], Xue, Liu and Ou-Yang introduce a model for this stage of transcription. We will refer to this model as the *XLO-Y model*. The main idea of this model is that, after RNA polymerase binds the promoter and forms the open complex, there are three competitive reaction pathways that can be followed, which are the abortive pathway, the scrunching pathway and the escape pathway.

The initial set configuration in this model is the RNA polymerase-promoter open complex, which is assumed to be 14 bp long. Using the first two complementary NTPs RNA polymerase performs the first polymerization reaction, starting to create the RNA, which has length 2 at this point, and keeping the open bubble length at 14 bp long. With the first polymerization reaction the active site of RNA polymerase becomes unavailable, so in this conformation it is not able to elongate the nascent RNA. At this point, the RNA polymerase-promoter complex needs to follow one of the competitive pathways to proceed with the transcription process. The three competitive pathways are the scrunching, abortive and escape pathways, and are described below as used in [31].

In the scrunching pathway, RNA polymerase, while still attached to the promoter, unwinds another base pair of the downstream DNA and scrunchs it past its active site, increasing the bubble length by 1 bp. Then it performs a polymerization reaction using the next complementary NTP.

In the abortive pathway RNA polymerase remains attached to the promoter, and for RNAs of length 3 or longer the scrunched DNA is "unscrunched" in a series of reversible reactions. When the RNA-DNA hybrid is of length 2 the short RNA segment that has just been created is irreversibly released with assistance of the first two incoming complementary NTPs. RNA polymerase performs the polymerization reaction and the complex returns to the conformation of a bubble of length 14 with a DNA-RNA hybrid of length 2.

In the escape pathway, RNA polymerase is able to "escape" the promoter: it releases the scrunched DNA, translocates 1 bp ahead in order to make its active site available for the next polymerization reaction, and only leaves a bubble of length 12 inside itself.

4.1.1 XLO-Y Reaction Rates

The RNA polymerase-promoter complex together with the nascent RNA form what is called the initial transcribing complex (ITS). After escape, when the RNA polymerase escapes the promoter and enters the elongation phase, the complex is called elongation complex (EC). Using similar notation to the one used in the literature for elongation [66, 67], these complexes are denoted by

$$P_m(M, N, n)$$

where

- m represents the pathway RNA polymerase is following, where
  - $m = 0$ represents the scrunching pathway,
  - $m = -1$ represents the abortive pathway, and
  - $m = +1$ represents the escape pathway;
- M is the length of the transcription bubble;
- N is the length of the nascent RNA, which is assumed always less or equal to 9;
- n is the length of the RNA-DNA hybrid.

Using similar notation, the state energy of each configuration of the ITC or EC is denoted by $\Delta G^m_{M,N,n}$ and it has three components: the energy of the transcription bubble, the energy of the RNA-DNA hybrid and the binding energy of the RNA polymerase to the promoter.

$$\Delta G^m_{M,N,n} = \Delta G^{\text{bubble}}_{M,N,n} + \Delta G^{\text{hybrid}}_{M,N,n} + \Delta G^{\text{binding}}_{M,N,n}$$

The reactions involved in the scrunching pathway are described by Michaelis-Menten enzyme kinetics in the presence of competitive inhibitors, using rapid equilibrium kinetics to describe the "unscrunching" reactions. The scrunching rate from state $P_0(M, N, n)$ to state $P_0(M + 1, N + 1, min(n + 1, 9))$ is given by

$$k_0^N = \frac{k_1 C}{C + K_{d_1}^N} \tag{4.1}$$

where

$$K_{d_1}^N = K_C \Big\{ 1 + \sum_{i=2}^{N-1} e^{-\beta(\Delta G^{-1}_{i+12,N,min(i,9)} - \Delta G^0_{N+13,N,min(N,8)})} + e^{-\beta(\Delta G^0_{N+12,N,min(N,9)} - \Delta G^0_{N+13,N,min(N,8)})} \Big\}$$

and

- $\beta = \frac{1}{k_B T}$, with k_B the Boltzmann's constant and T the absolute temperature,
- C is the concentration of the complementary NTP,
- K_C is the equilibrium dissociation constant for the complementary NTP,
- k_1 represents the polymerization rate

The reaction rates in the abortive pathway are also obtained using the same rapid equilibrium kinetics, except that the abortive pathway contains a two-substrate enzymatic reaction. The abortive rate from state $P_0(M, N, n)$ to state $P_0(14, 2, 2)$ is given by

$$k_{-1}^N = \frac{k_2 AB}{AB + K_B A + K_B K_{d_2}^N} \tag{4.2}$$

with

$$\begin{aligned} K_{d_2}^N &= K_A \Big\{ e^{-\beta(\Delta G^0_{N+13,N,min(N,8)} - \Delta G^0_{14,N,2})} + \sum_{i=2}^{N-1} e^{-\beta(\Delta G^{-1}_{i+12,N,min(i,9)} - \Delta G^0_{14,N,2})} \\ &\quad + e^{-\beta(\Delta G^0_{N+12,N,min(N,9)} - \Delta G^0_{14,N,2})} \Big\} \end{aligned}$$

where

- A and B are the concentrations of the first and second complementary NTPs, respectively
- K_A and K_B are the equilibrium dissociation constants for the first and second complementary NTPs, respectively
- k_2 represents the polymerization rate.

Since no polymerization reaction is involved in the escape reaction, Michaelis-Menten kinetics are not used in the escape pathway, and instead the escape reaction is treated as an irreversible reaction and the escape rate is calculated using Arrhenius kinetics. The escape rate from state $P_0(M, N, n)$ to state $P_{+1}(12, N, min(n, 8))$ is given by

$$k_{+1}^N = \frac{k_3 e^{-\beta \Delta G^{+1}_{12,N,min(N,8)}}}{V^N} \tag{4.3}$$

where k_3 is the Arrhenius pre-factor constant and

$$V^N = e^{-\beta \Delta G^0_{N+12,N,min(N,9)}} + e^{-\beta \Delta G^0_{N+13,N,min(N,8)}} + \sum_{i=2}^{N-1} e^{-\beta \Delta G^{-1}_{i+12,N,min(i,9)}} \tag{4.4}$$

4.1.2 Parameters

There are various parameters involved in this model.

- The DNA-DNA and DNA-RNA hybrid energies are calculated using the nearest-neighbor model with energy values at 37^oC [71, 72], with ionic conditions considered.
- The values for the polymerization rates k_1 and k_2, and the equilibrium dissociation constant K_C are taken from [66] as:
  - $k_1 = k_2 = 24.7\ \text{s}^{-1}$ and
  - $K_C = 15.6\ \mu$M.
- The values for the equilibrium dissociation constants for the first and second complementary NTPs are taken from [68] as:
  - $K_A = 1800\ \mu$M and
  - $K_B = 31\ \mu$M.
- The Arrhenius constant for the escape reaction is used as $k_3 = 0.8\ \text{s}^{-1}$ [13].
- The energy difference between the binding energies of the RNAP-DNA interaction during scrunching and escape pathways is denoted by

$$\Delta\Delta G^{binding} = \Delta G^{binding}_{+1} - \Delta G^{binding}_{0}$$

  and the value used is 3.5 kcal/mol.

Table 4.1 summarizes the list of parameters and its values, as used in [31].

k_1	$24.7\ \text{s}^{-1}$
k_2	$24.7\ \text{s}^{-1}$
K_C	$15.6\ \mu$M
K_A	$1800\ \mu$M
K_B	$31\ \mu$M
k_3	$0.8\ \text{s}^{-1}$
$\Delta\Delta G$	3.5 kcal/mol

Table 4.1: XLO-Y parameters

4.1.3 XLO-Y Model Results

In [31] Xue, Liu and Ou-Yang define the abortive probability at DNA site i by

$$P_i = \frac{X_i}{T - \sum_{j=2}^{i-1}(X_j + Y_j)}, \tag{4.5}$$

where X_j and Y_j represent, respectively, the number of transcripts that abort or escape at DNA site j in Monte Carlo simulation, and $T = \sum_{i=2}^{\infty}(X_i + Y_i)$ is the total number of times RNA polymerase leaves the scrunching pathway. The maximum size of abortive transcript, MSAT, is assigned as the maximum site (ie, RNA size) that has nonzero abortive probability, and the abortive-productive ratio, APR, is the average number of abortive products per successful escape event.

	MSAT		APR	
Sequence	Experiment	Xue	Experiment	Xue
T5N25	10	10	31 ± 5	31
T5N25anti	15	12	174 ± 27	102
T7A1	8	8	7 ± 2	12

Table 4.2: MSATs and APRs comparison

The sequences for these three promoters are listed in Figure 4.1.

	−40	−35	−10	+1	+20
T5$N25$	TTTAT**TTGCTT**TCAGGAAAATTTTTCTG**TATAAT**AGATTC	**A**TAAATTTGAGAGAGGAGTT			
T5$N25anti$	TTTAT**TTGCTT**TCAGGAAAATTTTTCTG**TATAAT**AGATTC	**A**TCCGGAATCCTCTTCCCGG			
T7$A1$	GAGTA**TTGACT**TAAAGTCTAACCTATAG**GATACT**TACAGCC**A**TCGAGAGGGACACGGCGAA				

Figure 4.1: The three promoter sequences used in [31] to compare the XLO-Y model to experimental data [73].

4.2 Modifications to the XLO-Y Model

While we agree with the main idea of the model (the three competitive pathways), there is room for improvements.

The main observation we have is on the hypothesis of NTP-assisted release in the abortive pathway used in [31]. This hypothesis does not seem to be realistic, as there is no biological evidence for it. Instead we look at the release of the short RNA segments as an irreversible reaction, and use Arrhenius kinetics to describe the abortive rate since, like in the escape

pathway, there is no polymerization reaction involved. We define the abortive rate from state $P_0(M, N, n)$ to state $P_0(M+1, N+1, min(n+1, 9))$ as

$$a_N = \frac{k_a e^{-\beta \Delta G^0_{14,0,0}}}{V^N} \tag{4.6}$$

where k_a is the Arrhenius pre-factor constant for the abortive reaction, and V^N is defined as (4.4):

$$V^N = e^{-\beta \Delta G^0_{N+12,N,min(N,9)}} + e^{-\beta \Delta G^0_{N+13,N,min(N,8)}} + \sum_{i=2}^{N-1} e^{-\beta \Delta G^{-1}_{i+12,N,min(i,9)}} \tag{4.7}$$

We also reconsider the escape rate. In [31] the assumption is that when escape occurs, the transcription bubble becomes 12 bp long, and it will remain the same length through elongation. This assumption is due to the release of the σ factor which has been thought to occur simultaneously with the transition to elongation [18, 19, 20]. This hypothesis has been questioned since studies have found that RNA polymerase can retain its σ^{70} subunit after the transition to elongation [21, 22]. The release of the σ factor may occur stochastically [23, 24, 25, 26]. We assume that when escape occurs, the σ factor is still associated to RNA polymerase, and therefore in this state the bubble length is back to 14 instead of 12. With this assumption we define the escape rate from state $P_0(M, N, n)$ to state $P_{+1}(14, N, min(n, 8))$ as

$$e_N = \frac{k_e e^{-\beta \Delta G^{+1}_{14,N,min(N,8)}}}{V^N} \tag{4.8}$$

where k_e is the Arrhenius pre-factor constant for the escape reaction, and V^N is given by (4.7).

We do not modify the scrunching rates. Just to be consistent with our notation, we define the scrunching rate as

$$s_N = k_0^N \tag{4.9}$$

where k_0^N is given by (4.1). The derivation of the formula for the scrunching rates is given in Appendix B

We observe that while the -35 and -10 regions are identical for promoters T5N25 and T5N25anti, they are both different for the promoter T7A1. See Figure 4.1. Since these regions directly affect the binding affinity for the RNA polymerase, and therefore its binding energy, it does not seem appropriate to use the same value of $\Delta\Delta G^{\text{binding}}$ for promoters that are not identical in those regions. We therefore do not use the promoter T7A1 in our comparisons. T5N25anti is a T5N25 ITS variant, that is, T5N25anti is obtained by replacing the ITS of

T5N25 from position +3 to +20 with an antisense mutation, i.e., A $\rightleftarrows$ C and G $\rightleftarrows$ T. In [76] the abortive initiation-promoter escape properties of 43 T5N25 promoter random-ITS variants are analyzed. See Table 4.3 for the sequences. Instead of using the data from [73] for T5N25 and T5N25anti, we use the data for the 43 sequences in [76]. Data for the 43 sequences presented in Table 4.3 was kindly provided by Lilian Hsu. For simplicity, from now on we will refer to T5N25 and T5N25anti as N25 and N25anti, respectively.

Promoter	ITS (+1 – +20)	PY (%)	APR	MSAT
DG146a	ATTAAAAAAC CTGCTAGGAT	8.0 ± 2.4	13 ± 4	20
N25/A1	ATCGAGAGGG ACACGGCGAA	7.1 ± 0.9	13 ± 2	19
DG122	ATAAAGGAAA ACGGTCAGGT	7.0 ± 1.1	14 ± 1	18
DG130a	ATATAGTGAA CAAGGATTAA	6.9 ± 0.5	14 ± 1	18
DG131a	ATAGGTTAAA AGCCAGACAT	5.1 ± 2.2	21 ± 9	16
N25	ATAAATTTGA GAGAGGAGTT	6.0 ± 1.9	18 ± 5	11
DG151a	ATCAGGATAC AAGAAGGTTT	6.0 ± 2.7	19 ± 9	16
DG161a	ATAAAAGTAC TCAGTTCAAA	5.1 ± 2.1	22 ± 10	15
DG159	ATAACTAGGG AAAATAATAT	4.6 ± 2.2	26 ± 16	18
DG121	ATACACCATA AAGAAACAGT	3.4 ± 1.5	33 ± 17	17
DG132a	ATTCTAGTGA AAATCCCCAT	3.8 ± 1.5	30 ± 12	16
DG115a	ATCCCGCTCA AGAGCAACAT	3.5 ± 0.2	28 ± 2	18
DG162	ATGTAAATAA GGTAGGCAAT	3.9 ± 1.1	27 ± 8	16
DG128a	ATCCCAGTAA GGAATGATAT	3.7 ± 1.4	30 ± 11	18
DG126	ATAAGCACAC GGATACCTTT	2.5 ± 0.7	40 ± 14	16
DG163a	ATTATACACG GTAATCGCTT	3.4 ± 1.4	34 ± 14	18
DG164	ATTAAGAAAA ATCTTCTATT	3.1 ± 0.6	34 ± 6	17
DG149	ATAGCGGATG GTAACAGAAT	2.9 ± 1.2	38 ± 11	14
DG165b	ATCATCTGAA ATCATAGTGT	3.1 ± 0.9	33 ± 14	16
DG169a	ATCCAGACGA ACTGGGGAAT	3.1 ± 0.3	31 ± 3	20
DG155	ATTAAAAATC CTTTCCTCTT	2.8 ± 0.4	38 ± 2	15
DG168a	ATCACGCAAC CGGACTAACT	2.7 ± 0.7	38 ± 10	16
DG127	ATCCTAGTAT ATGGAACTGT	2.7 ± 1.2	40 ± 16	14
DG135a	ATAATGCTGT GAACGCGAGT	2.2 ± 0.6	53 ± 16	20
DG160a	ATATACTAGC AGCACCAATT	2.4 ± 1.0	50 ± 17	15
DG133	ATATCGAATT ACTCAGATAT	1.8 ± 0.3	58 ± 14	16

DG147a	ATAATGGTCG GTTACACGAT	1.8 ± 1.0	70 ± 43	19
DG125a	ATATCGTTCC CTTGACCCAT	1.3 ± 0.1	76 ± 6	16
N25anti	ATCCGGAATC CTCTTCCCGG	1.4 ± 1.0	68 ± 31	15
DG156-3	ATCGCCGATA AATACGTAGT	1.4 ± 0.6	76 ± 25	15
DG138a	ATCTTCTTCG TAACTGGAGT	0.9 ± 0.3	121 ± 40	16
DG142	ATGATTTCAT CTGACTCTAT	0.9 ± 0.1	121 ± 5	16
DG170a	ATTACTGCAC ATTAATGAAT	0.8 ± 0.1	118 ± 25	16
DG167	ATTACATCTG CCGCCTTCCT	0.9 ± 0.5	151 ± 95	20
DG166	ATCTAATCTC TGATAATATT	0.8 ± 0.3	142 ± 60	17
DG152a	ATTACTATGC CCCATATCCT	0.8 ± 0.3	144 ± 52	15
DG148	ATAATTGTAC ATTTGAAACT	1.0 ± 0.5	135 ± 82	17
DG145	ATAACCCTTG ACTCCGAAAT	0.5 ± 0.2	202 ± 62	15
DG141	ATACATTATC AACGCATGCT	0.6 ± 0.2	169 ± 65	14
DG124a	ATCGCAACCT CCTAAATGAT	0.4 ± 0.2	205 ± 51	15
N25/A1anti	ATATCTCTTT CACATTATCC	0.4 ± 0.1	255 ± 52	16
DG154a	ATGGTTCATT TTTCCACACT	0.5 ± 0.3	217 ± 101	17
DG137a	ATCGCTCTAC TAAATGTCTT	0.3 ± 0.1	386 ± 27	15

Table 4.3: N25 promoter random-ITS variants constructed in [76]: sequences and properties.

While the expected range for $\Delta\Delta G$ is $15 - 18$ kcal/mol [77], the value of 3.5 kcal/mol used in [31] is extremely low. As we will show in Section 4.2.4 our computations also suggest a value lower than the expected range may be more appropriate for the model.

We also observe that the abortive probabilities (4.5), computed in [31] to be compared to the abortive profiles in [73], are slightly different from the abortive probabilities defined in [73]. The denominator in (4.5) corresponds to the number of transcripts that aborted or escaped at or after position i, while the corresponding quantity in [73] would be the total number of full length transcripts plus the number of transcripts that aborted at or after position i. In Section 4.2.2 we describe how we calculate and compare the abortive probabilities in the abortive profiles.

There are other simple modifications we make in order to optimize parameters. In [31] a single value is used for the polymerization rate, $k_1 = k_2 = 24.7s^{-1}$ [66], and a single value is used for the NTP-dissociation constant, $K_C = 15.6\mu M$ [66] (except for the first and second

NTPs). Instead we use four NTP-specific values for both the polymerization rate and the NTP-dissociation constant, based on elongation studies in [78]. The values for the NTP-specific polymerization rates as in [78] are $k_A = 50 \pm 6$ s^{-1}, $k_U = 18 \pm 1$ s^{-1}, $k_G = 36 \pm 5$ s^{-1}, and $k_C = 33 \pm 6$ s^{-1}, while the values for the NTP-dissociation constants are $K_A = 38 \pm 7$ kcal/mol, $K_U = 24 \pm 4$ kcal/mol, $K_G = 62 \pm 18$ kcal/mol and $K_C = 7 \pm 4$ kcal/mol. We use the average value for each of these quantities.

We also note that while the APR and MSAT are very important parameters that one should be interested in matching when trying to reasonably describe the behavior of specific promoters, an overall agreement with the whole abortive profile is expected. Although the XLO-Y model provides a good agreement on APR and MSAT, the comparison with the abortive profiles is less satisfactory.

As in [31], we also estimate the DNA-DNA and DNA-RNA hybrid energies using the nearest-neighbor model with energy values at 37^oC. For the DNA-DNA values we use the unified nearest-neighbor values from [79], with the salt correction done by using the empirical equation given in [79]:

$$\Delta G_{37}(\text{polymer NN}, [\text{Na}^+]) = \Delta G_{37}(\text{unified NN}, [\text{Na}^+] = 1M) - 0.175\ln[\text{Na}^+] - 0.20$$

We believe these values are the same as the ones used in [31]. For the DNA-RNA we use values from [72] also taking the ionic conditions into consideration. As there is no published formula for the salt correction for DNA-RNA hybrids, we predicted the corrections using the HYTHER server from John SantaLucia group [80]. We estimate the salt correction results in values that are approximately 90% of those in [72].

4.2.1 Probabilities

We define the probability of the reactions a_N, s_N and e_N, given respectively by (4.6), (4.9) and (4.8), to be

$$\begin{aligned} P_a(N) &= \frac{a_N}{a_N + s_N + e_N} \times 100\% \\ P_s(N) &= \frac{s_N}{a_N + s_N + e_N} \times 100\% \\ P_e(N) &= \frac{e_N}{a_N + s_N + e_N} \times 100\% \end{aligned} \tag{4.10}$$

We refer to the probabilities in (4.10) as abortive, scrunching and escape probability at position N, respectively. For simplicity, we will also refer to the correspondent rates as abortive,

scrunching and escape rates at position N.

We can also define the probabilities in (4.10) in terms of the number of times the abortive, scrunching and escape reactions occurred at each position. For that, let $a(N), s(N)$ and $e(N)$ be the number of abortive, scrunching and escape reactions, respectively, at position N. We can then write

$$P_a(N) = \frac{a(N)}{a(N) + s(N) + e(N)} \times 100\%$$
$$P_s(N) = \frac{s(N)}{a(N) + s(N) + e(N)} \times 100\%$$
$$P_e(N) = \frac{e(N)}{a(N) + s(N) + e(N)} \times 100\%.$$

We need to impose a restriction on how long scrunching can happen. Let M be the RNA length for which we assume only abortive and escape reactions can occur, that is, scrunching can only occur while the RNA length is less than M.

4.2.2 Comparing Abortive Probabilities

We want to compare the results of the model with experimental data. The experimental data is usually presented in the form of an abortive profile. Let P_n be the abortive probability represented in the abortive probability profiles. This quantity is defined in [81] as

$$P_N = \frac{X_N}{100\% - \sum_{i=2}^{N-1} X_i} \times 100\%, \tag{4.11}$$

where X_i is the abortive yield, in percent, of the ith RNA species, i.e., the sum of abortive RNAs of length i as a percentage of the total number of RNAs that were produced.

Using the abortive, scrunching and escape rates at position N, we defined the abortive, scrunching and escape probabilities $P_a(N), P_s(N)$ and $P_e(N)$. In order to compare results with the experimental data we need to express P_N in terms of the probabilities $P_a(N), P_s(N)$ and $P_e(N)$.

The abortive yield X_i can be expressed in terms of the numbers $a(i)$ and $e(i)$ as

$$X_i = \frac{a(i)}{\sum_{j=2}^{M}(a(j) + e(j))}$$

We can then rewrite the expression (4.11) for P_N in terms of the numbers $a(N)$ and $e(N)$

$$P_N = \frac{a(N)}{\sum_{i=N}^{M} a(i) + \sum_{i=2}^{M} e(i)} \times 100\%.$$

Notice that

$$\begin{aligned} &a(N) + s(N) + e(N) = s(N-1), \text{ for } N = 3, \ldots, M-1 \\ &a(M) + e(M) = s(M-1). \end{aligned}$$

Therefore

$$\begin{aligned} P_2 &= \frac{a(2)}{a(M) + e(M) + \sum_{i=2}^{M-1}(a(i) + e(i))} \\ &= \frac{a(2)}{s(M-1) + \sum_{i=2}^{M-1}(a(i) + e(i))} \\ &= \frac{a(2)}{s(M-1) + a(M-1) + e(M-1) + \sum_{i=2}^{M-2}(a(i) + e(i))} \\ &= \frac{a(2)}{s(M-2) + \sum_{i=2}^{M-2}(a(i) + e(i))} \\ &\quad \vdots \\ &= \frac{a(2)}{s(2) + a(2) + e(2)} \\ &= P_a(2) \end{aligned}$$

and for $N = 3, \ldots, M$

$$\begin{aligned} P_N &= \frac{a(N)}{a(M) + e(M) + \sum_{i=N}^{M-1} a(i) + \sum_{i=2}^{M-1} e(i)} \\ &= \frac{a(N)}{s(M-1) + \sum_{i=N}^{N-1} a(i) + \sum_{i=2}^{N-1} e(i)} \\ &= \frac{a(N)}{s(M-1) + a(M-1) + e(M-1) + \sum_{i=N}^{M-2} a(i) + \sum_{i=2}^{M-2} e(i)} \\ &= \frac{a(N)}{s(M-2) + \sum_{i=N}^{M-2} a(i) + \sum_{i=2}^{M-2} e(i)} \\ &\quad \vdots \\ &= \frac{a(N)}{s(N-1) + \sum_{i=2}^{N-1} e(i)}. \end{aligned}$$

Notice that $P_N = P_a(N)$ only when $N = 2$.

For $N = 3, \ldots, M$ we have

$$P_N = \frac{P_a(N)P_s(N-1)P_s(N-2)\ldots P_s(2)}{(1-P_2)(1-P_3)\ldots(1-P_{N-1})},$$

or, writing recursively,

$$P_N = \frac{P_a(N)P_s(N-1)P_{N-1}}{P_a(N-1)(1-P_{N-1})}. \tag{4.12}$$

To prove (4.12), first notice that, for $N = 3, \ldots, M$, we have

$$\begin{aligned} 1 - P_{N-1} &= \frac{s(N-2) + \sum_{i=2}^{N-2} e(i) - a(N-1)}{s(N-2) + \sum_{i=2}^{N-2} e(i)} \\ &= \frac{a(N-1) + s(N-1) + e(N-1) + \sum_{i=2}^{N-2} e(i) - a(N-1)}{s(N-2) + \sum_{i=2}^{N-2} e(i)} \\ &= \frac{s(N-1) + e(N-1) + \sum_{i=2}^{N-2} e(i)}{s(N-2) + \sum_{i=2}^{N-2} e(i)} \end{aligned}$$

and then

$$\frac{P_{N-1}}{1-P_{N-1}} = \frac{a(N-1)}{s(N-1) + e(N-1) + \sum_{i=2}^{N-2} e(i)}.$$

Now

$$\begin{aligned} P_N &= \frac{a(N)}{s(N-1) + \sum_{i=2}^{N-1} e(i)} = \frac{\frac{a(N)}{a(N)+s(N)+e(N)}}{\frac{s(N-1)+\sum_{i=2}^{N-1} e(i)}{a(N)+s(N)+e(N)}} = \frac{P_a(N)}{\frac{s(N-1)+\sum_{i=2}^{N-1} e(i)}{s(N-1)}} \\ &= \frac{P_a(N)s(N-1)}{s(N-1) + \sum_{i=2}^{N-1} e(i)} = \frac{\frac{P_a(N)s(N-1)}{a(N-1)+s(N-1)+e(N-1)}}{\frac{s(N-1)+\sum_{i=2}^{N-1} e(i)}{a(N-1)+s(N-1)+e(N-1)}} = \frac{P_a(N)P_s(N-1)}{\frac{s(N-1)+\sum_{i=2}^{N-1} e(i)}{a(N-1)+s(N-1)+e(N-1)}} \\ &= P_a(N)P_s(N-1)\frac{a(N-1) + s(N-1) + e(N-1)}{s(N-1) + \sum_{i=2}^{N-1} e(i)} \\ &= P_a(N)P_s(N-1)\frac{\frac{a(N-1)+s(N-1)+e(N-1)}{a(N-1)}}{\frac{s(N-1)+\sum_{i=2}^{N-1} e(i)}{a(N-1)}} = \frac{P_a(N)P_s(N-1)}{P_a(N-1)}\frac{a(N-1)}{s(N-1) + \sum_{i=2}^{N-1} e(i)} \\ &= \frac{P_a(N)P_s(N-1)P_{N-1}}{P_a(N-1)(1-P_{N-1})} \end{aligned}$$

and (4.12) holds.

The percentage of full length transcripts, which will be denoted by P_{FL}, corresponds to the percentage of escape reactions over the positions 2 through M, i.e.,

$$P_{\mathrm{FL}} = \frac{\sum_{i=2}^{M} e(i)}{\sum_{i=2}^{M}(a(i) + e(i))},$$

where $a(i)$ and $e(i)$ are the number of abortive and escape reactions, respectively, occurred at

position i.

We can express P_{FL} in terms of the escape and scrunching probabilities $P_e(N)$ and $P_s(N)$. We have

$$
\begin{aligned}
P_{FL} &= \frac{\sum_{i=2}^{M-1} e(i) + e(M)}{\sum_{i=2}^{M-1}(a(i)+e(i)) + a(M) + e(M)} = \frac{\sum_{i=2}^{M-1} e(i) + e(M)}{\sum_{i=2}^{M-1}(a(i)+e(i)) + s(M-1)} \\
&= \frac{\sum_{i=2}^{M-1} \frac{e(i)}{a(M)+e(M)} + \frac{e(M)}{a(M)+e(M)}}{\sum_{i=2}^{M-1} \frac{a(i)+e(i)}{a(M)+e(M)} + \frac{s(M-1)}{a(M)+e(M)}} = \frac{\sum_{i=2}^{M-1} \frac{e(i)}{s(M-1)} + P_e(M)}{\sum_{i=2}^{M-1} \frac{a(i)+e(i)}{s(M-1)} + 1} \\
&= \frac{\sum_{i=2}^{M-2} \frac{e(i)}{s(M-1)} + \frac{e(M-1)}{s(M-1)} + P_e(M)}{\sum_{i=2}^{M-2} \frac{a(i)+e(i)}{s(M-1)} + \frac{a(M-1)+e(M-1)}{s(M-1)} + 1} \\
&= \frac{\sum_{i=2}^{M-2} \frac{\frac{e(i)}{a(M-1)+s(M-1)+e(M-1)}}{P_s(M-1)} + \frac{P_e(M-1)}{P_s(M-1)} + P_e(M)}{\sum_{i=2}^{M-2} \frac{\frac{a(i)+e(i)}{a(M-1)+s(M-1)+e(M-1)}}{P_s(M-1)} + \frac{1-P_s(M-1)}{P_s(M-1)} + 1} \\
&= \frac{\sum_{i=2}^{M-2} \frac{e(i)}{a(M-1)+s(M-1)+e(M-1)} + P_e(M-1) + P_s(M-1)P_e(M)}{\sum_{i=2}^{M-2} \frac{a(i)+e(i)}{a(M-1)+s(M-1)+e(M-1)} + 1} \\
&= \frac{\sum_{i=2}^{M-2} \frac{e(i)}{s(M-2)} + P_e(M-1) + P_s(M-1)P_e(M)}{\sum_{i=2}^{M-2} \frac{a(i)+e(i)}{s(M-2)} + 1} \\
&= \frac{\sum_{i=2}^{M-3} \frac{e(i)}{s(M-2)} + \frac{e(M-2)}{s(M-2)} + P_e(M-1) + P_s(M-1)P_e(M)}{\sum_{i=2}^{M-3} \frac{a(i)+e(i)}{s(M-2)} + \frac{a(M-2)+e(M-2)}{s(M-2)} + 1} \\
&= \frac{\sum_{i=2}^{M-3} \frac{\frac{e(i)}{a(M-2)+s(M-2)+e(M-2)}}{P_s(M-2)} + \frac{P_e(M-2)}{P_s(M-2)} + P_e(M-1) + P_s(M-1)P_e(M)}{\sum_{i=2}^{M-3} \frac{\frac{a(i)+e(i)}{a(M-2)+s(M-2)+e(M-2)}}{P_s(M-2)} + \frac{1-P_s(M-2)}{P_s(M-2)} + 1} \\
&= \frac{\sum_{i=2}^{M-3} \frac{e(i)}{s(M-3)} + P_e(M-2) + P_s(M-2)P_e(M-1) + P_s(M-2)P_s(M-1)P_e(M)}{\sum_{i=2}^{M-3} \frac{a(i)+e(i)}{s(M-3)} + 1} \\
&= \dots \\
&= \frac{\frac{e(2)}{s(2)} + P_e(3) + P_s(3)P_e(4) + \dots + P_s(3)P_s(4)\dots P_s(M-1)P_e(M)}{\frac{a(2)+s(2)}{s(2)} + 1} \\
&= P_e(2) + P_s(2)P_e(3) + P_s(2)P_s(3)P_e(4) + \dots + P_s(2)P_s(3)\dots P_s(M-1)P_e(M)
\end{aligned}
$$

4.2.3 Promoter Clearance as a Markov Chain

For each $2 \leq N \leq M$, we defined the scrunching, abortive and escape rates when RNA length is N as the transition rates between states

$$\begin{aligned} P_0(12+N, N, min(N,9)) &\xrightarrow{s_N} P_0(13+N, N+1, min(N+1,9)) \\ P_0(12+N, N, min(N,9)) &\xrightarrow{a_N} P_0(14,0,0) \\ P_0(12+N, N, min(N,9)) &\xrightarrow{e_N} P_{+1}(14, N, min(N,9)) \end{aligned}$$

and, then, based on these rates, we defined the probabilities $P_s(N), P_a(N), P_e(N)$ of a scrunching, abortive or escape reaction occur when RNA length is N. Therefore, when RNA length is N, for each $2 \leq N \leq M$, with probability $P_s(N)$ the RNA length will be increased by 1 while RNA polymerase is still physically attached to the promoter; with probability $P_a(N)$ an abortive reaction will occur, an abortive transcript of length N will be released and then RNAP-promoter will return to the original open-bubble conformation; and with probability $P_e(N)$ RNA polymerase will break the bounds with the promoter DNA and will enter the elongation phase. We summarize this information with the diagram in Figure 4.2.

To simplify the notation, we rewrite the diagram in Figure 4.2 by representing each state by its RNA length. To avoid confusion, the escape states with RNA length N will be represented by E^N. We rewrite the diagram of Figure 4.2 as

$$\begin{array}{ccccccccc} 0 & & 0 & & & 0 & & 0 & \\ \uparrow_{P_a(2)} & & \uparrow_{P_a(3)} & & & \uparrow_{P_a(N)} & & \uparrow_{P_a(N+1)} & \\ 2 & \xrightarrow{P_s(2)} & 3 & \xrightarrow{P_s(3)} & \ldots & N & \xrightarrow{P_s(N)} & N+1 & \rightarrow \\ \downarrow_{P_e(2)} & & \downarrow_{P_e(3)} & & & \downarrow_{P_e(N)} & & \downarrow_{P_e(N+1)} & \\ E^2 & & E^3 & & & E^N & & E^{N+1} & \end{array}$$

$$
\begin{array}{ccccccccc}
P_0(14,0,0) & & P_0(14,0,0) & & P_0(14,0,0) & & P_0(14,0,0) & & \\
\uparrow_{P_a(2)} & & \uparrow_{P_a(3)} & & \uparrow_{P_a(N)} & & \uparrow_{P_a(N+1)} & & \\
P_0(14,2,2) & \xrightarrow{P_s(2)} & P_0(15,3,3) & \xrightarrow{P_s(3)} & \ldots \quad P_0(12+N,N,min(N,9)) & \xrightarrow{P_s(N)} & P_0(13+N,N+1,min(N+1,9)) & \rightarrow & \ldots \\
\downarrow_{P_e(2)} & & \downarrow_{P_e(3)} & & \downarrow_{P_e(N)} & & \downarrow_{P_e(N+1)} & & \\
P_{+1}(14,2,2) & & P_{+1}(15,3,3) & & P_{+1}(14,N,min(N,8)) & & P_{+1}(14,N+1,min(N+1,8)) & &
\end{array}
$$

Figure 4.2: Transition probabilities between states.

After an abortive reaction the system returns to the initial open bubble configuration $P_0(14,0,0)$. Then the first two complementary nucleotides can bind, RNA polymerase performs a polymerization reaction, and the system returns to the configuration with RNA of length 2, $P_0(14,2,2)$. We assume then that the probability of going from state $P_0(14,0,0)$ to state $P_0(14,2,2)$ is 1, and we can look to $P_a(N)$ as the probability going from the state $P_0(12+N,N,min(N,9))$ to state $P_0(14,2,2)$. We again rewrite the diagram of Figure 4.2 as:

$$\begin{array}{ccccccccc} 2 & & 2 & & & 2 & & 2 & \\ \uparrow_{P_a(2)} & & \uparrow_{P_a(3)} & & & \uparrow_{P_a(N)} & & \uparrow_{P_a(N+1)} & \\ 2 & \xrightarrow{P_s(2)} & 3 & \xrightarrow{P_s(3)} & \dots & N & \xrightarrow{P_s(N)} & N+1 & \rightarrow \\ \downarrow_{P_e(2)} & & \downarrow_{P_e(3)} & & & \downarrow_{P_e(N)} & & \downarrow_{P_e(N+1)} & \\ E^2 & & E^3 & & & E^N & & E^{N+1} & \end{array}$$

For each $N \geq 2$ the probabilities $P_s(N), P_a(N)$ and $P_e(N)$ are the probabilities of taking one of the three possible paths, assuming the process is at state N. Clearly they do not depend on how the process reached this particular state. Also, since after an escape event the system enters the elongation phase instead of returning to any of the previous possible states, we can assume that if the system reaches the escape state it remains at that state with probability 1. Therefore this is a stationary Markov chain with state space $S = \{2,3,4,\dots,M,E^2,E^3,\dots,E^M\} = \{T,E^2,E^3,\dots,E^M\}$. (See Appendix C for basic definitions and properties of Markov chains [82, 83].) The escape states $E^2,E^3,\dots,E^M$ are clearly closed. The states $T = \{2,3,\dots,M\}$ are transient. The transition probabilities matrix is given by

$$P = \begin{pmatrix} Q & R \\ 0 & I \end{pmatrix}$$

where the matrix Q represents the transition probabilities between the states in T

$$Q = \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ \vdots \\ M-1 \\ M \end{array} \begin{array}{c} \begin{array}{cccccc} 2 & 3 & 4 & \dots & M-1 & M \end{array} \\ \begin{pmatrix} P_a(2) & P_s(2) & 0 & \dots & 0 & 0 \\ P_a(3) & 0 & P_s(3) & \dots & 0 & 0 \\ P_a(4) & 0 & 0 & & 0 & 0 \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ P_a(M-1) & 0 & 0 & \dots & 0 & P_s(M-1) \\ P_a(M) & 0 & 0 & \dots & 0 & 0 \end{pmatrix} \end{array}$$

while the matrix R represents the transition probabilities from states in T to the escape states $E^2, E^3, \ldots, E^M$:

$$R = \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{ccccc} E^2 & E^3 & E^4 & \ldots & E^M \end{array} \\ \begin{pmatrix} P_e(2) & 0 & 0 & \ldots & 0 \\ 0 & P_e(3) & 0 & \ldots & 0 \\ 0 & 0 & P_e(4) & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & P_e(M) \end{pmatrix} \end{array}$$

and 0 is the $(M-1) \times (M-1)$ null matrix, and I is the $(M-1) \times (M-1)$ identity matrix.

There are several quantities of interest that describe the promoter clearance process, and they can be easily computed with the Markov chain approach:

- Number of abortive transcripts at each position until escape

  For each $2 \leq N \leq M$, in order to compute the number of abortive transcripts of length n produced before escape we need to estimate the number of visits to state 2 coming from state N. By considering the set of states we are considering at the moment, we can calculate the expected number of visits to state 2, but it is not possible to know from which state each visit is coming from. So we will add additional states $T_1, T_2, \ldots, T_M$ representing transition states between N and 2, for $N = 2, 3, \ldots, M$. The new matrices that compose the transition probability matrix P are

$$\bar{Q} = \begin{array}{c} \\ 2 \\ T_2 \\ 3 \\ T_3 \\ \vdots \\ M \\ T_M \end{array} \begin{array}{c} \begin{array}{cccccccc} 2 & T_2 & 3 & T_3 & 4 & \ldots & M & T_M \end{array} \\ \begin{pmatrix} 0 & P_a(2) & P_s(2) & 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & P_a(3) & P_s(3) & \ldots & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & P_a(20) \\ 1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \end{pmatrix} \end{array}$$

and

$$\bar{R} = \begin{array}{c} \\ 2 \\ T_2 \\ 3 \\ T_3 \\ \vdots \\ M \\ T_M \end{array} \begin{array}{c} \begin{array}{cccc} E^2 & E^3 & \dots & E^M \end{array} \\ \begin{pmatrix} P_e(2) & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & P_e(3) & \dots & 0 \\ 0 & 0 & & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & P_e(M) \\ 0 & 0 & \dots & 0 \end{pmatrix} \end{array}$$

in addition to the null matrix 0 and the identity matrix I.

To calculate the number of abortive transcripts of length N we compute the expected number of visits to state T_N by

$$a(N) = \sum_{j=0}^{\infty} \bar{Q}^j_{1,2N} = (I - \bar{Q})^{-1}_{1,2N}.$$

Notice that with these values we build the abortive profile.

- Probability of escape at each position

  For each $2 \leq N \leq M$ the probability of escape to state e_N is given by $F(1, N-1)$, where

$$F = \sum_{j=0}^{\infty} \bar{Q}^j \bar{R} = (I - \bar{Q})^{-1} \bar{R}.$$

- APR

  The abortive:productive ratio is by definition the total number of abortive transcripts divided by the total number of full length transcripts. Since in our case the number of full length transcripts is 1, we simply have

$$\text{APR} = \text{total \# of abortive transcripts} = \sum_{N=2}^{M} a(N)$$

- MSAT

  Since we can calculate the expected number of abortive transcripts of each length N, we can define the maximum size of abortive transcript as the largest value of N, $2 \leq N \leq M$

that has expected number of abortive transcripts larger than 1:

$$\text{MSAT} = \max_{2 \leq N \leq M}\{N | a(N) \geq 1\}$$

- PY

  The productive yield is the total number of full length transcripts divided by the total number of transcripts produced. Since it is assumed that only one full length transcript is produced, we have

  $$\text{PY} = \frac{1}{\sum_{N=2}^{M} a(N) + 1}.$$

- Abortive Profile

  The abortive probabilities (4.11) for the abortive profiles defined in [81] can be computed here as

  $$P_N = \frac{a(N)}{\sum_{i=N}^{M} a(i) + 1}.$$

Remark 4.2.1 The Markov approach and the approach from Section 4.2.2 result in the same abortive profiles.

4.2.4 Comparison to Data

As described in the beginning of this chapter, we will use the 43 sequences from Table 4.3 to compare experimental data from [76] to the results obtained with the model.

We introduce the changes one at time to observe the impact of each modification to the original model. We compute the abortive profiles using the expressions derived in Section 4.2.2.

The following notation is used to describe which version of the model we are using.

- Setup I: model using the XLO-Y rates given by (4.1), (4.2) and (4.3)
- Setup II: model using the XLO-Y scrunching and escape rates, and the modified abortive rate given by (4.6)
- Setup III: model using the XLO-Y scrunching and abortive rates, and with the modified escape rate given by (4.8)
- Setup IV: model using the XLO-Y rates and $\Delta\Delta G = 5$ kcal/mol
- Setup V: model using the XLO-Y rates and $\Delta\Delta G = 8$ kcal/mol
- Setup VI: model using the XLO-Y rates and $\Delta\Delta G = 15$ kcal/mol

- Setup VII: model using the XLO-Y scrunching rate, and the modified abortive and escape rates
- Setup VIII: model using the XLO-Y scrunching rate, modified abortive and escape rates , and $\Delta\Delta G = 5$ kcal/mol
- Setup IX: model using the XLO-Y scrunching rate, modified abortive and escape rates , and $\Delta\Delta G = 8$ kcal/mol
- Setup X: model using the XLO-Y scrunching rate, modified abortive and escape rates , and $\Delta\Delta G = 15$ kcal/mol
- Setup XI: model using the XLO-Y scrunching rate, modified abortive and escape rates , $\Delta\Delta G = 5$ kcal/mol and NTP-specific values for the polymerization rates and NTP-dissociation constants

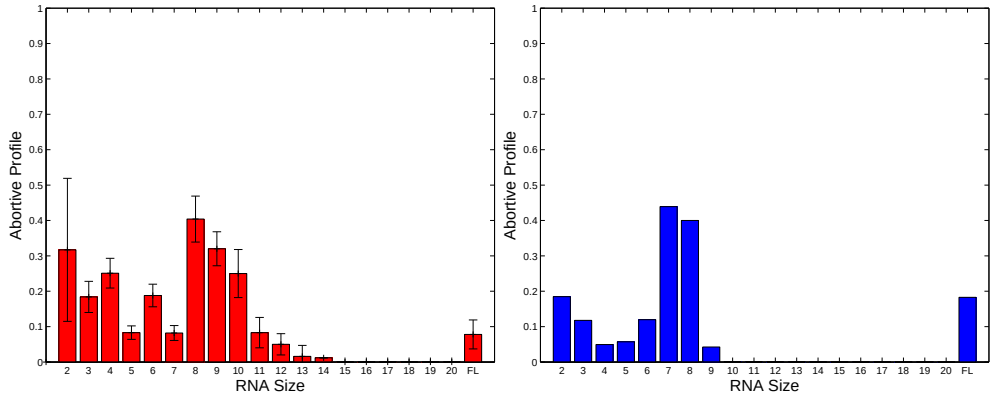
Figures 4.3 – 4.12 show the abortive profiles for each one of the setups described above for sequences N25, N25anti, DG146a, DG149a and DG137a.

We observe that while Setup I seems to reasonably match the experimental data for promoter N25 in Figure 4.3b, the same is not true for the other promoters. By comparing the profiles for Setups I and II we notice that the use of the modified abortive rates results in a decrease in production of abortive transcripts, and therefore an increase of full length transcripts. When we compare the profiles for Setups I and III we see that the use of the modified escape rates results in an increase of abortive transcripts and in production of longer transcripts, and therefore a decrease of full length transcripts.

It is clear that the increase of $\Delta\Delta G$ in setups IV, V and VI results in an increase of the length and percentage of abortive transcripts produced, and therefore a decrease in the percentage of full length transcripts. This is not surprising. We expect that as the binding affinity of the RNA polymerase to the promoter is increased, it will be more difficult to escape the promoter.

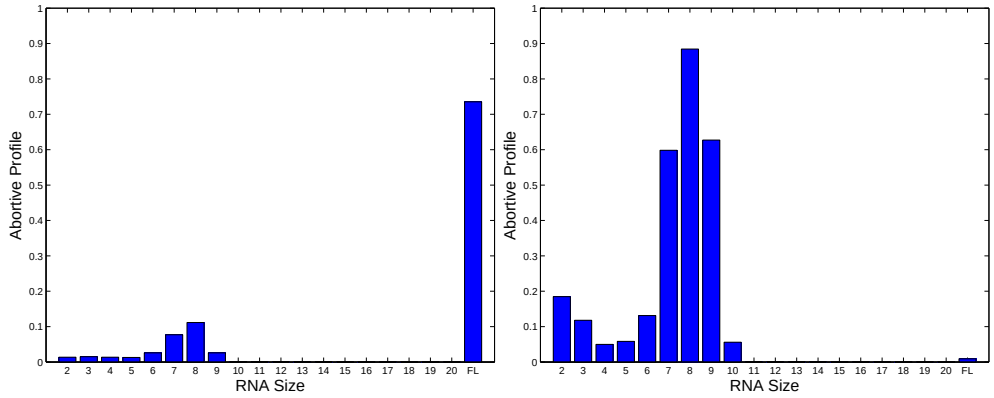
From setups I – VI we conclude that we cannot match the data either by using the Xue, Liu and Ou-Yang rates (setup I) or by modifying only the abortive rates (setup II), only the escape rates (setup III), or only the value used for $\Delta\Delta G$ (setups IV, V and VI). We then combine these three modifications in the Setups VII, VIII, IX and X, and again we are unable to match the experimental abortive profiles. The NTP-specific parameter values in setup XI do not solve our problem.

Being unable to match the experimental data by using modified abortive rates, modified escape rates, increased $\Delta\Delta G$ and NTP-specific values for the polymerization rates and NTP-dissociation constant, we look into the only other parameters we see as a possibility to fit the



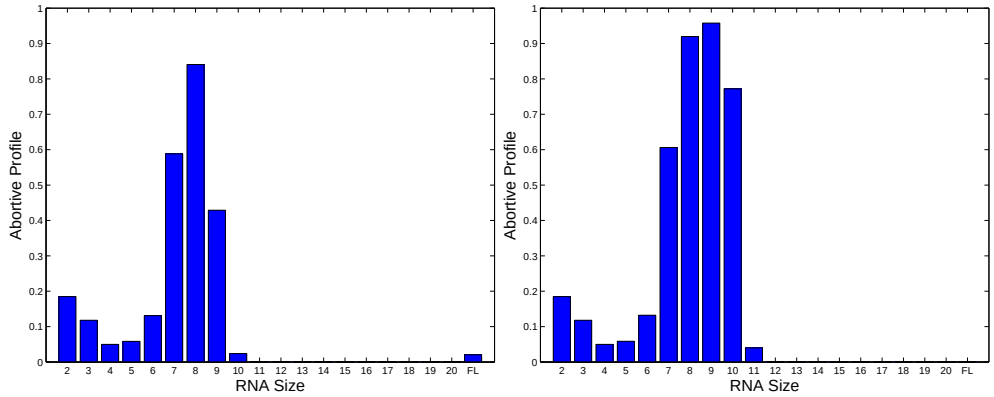
(a) N25 - Experimental Data

(b) N25 - Setup I



(c) N25 - Setup II

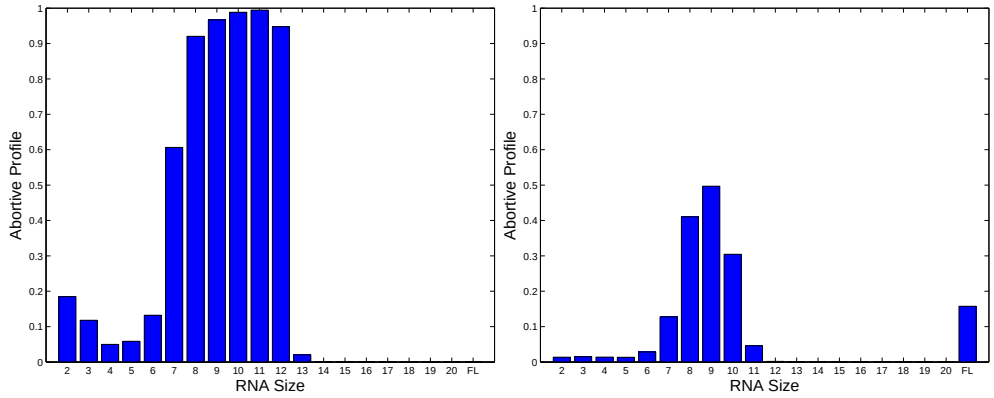
(d) N25 - Setup III



(e) N25 - Setup IV

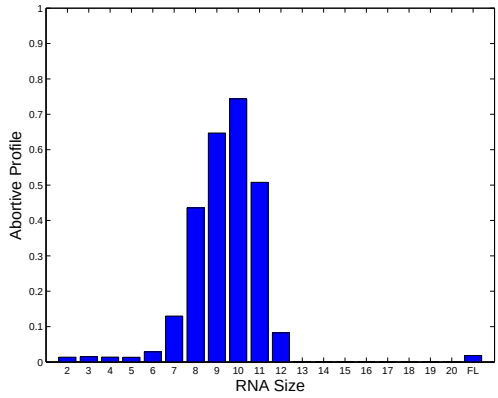
(f) N25 - Setup V

Figure 4.3: Abortive Profiles for N25

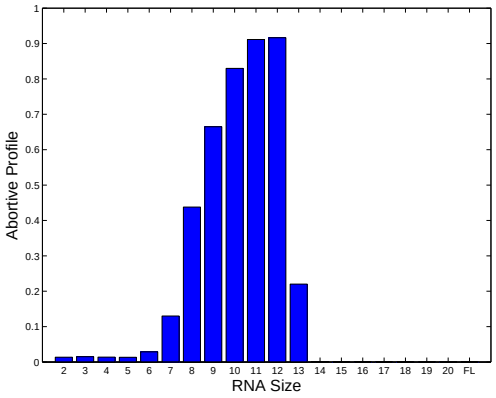


(a) N25 - Setup VI

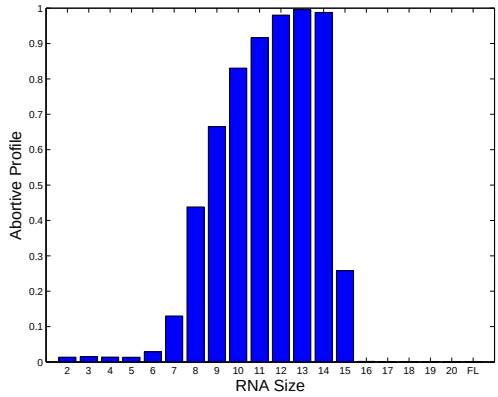
(b) N25 - Setup VII



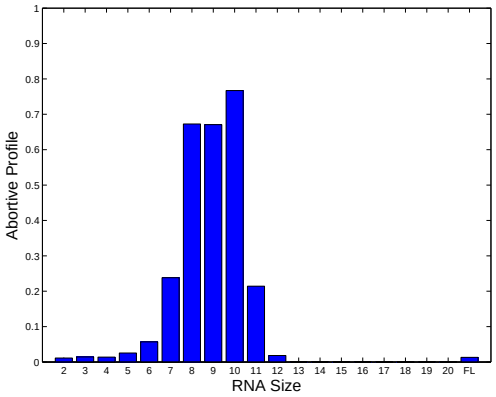
(c) N25 - Setup VIII



(d) N25 - Setup IX

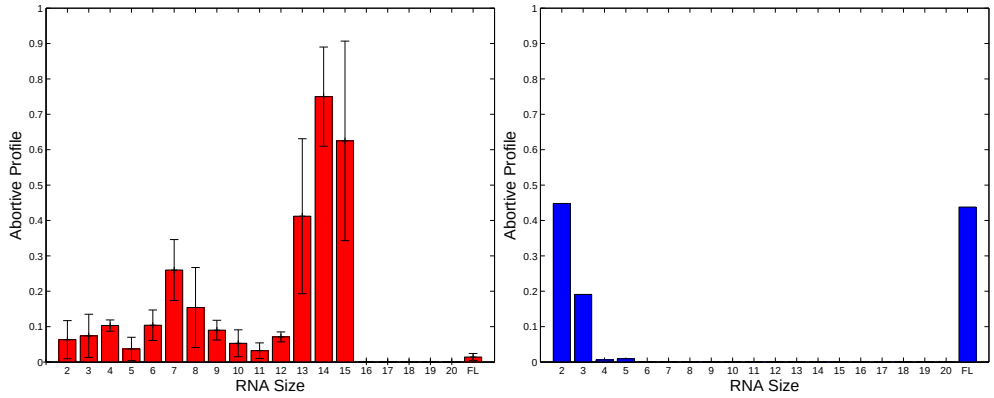


(e) N25 - Setup X



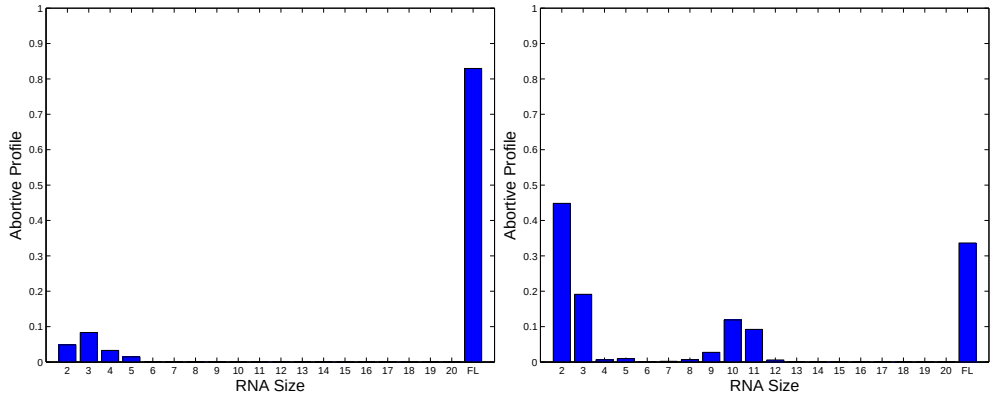
(f) N25 - Setup XI

Figure 4.4: Abortive Profiles for N25



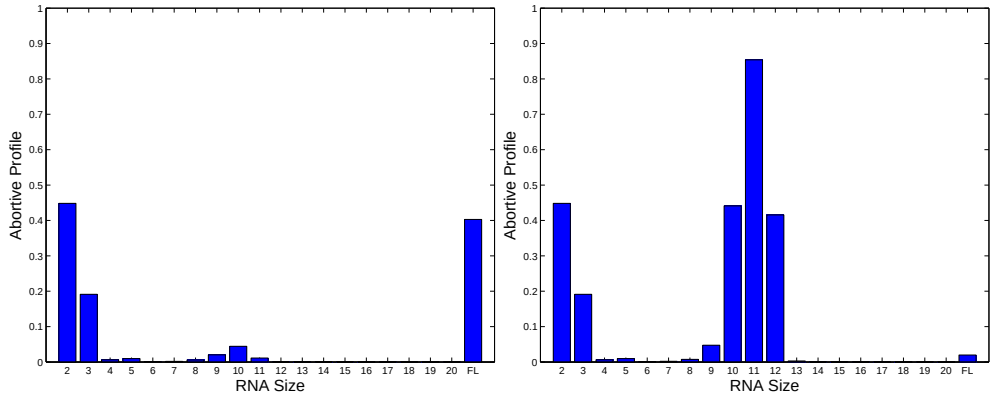
(a) N25anti - Experimental Data

(b) N25anti - Setup I



(c) N25anti - Setup II

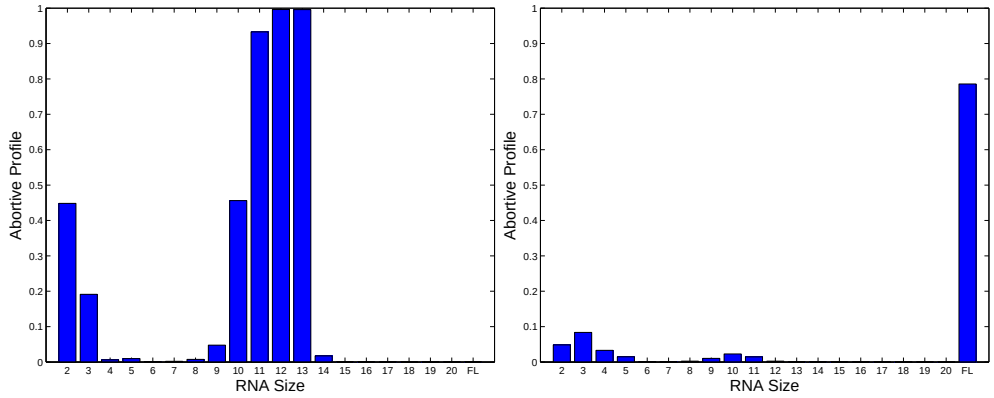
(d) N25anti - Setup III



(e) N25anti - Setup IV

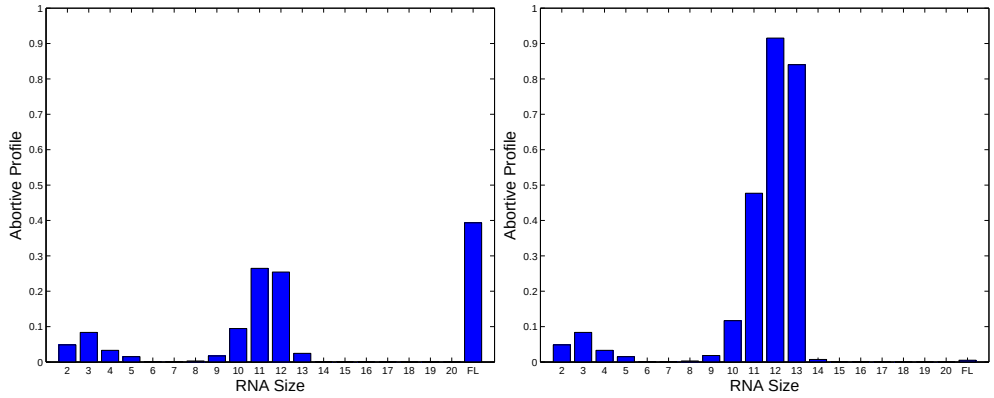
(f) N25anti - Setup V

Figure 4.5: Abortive Profiles for N25anti



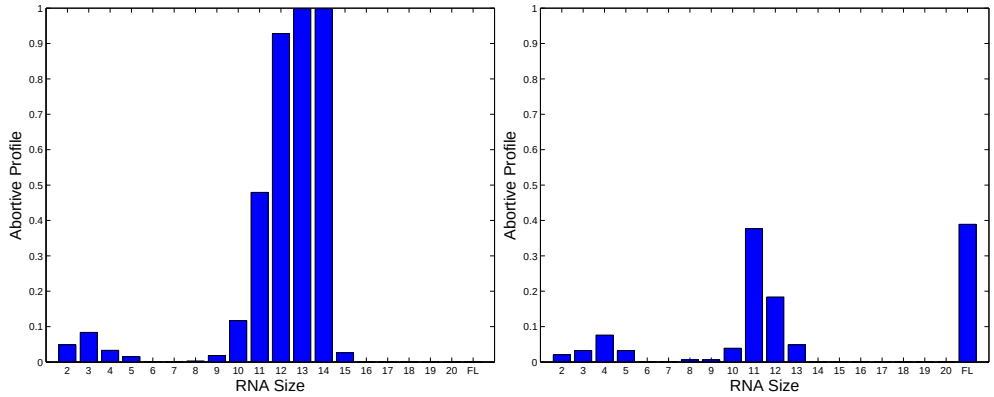
(a) N25anti - Setup VI

(b) N25anti - Setup VII



(c) N25anti - Setup VIII

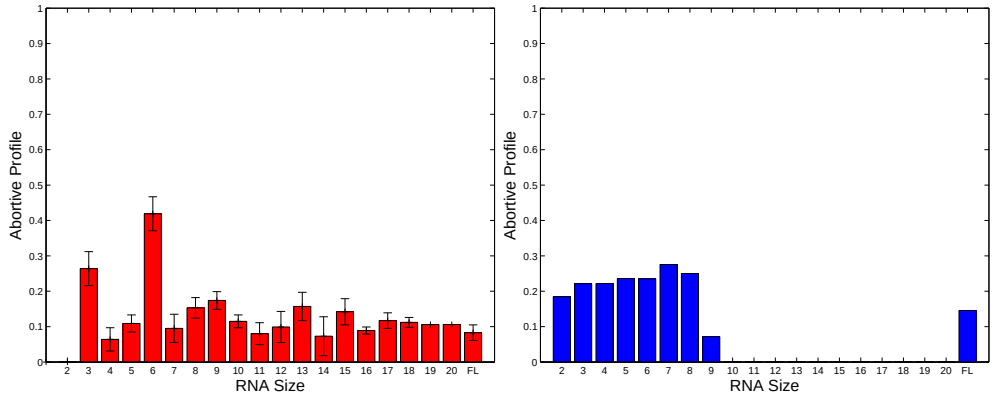
(d) N25anti - Setup IX



(e) N25anti - Setup X

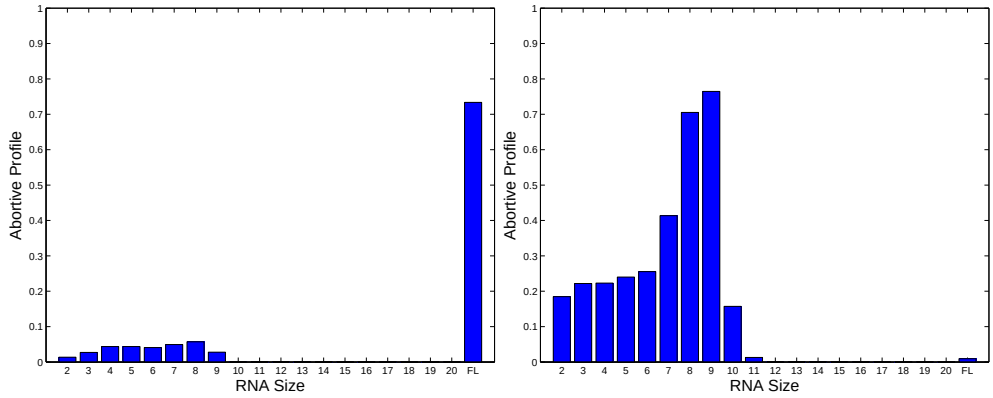
(f) N25anti - Setup XI

Figure 4.6: Abortive Profiles for N25anti



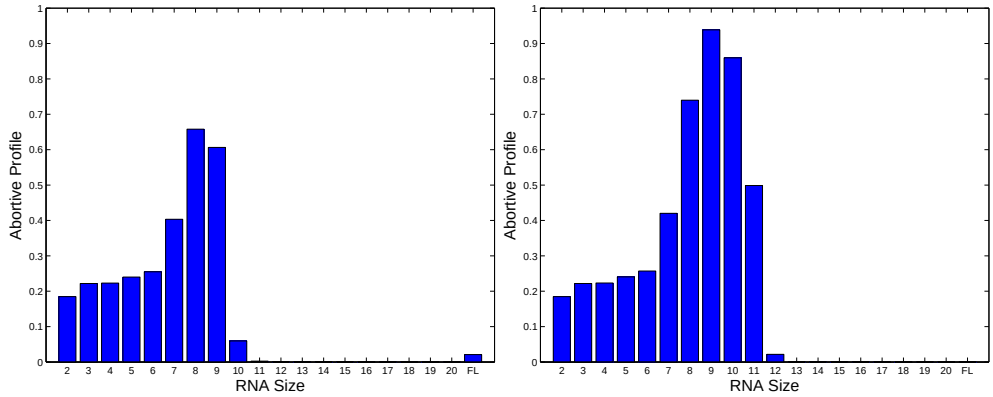
(a) DG146a - Experimental Data

(b) DG146a - Setup I



(c) DG146a - Setup II

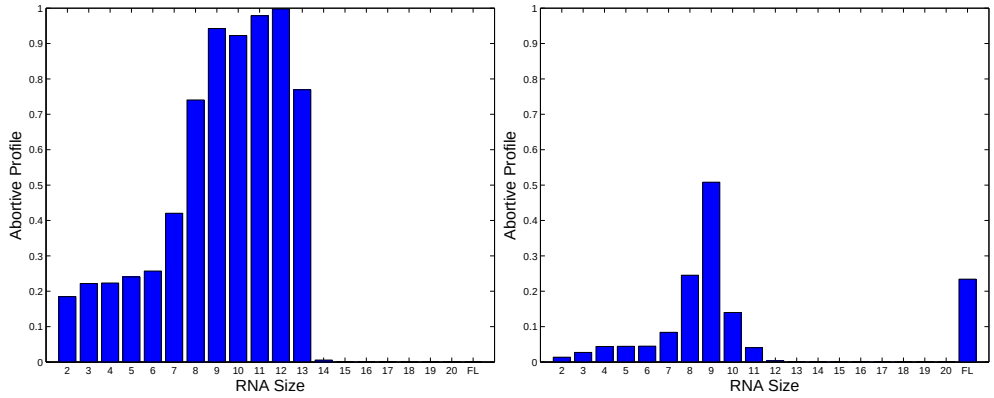
(d) DG146a - Setup III



(e) DG146a - Setup IV

(f) DG146a - Setup V

Figure 4.7: Abortive Profiles for DG146a



(a) DG146a - Setup VI

(b) DG146a - Setup VII

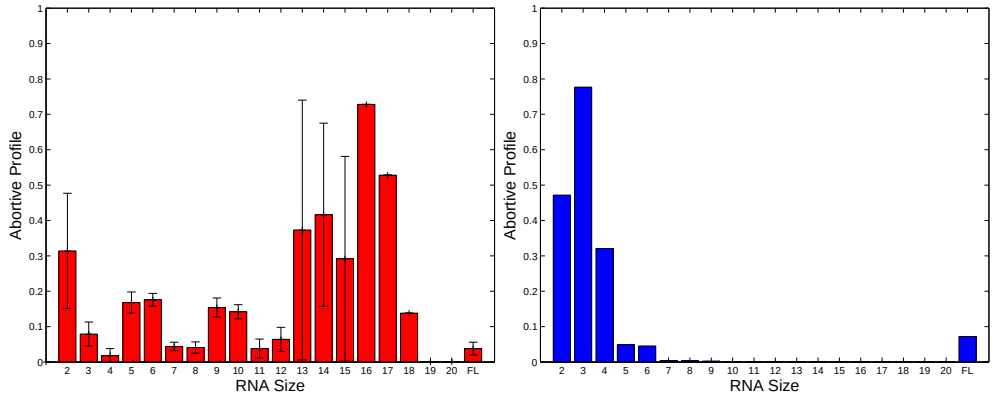
(c) DG146a - Setup VIII

(d) DG146a - Setup IX

(e) DG146a - Setup X

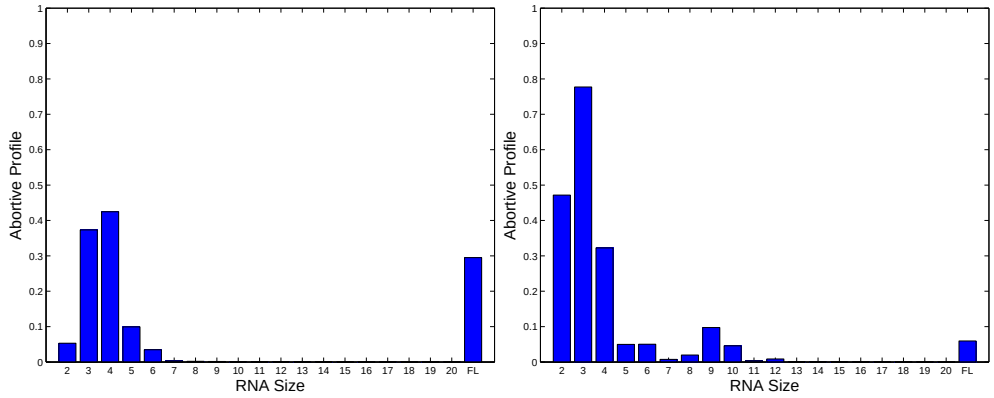
(f) DG146a - Setup XI

Figure 4.8: Abortive Profiles for DG146a



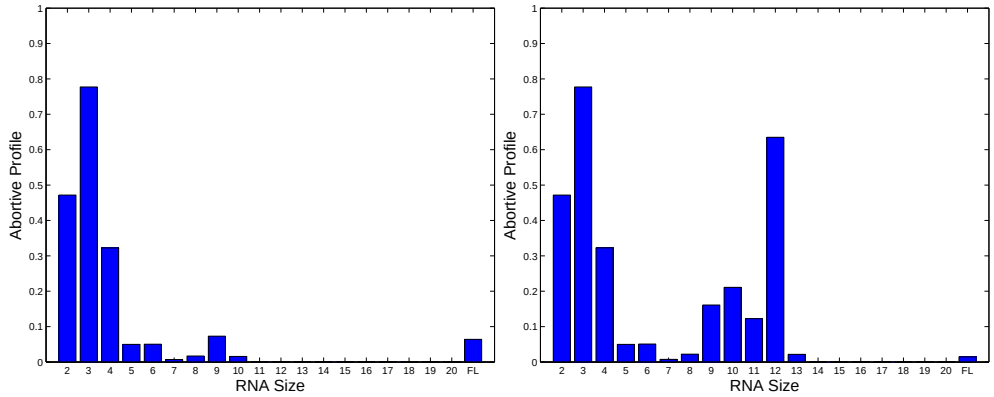
(a) DG149a - Experimental Data

(b) DG149a - Setup I



(c) DG149a - Setup II

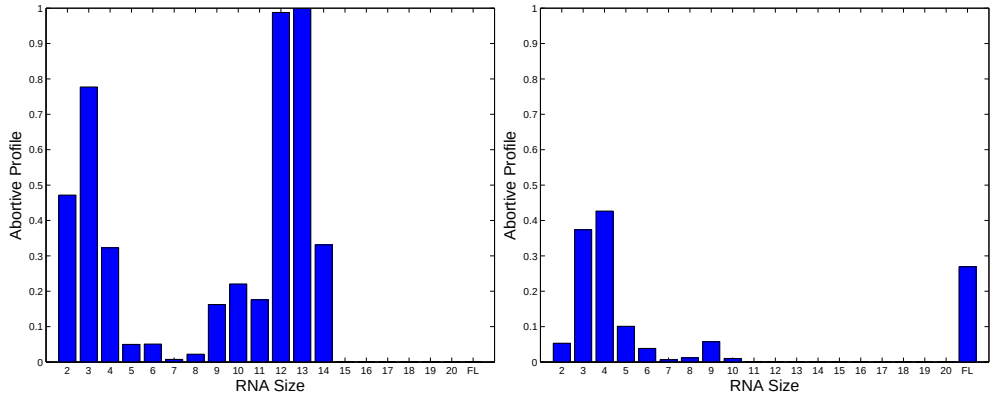
(d) DG149a - Setup III



(e) DG149a - Setup IV

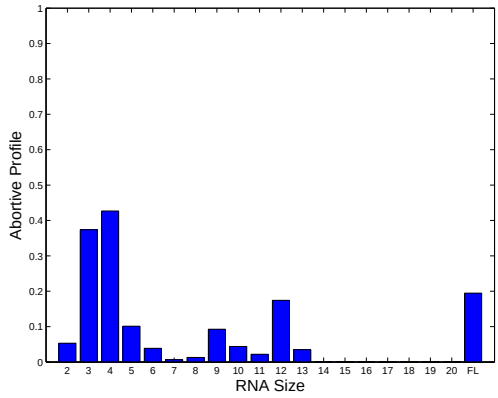
(f) DG149a - Setup V

Figure 4.9: Abortive Profiles for DG149a

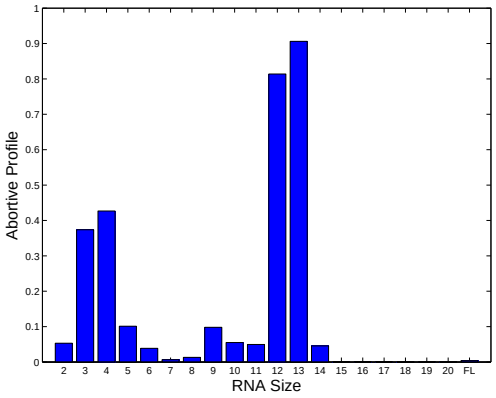


(a) DG149a - Setup VI

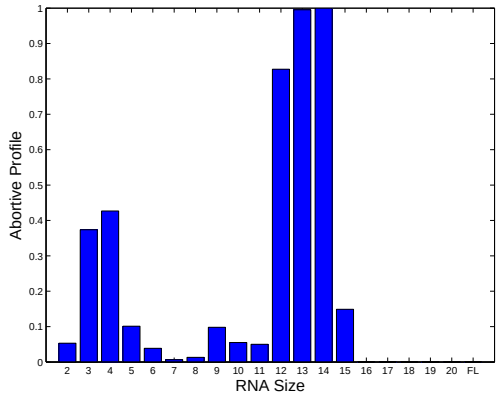
(b) DG149a - Setup VII



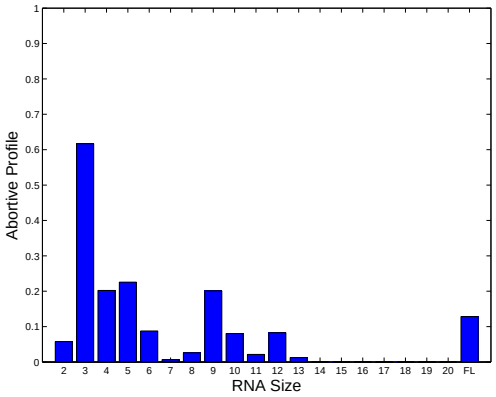
(c) DG149a - Setup VIII



(d) DG149a - Setup IX

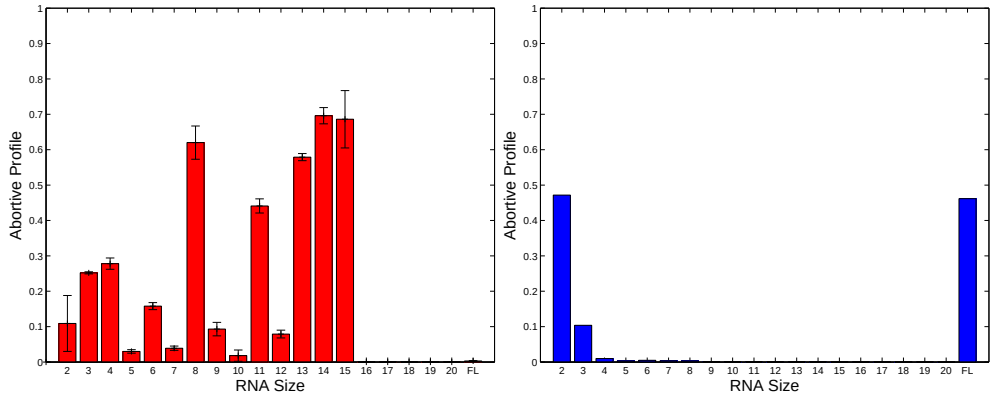


(e) DG149a - Setup X



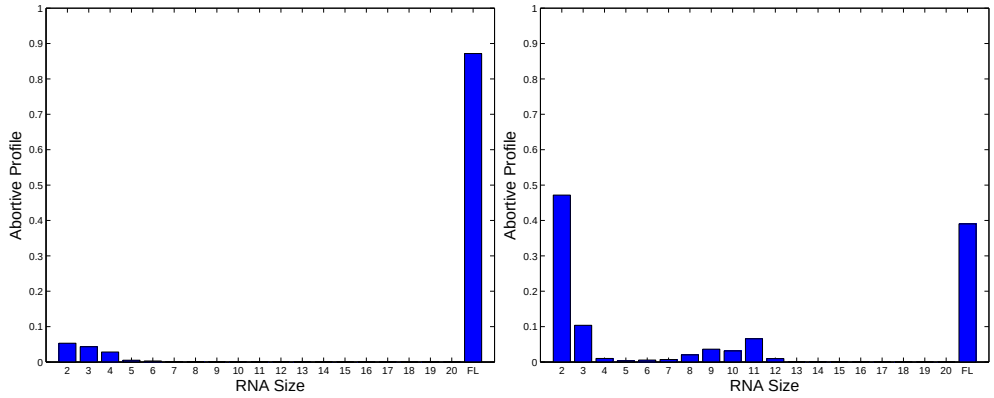
(f) DG149a - Setup XI

Figure 4.10: Abortive Profiles for DG149a



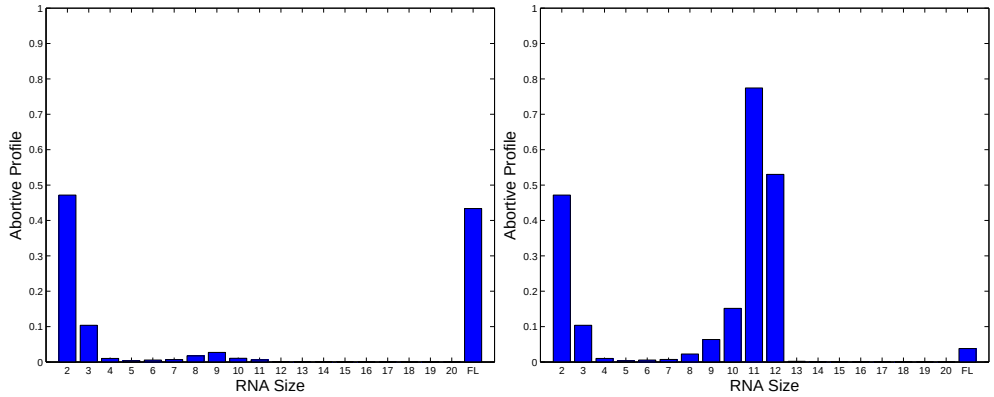
(a) DG137a - Experimental Data

(b) DG137a - Setup I



(c) DG137a - Setup II

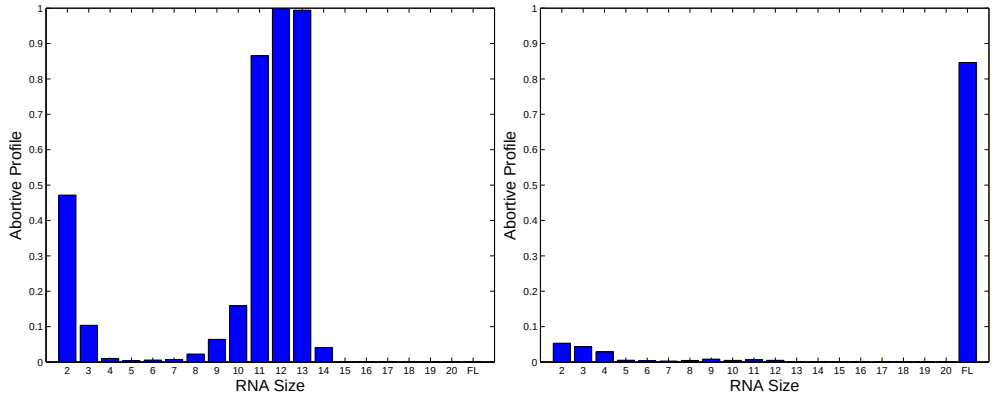
(d) DG137a - Setup III



(e) DG137a - Setup IV

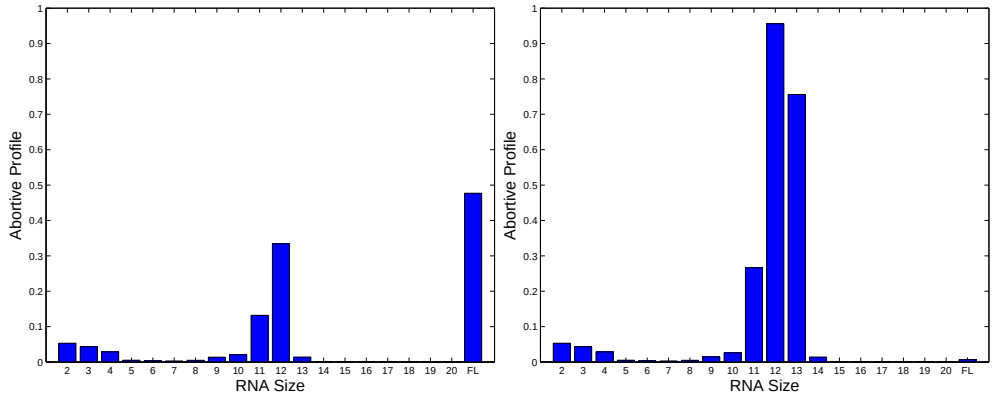
(f) DG137a - Setup V

Figure 4.11: Abortive Profiles for DG137a



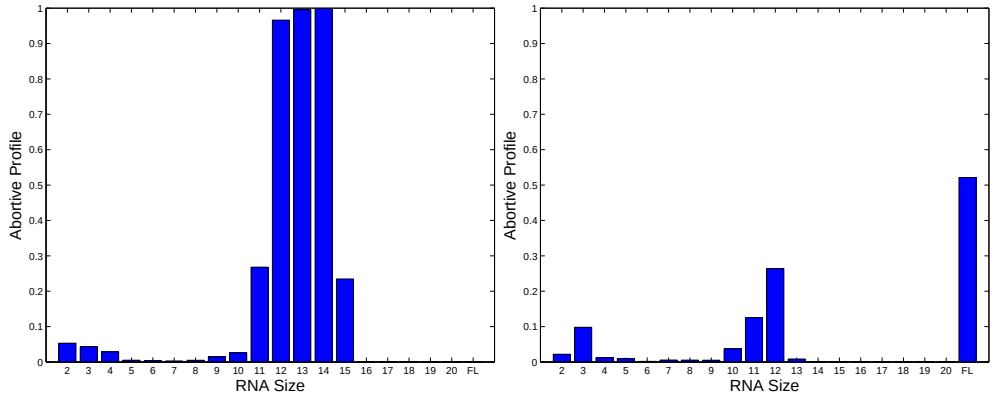
(a) DG137a - Setup VI

(b) DG137a - Setup VII



(c) DG137a - Setup VIII

(d) DG137a - Setup IX



(e) DG137a - Setup X

(f) DG137a - Setup XI

Figure 4.12: Abortive Profiles for DG137a

model: the Arrhenius constants k_a and k_e used in the abortive and escape rates, respectively.

4.2.5 Trying to Fit the Data

Since we believe the modified abortive and escape rates make more sense and the NTP-specific parameters are more accurate, from now on we will focus on the model with these modifications. We call *Modified Model* the model represented by Setup XI, that is, the model with abortive, escape and scrunching rates given by (4.6), (4.8) and (4.9), respectively, with polymerization rate and NTP-dissociation constants that are NTP-specific, and with $\Delta\Delta G = 5$ kcal/mol.

We now try to use the Arrhenius constants k_a and k_e to fit the Modified Model to the data. Recall the value we have been using so far is $k_a = k_e = 0.8$ s^{-1}, which is the same used in [31] for the Arrhenius constant in the escape rate. Also recall that in the abortive profiles the percentage P_n of abortive transcripts of length n is given by

$$P_n = \frac{P_a(n)P_s(n-1)P_s(n-2)\dots P_s(2)}{(1-P_2)(1-P_3)\dots(1-P_{n-1})}$$

where $P_a(.)$ and $P_s(.)$ represent the abortive and escape probabilities given by (4.10). Since P_n is given by the experimental data values, and we also have available the standard deviation for each n, we can find the range for values of the Arrhenius constants that would result in matching the data.

We start by fixing $k_e = 0.8$ s^{-1} as in [31] and using k_a to fit the data. Table 4.4 shows the values required in order to match the data for sequences N25 and N25anti using the Modified Model.

Table 4.4: Abortive Arrhenius constants

Position	N25	N25anti
2	9.21 – 76.41	0.35 – 5.1
3	9.02 – 16.36	0.32 – 3.72
4	15.03 – 23.57	0.93 – 1.31
5	2.09 – 3.47	0.1 – 1.8
6	2.45 –3.73	48.43 – 128.47
7	0.17 – 0.29	178.16 – 447.44
8	0.2 – 0.33	5.88 – 50.1
9	0.14 – 0.2	8.48 – 17.16
10	0.02 – 0.04	0.29 – 1.91
11	0.005 – 0.017	0.012 – 0.064
12	0.0021 – 0.0088	0.0334 – 0.0513
13	0 – 0.01	0.015 – 0.1

We are able to match the data for each of the 43 sequences, but using a different abortive Arrhenius constant at each position for each different sequence. Moreover, the values needed often are out of the usual range $[0, 1]$ for Arrhenius constants [77]. We do not believe this is an acceptable way to match the data. We then try to use both abortive and escape constants to fit the data. If we vary the Arrhenius constant for the escape reaction in the interval $[0, 1]$, the results for the abortive constant are basically unaltered. We also try, without success, to use the same approach for higher values of $\Delta\Delta G$. We try to find the best-fitting values for the Arrhenius constants minimizing the error. That does not work either. Therefore we cannot match the data by fitting these parameters.

It appears we are missing something.

4.3 Secondary Structure in the Scrunched DNA

In Section 4.2 we introduced improvements to the model of Xue, Liu and Ou-Yang, but we were unable to reasonably match the data in [76], even when we tried to fit parameters. Our approximations do not seem unreasonable, so we search for a more fundamental change to the model.

As described before, during the scrunching process RNA polymerase pulls the DNA and bulges of scrunched DNA occur in both strands. There is no doubt that portions of the accumulated single stranded DNA will be complementary (see Figure 4.13). From a theoreticaI point of view, it seems possible and likely that secondary structure will occur when there is complementarity, since secondary structure formation would result in a more favorable conformation. From a biological point of view there is no evidence that this will happen, but also no evidence of how this would be prevented.

We introduce this new feature to the Modified Model. If this feature has an effect in the model, we expect it to be non-uniform since it will be highly sequence-dependent.

The appearance of secondary structure in the scrunched DNA does not change the formulas for the scrunching, abortive and escape rates. The only difference is that each configuration has an additional energy contribution. Therefore for each configuration $P_m(M, N, n)$ we now have

$$\Delta G^m_{M,N,n} = \Delta G^{\text{bubble}}_{M,N,n} + \Delta G^{\text{hybrid}}_{M,N,n} + \Delta G^{\text{binding}}_{M,N,n} + \Delta G^{\text{fold}}_{M,N,n}$$

where $\Delta G^{\text{fold}}_{M,N,n}$ will be calculated as

$$\Delta G^{\text{fold}}_{M,N,n} = \Delta G^{\text{helix}}_{M,N,n} + \Delta G^{\text{init}}$$

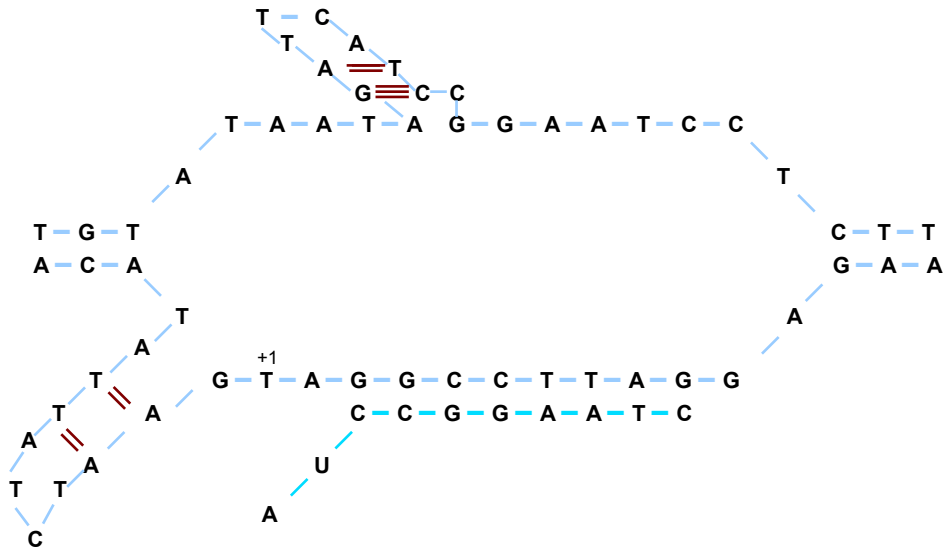


Figure 4.13: Possible complementarity in the scrunched bulges of DNA.

We use $\Delta G^{\text{init}} = -1$ kcal/mol and we use the software MFold [84] to predict the secondary structure in each scrunched strand of DNA and $\Delta G^{\text{helix}}_{M,N,n}$.

4.3.1 MFold

The MFold web server and software is used for prediction of secondary structure of single stranded nucleic acids. The MFold software for RNA folding was developed in the late 1980s by M. Zucker [85]. DNA folding prediction with the mfold software began in 1996, when DNA specific parameters were added to the MFold package through a collaboration with the SantaLucia group [79, 86, 87, 88, 89, 90]. In 1999, corrections for $[Na^+]$ and $[Mg^{++}]$ were incorporated into the MFold package [91]. The MFold web server was first created in 1995 and DNA folding parameters were added in 1996.

We use the software Mfold to predict scrunched DNA secondary structure. Single stranded DNA starts accumulating within the transcription bubble with the first scrunching reaction. Bulges of scrunched DNA will form in both DNA strands. The bulges are not symmetric: we assume the bulge on the 5′-to-3′ (coding) strand forms just after position -6, while the bulge on the 3′-to-5′ (template) strand forms just after position -11.

To predict secondary structure we will input the following information:

- sequence to be folded
- whether the sequence is linear or circular

- constraints, if any, to force or prohibit specific base pair formation
- folding temperature
- ionic conditions

Only the portions of the sequence correspondent to the accumulated bulges during scrunching are free to fold. Assume we want to fold the sequence in Fgure 4.14. In order to focus only

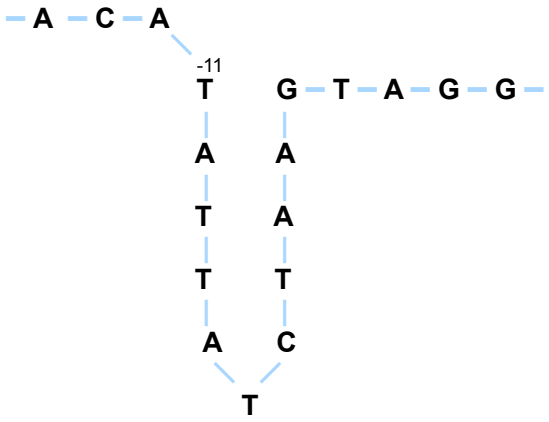


Figure 4.14: Example of a sequence to be folded

on the specific stretch of the sequence that we want to fold, we disregard the specific sequence composition for the part that will not be folded. We assume that the sequence to be folded is surrounded by "dummy" nucleotides. More specifically, we substitute the sequence that will not be folded by Ts. We then tried two approaches to describe our sequences to be folded: linear and circular DNA. Figure 4.15 shows the linear and circular approaches to the problem.

We tried both approaches with different number of dummy nucleotides inserted and there was no significant difference, so we choose to use the linear approach, adding 10 dummy nucleotides on each side of the scrunched sequence.

As we are comparing our results with experimental data, we use temperature and ionic conditions to be consistent with the ones used in [76]. The temperature used is 37^oC. Ionic conditions in [76] are $[Mg^{++}] = 10mM$ and $[K^+] = 200$ mM. Therefore we use $[Mg^{++}] = 0.01M$ and $[Na^+] = 0.2M$. (For the purposes of folding, Na^+ can be considered equivalent to K^+ as they have equivalent effects on nucleic acid thermal stability.)

Note that from a state with a scrunched DNA with n bases to the next state with $n + 1$ scrunched bases, we do not give preference to secondary structures already formed, instead we use the most energetically favorable structure.

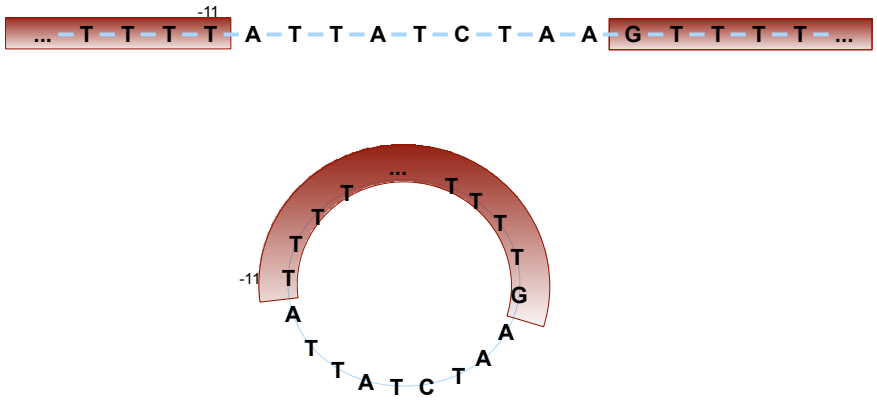


Figure 4.15: Linear vs. circular approach to fold scrunched DNA. Red boxes show the nucleotides for which base pairing will be prohibited.

Example 4.3.1 Let us consider the sequence in Figure 4.14. The sequence correspondent to the scrunched DNA in this case is

$$\text{ATTATCTAA}$$

As explained above, instead of using the sequence

$$\ldots \text{ACAT} \underbrace{\text{ATTATCTAA}}_{\text{to be folded}} \text{GATAGG} \ldots$$

as the input for MFold, we will use

$$\overbrace{\underbrace{\text{TTTTTTTTTT}}_{\text{cannot base pair}}}^{\text{dummy nucleotides}} \text{T} \overbrace{\text{ATTATCTAA}}^{\text{to be folded}} \text{G} \overbrace{\underbrace{\text{TTTTTTTTTT}}_{\text{cannot base pair}}}^{\text{dummy nucleotides}}$$

with the constraints that base pairing is only allowed for the sequence correspondent to the scrunched DNA. In this case our sequence corresponds to a state with 9 scrunched bases. Figure 4.16 has the resulting secondary structure we obtain from MFold. The helix energy contribution of the two base pairs is -0.93 kcal/mol.

We then use MFold to predict the secondary structure for the sequence correspondent to the state with 10 scrunched bases:

$$\overbrace{\underbrace{\text{TTTTTTTTTT}}_{\text{cannot base pair}}}^{\text{dummy nucleotides}} \text{T} \overbrace{\text{ATTATCTAAG}}^{\text{to be folded}} \text{A} \overbrace{\underbrace{\text{TTTTTTTTTT}}_{\text{cannot base pair}}}^{\text{dummy nucleotides}}$$

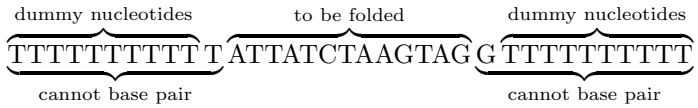
Figure 4.17 has the resulting secondary structure obtained. We see the result is the same

Figure 4.16: Mfold resulting plot for sequence correspondent to a state with 9 scrunched bases.

secondary structure predicted for the state with 9 scrunched bases, and therefore the energy contribution will be −0.93 kcal/mol.

Figure 4.17: Mfold resulting plot for sequence correspondent to a state with 10 scrunched bases.

The correspondent sequence input for the state with 13 scrunched bases is



and Figure 4.18 shows the two secondary structures we obtain with MFold. The secondary structure in Figure 4.18a has a helix energy contribution of −1.21 kcal/mol while the structure represented in Figure 4.18b has a contribution of −0.93 kcal/mol. The structure in Figure 4.18a is therefore the most favorable conformation.

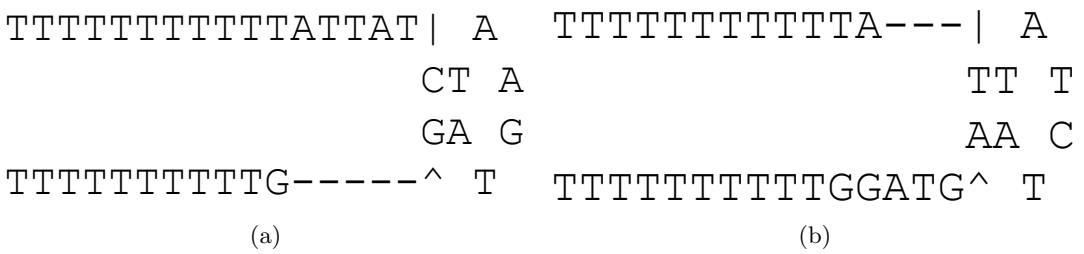


Figure 4.18: Mfold resulting plots for sequence correspondent to a state with 13 scrunched bases.

4.3.2 New Comparison to Data

Figure 4.19 shows the abortive profiles obtained using secondary structure in the scrunched DNA in comparison to the profiles obtained with the Modified Model and the experimental data. See Appendix D for additional comparisons.

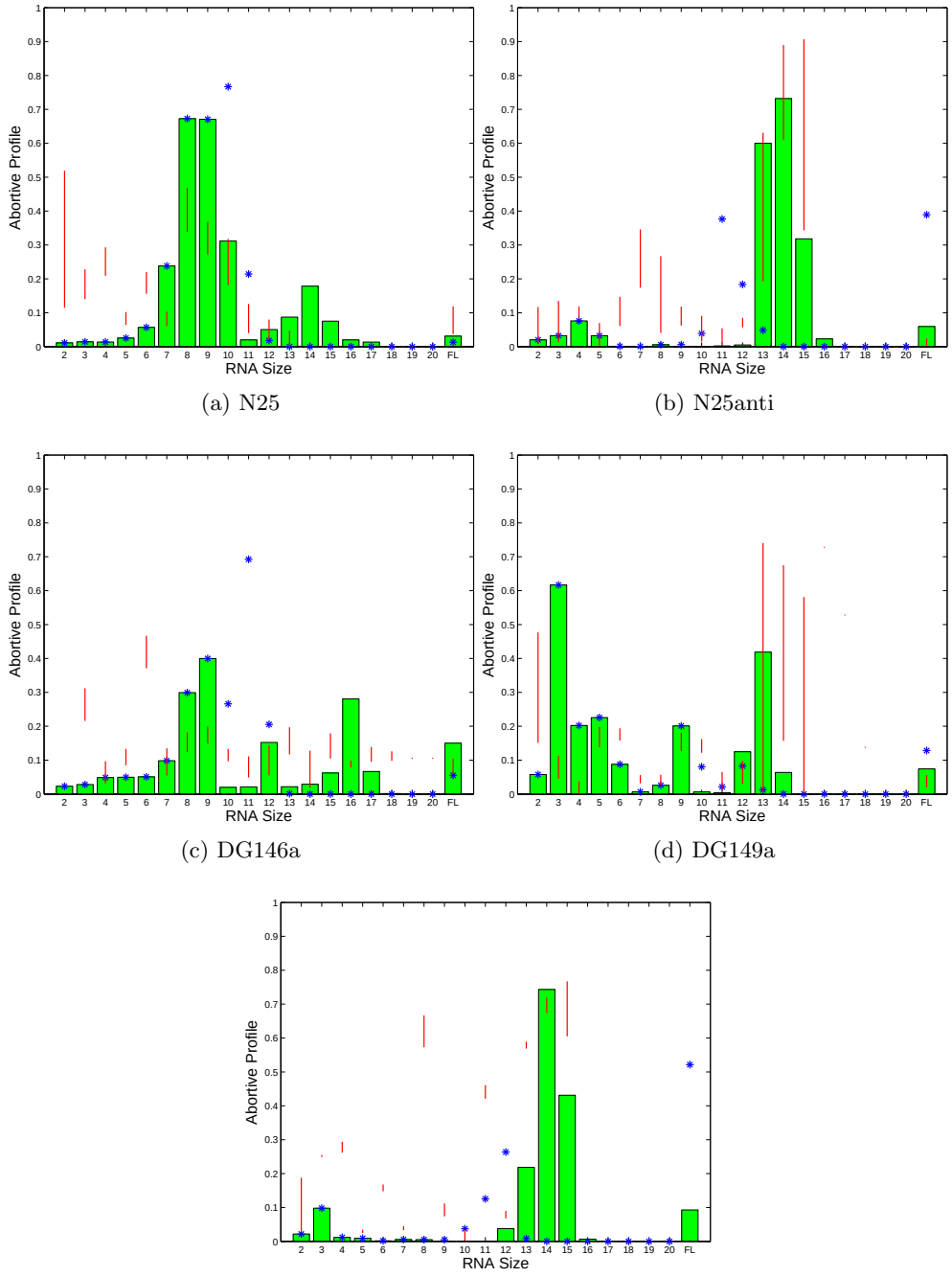
We see that the secondary structure starts to play a role in the model when the RNA length is 10, which corresponds to the states with 8 scrunched bases. This makes sense since enough DNA needs to accumulate in order to secondary structure to appear. We observe that by using the additional energy we are able to produce longer transcripts. We also notice an improvement in the percentage of full length transcripts produced. While the comparison to experimental data is not as satisfactory as we would like, we believe there is an overall improvement in the model.

4.4 Discussion

Our starting point for modeling promoter clearance is the model by Xue, Liu and Ou-Yang [31]. Although we agree with the main idea of their model, we do not agree with the approach used to model the rates. We then introduce some modifications to the XLO-Y model:

- modified abortive rates in order to avoid the NTP-assisted release hypothesis
- modified escape rates using transcription bubble of length 14 in the escape rates
- increased $\Delta\Delta G$
- NTP-specific values for the polymerization rates and NTP-dissociation constants

In Section 4.2.4 we see that the isolated introduction of each of the modifications above does not result in a good match to experimental data, and neither does the use of all the modifications.



(a) N25

(b) N25anti

(c) DG146a

(d) DG149a

(e) DG137a

Figure 4.19: Comparison of abortive profiles obtained using secondary structure in the scrunched DNA (bar plot), modified model (blue asterisk) and experimental data (red line).

In Section 4.2.5 we try to fit parameters, but the only way we can match the data is by using values for the Arrhenius constants that are not only different from sequence to sequence, but also from position to position.

In Section 4.3 we introduce another modification to the model:

- secondary structure of the scrunched DNA

While we are aware that there is no evidence for the secondary structure formation on the scrunched DNA, we are also unaware of anything which would prevent it. Clearly the only way to confirm or rule this out is experimentally. The introduction of this feature clearly has a positive impact to the model: the model now has the ability to produce longer abortive transcripts and we have an improved prediction of the percentage of full length transcripts produced. While we believe this resulted in an overall improvement to the model, we still cannot satisfactorily reproduce the abortive profiles for the experimental data we have been using.

There are other modifications we consider introducing to the model:

- DNA bending

  We believe there is an energy cost for the DNA bending during scrunching. Therefore we would have another energy contribution for each state with scrunched DNA. We do not believe this will have a big impact on the model.

- re-usage of short abortive transcripts

  Throughout our model we assume the release of abortive transcripts is an irreversible reaction. We are aware that, after being released, abortive transcripts of length 2 and 3 can be used to start transcription again. We cannot incorporate this fact to the model presented here, as we would have to be able to keep track of the continuous changes in the concentrations of these short transcripts. We believe this modification would have an impact in the model, specially for promoters that produce high percentage of abortive transcripts of length 2 and 3. And we believe the impact would be more evident for the initial positions of the abortive profiles in contrast to the secondary structure on the scrunched DNA. Therefore, in theory, the combination of these two features may provide a better chance to match the experimental data.

Appendices

Appendix A
Statistical Thermodynamics

Suppose we are investigating a system that has volume V, contains N molecules and is immersed in a large heat bath at temperature T. An ensemble is a collection of a very large number $\mathcal{M}$ of systems, each constructed to be a replica on a thermodynamic level of the actual system whose properties we are investigating.

At any instant of time, in an ensemble constructed by replication of a given thermodynamic system in a given environment, many different energy states are represented in the various systems of the ensemble. The ensemble average of the energy is then the average over these instantaneous values of the energy.

First Postulate: The time average of a mechanical variable M in the thermodynamic system of interest is equal to the ensemble average of M, in the limit as $\mathcal{M} \to \infty$, provided that the systems in the ensemble replicate the thermodynamic state and environment of the system of interest.

Second Postulate: In an essemble representative of an isolated thermodynamic system, the systems of the ensemble are distributed uniformly, with equal probability or frequency, over the possible energy states.

The first postulate tells us that the time average on the actual system may be replaced by an instantaneous average over a large number of representative systems. The second postulate says that if a system is selected at random from the ensemble, the probability that it will be found in a particular energy state is the same for all possible quantum states. An implication of the two postulates is that the single isolated system of interest spends equal amounts of time, over a long period of time, in each of the available states.

Since $\mathcal{M}$ is extremely large, the energy levels for the system in the Second Postulate will be so close together as to be pratically continuous, and furthermore, each of these levels will have an extremely large degeneracy (i.e., number of energy states). The number of energy states associated with an energy level E for a system with N molecules and volume V will be denoted by $\Omega(N, V, E)$. Thus the number of "possible energy states" referred to in the Second Postulate

is Ω.

We can look at the ensemble itself as an isolated system with volume $\mathcal{M}V$, number of molecules $\mathcal{M}N$ and a total energy that will be denoted by E_t.

All the possible energy states for such a system can be listed in increasing order of the energy value, $E_1, E_2, \ldots, E_j, \ldots$ When degeneracy occurs, several successive E_j's will have the same value. In the notation introduced above, the energy value E occurs Ω successive times in the list.

Since each system in the ensemble has the same N and V, all systems have the same set of energy states, represented by $E_1, E_2, \ldots, E_j, \ldots$ Let n_j be the number of systems found in state E_j. The set of numbers $n_1, n_2, \ldots$ is called a distribution. There are, of course, many possible distributions that might be observed, but obviously all must satisfy

$$\sum_j n_j = \mathcal{M} \tag{A.1}$$

$$\sum_j n_j E_j = E_t. \tag{A.2}$$

The number of states of the supersystem, $\Omega_t(n)$, consistent with a given distribution $n = (n_1, n_2, \ldots)$ is given by

$$\Omega_t(n) = \frac{(n_1 + n_2 + \ldots)!}{n_1! n_2! \ldots} = \frac{\mathcal{M}!}{\prod_j n_j!} \tag{A.3}$$

The objective here is to find the probability of observing a given energy state E_j in a system selected from a ensemble. For a particular distribution $n = (n_1, n_2, \ldots)$ this probablility is $n_j/\mathcal{M}$ for state E_j. But, in general, given $N, V, \mathcal{M}$ and E_t there are many possible distributions. What is needed is the over-all probability, that is, an average of $n_i/\mathcal{M}$ over these distributions. The probability of observing a given energy state E_j in an arbitrary system of a canonical ensemble is

$$P_j = \frac{1}{\mathcal{M}} \frac{\sum_n \Omega_t(n) n_j(n)}{\sum_n \Omega_t(n)}, \tag{A.4}$$

where $n_j(n)$ represents the value of n_j in the distribution n. The sum is over all distributions satisfying Eqs. (A.1) and (A.2). Of course, $\sum_j P_j = 1$. Then the desired ensemble averages of, for example, the energy and pressure are

$$\overline{E} = \sum_j P_j E_j \tag{A.5}$$

$$\overline{p} = \sum_j P_j p_j,$$

where p_j is the pressure in state E_j and is defined by

$$p_j = -(\frac{\partial E_j}{\partial V})_N.$$

That is, $-p_j dV = dE_j$ is the work that has to be done on the system, when in state E_j, in order to increase the volume by dV.

In principle, Eq. (A.4) tells us all we need to know to calculate ensemble averages of mechanical variables. But in practice, a much more explicit expression for P_j is necessary. In any particular case we are given $\mathcal{M}$, the E_j (determined by N and V) and E_t (determined by $\mathcal{M}$, N, V and T). There are too many possible distributions n consistent with the restrictions of Equations (A.1) and (A.2). For each of these distributions we can calculate from Eq. (A.3) the weight $\Omega_t(n)$ to be used in obtaining averages. Because of large numbers involved, the most probable distribution, and distributions that differ only negligibly from the most probable distribution, completely dominate the computation of the average in Eq. (A.6). In practice this means that, in the limit as $\mathcal{M} \to \infty$, we can regard all other weights $\Omega_t(n)$ as negligible compared with $\Omega_t(n^*)$.

Naturally, as we let $\mathcal{M} \to \infty$, holding N, V and T fixed, each $n_j \to \infty$ also. But all ensemble averages depend only on the ratio $n_j/\mathcal{M}$, which remains finite. Eq. (A.4) becomes, then,

$$P_j = \frac{1}{\mathcal{M}} \frac{\Omega_t(n^*) {n_j}^*}{\Omega_t(n^*)} = \frac{{n_j}^*}{\mathcal{M}}, \tag{A.6}$$

where ${n_j}^*$ is the value of n_j in the most probable distribution, n^*. Equation (A.6) tells us that in the computation of P_j we can replace the mean value n_j by the value of n_j in the most probable (largest Ω_t) distribution. This leads us to a purely mathematical question: Which of the possible sets of n_j's satisfying Eqs. (A.1) and (A.2) gives us the largest Ω_t? This problem is solved using Lagrange multipliers.

First observe that the distribution giving the largest Ω_t is also the distribution giving the largest $\ln \Omega_t$, since $\ln x$ increases monotonically with x. From Equation (A.3) and the use of Stirling's approximation we have

$$\begin{aligned}
\ln \Omega_t(n) &= \ln((\sum_j n_j)!) - \ln(\prod_j n_j!) \\
&\approx (\sum_j n_j) \ln(\sum_j n_j) - \sum_j n_j - \sum_j \ln(n_j!) \\
&\approx (\sum_j n_j) \ln(\sum_j n_j) - \sum_j n_j - \sum_j (n_j \ln n_j - n_j)
\end{aligned}$$

$$= (\sum_j n_j) \ln(\sum_j n_j) - \sum_j (n_j \ln n_j)$$

According to the method of Lagrange multipliers, the set of n_j's leading to the maximum value of $\ln \Omega_t(n)$, subject to the conditions (A.1) and (A.2) is found from the equations

$$\nabla(\ln \Omega_t(n^*)) = \alpha \nabla(\sum_i {n_i}^*) + \beta \nabla(\sum_i ({n_i}^* E_i)),$$

that is, for $j = 1, 2, ...$

$$\frac{\partial}{\partial n_j}[\ln \Omega_t(n^*) - \alpha \sum_i {n_i}^* - \beta \sum_i ({n_i}^* E_i)] = 0$$

where α and β are the undetermined multipliers. On carrying out the differentiation, we find

$$\ln(\sum_i n_i) + (\sum_i n_i) \frac{1}{\sum_i n_i} - (\ln {n_j}^* + {n_j}^* \frac{1}{n_j}) - \alpha - \beta E_j = 0$$

or, for $j = 1, 2, ...$

$$\begin{aligned}
&\ln(\sum_i n_i) - \ln {n_j}^* - \alpha - \beta E_j = 0 \\
&\ln {n_j}^* = \ln(\sum_i n_i) - \alpha - \beta E_j \\
&{n_j}^* = \sum_i n_i - \alpha - \beta E_j \\
&{n_j}^* = \mathcal{M} e^{-\alpha} e^{-\beta E_j}
\end{aligned} \tag{A.7}$$

This is the most probable distribution, expressed in terms of α and β.

Substituting n_j^* in Equations (A.1) and (A.2), we obtain

$$\begin{aligned}
&\sum_j n_j^* = \sum_j (\mathcal{M} e^{-\alpha} e^{-\beta E_j}) = \mathcal{M} e^{-\alpha} \sum_j e^{-\beta E_j} = \mathcal{M} \\
&e^{-\alpha} \sum_j e^{-\beta E_j} = 1 \\
&e^{\alpha} = \sum_j e^{-\beta E_j}.
\end{aligned}$$

Also, for $j = 1, 2, ...$

$$P_j = \frac{n_j^*}{\mathcal{M}} = \frac{\mathcal{M} e^{-\alpha} e^{-\beta E_j}}{\mathcal{M}} = e^{-\alpha} e^{-\beta E_j} = \frac{e^{-\beta E_j}}{\sum_i e^{-\beta E_i}} \tag{A.8}$$

Combining Eqs. (A.8) and (A.5) we have

$$\overline{E} = \sum_j P_j E_j = \sum_j E_j \frac{e^{-\beta E_j}}{\sum_i e^{-\beta E_i}} = \frac{\sum_j E_j e^{-\beta E_j}}{\sum_i e^{-\beta E_i}} \tag{A.9}$$

Notice that, by the First Postulate, we can associate the thermodynamic pressure p and energy E with the statistical-mechanical ensemble averages $\overline{p}$ and $\overline{E}$, respectively. Then from Eq. (A.9), holding N constant, we have:

$$\nabla\overline{E} = \sum_j E_j \nabla P_j + \sum_j P_j \nabla E_j$$

Now notice that

$$\nabla E_j = (\frac{\partial E_j}{\partial V})_N \nabla V$$

Let $Q = \sum_j e^{-\beta E_j}$. Then

$$\sum_j (\ln P_j + \ln Q) = \sum_j \ln(P_j Q) = \sum_j \ln(\frac{e^{-\beta E_j}}{\sum_i e^{-\beta E_i}} (\sum_j e^{-\beta E_j})) = \sum_j -\beta E_j$$

thus

$$\nabla\overline{E} = -\frac{1}{\beta} \sum_j (\ln P_j + \ln Q) \nabla P_j + \sum_j P_j (\frac{\partial E_j}{\partial V})_N \nabla V.$$

Since $\sum_j P_j = 1$, we have $\sum_j \nabla P_j = 0$. Also

$$\begin{aligned}
\nabla(\sum_j P_j \ln P_j) &= \sum_j (\nabla(P_j) \ln P_j) + \sum_j (P_j \nabla(\ln P_j)) \\
&= \sum_j (\nabla(P_j) \ln P_j) + \sum_j (P_j \frac{1}{P_j} \nabla P_j) \\
&= \sum_j (\nabla(P_j) \ln P_j) + \sum_j \nabla P_j \\
&= \sum_j (\nabla(P_j) \ln P_j).
\end{aligned}$$

Therefore we have

$$\nabla\overline{E} = \frac{-1}{\beta} \nabla(\sum_j (P_j \ln P_j)) + \sum_j P_j (\frac{\partial E_j}{\partial V})_N \nabla V,$$

or

$$\frac{-1}{\beta} \nabla(\sum_j (P_j \ln P_j)) = \nabla\overline{E} + \overline{p} \nabla V. \tag{A.10}$$

Since we already have the associations with thermodynamics $E \leftrightarrow \overline{E}$ and $p \leftrightarrow \overline{p}$, and since

in thermodynamics (N constant)

$$T\nabla S = \nabla E + p\nabla V,$$

we can deduce from Eq. (A.10) the further association

$$T\nabla S \leftrightarrow \frac{-1}{\beta}\nabla(\sum_j (P_j \ln P_j)) \tag{A.11}$$

Note that from Eq. (A.10) and $\nabla E = \nabla Q^* - \nabla W$ we have

$$\nabla Q^* = T\nabla S \leftrightarrow \sum_j E_j \nabla P_j$$
$$\nabla W = p\nabla V \leftrightarrow \sum_j P_j \nabla E_j,$$

where Q^* and W are heat absorbed and work done by the system, respectively. These relations provide us, in a general way, with the molecular interpretation of the thermodynamic concepts of heat and work. We see that when a closed thermodynamic system increases its energy infinitesimally by the absortion of heat from its surroundings, this is accomplished not by changing the energy levels of the system but rather by a shift in the fraction of time the system spends in the various energy states. The converse statement can be made about the work term.

From Eq. (A.11)

$$\nabla S \leftrightarrow \frac{1}{\beta T}\nabla G$$

or

$$\nabla S \leftrightarrow \phi(G)\nabla G = \nabla f(G), \tag{A.12}$$

where $\phi(G) = \frac{1}{\beta T}$, $G = -\sum_j P_j \ln P_j$ and $f(G)$ is obtained by integrating $\phi(G)$. Then

$$S = f(G) + c, \tag{A.13}$$

where c is a constant independent of G and therefore independent of the variables on which G depends (e.g., β and V, with N constant). Experimental information about the entropy always involves a difference in entropy between two states (e.g., the entropy change ΔS between T_1 and T_2 at constant N and V), never an absolute value. The constant c in Eq. (A.13) always cancels on taking such a difference. Hence its value is completely arbitrary from an operational point of view. But for convenience and simplicity, we adopt the particular choice $c = 0$.

Up to this point we have that $S \leftrightarrow f(G)$, but we do not know the function f. To settle this matter we make use of a thermodynamic property of entropy, namely its additivity. Specifically, suppose we have two thermodynamic systems A and B at the same temperature and with entropies S_A and S_B, respectively. Then if we regard the combined systems as a new system AB, we have $S_{AB} = S_A + S_B$. This relationship suffices to determine f, as we now show.

We first investigate whether the statistical-mechanical quantity G is additive in the above sense. For this purpose we form an ensemble of $\mathcal{M}$ systems AB representative of a thermodynamic AB system at temperature T. Heat can flow through all interior walls of the ensemble. The A part of the thermodynamic system is characterized further by N^A and V^A, and the B part by N^B and V^B . In general, the types of molecules may be different in A and B. We have two sets of energy states for the separate systems, ${E_1}^A$, ${E_2}^A$, ... and ${E_1}^B$, ${E_2}^B$, ... If ${n_j}^A$ stands for the number of A systems in the ensemble in state ${E_j}^A$, with similar meaning for ${n_j}^B$, then the number of states of the whole ensemble, or supersystem, consistent with given distributions n^A and n^B is

$$\Omega_t(n^A, n^B) = \frac{(\sum_j {n_j}^A)!}{\prod_j {n_j}^A!} \times \frac{(\sum_j {n_j}^B)!}{\prod_j {n_j}^B!},$$

since the A and B systems are independent of each other (except for energy exchange through the walls). The distributions of interest must satisfy the equations

$$\begin{aligned} \sum_j {n_j}^A &= \mathcal{M} \\ \sum_j {n_j}^B &= \mathcal{M} \\ \sum_j ({n_j}^A {E_j}^A + {n_j}^B {E_j}^B) &= E_t \end{aligned} \tag{A.14}$$

The argument here on is essentially the same as before. We want to find the distribution that maximizes the number $\Omega_t(n^A, n^B)$, subject to the restrictions given in (A.14). Again we will use Lagrange multipliers. The distribution, that we will call again n^*, giving the maximum $\Omega_t(n^A, n^B)$ will give also the maximum $\ln \Omega_t(n^A, n^B)$ and will be given by

$$\nabla \ln \Omega_t(n^A, n^B) = \alpha_A \nabla(\sum_j {n_j}^A) + \alpha_B \nabla(\sum_j {n_j}^B) + \beta \nabla(\sum_j ({n_j}^A {E_j}^A + {n_j}^B {E_j}^B)),$$

or, for $j = 1, 2, ...$

$$\frac{\partial}{\partial {n_j}^A}(\ln \Omega_t(n^A, n^B)) = \alpha_A \frac{\partial}{\partial {n_j}^A}(\sum_j {n_j}^A) + \alpha_B \frac{\partial}{\partial {n_j}^A}(\sum_j {n_j}^B) + \beta \frac{\partial}{\partial {n_j}^A}(\sum_j ({n_j}^A {E_j}^A + {n_j}^B {E_j}^B))$$

$$\frac{\partial}{\partial {n_j}^B}(\ln \Omega_t(n^A, n^B)) = \alpha_A \frac{\partial}{\partial {n_j}^B}(\sum_j {n_j}^A) + \alpha_B \frac{\partial}{\partial {n_j}^B}(\sum_j {n_j}^B) + \beta \frac{\partial}{\partial {n_j}^B}(\sum_j ({n_j}^A {E_j}^A + {n_j}^B {E_j}^B))$$

Continuing the differentiation (and using Stirling's approximation) we will have

$${n_j}^A* = \mathcal{M} e^{-\alpha_A} e^{-\beta E_j^A}$$

$${n_j}^B* = \mathcal{M} e^{-\alpha_B} e^{-\beta E_j^B}.$$

Using the fact that $\sum_j {n_j}^A = \sum_j {n_j}^B = \mathcal{M}$, for $j = 1, 2, ...$ we obtain

$$e^{\alpha_A} = \sum_j e^{-\beta {E_j}^A}$$

$$e^{\alpha_B} = \sum_j e^{-\beta {E_j}^B}$$

For the probability that the thermodynamic system AB has its A part in state ${E_i}^A$ and its B part in state ${E_j}^B$, we find

$$P_{ij} = \frac{e^{-\beta {E_i}^A} e^{-\beta {E_j}^B}}{Q_A Q_B} = {P_i}^A {P_j}^B, \tag{A.15}$$

where

$$Q_A = \sum_j e^{-\beta {E_j}^A} \text{ and } Q_B = \sum_j e^{-\beta {E_j}^B}.$$

We deduce from equation (A.15) that if two systems are in thermal contact with each other, they have the same β. This suggests a close connection between β and T, which we verify below.

For the combined system AB,

$$G_{AB} = -\sum_{i,j} P_{ij} \ln P_{ij} = -\sum_{i,j} {P_i}^A {P_j}^B (\ln {P_i}^A + \ln {P_j}^B)$$

$$= -\sum_i {P_i}^A \ln {P_i}^A - \sum_j {P_j}^B \ln {P_j}^B = G_A + G_B. \tag{A.16}$$

That is, G s additive. Also, since $S_{AB} = S_A + S_B$, we have

$$f(G_{AB}) = f(G_A) + f(G_B).$$

Then, from equation (A.16),

$$f(G_A + G_B) = f(G_A) + f(G_B).$$

The question now is: Given that

$$f(x + y) = f(x) + f(y), \tag{A.17}$$

what is the function f? Differentiating equation (A.17) with espect to x and y we have

$$\frac{\partial f(x+y)}{\partial x} = \frac{df(x+y)}{d(x+y)}\frac{\partial(x+y)}{dx} = \frac{df(x+y)}{d(x+y)} = \frac{df(x)}{dx}$$
$$\frac{\partial f(x+y)}{\partial y} = \frac{df(x+y)}{d(x+y)}\frac{\partial(x+y)}{dy} = \frac{df(x+y)}{d(x+y)} = \frac{df(y)}{dy}$$

Hence

$$\frac{df(x)}{dx} = \frac{df(x)}{dy}.$$

This is only possible if the function f is a constant, say k. Then

$$\frac{df(x)}{dx} = k, \;\; f(x) = kx + a,$$

where a is another constant. But we have to choose $a = 0$ in order to satisfy equation (A.17). Therefore, finally, we have found that $f(x) = kx$, and that

$$S \leftrightarrow f(G) = kG = -k\sum_j P_j \ln P_j.$$

Also, from equation (A.12),

$$\frac{1}{\beta T} = \phi(G) = k,$$

or

$$\beta = \frac{1}{kT}.$$

The value of k can be obtained once and for all by comparing statistical-mechanical and experimental values of the same property, on any convenient system. The numerical value of k

depends, of course, on the absolute temperature scale employed. With the conventional kelvin temperature scale, $k = 1.38044 \times 10^{-16}$ erg/K.

Summarizing, the probability that the system is in any particular energy state E_j is

$$P_j(N, V, T) = \frac{e^{-E_j(N,V)/kT}}{Q(N, V, T)}$$

where

$$Q(N, V, T) = \sum_j e^{-E_j(N,V)/kT}.$$

Appendix B

Derivation of scrunching rates

The scrunching rate at position N is the rate of transition from state $P_0(12+N, N, n)$ to state $P_0(13+N, N, min(n, 8))$, where $n \leq 9$. The scrunching rate is defined by

$$s_N = \frac{k_1 C}{C + K_{d_1}^N} \tag{B.1}$$

where

$$K_{d_1}^N = K_C \Big\{ 1 + \sum_{i=2}^{N-1} e^{-\beta(\Delta G^{-1}_{i+12,N,min(i,9)} - \Delta G^0_{N+13,N,min(N,8)})} + e^{-\beta(\Delta G^0_{N+12,N,min(N,9)} - \Delta G^0_{N+13,N,min(N,8)})} \Big\}$$

The reactions involved in this transition are

- the "unscrunching" reactions in the abortive pathway

$$P_0(12+N, N, n) \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} P_{-1}(11+N, N, n^*) \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} P_{-1}(10+N, N, n^{**})$$
$$\underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} \dots \underset{k_{-(N-3)}}{\overset{k_{N-3}}{\rightleftharpoons}} P_{-1}(15, N, 3) \underset{k_{-(N-2)}}{\overset{k_{N-2}}{\rightleftharpoons}} P_{-1}(14, N, 2)$$

where

$$n^* = \begin{cases} n & \text{if } N > 9 \\ n-1 & \text{if } N \leq 9 \end{cases}$$

and

$$n^{**} = \begin{cases} n^* & \text{if } N > 10 \\ n^*-1 & \text{if } N \leq 10 \end{cases}$$

- the scrunching reaction

$$P_0(12+N,N,n) \underset{k_{-(N-1)}}{\overset{k_{N-1}}{\rightleftharpoons}} P_0(13+N,N,min(n,8))$$

- the NTP binding reaction

$$P_0(13+N,N,min(n,8)) + NTP \underset{k_{-N}}{\overset{k_N}{\rightleftharpoons}} P_0(13+N,N,min(n,8)) * NTP$$

- and the polymerization reaction

$$P_0(13+N,N,min(n,8)) * NTP \xrightarrow{k_{N+1}} P_0(13+N,N+1,min(n+1,9))$$

To simplify the notation we will denote a state $P_*(M,N,n)$ by (M,N,n) and the concentration $[P_*(M,N,n)]$ simply by $[M,N,n]$.

The $N+2$ equations that describe the reactions above are given by the system (B.2)

On the system of equations (B.2) we notice the conservation law

$$[13+N,N,min(n,8)] + [(13+N,N,min(n,8)) * NTP] + [13+N,N+1,min(n+1,9)]$$
$$+ [14,N,2] + [15,N,3] + [16,N,4] + \ldots + [11+N,N,n^*] + [12+N,N,n] = const := L$$

Therefore

$$\begin{aligned}[(13+N,N,min(n,8)) * NTP] = L - ([14,N,2] + [15,N,3] + [16,N,4] + \ldots \\ + [11+N,N,n^*] + [12+N,N,n] + [13+N,N,min(n,8)] \\ + [13+N,N+1,min(n+1,9)])\,.\end{aligned}$$

$$
\begin{cases}
\dfrac{d[14,N,2]}{dt} = k_{N-2}[15,N,3] - k_{-(N-2)}[14,N,2] \\
\dfrac{d[15,N,3]}{dt} = k_{N-3}[16,N,4] + k_{-(N-2)}[14,N,2] - (k_{-(N-3)} + k_{N-2})[15,N,3] \\
\quad \vdots \\
\dfrac{d[11+N,N,n^*]}{dt} = k_1[12+N,N,n] + k_{-2}[10+N,N,n^{**}] \\
\qquad\qquad - (k_{-1} + k_2)[11+N,N,n^*] \\
\dfrac{d[12+N,N,n]}{dt} = k_{-1}[11+N,N,n^*] + k_{-(N-1)}[13+N,N,min(n,8)] \\
\qquad\qquad - (k_1 + k_{N-1})[12+N,N,n] \\
\dfrac{d[13+N,N,min(n,8)]}{dt} = k_{N-1}[12+N,N,n] + k_{-N}[(13+N,N,N) * NTP] \\
\qquad\qquad - (k_{-(N-1)} + k_N[NTP])[13+N,N,min(n,8)] \\
\dfrac{d[(13+N,N,min(n,8)) * NTP]}{dt} = k_N[13+N,N,min(n,8)][NTP] \\
\qquad\qquad - (k_{-N} + k_{N+1})[(13+N,N,N) * NTP] \\
\dfrac{d[13+N,N+1,min(n+1,9)]}{dt} = k_{N+1}[(13+N,N,N) * NTP]
\end{cases}
\tag{B.2}
$$

Assuming $k_{N+1} << k_1$ we have $k_{N+1}/k_1 \approx 0$. Dividing all the equations through by k_1 and letting $k_{N+1}/k_1 = 0$ we will get

$$[13+N, N+1, min(n+1,9)] = const := M.$$

Rewriting the third last equation from (B.2) we have

$$
\begin{aligned}
\frac{d[13+N,N,min(n,8)]}{dt} &= k_{N-1}[12+N,N,n] - (k_{-(N-1)} + k_N[NTP])[13+N,N,min(n,8)] \\
&\quad + k_{-N}(L-M) - k_{-N}\left([14,N,2] + [15,N,3] + [16,N,4] + \ldots + [11+N,N,n^*]\right. \\
&\quad \left. + [12+N,N,n] + [13+N,N,min(n,8)]\right) \\
&= k_{-N}(L-M) - k_{-N}[14,N,2] - k_{-N}[15,N,3] - \ldots - k_{-N}[11+N,N,n^*] \\
&\quad + (k_{N-1} - k_{-N})[12+N,N,n] - (k_{-(N-1)} + k_{-N} + k_N[NTP])[13+N,N,min(n,8)]
\end{aligned}
$$

The new system of N equations is

$$\begin{cases} \dfrac{d[14,N,2]}{dt} = k_{N-2}[15,N,3] - k_{-(N-2)}[14,N,2] \\ \dfrac{d[15,N,3]}{dt} = k_{N-3}[16,N,4] + k_{-(N-2)}[14,N,2] - (k_{-(N-3)} + k_{N-2})[15,N,3] \\ \quad \vdots \\ \dfrac{d[11+N,N,n^*]}{dt} = k_1[12+N,N,n] + k_{-2}[10+N,N,n^{**}] \\ \qquad\qquad - (k_{-1}+k_2)[11+N,N,n^*] \\ \dfrac{d[12+N,N,n]}{dt} = k_{-1}[11+N,N,n^*] + k_{-(N-1)}[13+N,N,min(n,8)] \\ \qquad\qquad - (k_1 + k_{N-1})[12+N,N,n] \\ \dfrac{d[13+N,N,min(n,8)]}{dt} = k_{-N}(L-M) - k_{-N}[14,N,2] - k_{-N}[15,N,3] - \dots \\ \qquad\qquad - k_{-N}[11+N,N,n^*] + (k_{N-1} - k_{-N})[12+N,N,n] + \\ \qquad\qquad - (k_{-(N-1)} + k_{-N} + k_N[NTP])[13+N,N,min(n,8)] \end{cases} \tag{B.3}$$

We need to find the equilibrium of this system (B.3). It follows from the first equation that

$$[14,N,2] = \frac{k_{N-2}}{k_{-(N-2)}}[15,N,3].$$

From the second equation of (B.3) we have

$$[16,N,4] = \frac{k_{-(N-3)}}{k_{N-3}}[15,N,3].$$

From the third equation of (B.3) we have

$$\begin{aligned} & k_{N-4}[17,N,5] + k_{-(N-3)}[15,N,3] - (k_{N-4}+k_{N-3})[16,N,4] \\ & \quad = k_{N-4}[17,N,5] + k_{-(N-3)}[15,N,3] - (k_{N-4}+k_{N-3})\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3] \\ & \quad = k_{N-4}[17,N,5] - k_{N-4}\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3] = 0 \end{aligned}$$

and then

$$[17,N,5] = \frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3].$$

From the fourth equation (B.3) we have

$$\begin{aligned}
&k_{N-5}[18,N,6]+k_{-(N-4)}[16,N,4]-(k_{-(N-5)}+k_{N-4})[17,N,5]\\
&\quad=k_{N-5}[18,N,6]+k_{-(N-4)}\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]\\
&\qquad-(k_{-(N-5)}+k_{N-4})\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]\\
&\quad=k_{N-5}[18,N,6]-k_{-(N-5)}\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]\\
&\quad=0
\end{aligned}$$

and then

$$[18,N,6]=\frac{k_{-(N-5)}}{k_{N-5}}\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3].$$

Repeat the process, until from the $(N-2)$th equation of (B.3) we get

$$\begin{aligned}
&k_1[12+N,N,n]+k_{-2}[10+N,N,n^{**}]-(k_{-1}+k_2)[11+N,N,n^*]\\
&\quad=k_1[12+N,N,n]+k_{-2}\frac{k_{-3}}{k_3}\dots\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]-(k_{-1}+k_2)\frac{k_{-2}}{k_2}\dots\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]\\
&\quad=k_1[12+N,N,n]-k_{-1}\frac{k_{-2}}{k_2}\dots\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]\\
&\quad=0
\end{aligned}$$

and then

$$[12+N,N,n]=\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\dots\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3].$$

From the $(N-1)$th equation we get

$$\begin{aligned}
&k_{-(N-1)}[13+N,N,min(n,8)]+k_{-1}[11+N,N,n^*]-(k_1+k_{N-1})[12+N,N,n]\\
&\quad=k_{-(N-1)}[13+N,N,min(n,8)]+k_{-1}\frac{k_{-2}}{k_2}\dots\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]\\
&\qquad-(k_1+k_{N-1})\frac{k_{-1}}{k_1}\dots\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]\\
&\quad=k_{-(N-1)}[13+N,N,min(n,8)]-k_{N-1}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\dots\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3]\\
&\quad=0
\end{aligned}$$

and then

$$[13+N,N,min(n,8)] = \frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}}[15,N,3].$$

From the Nth equation of (B.3) we have

$$\begin{aligned}
&k_{-N}(L-M) - k_{-N}[14,N,2] - k_{-N}[15,N,3] - \ldots - k_{-N}[11+N,N,n^*] \\
&\quad + (k_{N-1}-k_{-N})[12+N,N,n] - (k_{-(N-1)} + k_{-N} \\
&\quad + k_N[NTP])[13+N,N,min(n,8)] \\
&= k_{-N}(L-M) - \Big(k_{-N}\frac{k_{N-2}}{k_{-(N-2)}} + k_{-N} + k_{-N}\frac{k_{-(N-3)}}{k_{N-3}} + k_{-N}\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} \\
&\quad + k_{-N}\frac{k_{-(N-5)}}{k_{N-5}}\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} + \ldots + k_{-N}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}} \\
&\quad - (k_{N-1}-k_{-N})\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}} \\
&\quad + (k_{-(N-1)} + k_{-N} + k_N[NTP])\frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}}\Big)[15,N,3] \\
&= k_{-N}(L-M) - \Big(k_{-N}\Big(1 + \frac{k_{N-2}}{k_{-(N-2)}} + \frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} \\
&\quad + \frac{k_{-(N-5)}}{k_{N-5}}\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} + \ldots + \frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}} \\
&\quad + \frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}}\Big) + k_N\frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}}[NTP]\Big)[15,N,3] \\
&= 0
\end{aligned}$$

and then

$$[15,N,3] = \frac{k_{-N}(L-M)}{D},$$

where

$$\begin{aligned}
D = k_{-N}\Big(&1 + \frac{k_{N-2}}{k_{-(N-2)}} + \frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-(N-5)}}{k_{N-5}}\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} \\
&+ \ldots + \frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}}\Big) \\
&+ k_N\frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\ldots\frac{k_{-(N-3)}}{k_{N-3}}[NTP].
\end{aligned}$$

To simplify notation, let

$$D = k_{-N}A + k_N B[NTP]$$

where

$$A = 1 + \frac{k_{N-2}}{k_{-(N-2)}} + \frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-(N-5)}}{k_{N-5}}\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}}$$
$$+ \ldots + \frac{k_{-2}}{k_2}\cdots\frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\cdots\frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\cdots\frac{k_{-(N-3)}}{k_{N-3}}$$

and

$$B = \frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\cdots\frac{k_{-(N-3)}}{k_{N-3}}.$$

Now we need to go back to the expression for $[(13+N, N, min(n,8)) * NTP]$. We have

$$\begin{aligned}
&[(13+N, N, min(n,8)) * NTP] \\
&\quad = (L-M) - ([14, N, 2] + [15, N, 3] + [16, N, 4] + \ldots + [11+N, N, n^*] \\
&\qquad + [12+N, N, n] + [13+N, N, min(n,8)]) \\
&\quad = (L-M) - \frac{k_{-N}(L-M)}{D}\Big(\frac{k_{N-2}}{k_{-(N-2)}} + 1 + \frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} \\
&\qquad + \frac{k_{-(N-5)}}{k_{N-5}}\frac{k_{-(N-4)}}{k_{N-4}}\frac{k_{-(N-3)}}{k_{N-3}} + \ldots + \frac{k_{-2}}{k_2}\cdots\frac{k_{-(N-3)}}{k_{N-3}} + \frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\cdots\frac{k_{-(N-3)}}{k_{N-3}} \\
&\qquad + \frac{k_{N-1}}{k_{-(N-1)}}\frac{k_{-1}}{k_1}\frac{k_{-2}}{k_2}\cdots\frac{k_{-(N-3)}}{k_{N-3}}\Big) \\
&\quad = (L-M) - \frac{k_{-N}(L-M)}{D}A \\
&\quad = (L-M)(1 - k_{-N}\frac{A}{D}) \\
&\quad = (L-M)(1 - k_{-N}\frac{A}{k_{-N}A + k_N B[NTP]}) \\
&\quad = (L-M)\frac{k_N B[NTP]}{k_{-N}A + k_N B[NTP]} \\
&\quad = (L-M)\frac{[NTP]}{[NTP] + \frac{k_{-N}}{k_N}\frac{A}{B}}
\end{aligned}$$

Therefore

$$\frac{d[13+N, N+1, min(n+1,9)]}{dt} = (L-M)\frac{k_{N+1}[NTP]}{[NTP] + \frac{k_{-N}}{k_N}\frac{A}{B}},$$

where

$$\begin{aligned}
\frac{A}{B} = 1 &+ \frac{k_{-(N-1)}}{k_{N-1}}\frac{k_1}{k_{-1}}\frac{k_2}{k_{-2}}\cdots\frac{k_{N-2}}{k_{-(N-2)}} + \frac{k_{-(N-1)}}{k_{N-1}}\frac{k_1}{k_{-1}}\cdots\frac{k_{N-3}}{k_{-(N-3)}} + \\
&+ \frac{k_{-(N-1)}}{k_{N-1}}\frac{k_1}{k_{-1}}\frac{k_2}{k_{-2}}\cdots\frac{k_{N-4}}{k_{-(N-4)}} + \ldots + \frac{k_{-(N-1)}}{k_{N-1}}\frac{k_1}{k_{-1}} + \frac{k_{-(N-1)}}{k_{N-1}}.
\end{aligned} \tag{B.4}$$

Now notice that

- the polymerization rate k_1 corresponds to k_{N+1}
- the dissociation constant for the next NTP K_C corresponds to $\frac{k_{-N}}{k_N}$
- the term $e^{-\beta(\Delta G^0_{N+12,N,min(N,9)} - \Delta G^0_{N+13,N,min(N,8)})}$ corresponds to $\frac{k_{-(N-1)}}{k_{N-1}}$
- the sum $\sum_{i=2}^{N-1} e^{-\beta(\Delta G^{-1}_{i+12,N,min(i,9)} - \Delta G^0_{N+13,N,min(N,8)})}$ corresponds to

$$\frac{k_{-(N-1)}}{k_{N-1}}\frac{k_1}{k_{-1}}\frac{k_2}{k_{-2}}\cdots\frac{k_{N-2}}{k_{-(N-2)}} + \frac{k_{-(N-1)}}{k_{N-1}}\frac{k_1}{k_{-1}}\cdots\frac{k_{N-3}}{k_{-(N-3)}} + \frac{k_{-(N-1)}}{k_{N-1}}\frac{k_1}{k_{-1}}\frac{k_2}{k_{-2}}\cdots\frac{k_{N-4}}{k_{-(N-4)}} + \ldots + \frac{k_{-(N-1)}}{k_{N-1}}\frac{k_1}{k_{-1}}$$

Therefore (B.1) and (B.4) are equivalent.

Appendix C
Markov Chains

Definition C.0.1 A *stochastic* or *random process* is a family of random variables $\{X(t)|t \in T\}$, where the parameter set T is a subset of the real line $\mathbb{R}$.

Definition C.0.2 A stochastic process $\{X(t)|t \in T\}$ is called a *discrete-time stochastic process* if the parameter set T is a countable set. If T is uncountable, then $\{X(t)|t \in T\}$ is called *continuous-time stochastic process.*

In the case of discrete-time stochastic processes it is common to write $\{X_t|t \in T\}$ instead of $\{X(t)|t \in T\}$.

Definition C.0.3 The set $S_{X(t)}$ of values that the random variables $X(t)$ can take is called **state space** of the stochastic process $\{X(t)|t \in T\}$. If $S_{X(t)}$ is countable, then $\{X(t)|t \in T\}$ is said to be a *discrete-state* process. If $S_{X(t)}$ is uncountable then $\{X(t)|t \in T\}$ is a *continuous-state* process.

Definition C.0.4 A stochastic process $\{X(t); t \in T\}$ is said to be *Markovian*, or to possess the *Markov property* if

$$P[X(t_{n+1}) \in A|X(t) = x_t, t \leq t_n] = P[X(t_{n+1}) \in A|X(t_n) = x_{t_n}] \tag{C.1}$$

for all events A and for all time instants $t_n < t_{n+1}$.

Equation (C.1) means that the probability that the process moves from state x_{t_n}, where it is at time t_n, to a state included in A at time t_{n+1} does not depend on the way the process reached x_{t_n} from x_{t_0}, where t_0 is the initial time.

When $\{X_n|n = 0, 1, \ldots\}$ is a discrete-time and discrete-state stochastic process, the Markov property implies that

$$P[X_{n+1} = j|X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0] = P[X_{n+1} = j|X_n = i]$$

for all states $i_0, \ldots, i_{n-1}, i, j$ and for any time $n \geq 0$. This also implies that

$$P[X_{n+1} = j | X_{n-1} = i, \ldots, X_0 = i_0] = P[X_{n+1} = j | X_{n-1} = i]$$

etc., which means that the transition probabilities between states depend only on the most recent information about the process that is available.

Definition C.0.5 A *Markov chain* is a discrete-time stochastic process that possesses the Markov property.

Definition C.0.6 A stochastic process $\{X_n | n = 0, 1, \ldots\}$ whose state space S_{X_n} is countable is a *stationary* (or *time-homogeneous*) *Markov chain* if

$$P[X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0] = P[X_{n+1} = j | X_n = i] = p_{i,j}$$

for all states $i_0, \ldots, i_{n-1}, i, j$ and for any time $n \geq 0$.

Definition C.0.7 The *one-step transition probability matrix* P of a Markov chain is given by

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & \ldots \end{array} \\ \begin{pmatrix} p_{0,0} & p_{0,1} & p_{0,2} & \ldots \\ p_{1,0} & p_{1,1} & p_{1,2} & \ldots \\ p_{2,0} & p_{2,1} & p_{2,2} & \ldots \\ \vdots & \vdots & & \ddots \end{pmatrix} \end{array}$$

Notice that since the process must be in one and only one state at time $n+1$ and the $p_{i,j}$'s are probabilities we then have

$$\sum_{j=0}^{\infty} p_{i,j} = 1 \quad \text{for all } i.$$

Suppose we have a stationary Markov chain with state space $\{0, 1, 2, \ldots\}$ and transition probability matrix P, and we are interested in the case when the process moves from state i to state j in n steps (or transitions). The probability of moving from state i to state j in n steps will be denoted by

$$p_{i,j}^{(n)} := P[X_{m+n} = j | X_m = i] \quad \text{for } m, n, i, j \geq 0$$

and the matrix of transition probabilities in n steps will be denoted by $P^{(n)}$.

Proposition C.0.8 (Chapman-Kolmogorov equations)

$$p_{i,j}^{(m+n)} = \sum_{k=0}^{\infty} p_{i,k}^{(m)} p_{k,j}^{(n)} \quad \textit{for } m, n, i, j \geq 0. \tag{C.2}$$

In matrix form, the various equations (C.2) are written as

$$P^{(m+n)} = P^{(m)} P^{(n)},$$

which implies

$$P^{(n)} = \underbrace{P^{(1)} P^{(1)} \dots P^{(1)}}_{\text{n times}} = P^n.$$

The probability of moving to state j, from initial state i, for the first time at the nth transition is denoted by

$$\rho_{i,j}^{(n)} := P[X_n = j, X_{n-1} \neq j, \dots, X_1 \neq j | X_0 = i] \quad \text{for } n \geq 1 \text{ and } i, j \geq 0.$$

The probabilities $p_{i,j}^{(n)}$ and $\rho_{i,j}^{(n)}$ are related by

$$p_{i,j}^{(n)} = \sum_{k=1}^{n} \rho_{i,j}^{(k)} p_{i,j}^{(n-k)}.$$

Definition C.0.9 The state j is *accessible* from state i if there exists an $n > 0$ such that $p_{i,j}^{(n)} > 0$. We denote this by $i \to j$.

Definition C.0.10 If state i is accessible from state j, and j is accessible from i, e say that the states i and i *communicate* and we write $i \leftrightarrow j$. In this case we say that i and j are in the same *class*.

Definition C.0.11 Let C be a subset of the state space if a Markov chain. We say that C is a *closed* set if, from any $i \in C$, the process always remains in C, that is,

$$P[X_{n+1} \in C | X_n = i \in C] = 1 \quad \text{for all } i \in C.$$

Definition C.0.12 A Markov chain is said to be *irreducible* if all the states communicate, that is, if the state space contains no closed set apart from the set of all states.

Proposition C.0.13 *If $i \to j$ or $j \to i$ for all pairs of states i and j of the Markov chain $\{X_n, n = 0, 1, \dots\}$, then the chain is irreducible.*

Definition C.0.14 The state i is said to be *recurrent* if

$$f_{i,i} := P[\cup_{n=1}^{\infty}\{X_n = i\}|X_0 = i] = 1.$$

If $f_{i,i} < 1$, we say that i is a *transient* state.

Notice that $f_{i,i}$ is the probability of an eventual return of the process to the initial state i. It is a particular case of

$$f_{i,j} := P[\cup_{n=1}^{\infty}\{X_n = j\}|X_0 = i]$$

which denotes the probability that, starting from state i, the process will eventually visit state j. We can write

$$f_{i,j} = \sum_{n=1}^{\infty} \rho_{i,j}^{(n)}.$$

Proposition C.0.15 *Suppose state j is recurrent and for $k \neq j$ we have $j \to k$. Then*

- *k is recurrent,*
- *$j \leftrightarrow k$,*
- *$f_{j,k} = f_{k,j} = 1$.*

Corollary C.0.16 *The state space S of a Markov chain may be decomposed as*

$$S = T \cup C_1 \cup C_2 \cup \ldots,$$

where T consists of transient states, $C_1, C_2, \ldots$ are closed, disjoint classes of recurrent states, and if $j \in C_\alpha$ then

$$f_{j,k} = \begin{cases} 1 & \text{if } k \in C_\alpha \\ 0 & \text{if } k \notin C_\alpha \end{cases}$$

Furthermore, if we relabel the states so that for $i = 1, 2, /ldots$ states in C_i have consecutive labels, then the transition matrix P can be rewritten as

$$P = \begin{pmatrix} Q & R \\ 0 & H \end{pmatrix}$$

where Q is the restriction of the matrix P to the states corresponding to T, R represent the transitions from states in T to states in $C_1, C_2, \ldots$, and H represents the transitions within the closed states.

Such decomposition is often called canonical decomposition of the state space S.

Proposition C.0.17 *If S is finite, not all states can be transient.*

Given a Markov chain with canonical decomposition $S = T \cup C_1 \cup C_2 \cup \ldots$, we are often interested in the problem of determining the probability that, starting from an element of T, the process will remain indefinitely in T or instead will enter one of the sets C_k, from where it cannot leave (often called absorption problem). For that let us use the following notation. Define

$$\tau = inf\{n \geq 0; X_n \notin T\}$$

to be the exit time of T. There are cases where $P[\tau = \infty | X_0 = i] > 0$, but assume that $P[\tau < \infty | X_0 = i] = 1$ for all i (and then X_τ is the first state hit outside T). Define for $i \in T$ and $k \notin T$

$$u_{i,k} = P[X_\tau = k | X_0 = i].$$

Once the chain leaves T, it will ht one of the closed recurrent classes and hence can never return to T. Thus we can interpret $u_{i,k}$ as the probability that the chain leaves T because of absorption at state k in the closed, recurrent class when the initial state is i. Quantities related to the absorption are easily computed from the $\{u_{i,k}\}$. For example, the probability that absorption takes place at class C_l is easily computed by summing the absorption probabilities corresponding to the states in C_l:

$$u_i(C_l) := P[X_\tau \in C_l | X_0 = i] = \sum_{k \in C_l} u_{i,k}.$$

For $i, j \in T$ and $n \geq 0$ we have

$$p_{i,j}^{(n)} = Q_{i,j}^{(n)},$$

and therefore $\sum_{n=0}^{\infty} Q_{i,j}^{(n)}$ is the expected number of visits to the transient state j starting from transient state i.

For $i \in T$ and $j \notin T$ we can decompose the event $[X_\tau = j]$ according to what happens at the first transition:

$$[X_\tau = j] = \cup_{k \in S} [X_\tau = j, X_1 = k].$$

This gives a recursion for the $u_{i,j}$'s. We have

$$\begin{aligned} u_{i,j} &= P[X_\tau = j | X_0 = i] = \sum_{k \in S} P[X_\tau = j, X_1 = k | X_0 = i] \\ &= \sum_{k \in T} P[X_\tau = j, X_1 = k | X_0 = i] + \sum_{k \notin T} P[X_\tau = j, X_1 = k | X_0 = i] \\ &= A + B. \end{aligned}$$

To analyze B, observe that if $k \notin T$ then the events $[X_\tau = j]$ and $[X_1 = k]$ are disjoint unless $j = k$, so we have $B = p_{i,j}$. For A we have that $\tau \geq 2$, and by conditioning on X_1 and using the Markov property,

$$\begin{aligned} A &= \sum_{k \in T} \sum_{n \geq 2} P[\tau = n, X_n = j, X_1 = k | X_0 = i] \\ &= \sum_{k \in T} \sum_{n \geq 2} P[X_2 \in T, \ldots, X_{n-1} \in T, X_n = j, X_1 = k | X_0 = i] \\ &= \sum_{k \in T} \sum_{n \geq 2} P[X_2 \in T, \ldots, X_{n-1} \in T, X_n = j | X_1 = k, X_0 = i] P[X_1 = k | X_0 = i] \\ &= \sum_{k \in T} \sum_{n \geq 2} p_{i,k} P[X_1 \in T, \ldots, X_{n-2} \in T, X_{n-1} = j | X_0 = k] \\ &= \sum_{k \in T} \sum_{n \geq 2} p_{i,k} P[\tau = n - 1, X_\tau = j | X_0 = k] \\ &= \sum_{k \in T} p_{i,k} P[X_\tau = j | X_0 = k] = \sum_{k \in T} p_{i,k} u_{k,j}. \end{aligned}$$

Since for $i, k \in T$ we have $p_{i,k} = Q_{i,k}$, by combining A and B, we get that

$$u_{i,j} = \sum_{k \in T} Q_{i,k} u_{k,j} + p_{i,j} \quad \text{for } i \in T \text{ and } j \notin T. \tag{C.3}$$

This recursion, of course, merely says that absorption by a recurrent state j can take place in two ways: either absorption is accomplished in one step (with probability $p_{i,j}$), or, if not in one step, then a transition must be made to an intermediate transition state k (probability $Q_{i,k}$) and then from k the chain must be absorbed by state j (probability $u_{k,j}$).

If we set $U = (u_{i,j}, i \in T, j \notin T)$, then in matrix notation (C.3) becomes

$$U = QU + R$$

which is the same as $U - QU = U(I - Q) + R$. If $I - Q$ has an inverse, we get the matrix

solution

$$U = (I = Q)^{-1}R.$$

The matrix $(I - Q)^{-1}$ arises frequently in absorption calculations and is known as the fundamental matrix. When the state space is finite (or when T is finite) $I - Q$ indeed has an inverse, which can be represented as

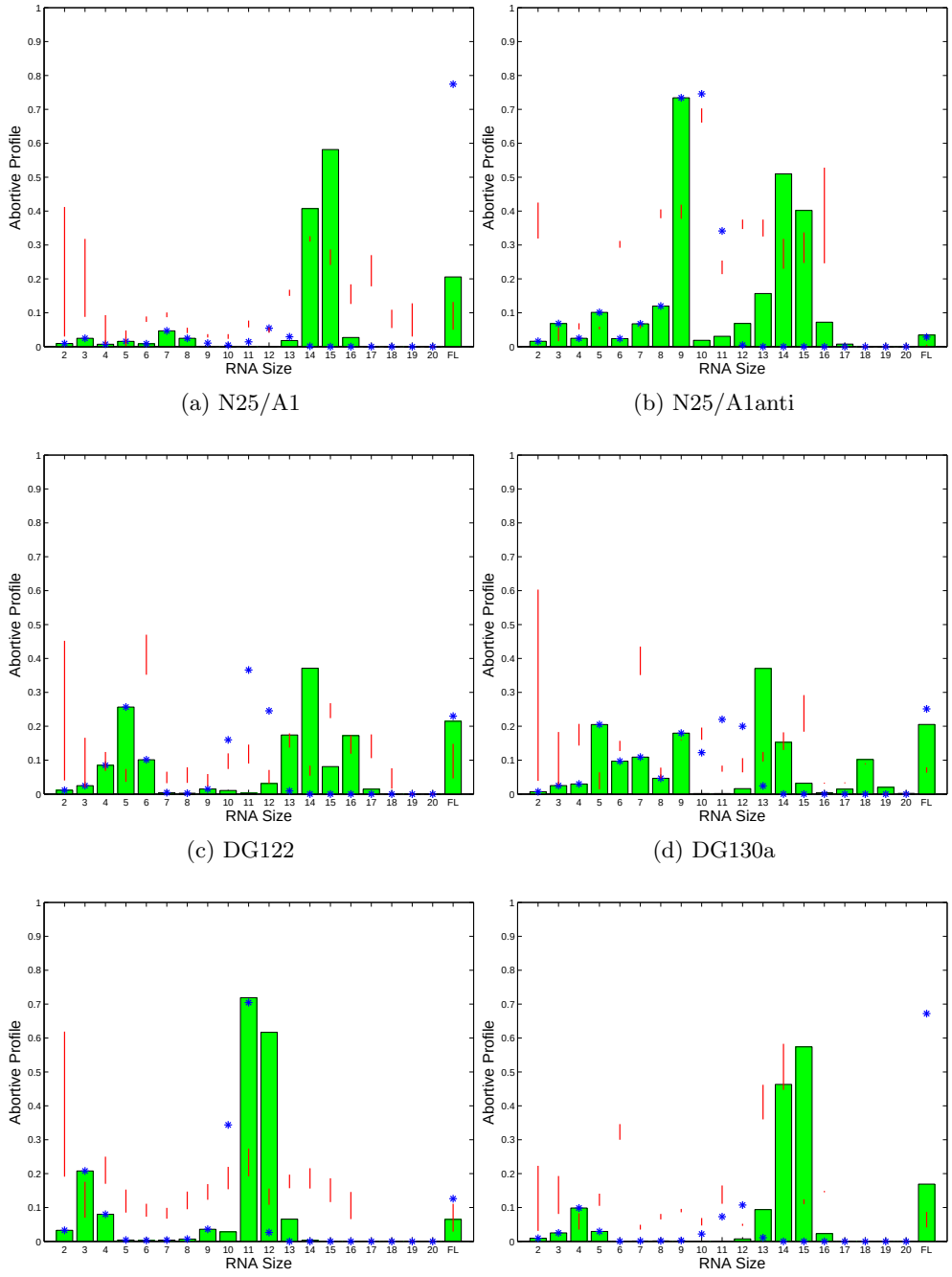
$$(I - Q)^{-1} = \sum_{n=0}^{\infty} Q^n$$

so that

$$U = \sum_{n=0}^{\infty} Q^n R.$$

Appendix D

Additional Figures



(a) N25/A1

(b) N25/A1anti

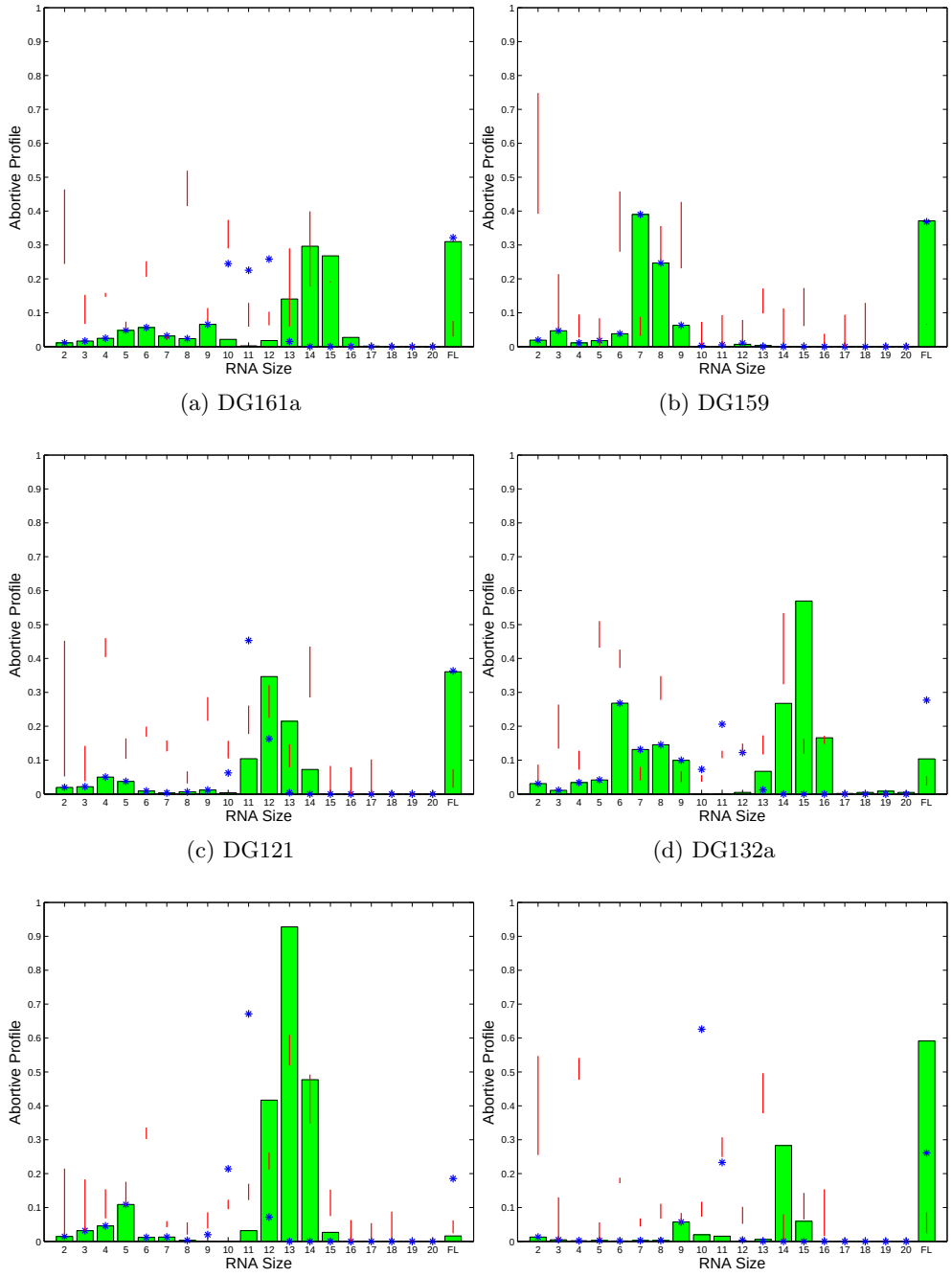
(c) DG122

(d) DG130a

(e) DG131a

(f) DG151a

Figure D.1: Comparison of abortive profiles obtained using secondary structure in the scrunched DNA (bar plot), modified model (blue asterisk) and experimental data (red line).

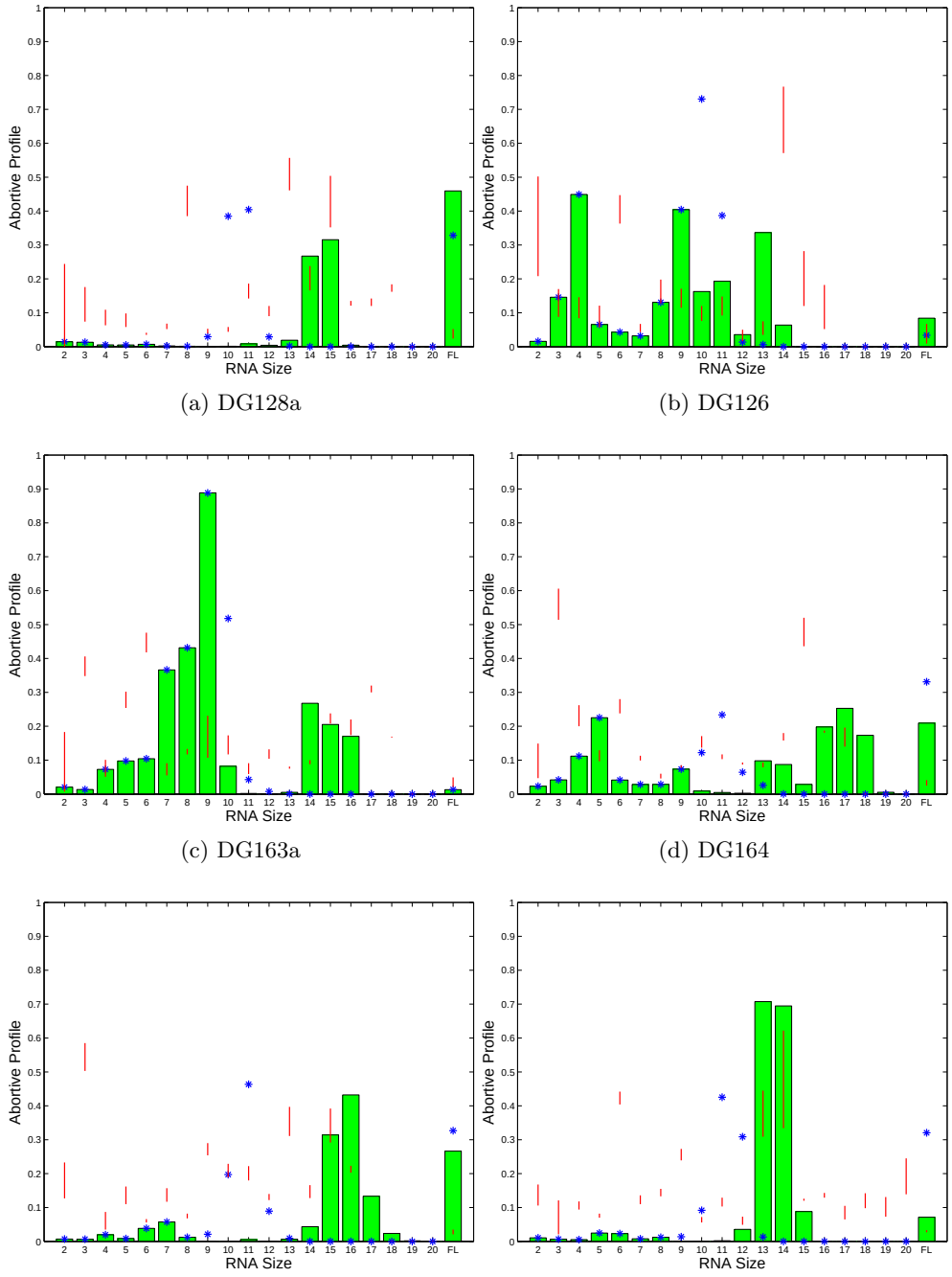


(a) DG161a (b) DG159

(c) DG121 (d) DG132a

(e) DG115a (f) DG162

Figure D.2: Comparison of abortive profiles obtained using secondary structure in the scrunched DNA (bar plot), modified model (blue asterisk) and experimental data (red line).

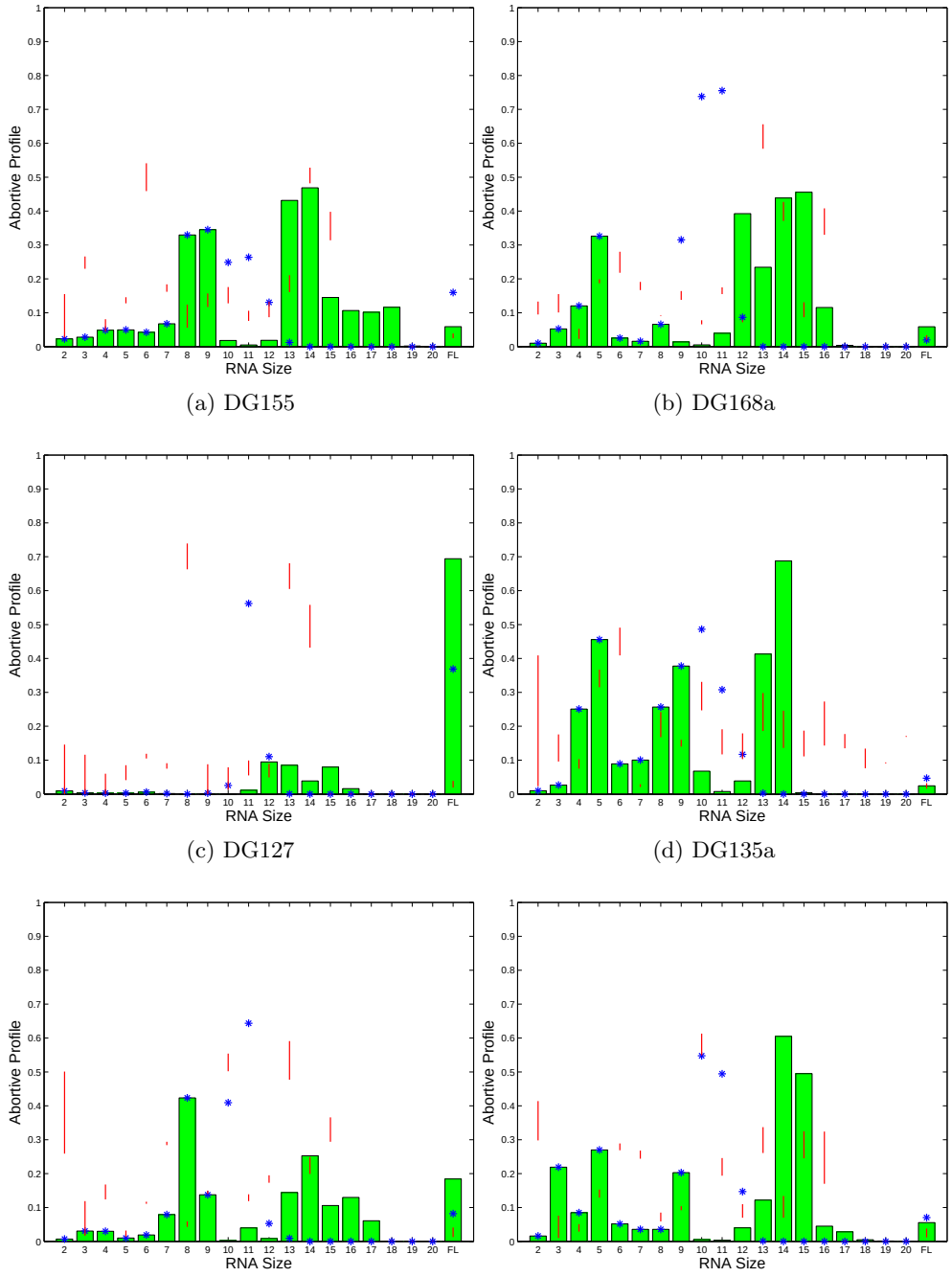


(a) DG128a (b) DG126

(c) DG163a (d) DG164

(e) DG165b (f) DG169a

Figure D.3: Comparison of abortive profiles obtained using secondary structure in the scrunched DNA (bar plot), modified model (blue asterisk) and experimental data (red line).

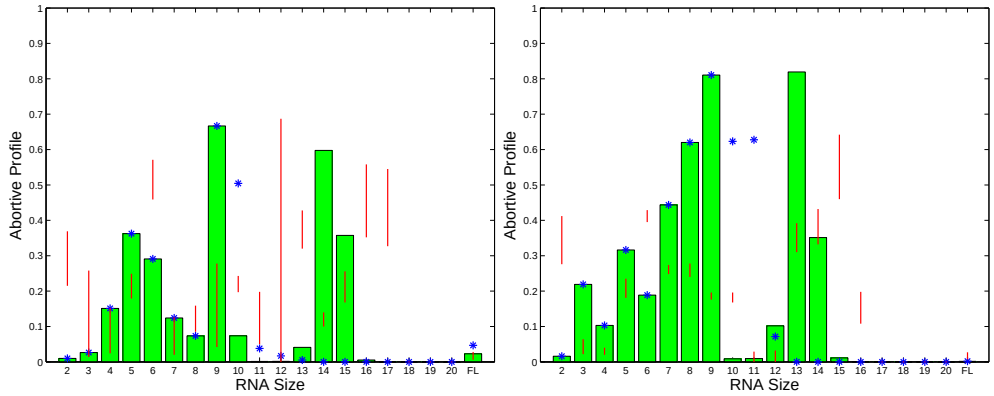


(a) DG155 (b) DG168a

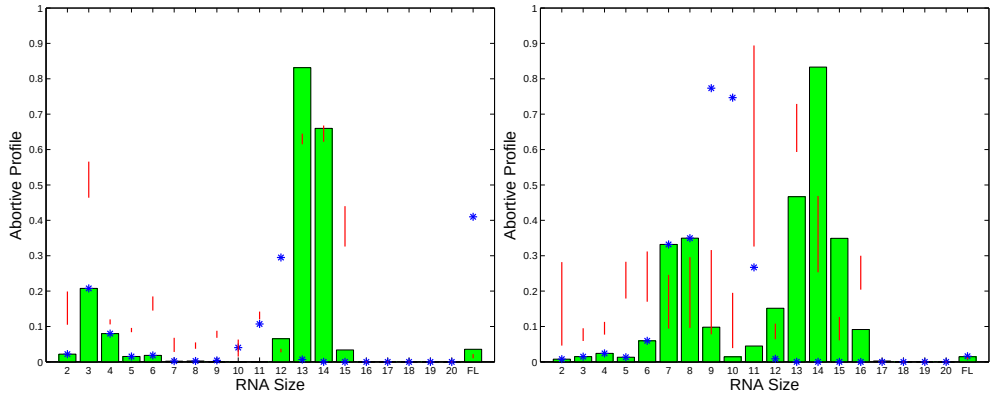
(c) DG127 (d) DG135a

(e) DG160a (f) DG133

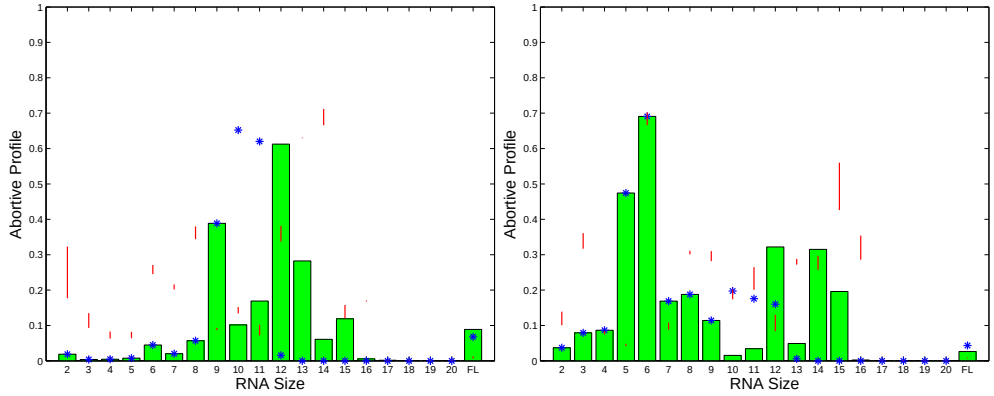
Figure D.4: Comparison of abortive profiles obtained using secondary structure in the scrunched DNA (bar plot), modified model (blue asterisk) and experimental data (red line).



(a) DG147a (b) DG125a

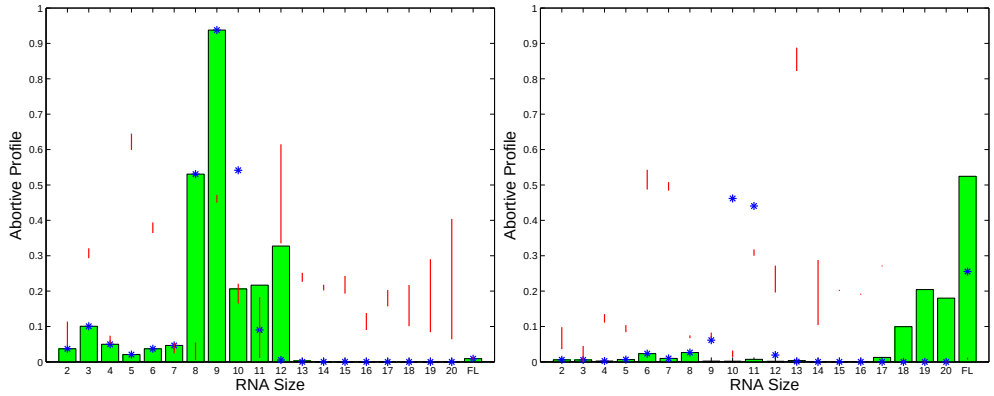


(c) DG156-3 (d) DG138a



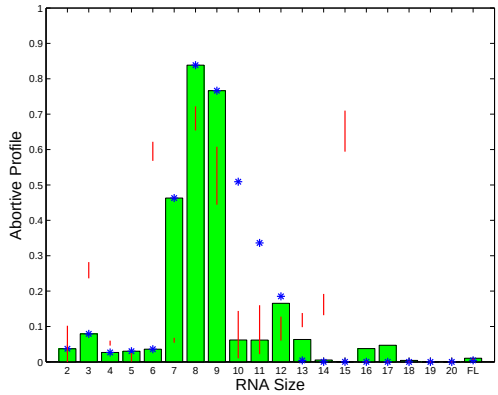
(e) DG142 (f) DG170a

Figure D.5: Comparison of abortive profiles obtained using secondary structure in the scrunched DNA (bar plot), modified model (blue asterisk) and experimental data (red line).

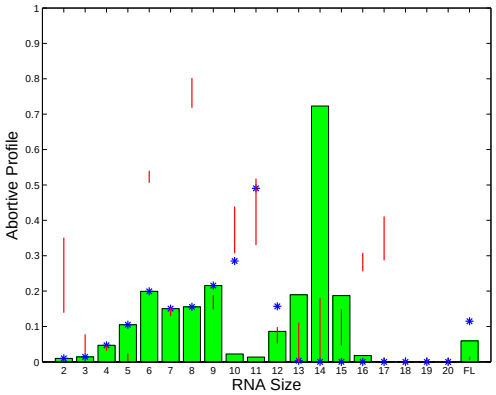


(a) DG167

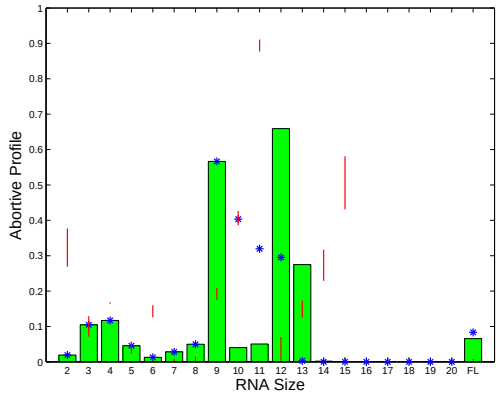
(b) DG166



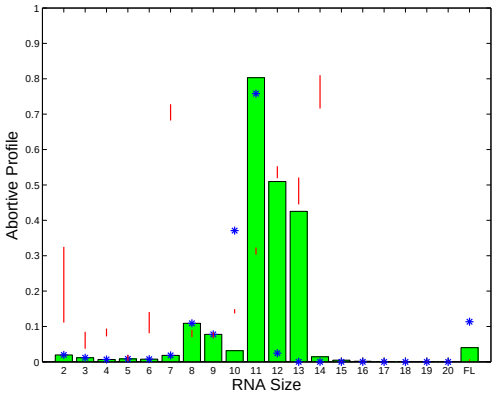
(c) DG152a



(d) DG148

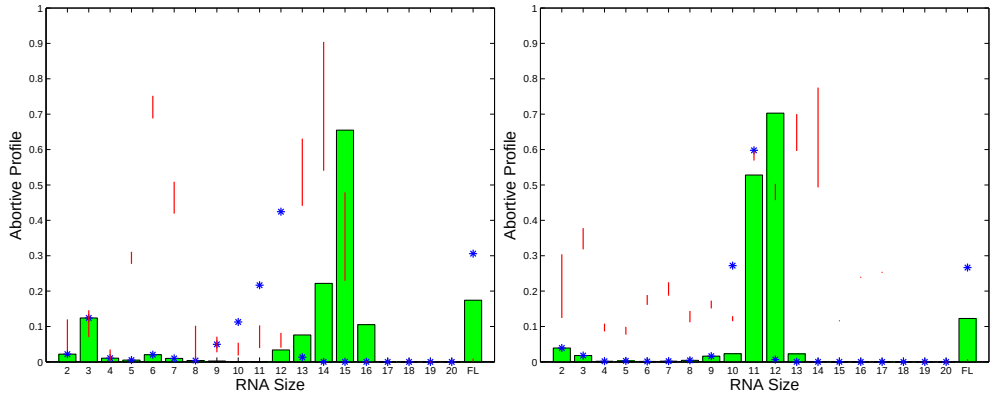


(e) DG145



(f) DG141

Figure D.6: Comparison of abortive profiles obtained using secondary structure in the scrunched DNA (bar plot), modified model (blue asterisk) and experimental data (red line).



(a) DG124a (b) DG154a

Figure D.7: Comparison of abortive profiles obtained using secondary structure in the scrunched DNA (bar plot), modified model (blue asterisk) and experimental data (red line).

References

[1] J. Archambault and J. Friesen, "Genetics of RNA polymerase I, II and III," *Microbiol Rev*, vol. 57, pp. 703–724, September 1993.

[2] L. Minakhin, S. Bhagat, A. Brunning, E. Campbell, S. Darst, R. Ebright, and K. Severinov, "Bacterial RNA polymerase subunit ω and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly," *Proc Natl Acad Sci USA*, vol. 98, pp. 892–897, January 2001.

[3] G. Zhang and S. Darst, "Structure of the *Escherichia coli* RNA polymerase α subunit amino-terminal domain," *Science*, vol. 281, pp. 262–266, July 1998.

[4] G. Zhang, E. Campbell, L. Minakhin, C. Richter, K. Severinov, and S. Darst, "Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 å resolution," *Cell*, vol. 98, pp. 811–824, September 1999.

[5] J. Watson, T. Baker, S. Bell, A. Gann, M. Levine, and R. Losick, *Molecular Biology of the Gene*. Benjamin Cummings, 2008.

[6] R. Ebright, "RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II," *J Mol Biol*, vol. 304, pp. 687–698, December 2000.

[7] S. Borukhov and K. Severinov, "Role of the RNA polymerase sigma subunit in transcription initiation," *Res Microbiol*, vol. 153, pp. 557–562, November 2002.

[8] A. Seshsayee, K. Sivaraman, and N. Luscombe, "An overview of prokaryotic transcription factors: a summary of function and occurrence in bacterial genomes," *Subcell Biochem*, vol. 52, pp. 7–23, 2011.

[9] A. Dombroski, W. Walter, M. Record, Jr, D. Siegele, and C. Gross, "Polypeptides containing highly conserved regions of transcription initiation factor sigma 70 exhibit specificity of binding to promoter DNA," *Cell*, vol. 70, no. 3, pp. 501–512, 1992.

[10] W. Ross, K. Gosink, J. Salomon, K. Igarashi, C. Zhou, A. Ishihama, K. Severinov, and R. Gourse, "A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase," *Science*, vol. 262, pp. 1407–1413, November 1993.

[11] S. Estrem, W. Ross, T. Gaal, Z. Chen, W. Niu, R. Ebright, and R. Gourse, "Bacterial promoter architecture: subsite structure of UP elements and interationc with the carboxy-terminal domain of the RNA polymerase α subunit," *Genes Dev*, vol. 13, pp. 2134–2147, August 1999.

[12] S. Haugen, W. Ross, M. Manrique, and L. Gourse, "Fine structure of the promoter–σ region 1.2 interaction," *Proc Natl Acad Sci USA*, vol. 105, pp. 3292–3297, March 2008.

[13] A. Revyakin, C. Liu, R. Ebright, and T. Strick, "Abortive initiation and productive initiation by RNA polymerase involve DNA-scrunching," *Science*, vol. 314, pp. 1139–1143, November 2006.

[14] A. Kapanidis, E. Margeat, S. Ho, E. Kortkhonjia, S. Weiss, and R. Ebright, "Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism," *Science*, vol. 314, pp. 1144–1147, November 2006.

[15] D. Johnston and W. McClure, "Modulation of chemical composition and other parameters of the cell by growth rate," in *Escherichia coli and Salmonella thyphymurium: Cellular and Molecular Biology* (F. N. et. al, ed.), vol. 2, 1996.

[16] L. Hsu, "Monitoring abortive initiation," *Methods*, vol. 47, pp. 25–36, January 2009.

[17] S. Goldman, R. Ebright, and B. Nickels, "Direct detection of abortive RNA transcripts in vivo," *Science*, vol. 324, pp. 927–928, May 2009.

[18] A. Carpousis and J. Gralla, "Interaction of rna polymerase with lacUV5 promoter dna during mRNA initiation and elongation. Footprint, methylation, and rifampicin-sensitivity changes accompanying transcription initiation," *J Mol Biol*, vol. 183, pp. 165–177, May 1985.

[19] D. Straney and D. Crothers, "A stressed intermediate in the formation of stably initiated RNA chains at the Escherichia coli lac UV5 promoter," *J Mol Biol*, vol. 193, pp. 267–278, January 1987.

[20] B. Krummel and M. Chamberlin, "RNA chain initiation by Eschericia coli RNA polymerase. Structural transitions of the enzyme in early ternary complexes," *Biochemistry*, vol. 28, pp. 7829–7842, September 1989.

[21] G. Bar-Nahum and E. Nudler, "Isolation and characterization of sigma(70)-retaining transcription elongation complexes from *Eschericia coli*," *Cell*, vol. 106, pp. 443–451, August 2001.

[22] J. Mukhopadhyay, A. Kapanidis, V. Mekler, E. Kortkhonjia, Y. Ebright, and R. Ebright, "Translocation of sigma(70) with RNA polymerase during transcription: fluorescent resonance energy transfer assay for movement relative to DNA," *Cell*, vol. 106, pp. 453–463, August 2001.

[23] B. Ring, W. Yarnell, and J. Roberts, "Function of E. coli RNA polymerase sigma factor sigma 70 in promoter proximal pausing," *Cell*, vol. 86, pp. 485–493, August 1996.

[24] R. Mooney and R. Landick, "Tethering sigma70 to RNA polymerase reveals high in vivo activity of sigma factors and sigma70-dependent pausing at promoter-distal locations," *Genes Dev*, vol. 17, pp. 2839–2851, November 2003.

[25] B. Nickels, J. Mukhopadhyay, S. Garrity, R. Ebright, and A. Hochschild, "The sigma 70 subunit of RNA polymerase mediates a promoter-proximal pause at the lac promoter," *Nat Struct Mol Biol*, vol. 11, pp. 544–550, June 2004.

[26] R. Mooney, S. Darst, and R. Landick, "Sigma and RNA polymerase: an on-again, off-again relationship?," *Mol Cell*, vol. 20, pp. 335–345, November 2005.

[27] G. Ackers, A. Johnson, and M. Shea, "Qualitative model for gene regulation by λ phage repressor," *Proc Natl Acad Sci*, vol. 79, pp. 1129–1133, February 1982.

[28] T. Hill, *Introduction to Statistical Thermodynamics*. Addison Wesley, 1960.

[29] T. Gedeon, K. Mischaikow, K. Patterson, and E. Traldi, "When activators repress and repressors activate: A qualitative analysis of the Shea-Ackers model," *Bull Math Biol*, vol. 70, pp. 1660–1683, August 2008.

[30] T. Gedeon, K. Mischaikow, K. Patterson, and E. Traldi, "Binding cooperativity in phage lambda is not sufficient to produce an effective switch," *Biophys J*, vol. 94, pp. 3384–3392, May 2008.

[31] X. Xue, F. Liu, and Z. Ou-Yang, "A kinetic model of transcription initiation by RNA polymerase," *J Mol Biol*, vol. 378, pp. 520–529, May 2008.

[32] C. Lartigue, J. Glass, N. Alperovich, R. Pieper, P. Parmar, C. H. 3rd, H. Smith, and J. Venter, "Genome transplantation in bacteria: changing one species to another," *Science*, vol. 317, pp. 632–638, August 2007.

[33] M. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, pp. 335–338, January 2000.

[34] T. Gardner, C. Cantor, and J. Collins, "Construction of a genetic toggle switch in *Escherichia coli*," *Nature*, vol. 403, pp. 339–342, January 2000.

[35] U. Alon, *An Introduction to Systems Biology*. Chapman & Hall/CRC, 2007.

[36] M. Ptashne and A. Gann, *Genes and Signals*. Cold Spring Harbor Laboratory Press, 2001.

[37] D. Angeli, J. Ferrell, and E. Sontag, "Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems," *Proc Natl Acad Sci*, vol. 101, pp. 1822–1827, February 2004.

[38] M. Shea and G. Ackers, "The O_R control system of bacteriophage lambda : A physical-chemical model for gene regulation," *J Mol Bio*, vol. 181, pp. 211–230, January 1985.

[39] L. Bintu, N. Buchler, H. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, "Transcriptional regulation by the numbers: models," *Curr Opin Genet Dev*, vol. 15, pp. 116–124, April 2005.

[40] M. Santillán and M. Mackey, "Why the lysogenic state of phage λ is so stable: A mathematical modeling approach," *Biophysical Journal*, vol. 86, pp. 75–84, January 2004.

[41] M. Santillán and M. Mackey, "Influence of catabolite repression and inducer exclusion on the bistable behavior of the lac operon," *Biophys J*, vol. 86, pp. 1282–1292, March 2004.

[42] J. Mallet-Paret and H. Smith, "The Poincaré-Bendixson theorem for monotone feedback systems," *J Dynam Diff Eq*, vol. 2, no. 4, pp. 367–421, 1990.

[43] T. Gedeon, *Cyclic feedback systems*, vol. 637 of *Memoirs of AMS*. Americal Mathematical Society, 1998.

[44] M. Ptashne, *A Genetic Switch*. Cold Spring Harbor Laboratory Press, 2004.

[45] L. Snyder and W. Champness, *Molecular Biology of Bacteria*. ASM Press, 2007.

[46] M. Nollman, N. Crisona, and P. Arimondo, "Thirty years of Escherichia coli DNA gyrase: From in vivo function to single-molecule mechanism," *Biochimie*, vol. 89, pp. 490–499, February 2007.

[47] O. Schroder and R. Wagner, "The bacterial DNA-binding protein H-NS represses ribosomal RNA transcription by trapping RNA polymerase in the initiation complex," *J Mol Biol*, vol. 298, pp. 737–748, May 2000.

[48] R. Schneider, A. Travers, T. Kuteleladze, and G. Muskhelishvili, "A DNA architectural protein couples cellular physiology and DNA topology in Escherichia coli," *Mol Microbiol*, vol. 34, pp. 953–964, December 1999.

[49] H. Bremer and P. Dennis, "Modulation of chemical composition and other parameters of the cell by growth rate," in *Escherichia coli and Salmonella thyphymurium: Cellular and Molecular Biology* (F. N. et. al, ed.), vol. 2, pp. 1553–1569, American Society for Microbiology, Washington DC, 1996.

[50] M. Lanzer and H. Bujard, "Promoters largely determine the efficiency of repressor," *Proc Natl Acad Sci*, vol. 85, pp. 8973–8977, 1988.

[51] R. Lutz and H. Bujard, "Independent and tight regulation of transcriptional units in *Escherishia coli* via the LacR/O, the TetR/O and AraC/I_1-I_2 regulatory elements," *Nucleic Acids Research*, vol. 25, no. 6, pp. 1203–1210, 1997.

[52] D. Hawley and W. McClure, "The effect of a lambda repressor mutation on the activation of transcription initiation from the lambda p_{RM} promoter," *Cell*, vol. 32, pp. 327–333, February 1983.

[53] M. Li, W. McClure, and M. Susskind, "Changing the mechanism of transcriptional activation by phage λ repressor," *Proc Natl Acad Sci USA*, vol. 94, pp. 3691–3696, April 1997.

[54] S. Busby and R. Ebright, "Transcription activation by catabolite activator protein (CAP)," *J Mol Biol*, vol. 293, pp. 199–213, October 1999.

[55] S. Roy, S. Garges, and S. Adhya, "Activation and repression of transcription by differential contact: two sides of a coin," *J Biol Chem*, vol. 273, pp. 14059–14062, June 1998.

[56] M. Santillán and M. Mackey, "Dynamic regulation of the tryptophan operon: A modeling study and comparison with experimental data," *Proc Natl Acad Sci USA*, vol. 98, pp. 1364–1369, February 2001.

[57] I. Dodd, K. Shearwin, and J. Egan, "Revisited gene regulation in bacteriophage λ," *Curr Opin Genet Dev*, vol. 15, pp. 145–152, April 2005.

[58] J. Roberts, C. Roberts, and D. Mount, "Inactivation and proteolytic cleavage of phage λ repressor in vitro in an ATP-dependent reaction," *Proc Natl Acad Sci USA*, vol. 74, pp. 2283–2287, June 1977.

[59] A. Bailone, A. Levine, and R. Devoret, "Inactivation of prophage λ repressor in vivo," *J Mol Biol*, vol. 131, pp. 553–572, July 1979.

[60] R. Sauer, M. Ross, and M. Ptashne, "Cleavage of the λ and P22 repressors by RecA protein," *J Biol Chem*, vol. 257, pp. 4458–4462, April 1982.

[61] M. Marr, J. Roberts, S. Brown, M. Klee, and G. Gussin, "Interactions among CII protein, RNA polymerase and the λp_{RE} promoter: contacts between RNA polymerase and the -35 region of p_{RE} are identical in the presence and absence of CII protein," *Nucleic Acid Res*, vol. 32, pp. 1083–1090, February 2004.

[62] P. Darling, J. Holt, and G. Ackers, "Coupled energetics of λ *cro* repressor self-assembly and site-specific DNA operator binding II: Cooperative interactions of *cro* dimers," *J Mol Biol*, vol. 302, pp. 625–638, September 2000.

[63] J. Little, D. Shepley, and D. Wert, "Robustness of a gene regulatory circuit," *EMBO J*, vol. 18, no. 15, pp. 4299–4307, 1999.

[64] C. Michalowski and J. Little, "Positive autoregulation of cI is a dispensable feature of the phage λ gene regulatory circuitry," *J Bacteriol*, vol. 187, pp. 6430–6442, September 2005.

[65] C. Michalowski, M. Short, and J. Little, "Sequence tolerance of the phage λ P_{RM} promoter: Implications for evolution of gene regulatory circuitry," *J Bacteriol*, vol. 186, pp. 7988–7999, December 2004.

[66] L. Bai, A. Shundrovsky, and M. Wang, "Sequence-dependent kinetic model for transcription elongation by RNA polymerase," *J Mol Biol*, vol. 344, pp. 335–349, November 2004.

[67] V. Tadigotla, D. Maoiléidigh, A. Sengupta, V. Epshtein, R. Ebright, E. Nudler, and A. Ruckenstein, "Thermodynamic and kinetic modeling of transcriptional pausing," *Proc Natl Acad Sci USA*, vol. 103, pp. 4439–4444, March 2006.

[68] W. McClure, C. Cech, and D. Johnston, "A steady state assay for the rna polymerase initiation reaction," *J Biol Chem*, vol. 253, pp. 8941–8948, December 1978.

[69] W. Smagowicz and K. Scheit, "A minimal mechanism for abortive initiation of transcription of T7 DNA," *Nucleic Acids Res*, vol. 9, pp. 5845–5854, November 1981.

[70] A. Revyakin, R. Ebright, and T. Strick, "Promoter unwinding and promoter clearance by RNA polymerase: detection by single-molecule DNA nano-manipulation," *Proc Natl Acad Sci USA*, vol. 101, pp. 4776–4780, April 2004.

[71] J. SantaLucia, Jr. and D. Hicks, "The thermodynamics of DNA structural motifs," *Annu Rev Biophys Biomol Struct*, vol. 33, pp. 415–440, 2004.

[72] N. Sugimoto, S. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, T. Ohmichi, M. Yoneyama, and M. Sasaki, "Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes," *Biochemistry*, vol. 34, pp. 11211–11216, September 1995.

[73] L. Hsu, V. Vo, C. Kane, and M. Chamberlin, "In vitro studies of transcript initiation by Escherichia coli RNA polymerase. 1. RNA chain initiation, abortive initiation, and promoter escape at three bacteriophage promoters," *Biochemitry*, vol. 42, pp. 3777–3786, April 2003.

[74] V. Vo, L. Hsu, C. Kane, and M. Chamberlin, "In vitro studies of transcript initiation by Escherichia coli RNA polymerase. 2. Formation and characterization of two distinct classes of initial transcribing complexes," *Biochemistry*, vol. 42, pp. 3787–3797, April 2003.

[75] V. Vo, L. Hsu, C. Kane, and M. Chamberlin, "In vitro studies of transcript initiation by Escherichia coli RNA polymerase. 3. Influences of individual DNA elements within the promoter recognition region on abortive initiation and promoter escape," *Biochemistry*, vol. 42, pp. 3798–3811, April 2003.

[76] L. Hsu, I. Cobb, J. Ozmore, M. Khoo, G. Nahm, L. Xia, Y. Bao, and C. Ahn, "Initial transcribed sequence mutations specifically affect promoter escape properties," *Biochemistry*, vol. 45, pp. 8841–8854, July 2006.

[77] R. Ebright, "Personal communication," 2010.

[78] L. Bai, R. Fullbright, and M. Wang, "Mechanochemical kinetics of transcription elongation," *Phys Rev Lett*, vol. 98, p. 068103, February 2007.

[79] J. SantaLucia, Jr., "A unified view of polymer, dumbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proc Natl Acad Sci USA*, vol. 95, pp. 1460–1465, February 1998.

[80] http://ozone3.chem.wayne.edu/.

[81] L. Hsu, "Quantitative parameters for promoter clearance," *Method Enzymol*, vol. 273, pp. 59–71, 1996.

[82] S. Resnick, *Adventures in Stochastic Processes*. Birkhauser, 1992.

[83] M. Lefebvre, *Applied Stochastic Processes*. Springer, 2000.

[84] M. Zucker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res*, vol. 31, pp. 3406–3415, July 2003.

[85] M. Zucker, "On finding all suboptimal foldings of an RNA molecule," *Science*, vol. 244, pp. 48–52, April 1989.

[86] H. Alawi and J. SantaLucia, Jr., "Thermodynamics and NMR of internal G.T mismatches in DNA," *Biochemistry*, vol. 36, pp. 10581–10594, August 1997.

[87] H. Alawi and J. SantaLucia, Jr., "Nearest-neighbor thermodynamic parameters for internal G.A mismatches in DNA," *Biochemistry*, vol. 37, pp. 2170–2179, February 1998.

[88] H. Alawi and J. SantaLucia, Jr., "Thermodynamics of internal C.T mismatches in DNA," *Nucleic Acids Res*, vol. 26, pp. 2694–2701, June 1998.

[89] H. Alawi and J. SantaLucia, Jr., "Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: Sequence dependence and pH effects," *Biochemistry*, vol. 37, pp. 9435–9444, June 1998.

[90] N. Peyret, P. Senevirtne, H. Allawi, and J. SantaLucia, Jr., "Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches," *Biochemistry*, vol. 38, pp. 3468–3477, March 1999.

[91] N. Peyret, *Prediction of nucleic acid hybridization: parameters and algorithms*. PhD thesis, Wayne State University, 2000.

Curriculum Vitae

Eliane Zerbetto Traldi

Education

2011 Ph. D. in Computational Biology and Molecular Biophysics, Rutgers University

2006 M. S. in Mathematics from Georgia Institute of Technology

1999 M. S. in Mathematics from Universidade de São Paulo, Brazil

1996 B. S. in Mathematics from Universidade de São Paulo, Brazil

Teaching/Academic Experience

Spring 2011 Teaching assistant, Department of Mathematics, Rutgers University

Fall 2006 – Fall 2010 Graduate assistant, BioMaPS Institute, Rutgers University

Fall 2005 – Summer 2006 Graduate research assistant, Department of Mathematics, Georgia Institute of Technology

Fall 2003 – Summer 2005 Teaching assistant, Department of Mathematics, Georgia Institute of Technology

Publications

1. *When activators repress and repressors activate: a qualitative analysis of the Shea-Ackers model*, (with T. Gedeon, K. Mischaikow and K. Patterson), Bull Math Biol. 2008 Aug;70(6):1660-83.

2. *Binding cooperativity in phage lambda is not sufficient to produce an effective switch*, (with T. Gedeon, K. Mischaikow and K. Patterson), Biophys J. 2008 May 1;94(9):3384-92.