Transcriptional Regulation by the Numbers

Thesis by

Hernan G. Garcia

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2011
(Defended December 13, 2010)

To my parents, Miguel Angel and Susana, and with love to Julia

Abstract

Recent decades have seen dramatic advances in our ability to make quantitative measurements of the level of gene expression in organisms of all types. The data resulting from these experiments has raised the need for quantitative models that go beyond the verbal and cartoon-level descriptions that have been so useful in developing a qualitative picture of the nature of gene expression. The improvement in our quantitative description of regulatory networks and our corresponding ability to rewire these networks at will has led many to argue for an analogy between biological regulatory networks and their electronic counterparts. In the electronic setting, we can predict the output current given knowledge of the input voltage and the parameters characterizing the circuit. However, this has so far been nothing more than a hopeful analogy since the input-output functions of most quantitative models of transcriptional regulation are based on phenomenological fits with little-to-no connection to the microscopic parameters of the system. This thesis sharpens this analogy by presenting an integrated approach to understanding transcriptional regulation in bacteria in terms of the microscopic parameters involved in the decision-making processes. This is achieved by a three-pronged approach consisting of theoretical models, *in vivo* measurements and single-molecule experiments *in vitro*.

The theoretical analysis is based upon two different families of models aimed at describing the output of several regulatory architectures as a function of their input parameters. Thermodynamic models of transcriptional regulation are used to predict the mean level of gene expression of several bacterial promoter architectures as a function of the concentration of the intervening regulatory proteins and their binding energies to DNA and to the associated transcriptional machinery. In recent years, however, an increasing body of work has been performed where levels of gene expression are quantified in single cells and sometimes even at the single molecule level. These measurements have revealed that "noise" in gene expression can play a significant role in decision-making processes in systems ranging from bacteria to mammalian cells. Stochastic models of transcriptional regulation predict this variability in gene expression as a function of the microscopic parameter of the system. Unlike thermodynamic models, however, the predictions from stochastic models are dependent on the rate constants describing the regulatory circuit of interest. A complete set of models that predict input-output functions of regulatory systems in bacteria as a function of not only equilibrium parameters, but also probabilities of transition between different regulatory states is presented.

The second half of the thesis complements the theoretical analyses by presenting several experiments aimed at testing the various predictions generated by these models. One of the experiments is carried out *in vivo* and aims to test the theoretical predictions for the input-output function of simple repression in terms of its microscopic parameters such as the concentration of repressor inside the cell and its binding energy to DNA. By quantifying the output level of gene expression as a function of the intracellular absolute concentration of repressor it is shown that our models can account for the level of gene expression as a function of the input parameters over several orders of magnitude. The simple repression motif is also explored experimentally using a second method based upon evaluating fluctuations in the partitioning of regulatory proteins during the cell division process. A third set of experiments performed at the single-molecule level *in vitro* show how a particular repressor protein binds to DNA at two different sites and loops the intervening DNA.

Acknowledgements

First and foremost I want to thank my committee members Michael Elowitz, Scott Fraser, Rob Phillips, David Politzer, Michael Roukes, Shimon Weiss and Jon Widom for taking the time to read these nearly 300 pages and think hard about the science in them. It's been a pleasure to interact with them over the last few years and I look forward to discussing their opinions about the work presented here.

These last seven years have been an amazing adventure. When I signed up for grad school I had no clue that it could be such an exciting experience! The fun that I have experienced is undeniably thanks to the fact that I found an advisor, Rob Phillips, who has also become my good friend. Over these last few years Rob has been an inexhaustible source of insight about how to do science, how to teach (which he views as all part of the same process) and, most importantly, how to bike down a mountain without breaking my collar bone. I think we've both learned a lot along the way and I look forward to learning even more in the future.

Rob's group has always been a very nurturing environment for the discussion of new ideas. This is in part because of the constant flow of external visitors, but mainly because of our rather eclectic combination of group members with different backgrounds and views on how to approach science. I have thoroughly enjoyed my discussions and interactions with all of them: Maja Bialecka, Seth Blumberg, Robert Brewster, James Boedicker, Yi-Ju Chen, Robert Sidney Cox, Paul Grayson, Lin Han, Christoph Haselwandter, Mandar Inamdar, Stephanie Johnson, Daniel Jones, Corinne Ladous (who introduced me to molecular biology!), Heun Jin Lee, Geoff Lovely, Martin Linden, Eric Peterson, Prashant Purohit, Effrosyni (Frosso) Seitaridou, Linda Song, Arbel Tadmor, Tristan Ursell, Dave van Valen, Franz Weinert, Paul Wiggins and Dave Wu.

In particular I'd like to thank Heun Jin Lee and Paul Wiggins. Each of them have been an infinite source of insights into both how to think about the big picture of a problem and how to be detailed and careful in order to come up with the best possible measurement.

Our neighbors in the Elowitz lab have been extremely generous to me. Without the help from people like Michael Elowitz, David Sprinzak and Jonathan Young my entry into the fascinating world of transcriptional regulation would have been a lot less smooth. I particularly appreciate their help when I was getting started with my experiments when I needed help regarding even the most menial things.

Another surprise that grad school had waiting for me was to become part of the book "Physical Biology of the Cell" together with Rob Phillips, Julie Theriot, Jane Kondev and Nigel Orme. This book gave me the opportunity to spend time learning and thinking in new ways about aspects of physical biology that I would have never touched had I just devoted myself to my research projects. More importantly, it has given

me the chance to interact with a passionate group of people which I'm very happy to call my friends and to have wonderful teaching experiences at Woods Hole and Cold Spring Harbor.

Jonathan Widom, with his research and constant advice, has been one of the main driving forces in several of the research projects I have undertaken so far. The project that that got me started in all of this still eludes us, which keeps me enthusiastic about further explorations into the mystery of sequence-dependent flexibility and *in vivo* DNA looping. I am confident that the next months we will get to the bottom of it with the enthusiastic help of James Boedicker.

In the last couple of months I have had the pleasure of interacting with Shimon Weiss and his group. It has given me the chance immerse myself into a new realm of experiments and experimental techniques. I certainly look forward to moving forward on that front in the next months and hopefully, as Shimon puts it, to starting a new field!

Finally, none of this would have been possible without my family and friends. My mom has had to endure having to see her son only once or twice a year. Still, she has never been anything but supportive of the path I've chosen in my life. My girlfriend Julia at this point wants me to finish this thesis more badly than I do; I wouldn't have been able to make it without her nurturing distractions. Finally, I need to thank my friends in the States who have also become part of my family: Nate, Lucia, Eric, Tris, Tristan, Jenn and Jenny. Grad school would not have been this much fun without them.

Contents

List of Figures

List of Tables

Chapter 1

Transcriptional Regulation by the Numbers: An Introduction

One of the great classes of mystery that has dominated biology over the last half century centers on how cells make different decisions and is sometimes referred to as regulatory biology. Examples range from bacteria making decisions to eat one type of sugar instead of another to the cells in an embryo deciding on their different developmental fates. Indeed, biology is full of examples where cells need to assess their environment, communicate with each other and integrate that information in order to make decisions. A big part of this signal integration and decision making process occurs at the level of gene expression. Jacob and Monod's realization that there were genes whose sole role was to regulate other genes was one of the great achievements of the biology of the last century [1]. Their finding that a gene can express a protein that regulates a second gene in a way that is dependent on the sugar content of their environment opened the door to a whole new approach to biology where the networks describing these interactions are mapped with exquisite detail [2].

In particular, the discovery of Jacob and Monod and subsequent workers resulted in the articulation of a simple model in which each gene is preceded by a regulatory region known as the promoter. This promoter serves as the binding site for RNA polymerase, the molecular machine that takes the genomic information in DNA and turns it into mRNA which is in turn converted into new proteins by the ribosome. This promoter region of the DNA is subject to control through two classes of regulatory protein known as repressors and activators which can bind the DNA and alter the probability of RNA polymerase binding.

Much effort has been put into mapping these regulatory networks. For example, we have an almost complete description of the interconnectivity between genes in *E. coli* [3]. We know to a reasonably high accuracy which gene regulates which other gene and if it does so in a positive (activator) or negative (repressor) fashion as exemplified by the graphical representation of the *E. coli* transcriptional network in figure 1.1. In a significant number of the cases the binding sites for the transcription factors in the vicinity of their target promoters have been mapped. Even though our current understanding of transcriptional regulation in higher organisms is much more limited than the bacterial case, significant progress has been made in, for

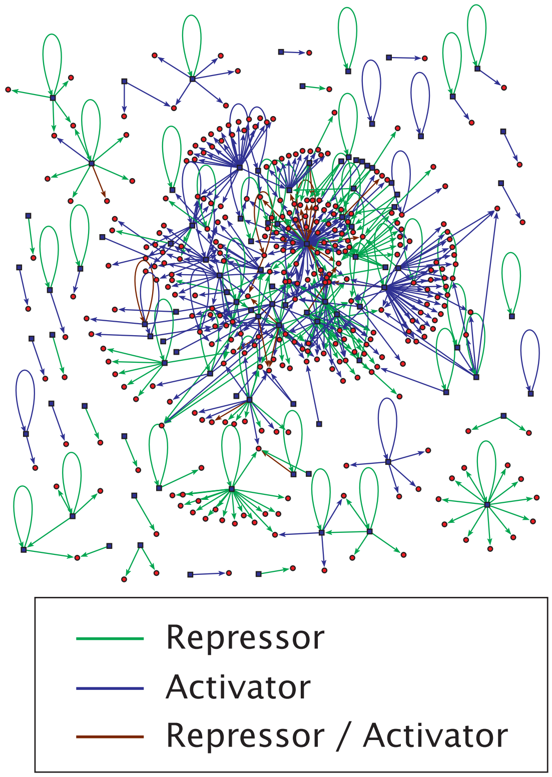


Figure 1.1: *E. coli* transcriptional network. Different genes in *E. coli* are shown together with their connections as activators, repressors or as being dually regulated. (Adapted from [8])

example, dissecting developmental networks. One of the most complete and fascinating examples is that of endomesoderm specification in the sea urchin *S. purpuratus* [4]. Figure 1.2 shows one of the latest versions of the understanding of this decision network. Here, the hierarchical organization of developmental decisions can be seen to span both space and time. Within such regulatory networks recurring modules defined as particular patterns of interconnectivity between the components of the network have been identified and characterized with respect to their functions [5–7]. In fact, a large portion of the community's effort has been concentrated in mapping new elements of networks and identifying new connection patterns [2].

The arrows in figures 1.1 and 1.2 indicate if a gene activates or represses a certain target gene. They are a necessary level of coarse graining in the description of transcriptional networks. However, the fact that these arrows have different quantitative meanings is often overlooked [9, 10]. This concept is exemplified in figure 1.3, where we show that the same regulatory motifs, namely activation and repression, can be realized in a variety of ways. For example, activation can be exerted by a single CRP molecule, by two independent CRP molecules that interact with RNA polymerase but not with each other or by lambda repressor, which multimerizes on the DNA in order to trigger activation. Repression can also be realized in a variety of different ways. The examples shown in figure 1.3 illustrate simple repression, a motif in which a single repressor binds in the vicinity of the promoter, DNA looping, a motif in which a transcription factor binds at two sites simultaneously and loops the intervening DNA and of multiple repressors binding near

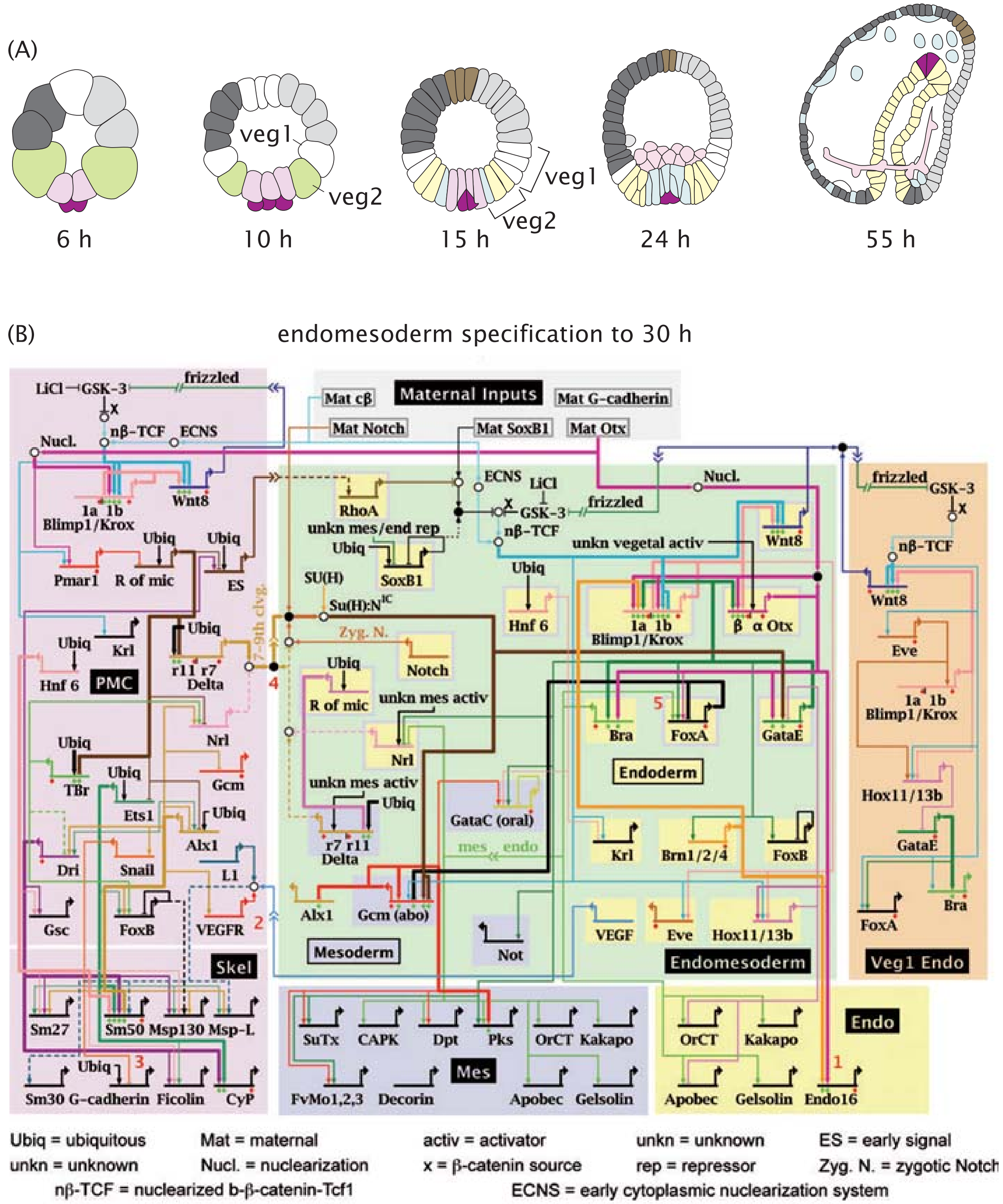


Figure 1.2: Genetic network associated with control of the developmental pathway of the sea urchin embryo. (A) Schematic of stages in the embryonic development of the sea urchin. (B) Genetic network associated with sea urchin development. (Adapted from [4])

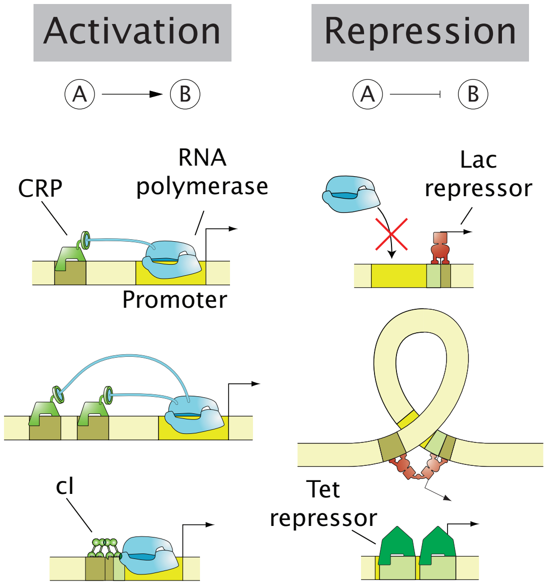


Figure 1.3: Arrows in network diagrams can have multiple realizations. Different realizations of the same network motif are shown for activation and repression. Though their qualitative features might be similar there will be quantifiable differences in their regulation functions.

the promoter.

Even within one realization of the repression motif, for example, simple repression, there is a huge richness in the DNA architecture involved. This richness is exemplified in figure 1.4. Figure 1.4(A) shows the number of different promoters that are regulated by the same repressor through this regulatory motif. Each regulated promoter has potentially a different binding sequence for the repressor and, hence, a different binding energy. The positioning of the binding is also a parameter that is widely exploited in *E. coli* as shown by figure 1.4(B). Finally, figure 1.4(C) shows that these repressors are present at a wide range of concentrations within the cell. It is interesting to note, however, that the concentrations used for figure 1.4(C) were obtained from a mass spectroscopy study [11] which is more likely to detect high concentration proteins accurately. As such figure 1.4(C) should not be viewed as a complete survey, but as a call to attention to the fact that the census of *E. coli* is still incomplete.

The detailed survey of regulatory architectures shown for the case of simple repression in figure 1.4 is just one example of the level of understanding that biologists have attained regarding regulatory architectures in bacteria. Together with the emergence of synthetic biology, where novel architectures can be created *de novo*, this has in part resulted in a view that we have a complete knowledge of how these regulatory circuits operate. There are many examples where analogies between these regulatory circuits and electronic circuits have been drawn [12, 13]. These hopeful analogies claim that our understanding of elements that make genetic circuits is comparable to our knowledge regarding electronic circuits. In electronic circuits our knowledge of the different elements that make them (resistors, transistors, etc.) allow us to predict the output voltage as a function of the input voltage by just looking at a map of the circuit.

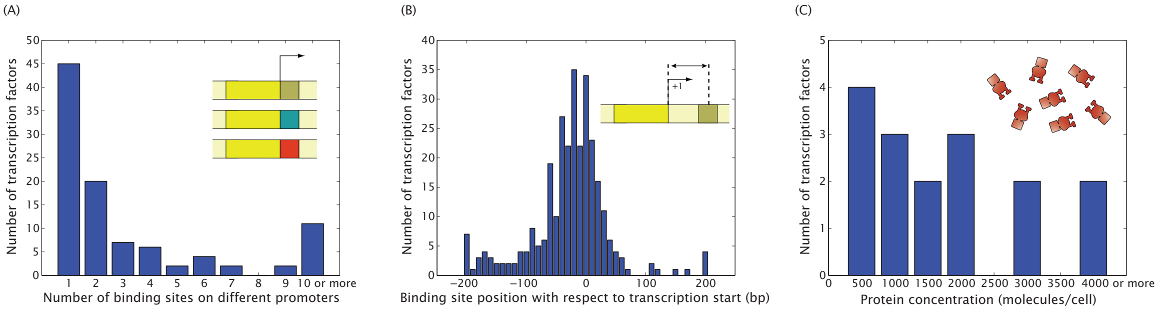


Figure 1.4: Different realizations of the simple repression motif in *E. coli.* (A) Histogram showing how many different promoters are regulated by the same repressor. (B) Histogram of the positions of simple repressor binding sites with respect to the transcription start. (C) Histogram of the concentrations at which various simple repressors are found in the cell. (A and B obtained from RegulonDB [3], C obtained from [11])

However, our knowledge about gene regulatory input-output functions is far from complete. It is certainly far from being anywhere comparable to our knowledge of their electronic counterparts. The most common description of regulatory response is usually cast using the phenomenological Hill function. For example, in the case of of repression it has the form

$$\text{gene expression level} = \frac{\alpha}{1 + \left([R]/K_d\right)^n} + \beta, \tag{1.1}$$

where n is the Hill coefficient which determines the sensitivity of the gene regulatory function, K_d is a dissociation constant and α and β are constants that determine the maximum and basal levels of expression. This description often provides a satisfactory fit to the data, but it often lacks a direct connection to the relevant microscopic parameters. For example, it is not uncommon to obtain non-integer values of n which are difficult to reconcile with simple models of transcription factor binding to the DNA. As a result it is a phenomenological fit and provides limited predictive power.

This thesis is in part an effort to circumvent the limitations of a description of gene regulatory circuits in terms of Hill functions such as the one shown in equation 1.1. The main challenge posed here is the recapitulation of the quantitative behavior of simple regulatory circuits in terms of their fundamental microscopic parameters. These efforts have been divided in a three-pronged approach: theoretical modeling of regulatory networks, reconstitution and characterization in an *in vitro* context and *in vivo* experimentation through the synthesis of regulatory motifs where the relevant parameters had been changed systematically. All three approaches are united by our search for relevant "knobs", parameters of the system that can be controlled both experimentally and theoretically. Figure 1.5 shows an example of the knobs we have identified for the regulatory architecture of simple repression. They can all be thought off as different inputs in the simple repression input-output functions. Though the results of our theoretical and experimental approaches can be clearly separated in sections as in this thesis, the experiments cannot be conceived without the theoretical background and the theoretical models were developed with constant attention to predictions about quantities that are actually measurable.

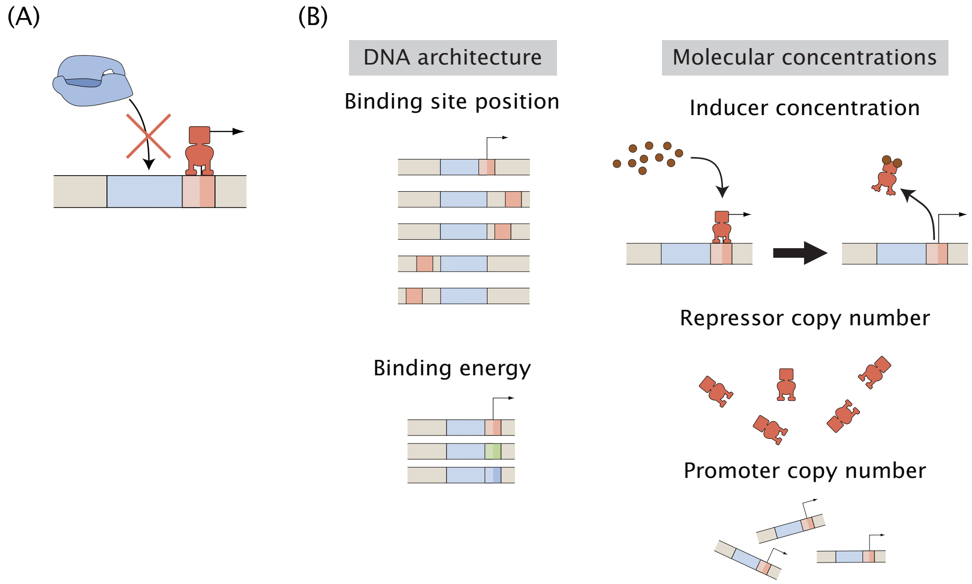


Figure 1.5: Knobs for the simple repression regulatory motif. (A) In simple repression a repressor binds to a site in the vicinity of the promoter and excludes RNA polymerase binding. (B) The different "knobs" or parameters that can be varied in this regulatory architecture are, on the DNA architecture front, the relative position of the binding site with respect to the promoter and the binding sequence of the repressor operator. On the front of molecular concentrations the concentrations of inducer, the concentration of repressor and the number of promoter copies within the cell can be controlled.

The thesis is divided in two parts. The first part describes our theoretical efforts. Here we dissect the mean response and variation of regulatory architectures in terms of microscopic parameters such as binding energies to DNA, interaction energies between molecular players and DNA mechanics. The second part describes our *in vitro* and *in vivo* approaches to testing those theoretical models developed in the first part. We show that approaching these problems through the prism of a theoretical background has allowed us to dissect some regulatory architectures and to contrast the predictions of our theoretical models with an unprecedented accuracy and precision.

1.1 A Roadmap to Part I: Theoretical Models of Transcriptional Regulation

Part I describes two approaches to dissecting transcriptional networks: thermodynamic models and stochastic models. Thermodynamic models have a rich history as tools to describe the mean transcriptional response in bacteria [14–16]. They predict the relative change in mean expression levels as a function of microscopic parameters such as the concentration of molecular players and their interaction energies with DNA and between each other. While incredibly useful and not fully explored experimentally, thermodynamic models only give us information about mean levels and binding energies (or, equivalently, equilibrium constants).

On the other hand, recent years have witnessed an explosion of new and varied experimental techniques that allow us to query the level of gene expression of single cells [17–19]. Being able to measure higher moments of the distribution of gene expression levels rather than only their mean value opens the door to dissecting regulatory networks in terms of rates of transitions between states. As such they are a further way of querying the microscopic properties of cells.

Chapter 2 is an introduction to thermodynamic models of transcriptional regulation. Here we conceive transcriptional regulation in bacteria as a competition for the real estate in the vicinity of the promoter. This competition results from repressors binding to DNA to exclude RNA polymerase from binding to it combined with the action of activators that can recruit RNA polymerase to the promoter. This approach is based on enumerating the possible states a promoter can be in and assigning a weight to each one as exemplified in figure 1.6(A). This chapter culminates with an expression for the fold-change in gene expression for several regulatory architectures. This is shown diagrammatically for the particular example of simple repression in figure 1.6(B). These expression are shown in table 1.1 and are a reference guide for researchers trying to model the response of the architectures addressed or arbitrary combinations of them. These results are described in detail in chapter 2 of this thesis.

Chapter 3 is a systematic exploration of the models developed in chapter 2. Here we dissect several regulatory motifs that have been previously experimentally characterized and generate predictions for their respective input-output functions. In essence, we use the regulatory architectures addressed in table 1.1 and use them to dissect regulatory circuits that have been investigated throughout the literature. The work described in this chapter for the first time dissected the quantitative response of eight regulatory motifs in terms of thermodynamic models of transcriptional regulation. Some of these predictions are shown as a collage in figure 1.7 while the results are described in detail in chapter 3 of the thesis. In general, these predictions consist in calculating the fold-change in gene expression as a function of the concentration of transcription factors and the relevant binding energies. Interestingly, not too many experiments had been done to date that allowed for a direct contrast of the theoretical models. This led us to the development of our experiment on the simple repression motif addressed in chapter 7. Both chapters 2 and 3 are the result of a very fruitful collaboration between the Phillips group, the Kondev group from Brandeis University and the Hwa group from UCSD. This led to the publication of two articles in [20, 21].

Our fascination with the connection between *in vivo* and *in vitro* DNA mechanics [22] led us to spend a considerable amount of effort in understanding repression by DNA looping in the context of the *lac* operon. Lac repressor has two binding heads which can bind to any of the three sites available in the wild-type *lac* operon, looping the intervening DNA and leading to repression. A regulatory architecture that exploits DNA mechanics is in itself extremely interesting as this is a strategy employed by both prokaryotes and eukaryotes alike [22]. Additionally, it is a beautiful opportunity to obtain information about the *in vivo* mechanical properties of the DNA.

In chapter 4 we present this study where we shed light on the role of DNA looping in bacteria through the quantitative dissection of this regulatory motif. DNA looping has been proposed to play a key role in

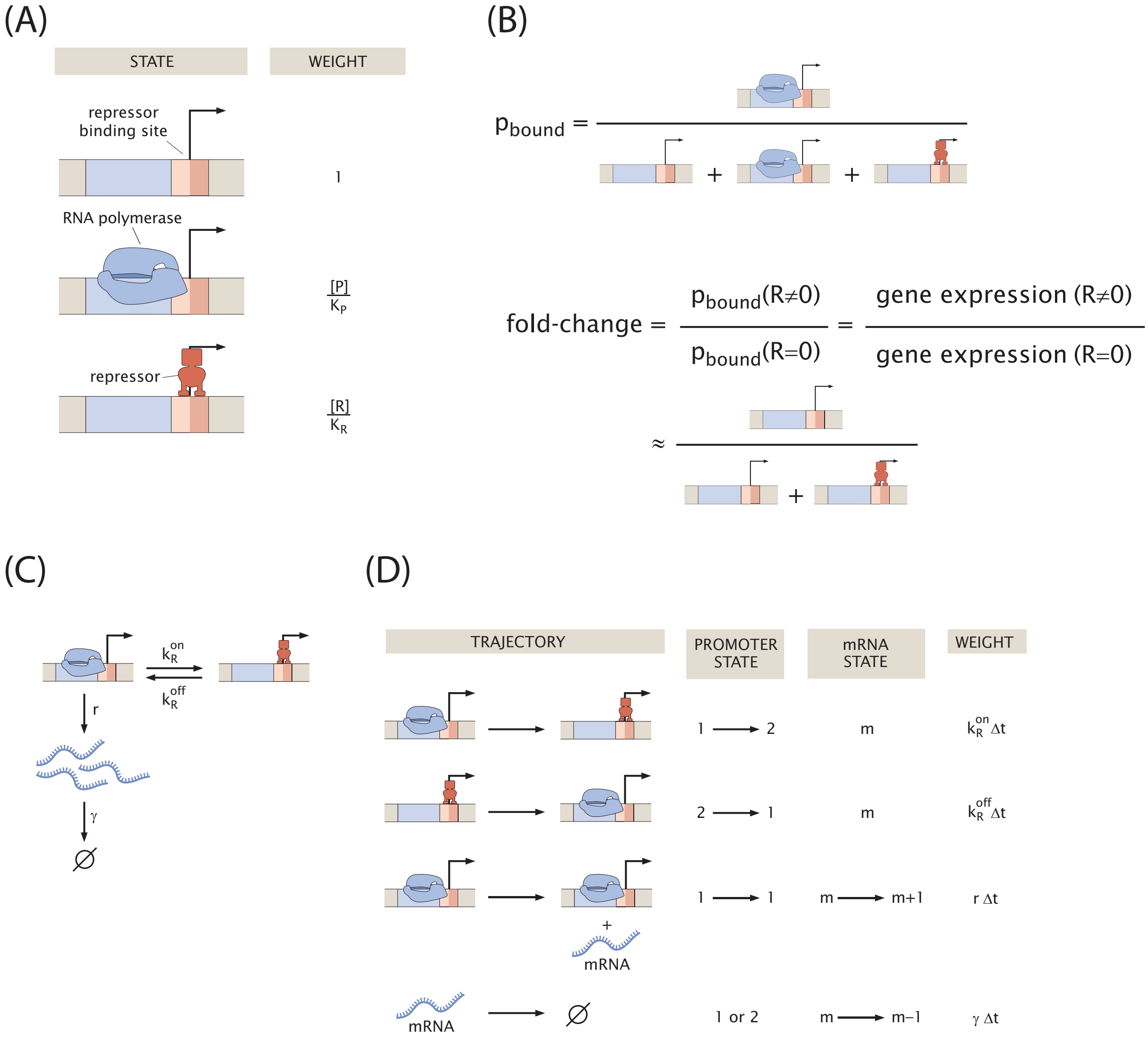


Figure 1.6: Thermodynamic and stochastic description of transcriptional regulation. (A) In the thermodynamic model approach the possible states of the system are enumerated and their corresponding weights assigned. In this case of simple repression the promoter can be either occupied by RNA polymerase or Lac repressor, but not by both simultaneously. (B) By assuming linearity between the promoter occupancy and the resulting level of gene expression the fold-change in gene expression can be calculated as the relative change in promoter occupancy in the presence and absence of the transcription factor. (C) An alternative view of transcriptional regulation is based on rate equations rather than state probabilities. In this simple model there is a rate of promoter switching between states (k_R^{on} and k_R^{off}), a mRNA production rate (r) and a degradation rate (γ). (D) In analogy to the states and weight from (A) we can understand this stochastic models as different trajectories of the system each with their own probability or weight.

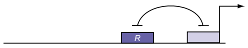
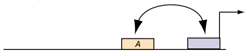
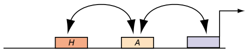
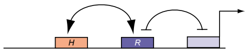
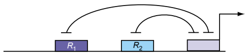
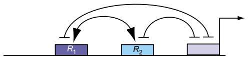
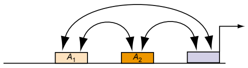
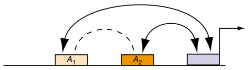
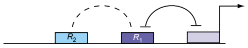
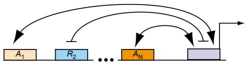
CASE	FOLD-CHANGE IN GENE EXPRESSION	
1. Simple repressor 	$(1+r)^{-1}$	$\left(1+\frac{[R]}{K_R}\right)^{-1}$
2. Simple activator 	$\frac{\left(1+a\,e^{-\beta\varepsilon_{ap}}\right)}{1+a}$	$\frac{\left(1+\frac{[A]}{K_A}\right)f}{1+\frac{[A]}{K_A}}$
3. Activator recruited by a helper (H) 	$\frac{1+a\frac{\left(1+h\,e^{-\beta\varepsilon_{ha}}\right)}{1+h}e^{-\beta\varepsilon_{ap}}}{1+a\frac{\left(1+h\,e^{-\beta\varepsilon_{ha}}\right)}{1+h}}$	$\frac{1+\frac{[H]}{K_H}+\frac{[A]}{K_A}f+\frac{[A]}{K_A}\frac{[H]}{K_H}f\,\omega}{1+\frac{[H]}{K_H}+\frac{[A]}{K_A}+\frac{[A]}{K_A}\frac{[H]}{K_H}\omega}$
4. Repressor recruited by a helper (H) 	$\left(1+\frac{1+h\,e^{-\beta\varepsilon_{hr}}}{1+h}r\right)^{-1}$	$\frac{1+\frac{[H]}{K_H}}{1+\frac{[H]}{K_H}+\frac{[R]}{K_R}+\frac{[R]}{K_R}\frac{[H]}{K_H}\,\omega}$
5. Dual repressors 	$(1+r_1)^{-1}(1+r_2)^{-1}$	$\left(1+\frac{[R_1]}{K_{R_1}}\right)^{-1}\left(1+\frac{[R_2]}{K_{R_2}}\right)^{-1}$
6. Dual repressors interacting 	$\left(1+r_1+r_2+r_1r_2e^{-\beta\varepsilon_{r_1r_2}}\right)^{-1}$	$\left(1+\frac{[R_1]}{K_{R_1}}+\frac{[R_2]}{K_{R_2}}+\frac{[R_1]}{K_{R_1}}\frac{[R_2]}{K_{R_2}}\,\omega\right)^{-1}$
7. Dual activators interacting 	$\frac{\left(1+a_1e^{-\beta\varepsilon_{a_1p}}+a_2e^{-\beta\varepsilon_{a_2p}}+a_1a_2e^{-\beta\left(\varepsilon_{a_1p}+\varepsilon_{a_2p}+\varepsilon_{a_1a_2}\right)}\right)}{1+a_1+a_2+a_1a_2e^{-\beta\left(\varepsilon_{a_1p}+\varepsilon_{a_2p}\right)}}$	$\frac{1+\frac{[A_1]}{K_{A_1}}f_1+\frac{[A_2]}{K_{A_2}}f_2+\frac{[A_1]}{K_{A_1}}\frac{[A_2]}{K_{A_2}}\,f_1f_2\omega}{1+\frac{[A_1]}{K_{A_1}}+\frac{[A_2]}{K_{A_2}}+\frac{[A_1]}{K_{A_1}}\frac{[A_2]}{K_{A_2}}\,\omega}$
8. Dual activators cooperating via looping 	$\frac{1+a_1e^{-\beta\varepsilon_{a_1p}}+a_2e^{-\beta\varepsilon_{a_2p}}+a_1a_2e^{-\beta\left(\varepsilon_{a_1p}+\varepsilon_{a_2p}+\varepsilon_{a_1a_2}\right)}}{1+a_1+a_2+a_1a_2e^{-\beta\left(\varepsilon_{a_1p}+\varepsilon_{a_2p}\right)}}$	$\frac{1+\frac{[A_1]}{K_{A_1}}f_1+\frac{[A_2]}{K_{A_2}}f_2+\frac{[A_1]}{K_{A_1}}+\frac{[A_2]}{K_{A_2}}f_1f_2\omega}{\left(1+\frac{[A_2]}{K_{A_2}}\right)\left(1+\frac{[A_1]}{K_{A_1}}\right)}$
9. Repressor with two DNA binding units and DNA looping 	$\left(1+r_m+\frac{r_m}{1+r_a}e^{-\beta\left(\varepsilon_{r_ad}+F_{loop}\right)}\right)^{-1}$	$\frac{1+\frac{[R]}{K_a}}{\left(1+\frac{[R]}{K_m}\right)\left(1+\frac{[R]}{K_a}\right)+\frac{[R][L]}{K_mK_a}}$
10. N non-overlapping activators and/or repressors acting independently on RNAP 	$F_{reg1}\times F_{reg2}\times\ldots\times F_{reg3}$	$F_{reg1}\times F_{reg2}\times\ldots\times F_{reg3}$

Table 1.1: Thermodynamic model predictions for several regulatory architectures. The thermodynamic models presented in chapter 2 lead to a prediction for the fold-change in gene expression as a function of key microscopic parameters for several regulatory architectures. The regulation factor is shown using a statistical mechanics notation (left) and a biochemical notation (right). In the statistical mechanics notation the lowercase letters x are defined as $x = X/N_{NS}\exp(-\beta\Delta\varepsilon_{xd})$, where X is the intracellular number of the transcription factor x is describing, N_{NS} is the number of non-specific sites and $\Delta\varepsilon_{xd}$ corresponds to the interaction energy of the transcription factor with the DNA. Interaction energies between proteins X and Y are denoted ε_{xy}. Please, see chapter 2 for more details.

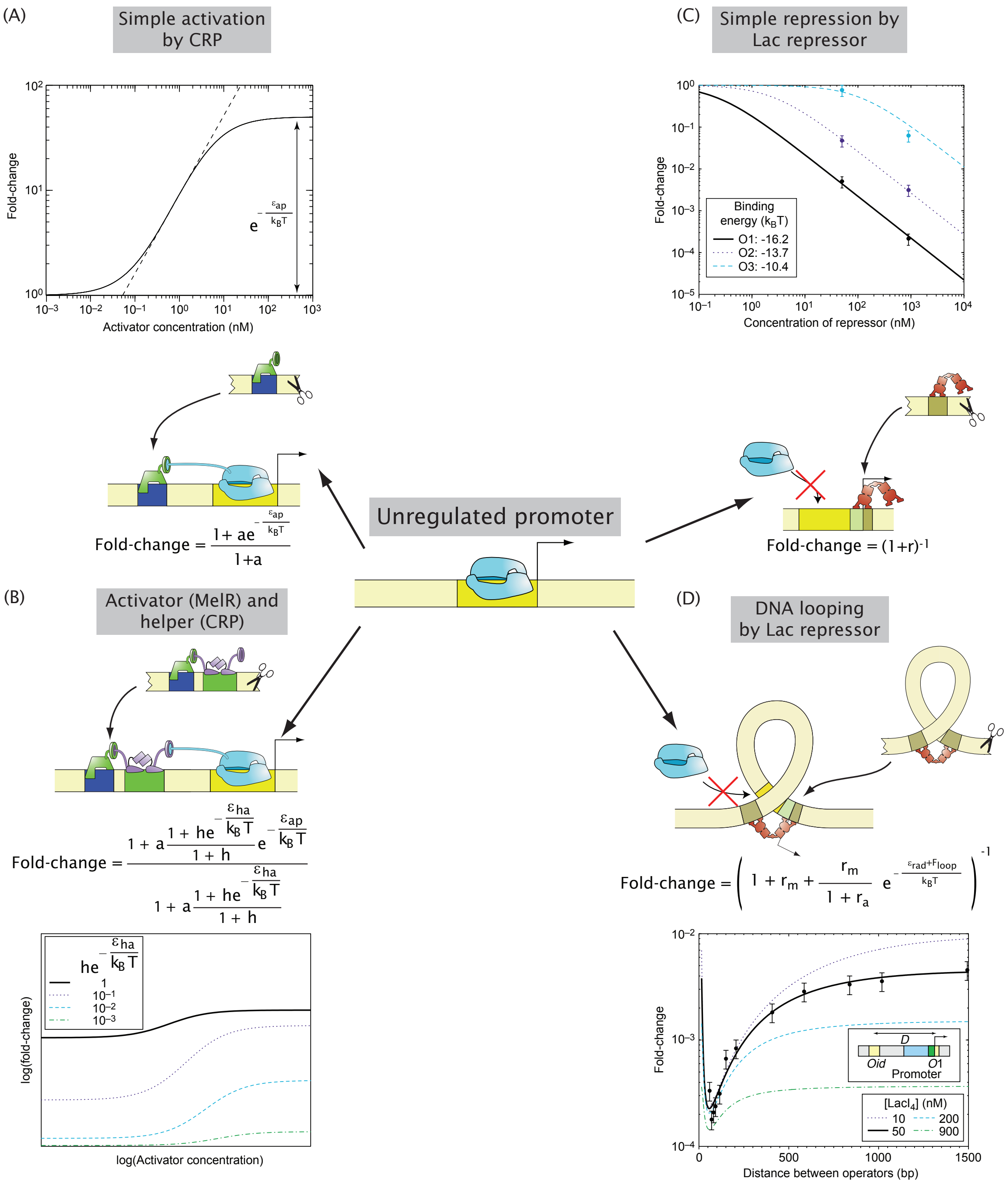

(A)
Simple activation by CRP
Fold-change
10^0
10^1
10^2
10^{-3} 10^{-2} 10^{-1} 10^0 10^1 10^2 10^3
Activator concentration (nM)
$e^{-\frac{\varepsilon_{ap}}{k_BT}}$
Fold-change = $\frac{1+ae^{-\frac{\varepsilon_{ap}}{k_BT}}}{1+a}$
(C)
Simple repression by Lac repressor
Fold-change
10^0 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5}
10^{-1} 10^0 10^1 10^2 10^3 10^4
Concentration of repressor (nM)
Binding energy (k_BT)
O1: -16.2
O2: -13.7
O3: -10.4
Fold-change = $(1+r)^{-1}$
Unregulated promoter
(B)
Activator (MelR) and helper (CRP)
Fold-change = $\frac{1+a\frac{1+he^{-\frac{\varepsilon_{ha}}{k_BT}}}{1+h}e^{-\frac{\varepsilon_{ap}}{k_BT}}}{1+a\frac{1+he^{-\frac{\varepsilon_{ha}}{k_BT}}}{1+h}}$
$he^{-\frac{\varepsilon_{ha}}{k_BT}}$
1
10^{-1}
10^{-2}
10^{-3}
log(fold-change)
log(Activator concentration)
(D)
DNA looping by Lac repressor
Fold-change = $\left(1+r_m+\frac{r_m}{1+r_a}e^{-\frac{\varepsilon_{rad}+F_{loop}}{k_BT}}\right)^{-1}$
Fold-change
10^{-2} 10^{-3} 10^{-4}
0 500 1000 1500
Distance between operators (bp)
D
Oid
O1
Promoter
[LacI$_4$] (nM)
10
200
50
900

Figure 1.7: Examples of regulatory architectures and their corresponding predictions. Some of the results from chapter 3 corresponding to several regulatory architectures are shown here for the case of (A) simple activation by CRP, (B) cooperative co-activation with MelR as an activator and CRP as a helper molecule, (C) simple repression by Lac repressor and (D) repression by DNA looping by Lac repressor. In all these cases we predict the fold-change in gene expression as a function of the concentration of the particular transcription factors involved and as a function of their interaction energies.

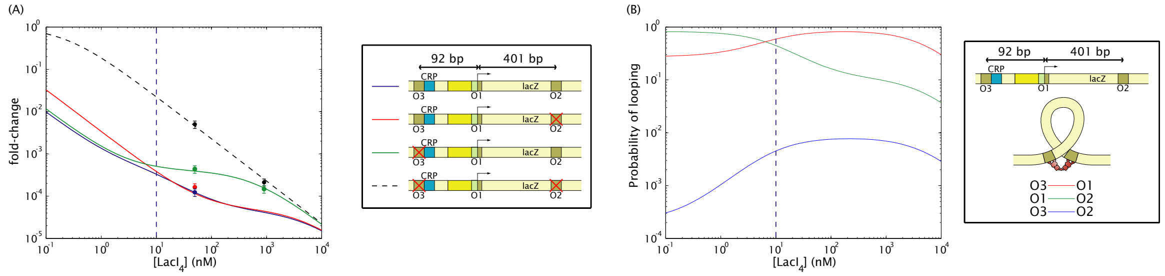


Figure 1.8: Quantitative dissection of the wild-type *lac* operon. We use our thermodynamic models to fit the fold-change in gene expression for various *lac* operon mutants measured by Oehler et al. [24] resulting in a prediction for the fold-change as a function of the concentration. Notice that both loops between O1 and O3 and between O1 and O2 can lead to similar levels of repression at the wild-type concentration of Lac repressor (denoted by the dashed vertical line). (B) Using the parameters obtained in (A) we can predict the probability of looping for all three of the possible loops in the *lac* operon as a function of the Lac repressor concentration. Interestingly, the two loops that contribute to repression have similar probabilities, whereas the O3-O2 loop, which is not capable of repressing the circuit, has a negligible probability. Notice how different concentrations of repressor can lead to one loop being more favorable than the other one. Why evolution would select for such a sophisticated architecture rather than a simple loop is a mystery the chapter leaves us with.

stabilizing levels of gene expression with respect to fluctuations in the concentration of transcription factors [23]. However, the situation of the wild-type *lac* operon is different due to the presence of multiple DNA loops that can lead to repression independently. The *lac* operon has three binding sites leading to three different loops, two of which can cause repression (O1-O2 and O3-O1). A cartoon of the *lac* operon can be found in the legend of figure 1.8(A).

By dissecting experiments on various *lac* operon mutants through our models we suggest that this operon is not making use of the isolation with respect to fluctuations. These conclusions are shown schematically in figure 1.8(A), where we show that rather than providing "robustness" against fluctuations in repressor concentration the role of the multiple loops in the *lac* operon seems to be related to redundancy. Each loop leads to the same level of repression independently. This analysis also leads to a calculation of the *in vivo* probability for the different loops in the operon as shown in figure 1.8(B). Here we show that though at wild-type concentrations the probabilities of the different loops are comparable, the concentration of Lac repressor can be used as a tuning parameter to favor the formation of one loop over the other.

The looping free energy we obtain can be used for more than the understanding of the *lac* operon. It can also be used as a way to access the *in vivo* mechanical properties of DNA. In chapter 4 and in figure 1.9 we show that the looping free energy has predictive power by considering the outcome of a recent experiment by Becker et al. [25] based on the looping free energy obtained from the Müller et al. experiment [26].

Most available experimental techniques are able to give us information about the *in vivo* properties of DNA on length scales larger than 1 kilobasepair [27]. In contrast DNA looping experiments can give us access to the mechanical response of DNA at shorter length scales. We contrast the looping free energy obtained with several polymer models of DNA shown in figure 1.10 and determine that at lengths between 150 bp and 1 kbp, the behavior of DNA is consistent with a self-avoiding random walk with a persistence

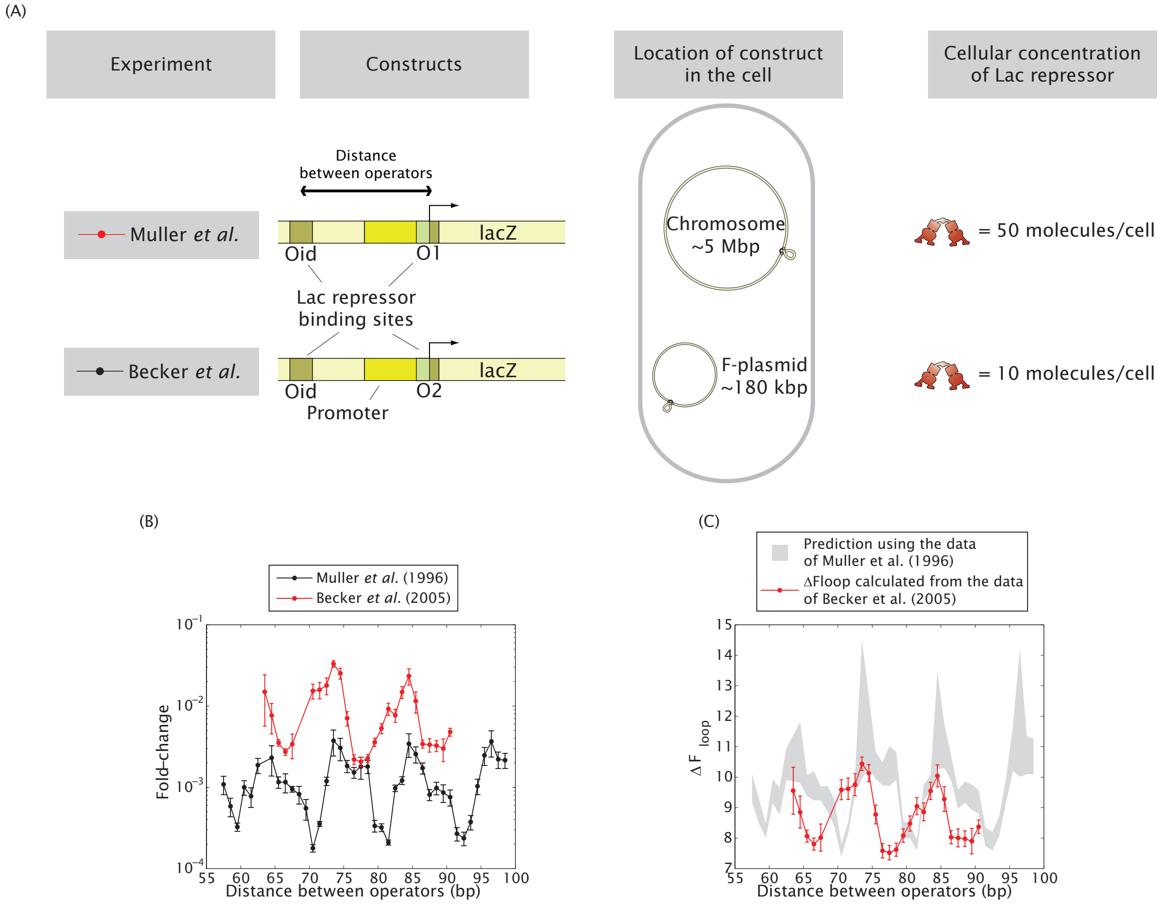


Figure 1.9: Predictions for the looping free energy. The models developed in chapter 4 state that the looping free energy is an intrinsic parameter describing the effective mechanical properties of the intervening DNA. If this is true, different experiments performed with different binding sites and concentrations of repressor, but the same distance between the repressor binding sites, should lead to the same looping free energy. (A) Müller et al. [26] and Becker et al. [25] each measured the level of gene expression in DNA looping using different experimental setups. (B) These differences in setup led to significant differences in the fold-change levels. (C) The looping free energy calculated from the Müller et al. experiment leads to an expectation about the free energy obtained from the Becker et al. measurements. In general terms, the two looping free energies are comparable. However, there are intriguing differences between the looping free energies that are discussed in chapter 4 in more detail.

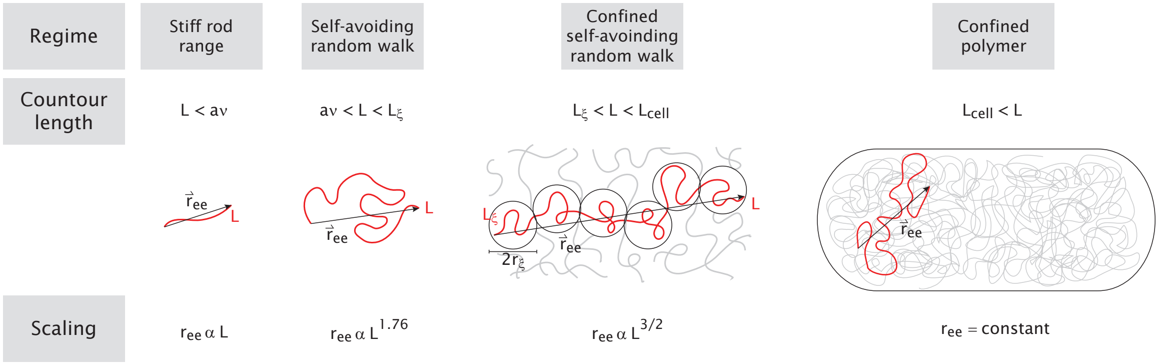


Figure 1.10: Different regimes of scaling for a semiflexible polymer confined to a cell. In chapter 4 we discuss the implications of the different polymer models for DNA. For contour lengths below the Kuhn length L_{Kuhn}, which is a measure of the stiffness of the DNA, the polymer behaves like a stiff rod. The end-to-end distance r_{ee} scales linearly with the corresponding contour length L. As we go up in scale the polymer behaves like a self-avoiding random walk up to a length scale defined by a mesh with a typical radius r_ξ. The corresponding contour length of polymer within that radius is defined as L_ξ. Because of the constrained volume given by the *E. coli* cell the Flory theorem states that for length scales beyond the mesh size, but smaller than the typical dimensions of the cell the effective polymer will behave like an entropic spring with a scaling $r_{ee} \propto L^{3/2}$. Finally, for contour lengths beyond the typical cell dimension, L_{cell}, the density of polymer monomers is uniform throughout the cell. As a result, the end-to-end distance becomes constant and independent of the contour length. In the chapter we show that the available data is consistent with a model of DNA as a self-avoiding random walk on length scales between 150 bp and 1 kbp.

length of about 25 bp.

One problem when comparing free energies obtained in *in vivo* and *in vitro* experiments is that they might not be defined with respect to the same zero of energy. Chapter 4 finalizes with a careful treatment of the contribution of non-specific binding to DNA looping which allows us to, for the first time, plot the outcome of *in vivo* and *in vitro* experiments (some of which are described in chapter 9) on the same figure. As a result in figure 1.18 we compare the results from the *in vivo* experiments by Müller et al., the *in vitro* measurements corresponding to chapter 9 and several other key experiments on DNA looping by Lac repressor and DNA mechanics at short length scales. Please, refer to the description of chapter 9 below for a detailed explanation of the conclusions stemming from figure 1.18. Chapter 4 is again the result of long-standing collaboration between the Phillips and Kondev groups and is about to be submitted to *Physical Biology*.

Chapter 5 explores precisely the same collection of architectures as in chapters 2 and 3, but now with the ambition of calculating the higher moments of the mRNA distribution. Because of recent efforts in single-cell measurements on gene expression, these moments have become amenable to direct experimental measurement. The modeling approach relies on the description of possible trajectories of the system and their corresponding weights rather than the description in terms of states from chapters 2 and 3. An example of a kinetic scheme analyzed is shown in figure 1.6(C) and the corresponding trajectories and weights are shown in figure 1.6(D). The results calculated in this chapter provide a suite of predictions for what we expect will result from such measurements and the results are summarized in table 1.2. This is the result of a project led by Alvaro Sanchez from the Kondev group at Brandeis in collaboration with the Phillips

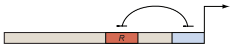
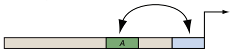
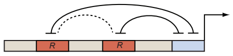
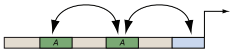
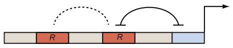
Promoter architecture	Fold-change in noise
1. Simple repression 	$1+\frac{r\,k_R^{on}}{\left(k_R^{\mathrm{off}}+k_R^{\mathrm{on}}\right)\left(\gamma\ +k_R^{\mathrm{off}}+k_R^{\mathrm{on}}\right)}$
2. Simple activation 	$1+\left(\frac{\left(\frac{r_2}{r_1}-1\right)^2 k_A^{off}k_A^{on}r_2}{\left(k_A^{off}+k_A^{on}\right)\left(\gamma\ +k_A^{off}+k_A^{on}\right)\left(k_A^{off}+\frac{r_2}{r_1}\,k_A^{on}\right)}\right)$
3. Dual repression 	$1+\frac{\left(r\,k_R^{on}\left(k_R^{2\,on}+2\,\Omega\,k_R^{off}\left(\gamma\ +2\,\Omega\,k_R^{off}\right)+k_R^{on}\left(\gamma\ +k_R^{off}+4\,\Omega\,k_R^{off}\right)\right)\right)}{\left(\left(2\,(k_R^{on})^2\ \ +\left(\gamma\ +k_R^{off}\right)\left(\gamma\ +2\,\Omega\,k_R^{off}\right)+k_R^{on}\left(3\,\gamma\ +4\,\Omega\,k_R^{off}\right)\right)\left((k_R^{on})^2+\Omega\,k_R^{off}\left(k_R^{off}+2\,k_R^{on}\right)\right)\right)}$
4. Cooperative activation 	$1+\frac{(r_2/r_1-1)^2 r_2\,\Omega\,k_A^{off}k_A^{on}}{\Omega\,k_A^{off}\left(k_A^{off}+k_A^{on}\right)+r_2/r_1\,k_A^{on}\left(\Omega\,k_A^{off}+k_A^{on}\right)}\left(\frac{1}{2\left(\gamma\ +k_A^{off}+k_A^{on}\right)}+\frac{2\,(k_A^{on})^3+(k_A^{on})^2\left(\gamma\ +6\,k_A^{off}\right)+\Omega\left(k_A^{off}\right)^2\left(\gamma\ +2\,\Omega\,k_A^{off}\right)+2\,k_A^{off}k_A^{on}\left(\gamma\ +k_A^{off}+2\,\Omega\,k_A^{off}\right)}{2\left(2\,(k_A^{on})^2+\left(\gamma\ +k_A^{off}\right)\left(\gamma\ +2\,\Omega\,k_A^{off}\right)+k_A^{on}\left(3\,\gamma\ +4\,\Omega\,k_A^{off}\right)\right)\left((k_A^{on})^2+\Omega\,k_A^{off}\left(k_A^{off}+2\,k_A^{on}\right)\right)}\right)$
5. Repression by DNA looping 	$1+r\,k_R^{on}k_l\frac{\left(k_R^{off}+k_R^{on}\right)\left(\gamma\ +k_R^{off}+k_R^{on}\right)\left(\gamma\ +2\left(k_R^{off}+k_R^{on}\right)\right)}{\left(\gamma\ +k_l+k_R^{off}+k_R^{on}\right)\left(k_l\,k_R^{on}+\left(k_R^{off}+k_R^{on}\right)^2\right)}$ $\frac{1+\left(k_lk_R^{off}+2\,(k_R^{on})^2+2\,k_R^{off}\left(\gamma\ +2\,k_R^{off}\right)+k_R^{on}\left(\gamma\ +5\,k_R^{off}\right)\right)\left(\left(k_R^{off}+k_R^{on}\right)+\left(\gamma^2+4\left(k_R^{off}+k_R^{on}\right)^2+\gamma\left(5\,k_R^{off}+4\,k_R^{on}\right)\right)\right)}{k_l(\gamma\ +2\,k_R^{on})+\left(\gamma\ +k_R^{off}+k_R^{on}\right)\left(\gamma\ +2\left(k_R^{off}+k_R^{on}\right)\right)}$

Table 1.2: Fold-change in noise for several regulatory architectures. The fold-change in noise (defined as variance/mean, also called the Fano factor) is shown for several regulatory architectures. Unlike the predictions of the thermodynamic models shown in table 1.1 the predictions from the stochastic models depend on rates of transition between promoter states. This gives access to a whole new set of microscopic parameters.

group. This work is currently under review at *PLoS Computational Biology*.

1.2 A Roadmap to Part II: Experimental Dissection of Gene Regulatory Motifs

As mentioned in the beginning of this chapter it is almost impossible to conceive of our experimental efforts in the absence of the theoretical models described in Part I. Hence, the order of presentation follows logically and also historically in terms of the work carried out in the thesis. In this section of the thesis we dissected gene regulatory motifs in terms of the "knobs" that we had identified through our theoretical explorations. We characterized the input-output relations of these networks both *in vivo* and *in vitro* as a function of parameters such as binding energies and concentrations of transcription factors. Surprisingly, in most of the cases the predictions from the theoretical models were found to correspond quite convincingly with experiment, showing that we could indeed largely compute the input-output function from first principles. Such successful theory-experiment interplay opens the door to a new generation of theoretical and experimental observations where increasingly complex regulatory motifs and interactions between such motifs are dissected.

The models of transcriptional regulation described in chapters 2 and 3 predict changes in gene expression that can span over several orders of magnitude. Being able to test such models will depend strongly on choosing the right reporters of gene expression. It is not only important to choose a technique with the right dynamic range. Since we want to characterize relative changes in gene expression, it is also of the utmost importance that the measuring technique is linear over the range assayed. In particular, the two most widespread techniques to measure gene expression are the enzymatic reporter β-galactosidase and fluorescent proteins. Figure 1.11 poses the question: does the absolute level of gene expression or fold-change in gene expression depend on the technique invoked to measure it? Chapter 6 is a systematic characterization and comparison of the enzymatic reporter β-galactosidase and the fluorescent reporter EYFP in this context. In order to set absolute bounds on their dynamic ranges we calibrate the *in vivo* fluorescence of YFP. Through this absolute calibration we determine that the reporters are linear related over four orders of magnitude and also that they are limited in different ways. EYFP is bounded on the low range to about 10 molecules per cell by the cellular autofluorescence. In contrast, β-galactosidase has no such limitation. However, when present inside the cell at high enough levels, over 20,000 molecules per cell, it can significantly affect cellular growth. The resulting calibration is summarized in figure 1.12, where the linearity can be clearly appreciated and where we have marked the limits of each reporter through the shaded regions. This chapter served as a fundamental guide when planning the *in vivo* experiments presented in this thesis. The fact that both reporters are interchangeable allowed us get the best of both worlds. This experimental work is currently being reviewed at *Biophysical Journal*.

One of the simplest yet richest regulatory motifs is that of simple repression (see figure 1.4, for example). Here, a single repressor binds to a site overlapping the promoter resulting in the exclusion of RNA polymerase from the DNA and the subsequent downregulation of the gene. In figure 1.5 we identified some of the "knobs" of this architecture that are relevant both experimentally and theoretically in the context of the thermodynamic models presented in chapters 2 and 3. In chapter 7 we dissect simple repression by tuning two of these knobs, namely the intracellular number of repressor and the binding energy. We generated several strains with different levels of Lac repressor and measured the fold-change in gene expression for constructs bearing binding sites of different strengths as shown in figure 1.13(A,B). Though it is relatively easy to generate strains with different concentrations of repressor, it is not that straightforward to predict the actual number of repressors each strain will bear. Using the binding energies obtained in chapters 3 and 4 and the prediction of the input-output function from the thermodynamic models we then predict the number of repressor in each strain. These predictions are shown in figure 1.13(C) and are validated by measuring the intracellular number of repressors directly through quantitative immunoblotting over two orders of magnitude. Such a direct census of Lac repressor had not been done since its purification in the 60s. This was a highly successful first excursion into the validation of our models. We show that we can account for the resulting changes in the output gene expression over more than three orders of magnitude.

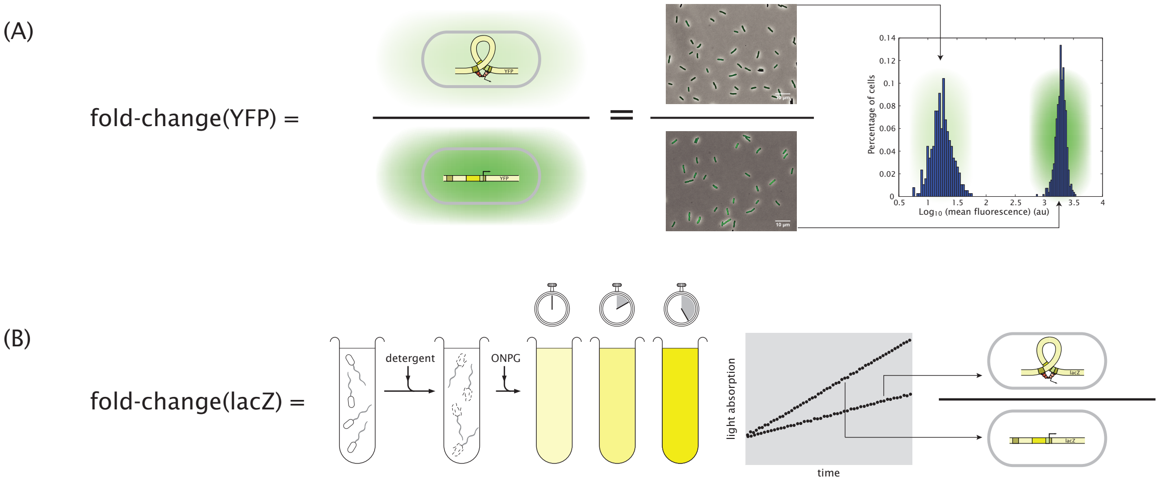


Figure 1.11: Different ways of quantifying the fold-change in gene expression. (A) Using fluorescent proteins the light emission of single cells can be quantified. In this case the fluorescence of a distribution of cells can be measured using microscopy. The mean of each distribution (corresponding to the presence and absence of transcription factor) can then be divided to obtain the fold-change. (B) Similarly, an enzyme such as β-galactosidase can be used to quantify the level of gene expression of the two strains making the fold-change. In this case a macroscopic volume of each sample is broken up using detergent and a substrate is subsequently added. This substrate becomes yellow when cleaved by the enzyme reporter. As a result, by monitoring the rate of change in the color of the reaction by spectrophotometry the amount of enzyme per cell can be calculated. Unlike fluorescence this an inherently bulk measurement. In chapter 6 we show that both reporter methods are complementary and necessary to span a significant dynamic and linear range of gene expression.

It is a direct example of the capabilities of the input-output functions generated by thermodynamic models and the door to addressing more complex regulatory architectures. This work is under review at *Proceedings of the National Academy of Sciences*.

The experiments presented in chapter 7 are a great way to characterize a regulatory network in terms of the thermodynamic models developed in chapters 2 and 3. This kind of analysis resulted in the estimation of the *in vivo* binding energies (or, equivalently equilibrium constants) of Lac repressor to DNA. However, if we want to obtain information about the *in vivo* rates involved in the regulation process we need to go beyond the mean level of gene expression obtained from bulk experiments. We also need to go beyond thermodynamic models, which deal exclusively with mean levels of expression, and adopt a description of regulation set by the stochastic models shown in chapter 5. Such inference of *in vivo* rates can only be done in the context of single cell experiments, where higher moments of the protein distribution than the mean are detected.

In chapter 8 we present a dissection of the simple repression motif at the single cell level. This experiment introduced by Rosenfeld et al. [28, 29] is based on a clever use of dilution and random fluctuations which results in a continuous titration of the repressor. Since the repressor in question is fused to a fluorescent protein its relative concentration can be tracked in single cells as a function of time. The production of this repressor fusion can be modulated by adding or removing a small inducer molecule, aTc, as shown in

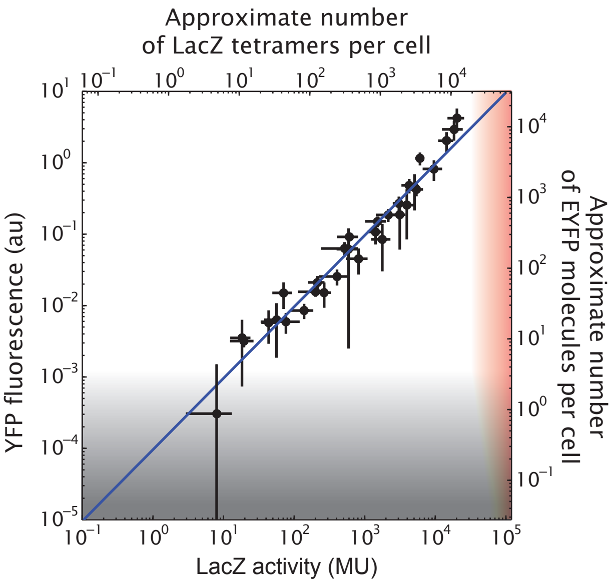
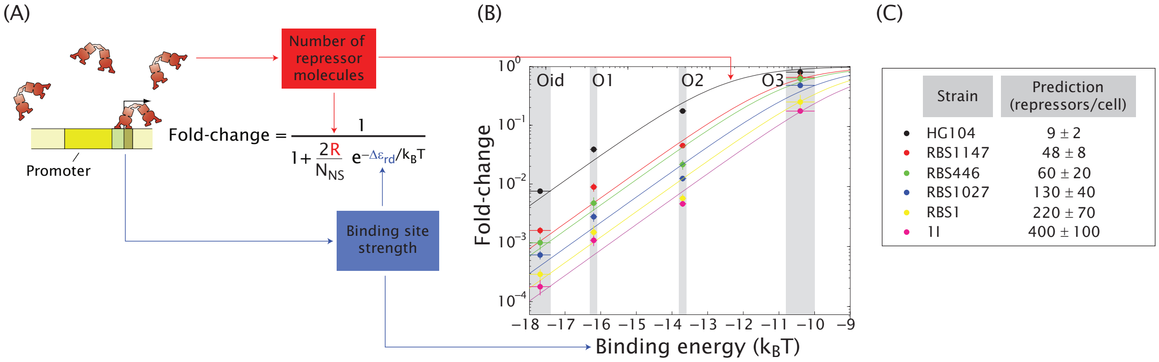


Figure 1.12: Comparison of β-galactosidase and EYFP as reporters of gene expression. By generating several strains expressing either EYFP or β-galactosidase from the same promoter we can compare the dynamic range of each reporter with respect to each other. In chapter 6 we perform this comparison and find them to be linear over four orders of magnitude which spans most of the relevant *in vivo* range of expression of bacterial promoters. We perform an absolute calibration of each reporter which allows us to set absolute bounds on their applicability. EYFP is limited to about 10 molecules/cell due to cell autofluroescence. On the other hand, β-galactosidase is reliable on the low end due to its low background, but starts affecting cell growth significantly when present at concentrations in excess of 20,000 enzymes/cell. These limitations are depicted as shaded areas in the plot.



	Strain	Prediction (repressors/cell)
●	HG104	9 ± 2
●	RBS1147	48 ± 8
●	RBS446	60 ± 20
●	RBS1027	130 ± 40
●	RBS1	220 ± 70
●	1I	400 ± 100

Figure 1.13: Probing the simple repression motif. (A) Knobs we control experimentally. The affinity of the binding sites can be varied by mutating the DNA sequence of the operators. We vary the concentration of repressor by generating several strains expressing different constitutive levels of repressor. However, this number of repressors is not known *a priori*. (B) Resulting fold-change in gene expression for the different values of the knobs. (C) Using the binding energies obtained in chapters 3 and 4 we can predict the number of Lac repressors in each strain, which we check through quantitative immunoblotting.

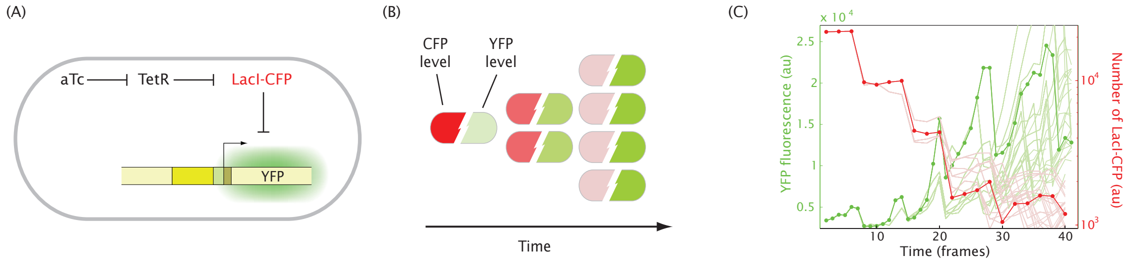


Figure 1.14: Characterizing a gene regulatory network through dilution and fluctuations. (A) The simple repression architecture expresses the fluorescent protein YFP. Its promoter is repressed by a LacI-CFP fusion. This repressor is in turn expressed off of a promoter that is regulated by Tet repressor. By adding aTc to the cells the production of LacI-CFP is induced. (B) Before starting the experiment aTc is removed resulting in the shutting down of LacI-CFP production. With subsequent cellular divisions the amount of LacI-CFP in each cell decreases due to the partitioning between daughter cells, resulting in a higher level of the regulated YFP gene. (C) Representative trace of fluorescence in the YFP (green) and CFP (red) channels as a function of time for a single cell. With each division event the total amount of LacI-CFP per cell halves on average. As the concentration of LacI-CFP decreases the rate of production of YFP increases. Notice that we are tracking a single cell lineage and that with each division new lineages are created. Those additional lineages are shown using a dimmer color. Also, it is clear from the figure that the partitioning of LacI-CFP between daughter cells is random. In chapter 8 we use these fluctuations to infer the relation between LacI-CFP fluorescence per cell and its absolute intracellular number.

figures 1.14(A,B). In our particular setup we have Lac repressor fused to the fluorescent protein CFP which is regulating the expression of YFP through simple repression. In figure 1.14(C) we show a representative trace of the total CFP and YFP fluorescence as a function of time. Notice that cell divisions lead to new lineages and that the partitioning in the number of repressors is random.

This dilution method is not only useful in getting a titration of the repressor. We can use the fluctuations in the binomial partitioning of LacI-CFP in order to relate the CFP fluorescence measured in arbitrary units (see, for example, the axis in figure 1.14(C)) to an absolute number of repressor molecules. As a result we can perform a similar characterization of the simple repression motif as that performed in chapter 7. The main difference in this new method is that we can now perform the assay at the single cell rather than bulk level. In figure 1.15 we show the fold-change in gene expression of single cells as a function of the intracellular concentration of the Lac repressor fusion for different choices of the Lac repressor binding site. Through our thermodynamic models we can obtain their respective *in vivo* binding energies. Interestingly, these binding energies are systematically lower than those found using bulk methods in chapter 7. This might reflect an unknown systematic error in one of the two methods for the absolute counting of repressors (fluctuations vs. immunoblots) or, more likely, it might be due to the fact that the LacI-CFP fusion used in this chapter behaves in a different way from wild-type Lac repressor, which is the protein used in chapter 7.

Finally, chapter 8 is our first attempt at contrasting the stochastic models developed in chapter 5 with single cell data on variability in transcriptional regulation. In figure 1.16(A) we show the fold-change in the variance of gene expression (measured with respect to a strain lacking the repressor) as a function of the

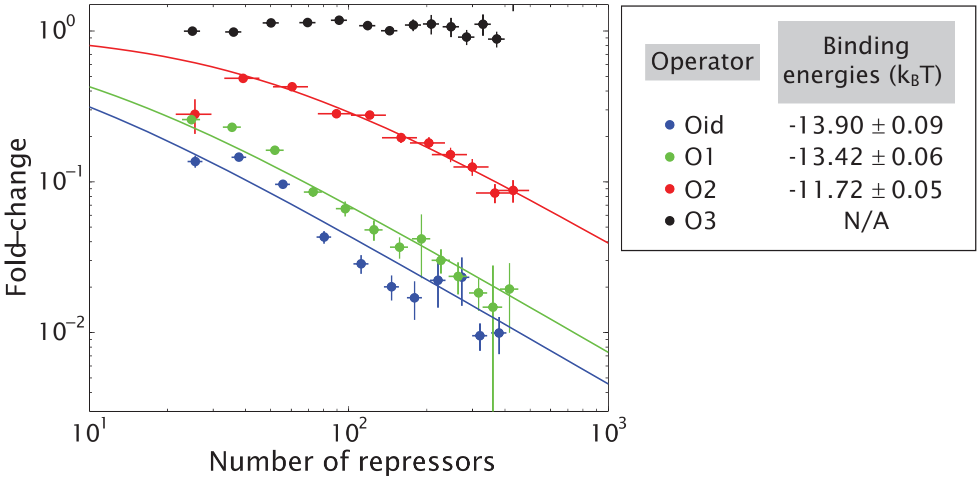


Figure 1.15: Single cell input-output function for simple repression. Through the dilution method shown in chapter 8 and figure 1.14 we quantify the single cell fold-change in gene expression as a function of the intracellular number of Lac repressor-CFP molecules. In analogy to the analysis developed in chapter 7 in a bulk context we can obtain the binding energies corresponding to each possible binding site sequence for this architecture. The fold-change corresponding to O3 is too low to determine the binding energy accurately. One intriguing outcome of this analysis is a systematic difference between the *in vivo* binding energies found through this approach and those found in chapter 7 using wild-type repressor rather than a fluorescent fusion. This suggests that either the binding activity of LacI-CFP is different than wild-type LacI or that one of the two methods (single cell dilution or immunoblots in bulk) is not accurately measuring the absolute number of molecules per cell.

fold-change in mean level of gene expression for several choices of the Lac repressor binding site. Though not very pronounced, there seems to be a systematic effect with stronger operators leading to higher variance for a given fold-change in mean gene expression. This observation is confirmed *qualitatively* in figure 1.16(B), where we show a theoretical prediction using the stochastic models developed in chapter 5. Such comparisons are very exciting as only a very limited number of examples where the effect of promoter architecture on expression noise have been shown in bacteria.

It is important to note that the experiments described in chapter 8 are a very recent development in the Phillips lab and that its details and our understanding of the data are still evolving significantly on a daily time scale. Nevertheless, given the various different dissections of a regulatory presented throughout this thesis we considered it relevant to include a snapshot of our current understanding of simple repression at the single cell level. This work is the result of a side-by-side collaboration with Linda Song in our lab and with the Kondev group at Brandeis University.

As pointed out in the description of chapter 4 in the previous section the mechanical properties of DNA are not well understood in the *in vivo* context. Further, they are not even completely clear in the *in vitro* setting, especially for short loops spanning only a couple of hundreds of base pairs, such as those found in bacteria [22, 30]. Even though the behavior of DNA mechanics is not well understood at the length scales involved in DNA looping by Lac repressor, no direct experiment to date had been done that satisfactorily addressed this issue in the *in vitro* context. Such an experiment requires quantifying the looping probability

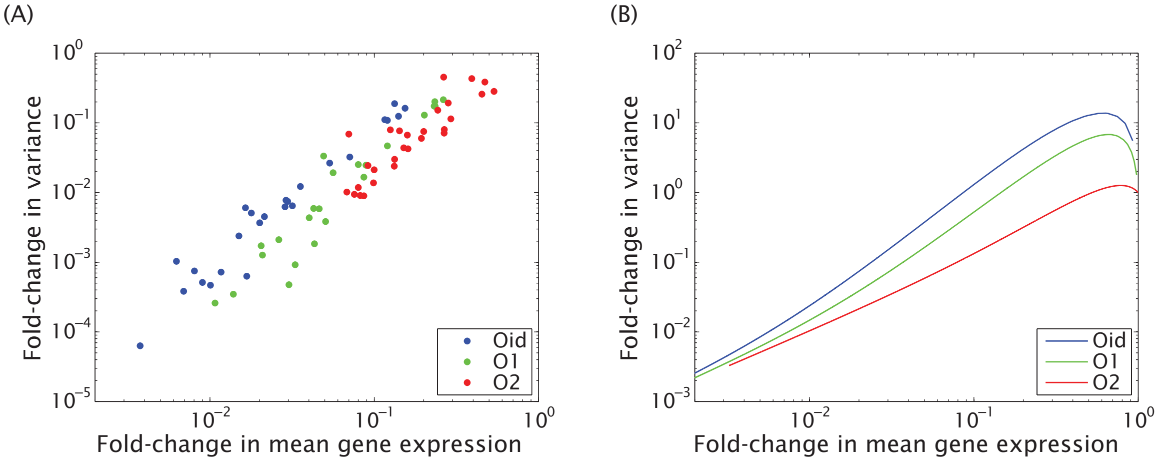


Figure 1.16: Variability in gene expression as a function of the promoter architecture. (A) The fold-change in variance of a strain expressing LacI-CFP measured with respect to a strain lacking the repressor is plotted as a function of the fold-change in mean gene expression for different realizations of the simple repression promoter architecture. A slight systematic trend, where binding sites with a higher affinity lead to higher noise in expression, is observed. (B) This behavior is reproduced *qualitatively* by using stochastic models of transcriptional regulations such as the ones presented in 5. In chapter 8 we discuss possible strategies to compare the theoretical prediction to the experimental data in quantitative detail.

of Lac repressor over several lengths of the loop. Chapter 9 is a careful and systematic *in vitro* dissection of DNA looping by Lac repressor at the single molecule level. This is performed using the Tethered Particle Method, shown diagrammatically in figure 1.17(A–C), where beads each bound to a single DNA molecule tethered to a surface are tracked. By measuring the excursion of the bead over time (figure 1.17(A,B)) the effective length of the DNA tethered can be inferred. This results in a probability distribution of the DNA molecule being in any of its discrete states exemplified by the histogram shown in figure 1.17(C). This single molecule technique has many advantages. First, unlike techniques such as nitrocellulose binding or gel retardation assay, it allows for a true equilibrium measurement of the system. Second, by monitoring single molecules we can distinguish multiple looping states which bulk techniques are unable to do. In 9 we show that even though some aspects of the thermodynamic models apply to this motif there are several gaps in our understanding related to the mechanical properties of the DNA-repressor complex. In fact, we show that theoretical simulations based on the known geometrical details of Lac repressor and on the canonical elastic view of DNA cannot account for the observed behavior.

This is summarized in figure 1.18. The data denoted as “TPM, Han et al.” corresponds to our *in vitro* results, whereas the curve “Monte Carlo simulation, Towles et al.” shows our theoretical expectations based on our knowledge of the geometrical details of Lac repressor and the mechanical properties of DNA. We also show that the *in vivo* data (“Müller et al.”) presents a much higher flexibility than either *in vitro* outcome. The result of this comparison is a still confusing picture: neither the *in vivo*, *in vitro* or the theoretical expectation agree with each other, and this casts doubt on our understanding of DNA mechanics at short length scales. In chapters 4 and 9 we discuss several reasons for the differences, ranging from a breakdown of the elastic model of DNA at short length scales to the role of *in vivo* DNA binding proteins such as HU, H-NS and IHF which decorate the chromosome and the effect of supercoiling. The experimental work presented here was almost exclusively performed by Lin Han, a graduate student in the Phillips lab with whom I had some overlap. My main contribution was related to data analysis and to contrasting it with theoretical models of transcriptional regulation and DNA mechanics. This work was done in collaboration with the Phil Nelson group at the University of Pennsylvania and led to the publication of two papers [31, 32].

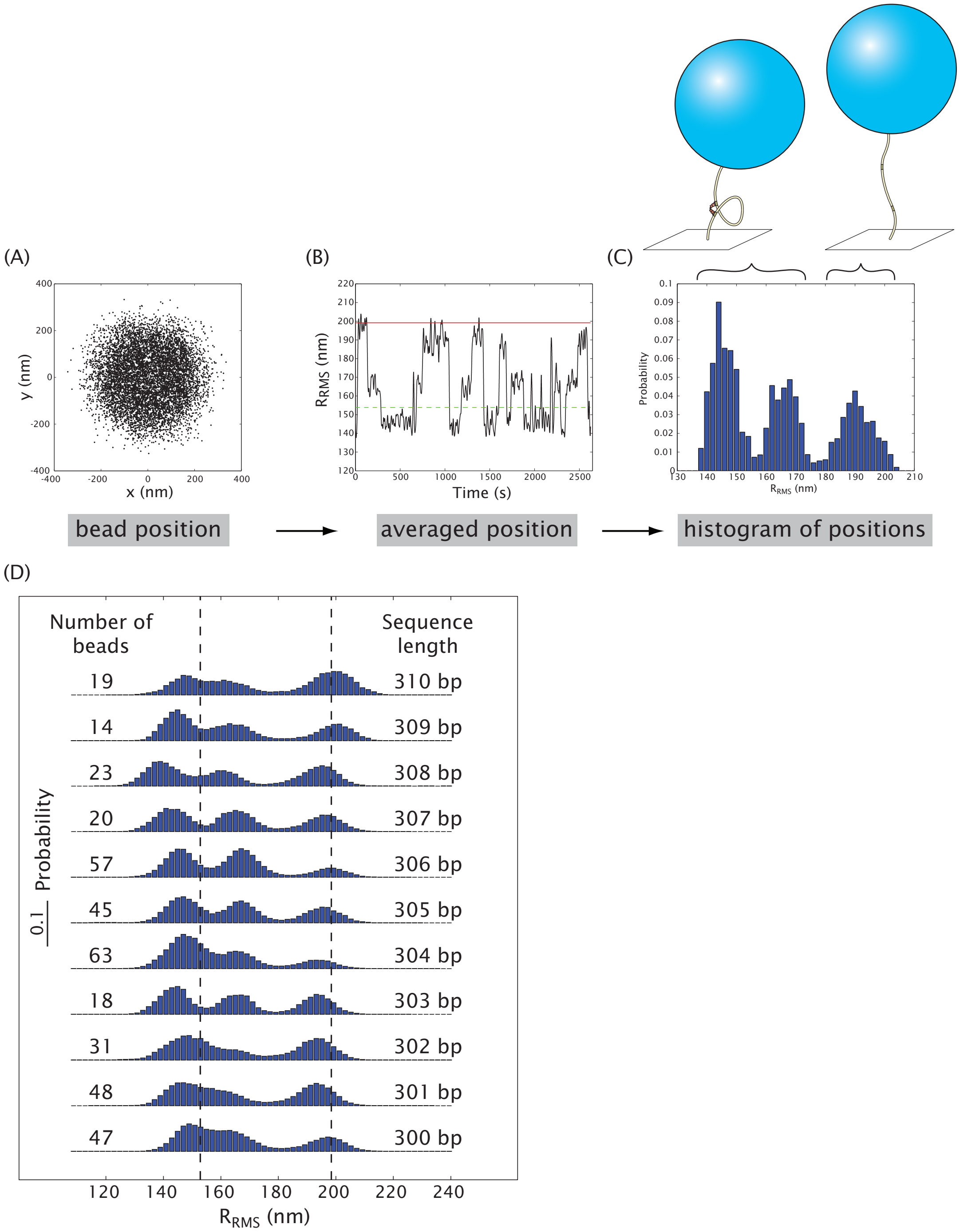


Figure 1.17: Tethered particle method to measure *in vitro* DNA looping. (A) A bead bound to a DNA molecule which is in turn tethered to the surface is tracked as a function of time. (B) The mean root square deviation from its center is calculated over a time window revealing the existence of multiple discrete states of the molecule. (C) The information in (B) corresponding to many molecules can be collapsed into a histogram which gives a measure of the probability of the DNA molecule being in each of its states. In this particular example, two looped configurations and one unlooped configuration are detected. (D) The process described in (A–C) can be repeated for DNA molecules with different loop lengths resulting in a direct measurement of the looping probability as a function of the distance between Lac repressor binding sites.

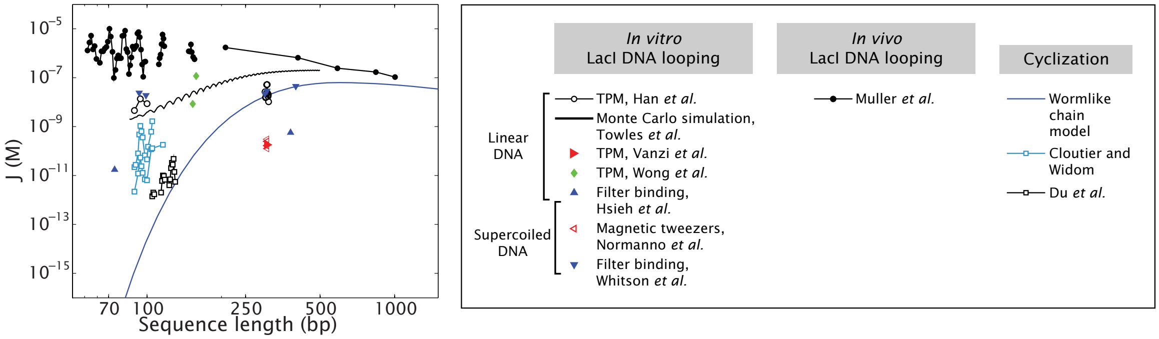


Figure 1.18: *In vivo* and *in vitro* experiments and calculations on DNA looping and DNA mechanics. The looping J-factor J related to the looping free energy F_{loop} through the definition $J = 1\ \mathrm{M}e^{-\beta F_{loop}}$ is shown for the *in vivo* data of Müller et al. [26] data based on the model for the non-specific looping reservoir presented in chapter 4. This prediction is superimposed with the looping J-factors derived from the the *in vitro* experiments presented in chapter 9 (Han et al., [31]) and a theoretical expectation (Towles et al., [32]). The results of several other *in vitro* looping experiments are shown. In some cases the looping J-factor was not reported explicitly and had to be estimated from the data. As such some of the *in vitro* values should be viewed as approximations. even in those cases where there was no direct measurement of J itself. From this plot it is clear that there is still a large spread in experimental data and that the experimental techniques need to improved upon. A particular example of this is the large effect due to supercoiling observed by Whitson et al. [33], which is not present in the more recent experiments by Normanno et al. [34]. Finally, we show results for DNA cyclization, where the propensity of DNA to form circles in the absence of any proteins is measured. Here, too, the results of Du et al. [35] agreeing with the theoretical expectation based on the wormlike chain model [36] are to be contrasted with the much higher flexibility obtained by Cloutier and Widom [37]. Other data sources are: Hsieh et al. [38], Vanzi et al. [39] and Wong et al. [40].

Bibliography

[1] F. Jacob, D. Perrin, C. Sanchez, J. Monod, and S. Edelstein. [the operon: A group of genes with expression coordinated by an operator. C.R.Acad. Sci. Paris 250 (1960) 1727-1729]. *C R Biol*, 328(6):514–20, 2005.

[2] U. Alon. *An introduction to systems biology: Design principles of biological circuits*. Chapman & hall/crc mathematical and computational biology series. Chapman & Hall/CRC, Boca Raton, FL, 2007.

[3] S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Trevino-Quintanilla, and J. Collado-Vides. Regulondb (version 6.0): Gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic Acids Res*, 36:D120–4, 2008.

[4] S. Ben-Tabou De-Leon and E. H. Davidson. Gene regulation: Gene control network in development. *Annu Rev Biophys Biomol Struct*, 36:191, 2007.

[5] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–7, 2002.

[6] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–8, 2002.

[7] I. S. Peter and E. H. Davidson. Modularity and design principles in the sea urchin embryo gene regulatory network. *FEBS Lett*, 583(24):3948–58, 2009.

[8] R. Dobrin, Q. K. Beg, A. L. Barabasi, and Z. N. Oltvai. Aggregation of topological motifs in the escherichia coli transcriptional regulatory network. *BMC Bioinformatics*, 5:10, 2004.

[9] M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A*, 99(16):10555–60, 2002.

[10] H. D. Kim, T. Shay, E. K. O'shea, and A. Regev. Transcriptional regulatory circuits: Predicting numbers from alphabets. *Science*, 325:429–432, 2009.

[11] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1):117–24, 2007.

[12] D. Endy. Foundations for engineering biology. *Nature*, 438(7067):449–53, 2005.

[13] C. A. Voigt. Genetic parts to program bacteria. *Curr Opin Biotechnol*, 17(5):548–57, 2006.

[14] G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A*, 79(4):1129–33, 1982.

[15] M. A. Shea and G. K. Ackers. The or control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol*, 181(2):211–30, 1985.

[16] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003.

[17] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: From theories to phenotypes. *Nat Rev Genet*, 2005.

[18] X. S. Xie, P. J. Choi, G. W. Li, N. K. Lee, and G. Lia. Single-molecule approach to molecular biology in living bacterial cells. *Annu Rev Biophys*, 37:417–44, 2008.

[19] D. R. Larson, R. H. Singer, and D. Zenklusen. A single molecule view of gene expression. *Trends Cell Biol*, 19(11):630–7, 2009.

[20] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[21] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[22] H. G. Garcia, P. Grayson, L. Han, M. Inamdar, J. Kondev, P. C. Nelson, R. Phillips, J. Widom, and P. A. Wiggins. Biological consequences of tightly bent DNA: The other life of a macromolecular celebrity. *Biopolymers*, 85(2):115–30, 2007.

[23] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[24] S. Oehler, M. Amouyal, P. Kolkhof, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. Quality and position of the three *lac* operators of *e. Coli* define efficiency of repression. *EMBO J*, 13(14):3348–55, 1994.

[25] N. A. Becker, J. D. Kahn, and L. J. Maher Iii. Bacterial repression loops require enhanced DNA flexibility. *J Mol Biol*, 349(4):716–30, 2005.

[26] J. MÜLler, S. Oehler, and B. MÜLler-Hill. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol*, 257(1):21–9, 1996.

[27] N. Naumova and J. Dekker. Integrating one-dimensional and three-dimensional maps of genomes. *J Cell Sci*, 123(Pt 12):1979–88, 2010.

[28] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–5, 2005.

[29] N. Rosenfeld, T. J. Perkins, U. Alon, M. B. Elowitz, and P. S. Swain. A fluctuation method to quantify in vivo fluorescence data. *Biophys J*, 91(2):759–66, 2006.

[30] J. P. Peters and L. J. Maher. DNA curvature and flexibility in vitro and in vivo. *Quarterly Reviews of Biophysics*, 43(1):23–63, 2010.

[31] L. Han, H. G. Garcia, S. Blumberg, K. B. Towles, J. F. Beausang, P. C. Nelson, and R. Phillips. Concentration and length dependence of DNA looping in transcriptional regulation. *PLoS One*, 4(5):e5621, 2009.

[32] K. B. Towles, J. F. Beausang, H. G. Garcia, R. Phillips, and P. C. Nelson. First-principles calculation of DNA looping in tethered particle experiments. *Phys Biol*, 6(2):25001, 2009.

[33] P. A. Whitson, W. T. Hsieh, R. D. Wells, and K. S. Matthews. Influence of supercoiling and sequence context on operator DNA binding with *lac* repressor. *J Biol Chem*, 262(30):14592–9, 1987.

[34] D. Normanno, F. Vanzi, and F. S. Pavone. Single-molecule manipulation reveals supercoiling-dependent modulation of *lac* repressor-mediated DNA looping. *Nucleic Acids Res*, 36(8):2505–13, 2008.

[35] Q. Du, C. Smith, N. Shiffeldrim, M. Vologodskaia, and A. Vologodskii. Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proc Natl Acad Sci U S A*, 102(15):5397–402, 2005.

[36] H. Yamakawa. *Helical wormlike chains in polymer solutions.* Springer, Berlin, New York, 1997.

[37] T. E. Cloutier and J. Widom. DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc Natl Acad Sci U S A*, 102(10):3645–50, 2005.

[38] W. T. Hsieh, P. A. Whitson, K. S. Matthews, and R. D. Wells. Influence of sequence and distance between two operators on interaction with the *lac* repressor. *J Biol Chem*, 262(30):14583–91, 1987.

[39] F. Vanzi, C. Broggio, L. Sacconi, and F. S. Pavone. Lac repressor hinge flexibility and DNA looping: Single molecule kinetics by tethered particle motion. *Nucleic Acids Res*, 34(12):3409–20, 2006.

[40] O. K. Wong, M. Guthold, D. A. Erie, and J. Gelles. Interconvertible lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biol*, 6(9):e232, 2008.

Part I

Theoretical Models of Transcriptional Regulation

Chapter 2

Introduction to Thermodynamic Models of Transcriptional Regulation

This chapter is a reproduction of reference [1] which was published together with the next chapter corresponding to reference [2].

The expression of genes is regularly characterized with respect to how much, how fast, when and where. Such quantitative data demands quantitative models. Thermodynamic models are based on the assumption that the level of gene expression is proportional to the equilibrium probability that RNA polymerase (RNAP) is bound to the promoter of interest. Statistical mechanics provides a framework for computing these probabilities. Within this framework, interactions of activators, repressors, helper molecules and RNAP are described by a single function, the regulation factor. This analysis culminates in an expression for the probability of RNA polymerase binding at the promoter of interest as a function of the number of regulatory proteins in the cell.

2.1 Introduction

The biological literature on the regulation and expression of genes is, with increasing frequency, couched in the language of numbers. Four key ways in which gene expression is characterized quantitatively are through measurement of: (i) the level of expression relative to some reference value; (ii) how fast a given gene is expressed after induction; (iii) the precise relative timing of expression of different genes; and (iv) the spatial location of expression. In the first section of this review we revisit particular examples of such measurements in the bacterial setting. These provide the motivation for the models that form the main substance of this and the companion article [2]. Through much of these reviews we call attention to particular revealing case studies rather than giving a thorough coverage of the literature.

2.1.1 How much, when and where?

One class of particularly well-characterized examples of gene expression levels includes cases associated with bacterial metabolism and the infection of bacteria by phage [3, 4]. This group will serve as the centerpiece of this and the companion article. In the classic case of the *lac* operon, several beautiful measurements have been taken. These characterize the extent to which the genes are repressed as a function of the strength of the operators, their spacing and the number of repressor molecules [5–7]. Similar measurements have been made for other genes implicated in bacterial metabolism, in addition to those tied to the decision between the lytic and lysogenic pathways after infection of *Escherichia coli* by phage λ [8–13]. A second way by which the regulatory status of a given system is quantified is by measuring when genes of interest are being expressed. The list of examples is long and inspiring, and several representative case studies can be found in the literature [14–16] . A third way in which an increasingly quantitative picture of gene expression is emerging is based on the ability to make precise statements about the spatial location of the expression of different genes. Here, too, the number of different examples that can be mustered to prove the general point is staggering [17–19]. The key point of these examples is to note the growing pressure head of quantitative *in vivo* data, which calls for more than a cartoon-level description of expression.

The physicochemical modeling of the type of quantitative data described above is still in its infancy. One class of models, which will serve as the basis of this article, comprises the so-called "thermodynamic models" [20–22]. The conceptual basis of this class of models is the idea that the expression level of the gene of interest can be deduced by examining the equilibrium probabilities that the DNA associated with that gene is occupied by various molecules; these include RNAP and a battery of transcription factors (TFs) such as repressors and activators. There is a long-standing tradition of using these ideas to unravel the dynamics of gene expression systems, particularly important examples being associated with the famed *lac* operon and phage λ systems [20, 22–26]. Importantly, the thermodynamic models can serve as input to more general chemical kinetic models.

The key aim of this and the accompanying article [2] is to show how the thermodynamic models yield a general conceptual picture of regulation using what we call the "regulation factor". Such arguments are useful because they enable direct comparison with quantitative experiments, such as those discussed above. The purpose of models is not just to "fit the data" (although such fits can reveal which mechanisms are operative) but also to provide a conceptual scheme for understanding measurements and, more importantly, for suggesting new experiments. It is also worth noting that when such models fall short it provides an opportunity to find out why and learn something new.

This article is, to a large extent, pedagogical and aims to demonstrate how a microscopic picture of the various states of the gene of interest can be mathematized using statistical mechanics. The companion article is built around the analysis of case-studies in bacterial transcription and centers specifically on how the activity of a given promoter is altered (the "fold-change" in promoter activity) by the presence of transcription factors.

2.2 Thermodynamic Models of Gene Regulation: The Regulation Factor

The fundamental tenet of the thermodynamic models for gene regulation is that we can replace the difficult task of computing the level of gene expression, as measured by the concentration of gene product ([protein]), with the more tractable question of the probability (p_{bound}) that RNAP occupies the promoter of interest. More precisely, these models are founded on the idea that the instantaneous disposition of the gene of interest can be established from the probability that various molecules — RNAP, activators, repressors and inducers — are bound to their relevant targets.

Such models are based on a variety of different assumptions, all of which can and should be evaluated critically. Perhaps the most glaring assumption is that of equilibrium itself. This assumption can be examined quantitatively on the basis of the relative rates of transcription factor binding, RNAP binding, open complex formation, transcript formation and translation itself. For example, if the rate for open complex formation is much smaller than the rates for RNAP binding and unbinding from the promoter, then the probability of finding the polymerase on the promoter will be given by its equilibrium value. A second key assumption of this class of models is the idea that the probability of promoter occupancy by RNAP is simply proportional to the level of expression of a given gene. The difficulty lies in the fact that there are several different mechanisms that can intervene between RNAP binding and the existence of a functional gene product. Despite these caveats, we argue that this class of models is both instructive and predictive and, in those cases where the models are found wanting, provides an opportunity to learn something.

In this review, we first analyze the probability that RNAP will be bound at the promoter of interest in the absence of any activators or repressors. This is followed by cases of increasing complexity that involve batteries of transcription factors. Although our preliminary discussion is focused on the statistical mechanics of polymerase binding, the framework is the same for generic protein-DNA and protein-protein interactions. For the purposes of this review, we make the simplified assumption that the key molecular players (RNAP and TFs) are bound to the DNA either specifically or non-specifically. This question has been addressed in the context of the λ-switch [26], for the *lac* repressor [23, 27] and for RNAP [28]. Stated differently, as a simplification, we will ignore the contribution of free polymerase in the cytoplasm, in addition to those RNAP molecules that are engaged in transcription on other promoters. Relaxing this assumption has no effect on the framework developed below. Hence, to evaluate the probability of promoter occupancy in this simple model, the reservoir of RNAPs will be the non-specifically bound molecules (as shown in figure 2.1(A)).

To evaluate the probability of polymerase binding (p_{bound}) we must sum the Boltzmann weights over all possible states of P polymerase molecules on DNA [29, 30]. P is the effective number of RNAP molecules available for binding to the promoter. Estimating this number *in vivo* is fraught with difficulty because many RNAPs are engaged in transcription at any given time and, as such, are not available for binding.

Fortunately, this problem is avoided when calculating the fold-change for all the cases of interest, as we do in the accompanying paper [2]. This is because, in these cases, the absence of activators results in a very small p_{bound} value and so P drops out of the problem.

We calculate p_{bound} by considering the distribution of P RNAP on the non-specific sites (N_{NS}), which make up the genome itself, and a single promoter. Then we distinguish two classes of outcomes (shown in figure 2.1(B)): all P RNAP molecules bound non-specifically, or one RNAP bound to the promoter and $P-1$ RNAP bound nonspecifically. Next, we count the number of different ways that these outcomes can be realized. Once these states have been enumerated, we weight each of them according to the Boltzmann law: if e is the energy of a state, its statistical weight is $\exp(-\beta\varepsilon)$. Finally, to compute the probability of promoter occupancy, we construct the ratio of the sum of the weights for the favorable outcome (i.e., promoter occupied) to the sum over all of the weights.

As noted above, this simple model includes two broad classes of microscopic outcomes: (i) those in which all P polymerase molecules are distributed among the nonspecific sites, and (ii) those in which the promoter is occupied and the remaining $P-1$ polymerase molecules are distributed among the non-specific sites. To evaluate the probabilities of these two eventualities we need to know the number of different ways that each outcome can be realized. The statistical question of how many ways there are to distribute P polymerase molecules among N_{NS} non-specific sites on the DNA is a classic problem in combinatorics, and the result is

$$\frac{N_{NS}!}{P!(N_{NS}-P)!}.$$

The overall statistical weight of these states is based not just on how many of them there are but also on their Boltzmann weights according to

$$Z(P) = \underbrace{Z(P)}_{\text{statistical weight-promoter unnoccupied}} + \underbrace{\frac{N_{NS}!}{P!(N_{NS}-P)!}}_{\text{number of arrangements}} \times \underbrace{e^{-\beta P \varepsilon_{pd}^{NS}}}_{\text{Boltzmann weight}}, \qquad (2.1)$$

where ε_{pd}^{NS} is an energy that represents the average binding energy of RNAP to the genomic background. The correct treatment of the genomic background requires explicit consideration of the distribution of binding energies of RNAP, and TFs, to different sites — both specific and non-specific — on the DNA. The question of how to treat this problem more generally than the simple-minded treatment given here can be found in [31, 32]. The total statistical weight can now be written as

$$\underbrace{Z_{tot}(P)}_{\text{total statistical weight}} = \underbrace{Z(P)}_{\text{promoter unoccupied}} + \underbrace{Z(P-1)e^{-\beta\varepsilon_{pd}^{S}}}_{\text{RNAP on promoter}}, \qquad (2.2)$$

where ε_{pd}^{S} is the binding energy for RNAP on the promoter (the S stands for specific). The states and corresponding weights, normalized by the weight of the promoter unoccupied states, $Z(P)$, are shown in figure 2.1(B).

To find the probability of RNAP being bound to the promoter of interest, we calculate

$$p_{bound} = \frac{Z(P-1)e^{-\beta\varepsilon^{S}_{pd}}}{Z_{tot}(P)}. \tag{2.3}$$

Note that the numerator in this case is the statistical weight of all microscopic states in which the promoter is occupied, and the denominator is the statistical weight of all microscopic states. If we now divide top and bottom by $Z(P-1)e^{-\beta\varepsilon^{S}_{pd}}$ and use the functional form given in equation 2.1, the probability of promoter occupancy is given by the simple form

$$p_{bound} = \frac{1}{1 + \frac{N_{NS}}{P}e^{\Delta\varepsilon_{pd}}}, \tag{2.4}$$

where we have introduced the notation $\Delta\varepsilon_{pd} = \varepsilon^{S}_{pd} - \varepsilon^{NS}_{pd}$ [33]. To obtain the last equation we made the simplifying assumption that $P \ll N_{NS}$. The results computed above can be depicted in graphical form (as shown in figure 2.1(C)) by plotting the probability of promoter occupancy as a function of the number of RNAP molecules for two different promoters. For this particular case we have used several rough estimates, explained in the figure legend, concerning the binding energies of RNAP molecules to specific and non-specific sites on the DNA in a typical bacterial cell. One interesting speculation is that the high probability of RNAP occupancy for the T7 promoter, even in the absence of transcription factors, could be related to the infection mechanism of T7 phage [34]. In contrast, it is also interesting to note the very low probability of occupancy of the *lac* promoter in this simple model in the absence of activation. We view equation 7.8 as characterizing the basal transcription rate in this simple model. In light of this result, the key conceptual outcome of the remainder of this review is the idea that the presence of transcription factors (activators and repressors, etc.) has the effect of altering equation 7.8 to the simple form

$$p_{bound} = \frac{1}{1 + \frac{N_{NS}}{PF_{reg}}e^{\beta\Delta\varepsilon_{pd}}}, \tag{2.5}$$

where we introduce the regulation factor, F_{reg}. The regulation factor should be seen as describing an effective increase (for $F_{reg} > 1$) or decrease (for $F_{reg} < 1$) of the number of RNAP molecules that are available to bind the promoter.

To illustrate precisely the idea of the regulation factor, we show how activators recruit [3] RNAP to the promoter of interest. The recruitment concept is illustrated in schematic form in figure 2.2(A), where it is seen that the activator molecule recruits the polymerase through favorable contacts characterized by an adhesive energy, ε_{ap}. The point of the schematic is to show how the various states of occupancy of the promoter and activator binding site can be assigned Boltzmann weights, which can then be used to compute their probabilities.

Once again, the first step in our analysis is to determine the total statistical weight. This is obtained

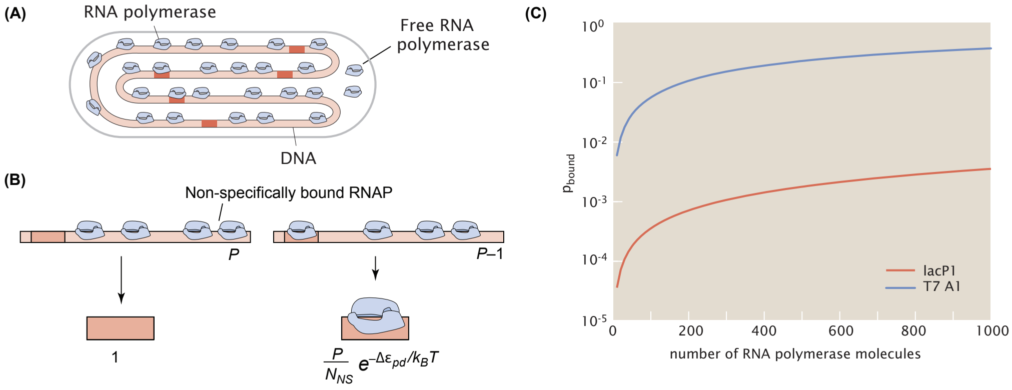


Figure 2.1: Probability of promoter occupancy. (A) Schematic showing how, in the simple model, the DNA molecule serves as a reservoir for the RNAP molecules, almost all of which are bound to DNA. (B) Illustration of the states of the promoter either with RNAP not bound or bound and the remaining polymerase molecules distributed among the non-specific sites. The statistical weights associated with these different states of promoter occupancy are also shown. (C) Probability of binding of RNAP to promoter as a function of the number of RNAP molecules for two different promoters. We assume the number of non-specific sites is $N_{NS} = 5 \times 10^6$, and calculate the binding energy difference using the simple relation $\Delta\varepsilon_{pd} = \beta^{-1} \ln\left(K_{pd}^{S}/K_{pd}^{NS}\right)$, where the equilibrium dissociation constants for specific binding (K_{pd}^{S}) and non-specific binding (K_{pd}^{NS}) are taken from *in vitro* measurements. In particular, making the simplest assumption that the genomic background for RNAP is given only by the non-specific binding of RNAP with DNA, we take $K_{pd}^{NS} = 10000$ nM [35], for the *lac* promoter $K_{pd}^{S} = 550$ nM [36] and for the T7 promoter, $K_{pd}^{S} = 3$ nM [37]. For the *lac* promoter, this results in $\Delta\varepsilon = -2.9\ k_BT$ and for the T7 promoter, $\Delta\varepsilon = -8.1\ k_BT$.

by summing the Boltzmann weights of all of the eventualities associated with the activators and polymerase molecules being distributed on the DNA (both non-specific sites and the promoter). As seen in figure 2.2(A), there are four classes of outcomes: (i) both the activator site and promoter unoccupied; (ii) just the promoter occupied by polymerase; (iii) just the activator site occupied by activator and (iv) both of the specific sites occupied. This is represented mathematically as

$$\begin{aligned} Z_{tot}(P,A) &= \underbrace{Z(P,A)}_{\text{empty sites}} + \underbrace{Z(P-1,A)e^{-\beta\varepsilon_{pd}^{S}}}_{\text{RNAP on promoter}} \\ &+ \underbrace{Z(P,A-1)e^{-\beta\varepsilon_{ad}^{S}}}_{\text{activator on specific site}} + \underbrace{Z(P-1,A-1)e^{-\beta(\varepsilon_{ad}^{S}+\varepsilon_{pd}^{S}+\varepsilon_{pa})}}_{\text{RNAP and activator bound specifically}} \end{aligned} \tag{2.6}$$

where the statistical weight for P polymerase molecules and A activator molecules distributed among N_{NS} nonspecific sites is given by

$$Z(P,A) = \underbrace{\frac{N_{NS}!}{P!A!(N_{NS}-P-A)!}}_{\text{number of arrangements}} \times \underbrace{e^{-\beta P\varepsilon_{pd}^{NS}} e^{-\beta A\varepsilon_{ad}^{NS}}}_{\text{weight of each state}} . \tag{2.7}$$

In figure 2.2(A) the weights of the four states are normalized by the weight of the empty state $Z(P,A)$. In equation 2.7 we use the notation ε_{xd} to characterize the binding energy of molecule X to DNA, and superscripts S and NS to signify specific or non-specific binding, respectively. $\Delta\varepsilon_{xd} = \varepsilon_{xd}^{S} - \varepsilon_{xd}^{NS}$ is the difference between the two. For the purposes of this simple model we have assumed that the reservoir for the activator molecules is the genomic DNA, although there is strong evidence that, in the case of the *lac* operon, many of the activators (cAMP receptor proteins; CRPs) are actually in the cytoplasm [38]. In contrast, as will be seen in the following paper [2], in our actual applications of thermodynamic models to real operons, the question of whether the reservoir is non-specific DNA or the cytoplasm never arises.

As usual, to compute the probability of interest, we construct the ratio of the sum of weights for all those outcomes that are favorable (i.e., polymerase bound to the promoter) to the sum of weights over the total set of outcomes $Z_{tot}(P, A)$. This results in a value of p_{bound} that adopts precisely the form described in equation 2.5. The regulation factor, $F_{reg}(A)$, is given by

$$F_{reg}(A) = \frac{1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{ad}} e^{-\beta\varepsilon_{ap}}}{1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{ad}}}, \tag{2.8}$$

where we have made the additional assumption that $N_{NS} \gg P, A$. Note that if the adhesive interaction between polymerase and activator goes to zero, the regulation factor itself goes to unity. Furthermore, for negative values of this adhesive interaction (i.e., activator and polymerase like to be near each other) the regulation factor is greater than one, which translates into an apparent increase in the number of polymerase molecules available for binding to the promoter. This claim can be seen more clearly if we define the fold-change in promoter activity as the ratio of the probability that RNAP is bound in the presence of transcription factors to the probability that it is bound in the absence of transcription factors: $fold - change = p_{bound}(P, A)/p_{bound}(P, A = 0)$. The fold-change is plotted in figure 2.2(B) for typical values of the adhesive interaction ε_{ap} and the other binding parameters, for the simple model in which the reservoir for CRP is assumed to be non-specific DNA.

Similar arguments can be made for the action of repressor molecules. Consider repression by R repressor molecules that can bind to an operator (with energy ε_{rd}^{S}) that overlaps with the promoter. By enumerating the different states with their associated weights in a way similar to that used in figure 2.2(A) and noting that the state where both the repressor and RNAP bind to their sites is not allowed, we can again derive the form for promoter occupation, equation 2.5, but this time with the regulation factor,

$$F_{reg}(R) = \frac{1}{1 + \frac{R}{N_{NS}} e^{-\beta\Delta_{rd}}}. \tag{2.9}$$

The above scheme can be extended further to describe co-regulation by two or more activators and/or repressors. For example, in the case of activation considered above, if the binding of the activator to its operator site is assisted itself by a helper protein, which might bind to an adjacent site [2], then the regulation factor still has the form given in equation 2.8 but with the number of activators, A, replaced by an effective

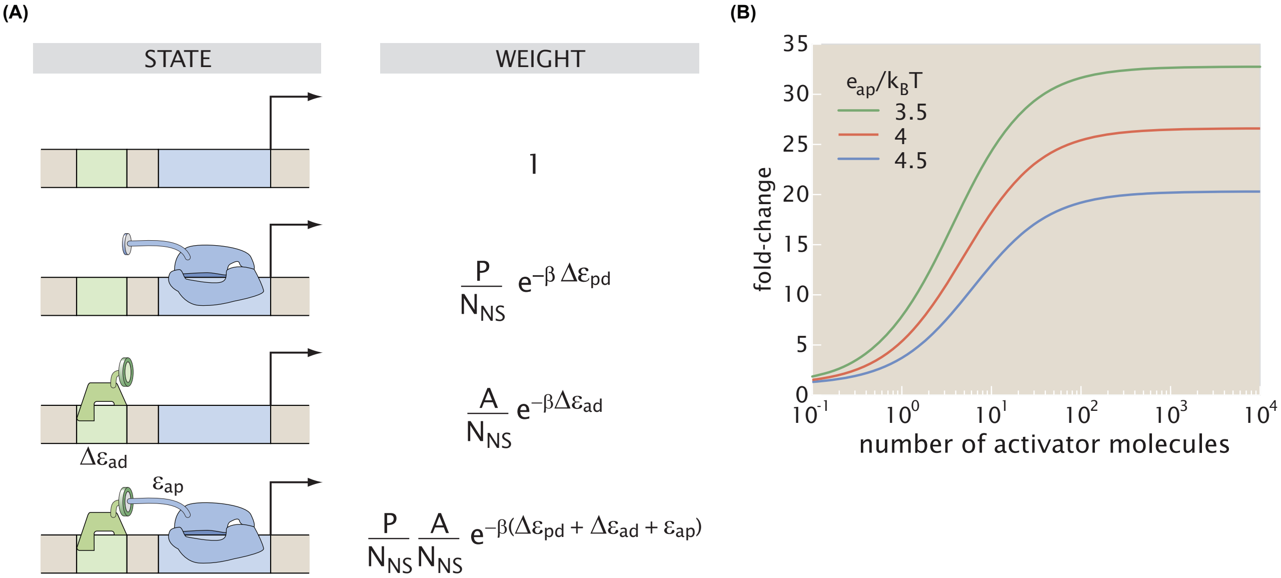


Figure 2.2: Statistical mechanics of recruitment. (A) Schematic showing the relationship between the various states of the promoter and its regulatory region, and their corresponding weights within the statistical mechanics framework. (B) Fold-change in promoter activity as a function of the number of activated (inducer-bound) CRP molecules, according to equations 7.8 and 2.8, for different values of the adhesive interaction energy between activator and RNAP. As in figure 2.1, $\Delta\varepsilon_{ad} = \beta^{-1}\ln\left(K_{ad}^{S}/K_{ad}^{NS}\right)$, with $K_{ad}^{NS} = 10000$ nM [39] and $K_{ad}^{S} = 0.02$ nM [40]. These *in vitro* numbers are chosen as a representative example to provide intuition for the action of activators. Applications to *in vivo* experiments are given in the accompanying paper [2]. Several different representative values of the adhesive interaction ε_{ad} that are consistent with measured activation are chosen to illustrate how activation depends upon this parameter.

number of activators

$$A' = A\frac{1 + \frac{H}{N_{NS}}e^{-\beta\varepsilon_{hd}}e^{-\beta\varepsilon_{ha}}}{1 + \frac{H}{N_{NS}}e^{-\beta\varepsilon_{hd}}}. \tag{2.10}$$

Note that the multiplicative factor in equation 2.10 has the same form as in equation 2.8 except that now the number of helper molecules, H, appears in the expression, and the interaction energy ε_{ha} refers to that between the helper molecules and activators. In fact, this is the generic expression describing the recruitment of one DNA binding protein by another, and it is not limited to activatorRNAP recruitment.

The introduction of the regulation factor enables a discussion of various regulatory motifs in a unified way, as made explicit by table 2.1. These examples will be discussed in the context of particular bacterial gene regulatory systems in the ensuing paper. The main point captured by this table is that the conceptual picture of thermodynamic models is identical regardless of regulatory motif and involves summing all of the relevant states. It culminates in the regulation factor which, as will be shown in the companion paper [2], is equal to the measurable fold-change of promoter activity.

As a final example, we consider the way in which DNA looping can play a role in dictating the regulation factor. Indeed, recent work by Vilar and Leibler [30] and Vilar and Saiz [41] and others [25, 42] has shown how the thermodynamic models can be applied to regulatory control by looping. In the accompanying paper [2], we apply these ideas to the particular question of how such regulation depends upon the distance between the two binding sites, but content ourselves here with a discussion of the conceptual basis. Two distinct looping scenarios are shown in figure 2.3. In case (A), a repressor molecule, which can bind to two distinct regions on the DNA, loops out the intervening region. The classic example of this mode of action is the Lac repressor. In case (B), one protein, such as CRP, favorably bends the DNA so that a second activator can contact RNAP, although paying a lower free energy cost than it would if it were acting alone. In both cases, the free energy cost associated with making a DNA loop is outweighed by the benefit of additional binding energy between the repressor and DNA [case (A)] and between the activator and RNAP [case (B)].

In summary, the statistical mechanical framework described here can be used to consider several different regulatory motifs [12, 13, 26, 29, 31, 32, 43], as showcased in table 2.1. In each of the cases considered in the table, the probability of promoter occupancy is given by equation 2.5, with the sole change from one case to the next being the form adopted by the regulation factor itself.

2.3 Conclusions and Future Prospects

We argue that as a result of the increasingly quantitative character of data on gene expression there is a corresponding need for predictive models. We have reviewed a series of general arguments about the way in which batteries of transcription factors work in generic ways to mediate transcriptional regulation. The models described here result in several important classes of predictions. The application of these ideas to particular bacterial scenarios forms the substance of the second article [2].

Though ideas like those presented here have the potential to serve as a quantitative framework for thinking

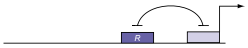
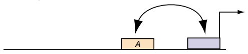
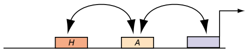
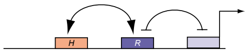
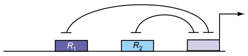
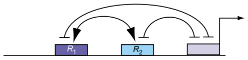
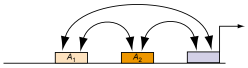
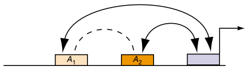
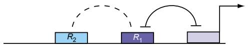
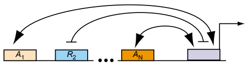
CASE	REGULATION FACTOR	
1. Simple repressor 	$(1+r)^{-1}$	$\left(1+\frac{[R]}{K_R}\right)^{-1}$
2. Simple activator 	$\frac{\left(1+a\,e^{-\beta\varepsilon_{ap}}\right)}{1+a}$	$\frac{\left(1+\frac{[A]}{K_A}\right)f}{1+\frac{[A]}{K_A}}$
3. Activator recruited by a helper (H) 	$\frac{1+a\frac{\left(1+h\,e^{-\beta\varepsilon_{ha}}\right)}{1+h}e^{-\beta\varepsilon_{ap}}}{1+a\frac{\left(1+h\,e^{-\beta\varepsilon_{ha}}\right)}{1+h}}$	$\frac{1+\frac{[H]}{K_H}+\frac{[A]}{K_A}f+\frac{[A]}{K_A}\frac{[H]}{K_H}f\,\omega}{1+\frac{[H]}{K_H}+\frac{[A]}{K_A}+\frac{[A]}{K_A}\frac{[H]}{K_H}\omega}$
4. Repressor recruited by a helper (H) 	$\left(1+\frac{1+h\,e^{-\beta\varepsilon_{hr}}}{1+h}r\right)^{-1}$	$\frac{1+\frac{[H]}{K_H}}{1+\frac{[H]}{K_H}+\frac{[R]}{K_R}+\frac{[R]}{K_R}\frac{[H]}{K_H}\,\omega}$
5. Dual repressors 	$(1+r_1)^{-1}(1+r_2)^{-1}$	$\left(1+\frac{[R_1]}{K_{R_1}}\right)^{-1}\left(1+\frac{[R_2]}{K_{R_2}}\right)^{-1}$
6. Dual repressors interacting 	$\left(1+r_1+r_2+r_1r_2e^{-\beta\varepsilon_{r_1r_2}}\right)^{-1}$	$\left(1+\frac{[R_1]}{K_{R_1}}+\frac{[R_2]}{K_{R_2}}+\frac{[R_1]}{K_{R_1}}\frac{[R_2]}{K_{R_2}}\,\omega\right)^{-1}$
7. Dual activators interacting 	$\frac{\left(1+a_1e^{-\beta\varepsilon_{a_1p}}+a_2e^{-\beta\varepsilon_{a_2p}}+a_1a_2e^{-\beta(\varepsilon_{a_1p}+\varepsilon_{a_2p}+\varepsilon_{a_1a_2})}\right)}{1+a_1+a_2+a_1a_2e^{-\beta\left(\varepsilon_{a_1p}+\varepsilon_{a_2p}\right)}}$	$\frac{1+\frac{[A_1]}{K_{A_1}}f_1+\frac{[A_2]}{K_{A_2}}f_2+\frac{[A_1]}{K_{A_1}}\frac{[A_2]}{K_{A_2}}\,f_1f_2\omega}{1+\frac{[A_1]}{K_{A_1}}+\frac{[A_2]}{K_{A_2}}+\frac{[A_1]}{K_{A_1}}\frac{[A_2]}{K_{A_2}}\,\omega}$
8. Dual activators cooperating via looping 	$\frac{1+a_1e^{-\beta\varepsilon_{a_1p}}+a_2e^{-\beta\varepsilon_{a_2p}}+a_1a_2e^{-\beta(\varepsilon_{a_1p}+\varepsilon_{a_2p}+\varepsilon_{a_1a_2})}}{1+a_1+a_2+a_1a_2e^{-\beta\left(\varepsilon_{a_1p}+\varepsilon_{a_2p}\right)}}$	$\frac{1+\frac{[A_1]}{K_{A_1}}f_1+\frac{[A_2]}{K_{A_2}}f_2+\frac{[A_1]}{K_{A_1}}+\frac{[A_2]}{K_{A_2}}f_1f_2\omega}{\left(1+\frac{[A_2]}{K_{A_2}}\right)\left(1+\frac{[A_1]}{K_{A_1}}\right)}$
9. Repressor with two DNA binding units and DNA looping 	$\left(1+r_m+\frac{r_m}{1+r_a}e^{-\beta\left(\varepsilon_{r_ad}+F_{loop}\right)}\right)^{-1}$	$\frac{1+\frac{[R]}{K_a}}{\left(1+\frac{[R]}{K_m}\right)\left(1+\frac{[R]}{K_a}\right)+\frac{[R][L]}{K_mK_a}}$
10. N non-overlapping activators and/or repressors acting independently on RNAP 	$F_{reg1}\times F_{reg2}\times\ldots\times F_{reg3}$	$F_{reg1}\times F_{reg2}\times\ldots\times F_{reg3}$

Table 2.1: Regulation factors for several different regulatory motifs. In the schematics of the motifs appearing in the first column, the inverted T symbol indicates repression, arrows represent activation, and a dashed line is for DNA looping. The second column gives the regulation factor in terms of the number of transcription factors (TFs) in the cell and their binding energies, and the third column provides a translation of the regulation factor into the language of concentrations and equilibrium dissociation constants (used in the following paper [2]). For an arbitrary TF we introduce the following notation: in the second column, x is the combination $\frac{X}{N_{NS}}e^{-\beta\Delta\varepsilon_{xd}}$, and $[X]$ in the third column denotes the concentration of transcription factor X. $K_X = [X]/x$ is the effective equilibrium dissociation constant of the TF and its operator sequence on the DNA. Furthermore, in the third column we introduce $f = e^{-\beta\varepsilon_{xp}}$ for the glue-like interaction of a TF and RNAP, and $\omega = e^{-\beta\varepsilon_{x_1x_2}}$ for the interaction between two TFs. In cases 8 and 9, F_{loop} is the free energy of DNA looping, ω in case 8 is defined as $e^{-\beta F_{loop}}$, while $[L]$ in case 9 is the combination $\frac{N_{NS}}{V_{cell}}e^{-\beta F_{loop}}$, V_{cell} being the volume of the cell.

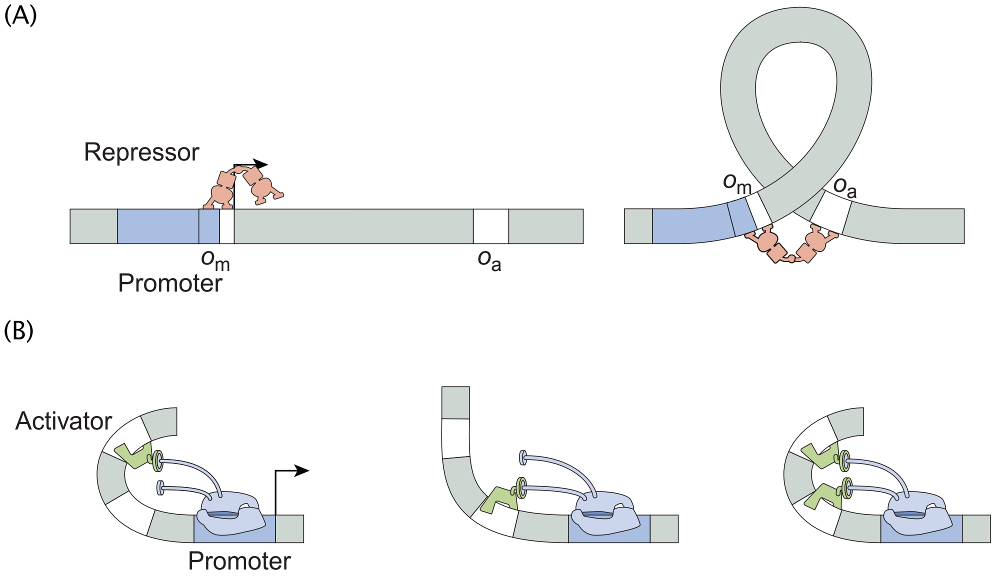


Figure 2.3: DNA bending in transcription regulation. (A) DNA looping enables Lac repressor to bind to the main and the auxiliary operators simultaneously, thereby increasing the weight of the states in which the promoter is unoccupied. This leads to stronger repression than in the single operator case. (B) DNA bending by the activator leads to cooperative binding of the two activators because the free energy cost of bending is paid only once. This leads to a boost in activation above that provided by independent binding of the two activators [44].

about transcriptional regulation, there are several outstanding issues. Some especially troubling features of these models are: (i) what are the precise conditions under which equilibrium assumptions are acceptable? (ii) When can the probability of RNAP binding at a promoter serve as a surrogate for gene expression itself? (iii) What is the role of fluctuations? (iv) These models pretend that the basal transcription apparatus is a single molecule that interacts with transcription factors, whereas the transcription apparatus is a complex that is itself probably subject to recruitment for its assembly. Despite these concerns, our view is that thermodynamic models have long demonstrated their utility and it will be of great interest to carefully explore their consequences experimentally. Case studies using the thermodynamic models are reviewed in the accompanying paper [2].

Bibliography

[1] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[2] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[3] M. Ptashne and A. Gann. *Genes and signals.* Cold Spring Harbor Laboratory Press, New York, 2002.

[4] M. Ptashne. *A genetic switch: Phage lambda revisited.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 3rd edition, 2004.

[5] G. R. Bellomy, M. C. Mossing, and M. T. J. Record. Physical properties of DNA in vivo as probed by the length dependence of the lac operator looping process. *Biochemistry*, 27(11):3900–6, 1988.

[6] S. Oehler, M. Amouyal, P. Kolkhof, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. Quality and position of the three *lac* operators of *e. Coli* define efficiency of repression. *EMBO J*, 13(14):3348–55, 1994.

[7] J. MÜLler, S. Oehler, and B. MÜLler-Hill. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol*, 257(1):21–9, 1996.

[8] D. H. Lee and R. F. Schleif. In vivo DNA loops in aracbad: Size limits and helical repeat. *Proc Natl Acad Sci U S A*, 86(2):476–80, 1989.

[9] D. E. Lewis and S. Adhya. *in vitro* repression of the gal promoters by galr and hu depends on the proper helical phasing of the two operators. *J Biol Chem*, 277(4):2498–504, 2002.

[10] A. Hochschild and M. Ptashne. Interaction at a distance between lambda repressors disrupts gene activation. *Nature*, 336(6197):353–7, 1988.

[11] J. K. Joung, D. M. Koepp, and A. Hochschild. Synergistic activation of transcription by bacteriophage lambda ci protein and e. Coli camp receptor protein. *Science*, 265(5180):1863–6, 1994.

[12] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci U S A*, 100(13):7702–7, 2003.

[13] T. Kuhlman, Z. Zhang, M. H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *escherichia coli*. *Proc Natl Acad Sci U S A*, 104(14):6043–8, 2007.

[14] S. Kalir, J. Mcclure, K. Pabbaraju, C. Southward, M. Ronen, S. Leibler, M. G. Surette, and U. Alon. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, 292(5524):2080–3, 2001.

[15] M. T. Laub, H. H. Mcadams, T. Feldblyum, C. M. Fraser, and L. Shapiro. Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, 290(5499):2144–8, 2000.

[16] M. N. Arbeitman, E. E. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White. Gene expression during the life cycle of drosophila melanogaster. *Science*, 297(5590):2270–5, 2002.

[17] S. Small, A. Blair, and M. Levine. Regulation of even-skipped stripe 2 in the drosophila embryo. *EMBO J*, 11(11):4047–57, 1992.

[18] E. H. Davidson. *Genomic regulatory systems: Development and evolution.* Academic Press, San Diego, CA, 2001.

[19] S. B. Carroll, J. K. Grenier, and S. D. Weatherbee. *From DNA to diversity: Molecular genetics and the evolution of animal design.* Blackwell Science, Malden, Mass., 2001.

[20] G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A*, 79(4):1129–33, 1982.

[21] M. A. Shea and G. K. Ackers. The or control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol*, 181(2):211–30, 1985.

[22] P. H. Von Hippel, A. Revzin, C. A. Gross, and A. C. Wang. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: i. the *lac* operon: Equilibrium aspects. *Proc Natl Acad Sci U S A*, 71(12):4808–12, 1974.

[23] S. M. Law, G. R. Bellomy, P. J. Schlax, and M. T. J. Record. *in vivo* thermodynamic analysis of repression with and without looping in *lac* constructs. Estimates of free and local *lac* repressor concentrations and of physical properties of a region of supercoiled plasmid DNA *in vivo*. *J Mol Biol*, 230(1):161–73, 1993.

[24] A. Ben-Naim. Cooperativity in binding of proteins to DNA. *J Chem Phys*, 107(23):10242–10252, 1997.

[25] I. B. Dodd, K. E. Shearwin, A. J. Perkins, T. Burr, A. Hochschild, and J. B. Egan. Cooperativity in long-range gene regulation by the lambda ci repressor. *Genes Dev*, 18(3):344–54, 2004.

[26] A. Bakk, R. Metzler, and K. Sneppen. Sensitivity of or in phage lambda. *Biophys J*, 86(1 Pt 1):58–66, 2004.

[27] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'conner, D. W. Noble, and P. H. Von Hippel. Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound *escherichia coli lac* repressor *in vivo*. *Proc Natl Acad Sci U S A*, 74(10):4228–32, 1977.

[28] W. Runzi and H. Matzura. *in vivo* distribution of ribonucleic acid polymerase between cytoplasm and nucleoid in escherichia coli. *J Bacteriol*, 125(3):1237–9, 1976.

[29] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003.

[30] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[31] U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc Natl Acad Sci U S A*, 99(19):12015–20, 2002.

[32] A. M. Sengupta, M. Djordjevic, and B. I. Shraiman. Specificity and robustness in transcription control networks. *Proc Natl Acad Sci U S A*, 99(4):2072–7, 2002.

[33] R. F. Bruinsma. physics of protein-DNA interaction. In H. Flyvbjerg, F. Julicher, P. Ormos, and F. David, editors, *Physics of bio-molecules and cells*. Springer-Verlag, 2002.

[34] I. J. Molineux. No syringes please, ejection of phage t7 DNA from the virion is enzyme driven. *Mol Microbiol*, 40(1):1–8, 2001.

[35] M. T. J. Record, W. Reznikoff, M. Craig, K. Mcquade, and P. Schlax. Escherichia coli rna polymerase (sigma70) promoters and the kinetics of the steps of transcription initiation. In F. C. Neidhardt, R. CURTISS III, J. L. INGRAHAM, E. C. C. LIN, K. BROOKS LOW, B. Magasanik, W. S. REZNIKOPP, M. Riley, M. SCHAECHTER, and H. E. UMBARGER, editors, *In escherichia coli and salmonella cellular and molecular biology*, pages 792–821. ASM Press, Washington, DC, 1996.

[36] M. Liu, G. Gupte, S. Roy, R. P. Bandwar, S. S. Patel, and S. Garges. Kinetics of transcription initiation at lacp1. Multiple roles of cyclic amp receptor protein. *J Biol Chem*, 278(41):39755–61, 2003.

[37] C. J. Dayton, D. E. Prosen, K. L. Parker, and C. L. Cech. Kinetic measurements of escherichia coli rna polymerase association with bacteriophage t7 early promoters. *J Biol Chem*, 259(3):1616–21, 1984.

[38] D. I. Cook and A. Revzin. Intracellular location of catabolite activator protein of escherichia coli. *J Bacteriol*, 141(3):1279–83, 1980.

[39] M. G. Fried and D. M. Crothers. Equilibrium studies of the cyclic amp receptor protein-DNA interaction. *J Mol Biol*, 172(3):241–62, 1984.

[40] P. Wong, S. Gladney, and J. D. Keasling. Mathematical model of the lac operon: Inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnol Prog*, 13(2):132–43, 1997.

[41] J. M. Vilar and L. Saiz. DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. *Curr Opin Genet Dev*, 15(2):136–44, 2005.

[42] R. R. Seabold and R. F. Schleif. Apo-arac actively seeks to loop. *J Mol Biol*, 278(3):529–38, 1998.

[43] E. Aurell, S. Brown, J. Johanson, and K. Sneppen. Stability puzzles in phage lambda. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65(5 Pt 1):051914, 2002.

[44] J. K. Joung, L. U. Le, and A. Hochschild. Synergistic activation of transcription by escherichia coli camp receptor protein. *Proc Natl Acad Sci U S A*, 90(7):3083–7, 1993.

Chapter 3

Applications of Thermodynamic Models of Transcriptional Regulation

This chapter is a reproduction of reference [1] which was published together with the previous chapter corresponding to reference [2].

With the increasing amount of experimental data on gene expression and regulation, there is a growing need for quantitative models to describe the data and relate them to their respective context. Thermodynamic models provide a useful framework for the quantitative analysis of bacterial transcription regulation. This framework can facilitate the quantification of vastly different forms of gene expression from several well-characterized bacterial promoters that are regulated by one or two species of transcription factors; it is useful because it requires only a few parameters. As such, it provides a compact description useful for higher-level studies (e.g., of genetic networks) without the need to invoke the biochemical details of every component. Moreover, it can be used to generate hypotheses on the likely mechanisms of transcriptional control.

3.1 Introduction

Biology is undergoing a transformation from a "component-centric" focus on the individual parts toward a "system-level" focus on how a limited number of parts work together to perform complex functions. For gene regulation, this theme has been discussed extensively in the context of simple genetic circuits [3–6] in addition to complex, developmental networks [7]. The functional properties of a genetic circuit often critically depend on the degree of cooperativity in the interactions between the molecular components [8]. For gene regulation, this cooperativity is dictated to a large extent by the architecture of the cis-regulatory region, [9] and the specific mechanism of transcriptional activation or repression [10], which is mediated through interactions among various transcription factors (TFs) and the RNA polymerase (RNAP) complex. Often, even qualitative features of a gene circuit (e.g., whether a circuit can be bistable or whether it can spontaneously oscillate) cannot be determined without quantitative knowledge of the transcriptional regulation of key genes in the circuit [5].

Predicting the expression level of genes directly from the underlying biochemistry and biophysics is a difficult task. This is due most notably to ignorance of many biochemical parameters, especially their relevant *in vivo* values. However, the thermodynamic model reviewed in the preceding article [2] yields several general mathematical forms for the dependence of the fold-change in gene expression on the concentration(s) of the TF(s) regulating transcription. These general forms contain only a few parameters characterizing the effective interactions between the molecular players. Thus, from a practical standpoint, it is expedient to quantify the transcriptional regulation of a gene by fitting expression data to the appropriate model function in order to obtain effective parameters that best describe the promoter [10,11]. This procedure might be useful even when the simplifying assumptions made by the thermodynamic models are not satisfied [2]. By analyzing gene expression data within the thermodynamic framework, one can elucidate whether an assumed set of interactions between TFs and RNAP can consistently explain the data. Failure of the analysis can suggest important missing ingredients, such as unknown mechanisms of cooperativity, whereas success can lead to predictions for new experiments (e.g., how operator deletion would affect gene expression).

There has been much recent progress in understanding the mechanistic aspect of bacterial gene regulation [10]. However, the systematic quantification of gene expression is still in its infancy. In this paper, we review several experimentally characterized cis-regulatory systems in bacteria. For each case, we provide what we believe to be the most appropriate form for the dependence of the promoter activity on the TF concentration(s). For each system, we show graphically how the expected form depends on the effective parameters. We hope to demonstrate how the thermodynamic models can provide a direct link between the arrangements of interactions in a promoter region and the quantitative characteristics of gene expression.

3.1.1 Quantitative characteristics of activation and repression

Our quantitative discussion focuses on several well characterized bacterial promoters controlled by one or two species of TFs. We use the results of the thermodynamic model listed in table 2.1 of the preceding chapter [2]. We make the additional simplifying assumption that the *in vivo* promoters are weak, so that even at full activation the equilibrium gene expression is still small (e.g., $< 10\%$ of the strongest promoters). Indeed, for a large number of bacterial promoters, the expression is small in the exponential growth phase when compared with the expression of the ribosomal genes, for example, which are fully turned on [11]. In this weak promoter limit, the fold-change in promoter activity (henceforth simply referred to as "fold-change") is given directly by the regulation factor (F_{reg}) listed in table 2.1. We will consider two types of activators: those activators that recruit RNAP to its promoter, and those that stimulate the transition rate of bound RNAP from a closed to an open complex. Even though the latter is a kinetic effect, its impact on the overall promoter activity (e.g., transcription initiation rate) can, nevertheless, be effectively described by the thermodynamic model in the weak promoter limit that we study.

3.1.2 Simple activation

The simplest example of activation involves the binding of an induced TF to a single operator site, and the subsequent recruitment of RNAP. This is the case with the *lac* promoter of E.coli, shown in figure 3.1(A) (in the absence of the *lac* repressor). The activating TF is a CRP (cAMP receptor protein) dimer in complex with the inducer cAMP [12, 13]. We will denote this complex by CRP_2^* and use $*$ to indicate the activated form of a TF. Case 2 in table 2.1 gives the mathematical form of the expected fold-change for this situation with $[A] = [\text{CRP}_2^*]$, and figure 3.1(B) plots its dependence on the induced dimer concentration. The two parameters of the model are the effective *in vivo* dissociation constant (K_A) between CRP and the operator, and the enhancement factor (f), which characterizes the degree of stimulation in transcription resulting from operator-bound CRP. These are readily revealed in a log-log plot of the relative promoter activity against the cellular concentration of the induced activator, $[\text{CRP}_2^*]$. As long as the range of $[\text{CRP}_2^*]$ probed is sufficiently broad, one can read the enhancement factor (f) from the graph as the maximal fold-change between full activation at saturating levels of $[\text{CRP}_2^*]$ and basal activity at low levels of $[\text{CRP}_2^*]$. One can also read off the effective dissociation constant (K_A) as the value of $[\text{CRP}_2^*]$ at half-activation. The steepness of the transition region — called the "sensitivity" (or "gain") in the literature [14] — plays an important role in the function of genetic circuits. Here, we quantify transcriptional sensitivity by the log-log slope (s) at the mid-point of the transition region. $s \leq 1$ for promoters containing a single operator, and s approaches 1 for only very large values of f. In contrast, functions such as amplification, bistability or spontaneous oscillation all require circuit components to have high sensitivity, with a value of $s > 1$ [8].

3.1.3 Cooperative activation

TFs often have domains that enable interaction with one another when bound to adjacent operator sites, and this interaction can result in cooperativity in transcriptional activation. The P_{RM} promoter of phage lambda, shown in figure 3.2(A), is such an example [3]. Binding of the dimeric lambda repressor cI to the operator O_{R2} (the "activator" site) stimulates transcription, and binding of cI to the upstream operator O_{R1} (the "helper" site) helps to recruit cI to I_{R2}. The expected fold-change (case 3 in table 2.1 with $[A] = [H] = [cI_2]$, $K_H = K_{R1}$ and $K_A = K_{R2}$) depends on the affinities K_{R1} and K_{R2} of cI to the two operators, the cooperative interaction (ω) between the two operatorbound cI dimers, and the enhancement factor f due to the O_{R2}-bound cI. It is shown in the log-log plot of figure 3.2(B) (thick solid line) as a function of $[cI_2]/K_{R2}$.

To quantify the possible role of the auxiliary operator O_{R1}, we also plot in figure 3.2(B) the fold-change for different ratios of K_{R1} and K_{R2}. Comparing these curves, it is clear that the auxiliary operator O_{R1} does not change the degree of full activation, given by f. The most significant feature of this dual-activator system is perhaps the increase in the log-log slope of the transition region (compared with the extreme cases) for intermediate values of K_{R2}/K_{R1}. In fact, for the realistic parameter of $K_{R2}/K_{R1} \approx 25$ (thick solid line in figure 3.2(B)), we have a sensitivity of $s \approx 0.93$. This is close to the maximum attainable for this system,

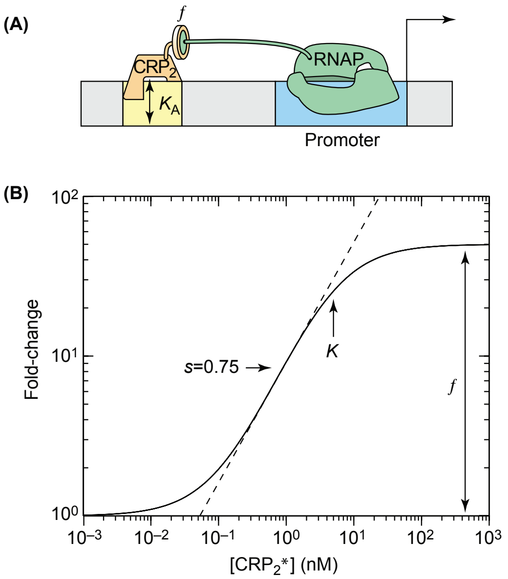


Figure 3.1: Simple activation. (A) Cis-regulatory architecture for transcriptional activation involving a single CRP operator, as found in the *lac* operon. The yellow box denotes the operator site and the blue box corresponds to the promoter. The DNA-binding affinity of the transcription factor for its operator is described by the *in vivo* dissociation constant K_A, which is the TF concentration at which the operator occupancy is half-maximal. The activator recruits RNAP through protein-protein interactions (schematically drawn as interacting protein subunits). (B) log-log plot of the fold-change in gene expression as a function of the induced CRP dimer concentration, $[\mathrm{CRP}_2^*]$. The maximum log-log slope in the transition region, which is defined as the sensitivity (s), is highlighted with the dashed line and is equal to 0.75. This plot was generated using $K_A = 5$ nM, $f = 50$. These parameter values were estimated from experiments similar to those of Setty et al. [15], who measured β-galactosidase activity as a function of extracellular cAMP concentration in E. coli MG1655 cells, but with the additional deletion of the *cyaA* gene which encodes adenyl cyclase [16]. The enhancement factor obtained is consistent with that of others [17]. The estimated value of the effective dissociation constant K_A is dependent on the literature values for several biochemical parameters concerning cAMP binding and transport, and is not expected to be accurate to within a factor of 2. (For comparison, previous *in vitro* measurement of the CRP-operator affinity has ranged from 0.001 nM to 50 nM depending on the ionic strength of the assay [18–20].)

with its small enhancement factor ($f \approx 11$), and is nearly double the maximum sensitivity ($s \approx 0.54$) for the promoter with $\mathrm{O_{R2}}$ only (thin solid line in figure 3.2(B). For TFs with larger values of ω and f, this cis-regulatory construct can, in principle, provide more sensitivity, with s approaching 2.

The same cis-regulatory design can be used to implement co-activation — one of the simplest forms of signal integration — if the two operators are targets of two distinct TF species. A possible example of this is the variant of *E. coli*s *melAB* promoter studied by Wade et al. [21] (see figure 3.3(A), where transcription is stimulated by an induced MelR dimer bound to the weak proximal operator, O_2. Meanwhile, CRP bound to the upstream operator O_1 helps recruit MelR but does not directly participate in activation. Assuming that the induction of MelR by melibiose results in an increase in MelR-operator binding affinity, we expect the form of the co-dependence to be given by case 3 in table 2.1, but with $[A] = [\mathrm{MelR}_2^*]$, $[H] = [\mathrm{CRP}_2^*]$ and $K_H = K_1$, $K_A = K_2$. The fold-change is plotted against the induced CRP concentration on the log-log plot of figure 3.3(B) for different concentrations of the induced MelR. To better visualize the co-dependence on CRP and MelR, it is useful to plot the fold-change as a three-dimensional plot; see figure 3.3(C). The transition region (the yellow band) is clearly dependent on both TFs. Consider a simplified situation where CRP and MelR can each take on two possible concentrations a pair of "low" and "high" values. Then it is possible to choose the pair of concentrations (e.g., those marked by the 4 open circles in figure 3.3(C)) such that the fold-change is large (the green region) only when both concentrations are high. This mimics a logical AND function of the two inputs [22]. It is also possible to choose the pair of concentrations as marked by the four solid circles such that the fold-change is large (the green region) unless both concentrations are "low". The latter choice mimics a logical OR function. The flexibility of this cis-regulatory scheme makes the shape of the fold-change readily evolvable [23] (e.g., between the AND/OR functions) by merely altering the operator sequences that encode the values of K_1 and K_2.

3.1.4 Synergistic activation

An alternative mechanism for co-activation is synergistic or dual activation [27–29], where two operator-bound TFs can simultaneously contact different subunits of RNAP and activate transcription. This mechanism is limited to TFs that can activate transcription at different locations relative to the core promoter. Prominent examples of such synergistic activation in the bacterial literature [27–33] all involve the activator CRP because it can recruit RNAP from multiple locations at varying distances upstream of the promoter [10, 34].

The synthetic promoter studied by Joung et al. [29] contained two operators: one for cI proximal to the core promoter (O_2) and the other for CRP at an upstream operator (O_1) (see figure 3.4(A)). The data from the study by Joung et al. support the model where each operator-bound activator can independently interact with RNAP and enhance transcription [29]. The expected fold-change is given by case 8 in table 2.1 (with $[A_1] = [\mathrm{CRP}_2^*]$, $[A_2] = [\mathrm{cI}_2]$, $K_{A1} = K_1$, $K_{A2} = K_2$ and $\omega = 1$) and shown in the log-log plot of figure 3.4(B) as a function of $[\mathrm{CRP}_2^*]$ for various cI concentrations. Note that, since $\omega = 1$, the dependence

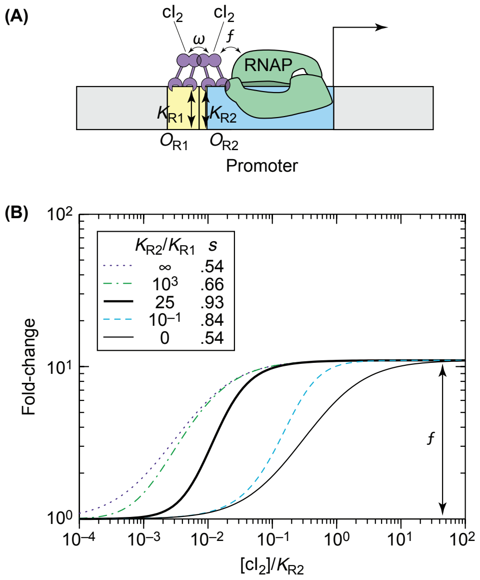


Figure 3.2: Enhanced sensitivity by cooperative activation. (A) Cis-regulatory architecture for cooperative transcriptional activation in phage lambda $\mathrm{P_{RM}}$ promoter. Here, we are considering $\mathrm{P_{RM}}$ alone without the upstream $\mathrm{P_R}$ promoter [3] or the upstream PL region, which affects $\mathrm{P_{RM}}$ activity through DNA looping [24]. We also neglect the operator $\mathrm{O_{R3}}$, which has very weak affinity to cI in the absence of P_L [24]. The yellow boxes denote the operator sites O_{R1}, O_{R2} and the blue box corresponds to the promoter. The DNA-binding affinity of $\mathrm{cI_2}$ for O_{R1} and O_{R2} is described by the dissociation constants K_{R1} and K_{R2}, respectively. The activator stimulates transcription and cI dimers interact with one another through intimate, cooperative interactions, both of which are indicated by overlapping protein-protein domains. (B) log-log plot of the fold-change in gene expression as a function of $\mathrm{cI_2}$ concentration for different ratios of K_{R2}/K_{R1}. The maximum log-log slopes (s) for the different curves are listed in the legend. The promoter with $K_{R2}/K_{R1} = 0$ corresponds to a deletion of O_{R1}, and the regulation function for this case (thin solid line) is identical to the single operator case shown in figure 3.1. If this promoter has a very small K_{R1} (i.e., strong O_{R1}), then the onset of full activation will be shifted to smaller cI concentrations (dotted line). The latter corresponds effectively to a stronger O_{R2} site, with dissociation constant K_{R2}/ω. These plots are generated using $f \approx 11$ [25] and $\omega \approx 100$ [26] as extracted from *in vitro* biochemical studies. The absolute *in vivo* values of the K values are not known (which is why the concentration is expressed in terms of $[\mathrm{cI_2}]/K_{R2}$). However, the ratio $K_{R2}/K_{R1} \approx 25$ (thick solid line) can be deduced from the *in vitro* results [26]. The transition region is steepest when $v \gg f$ and $K_{R2}/K_{R1} \approx f$. We note that the parameters for P_{RM} are nearly optimal for enhanced sensitivity.

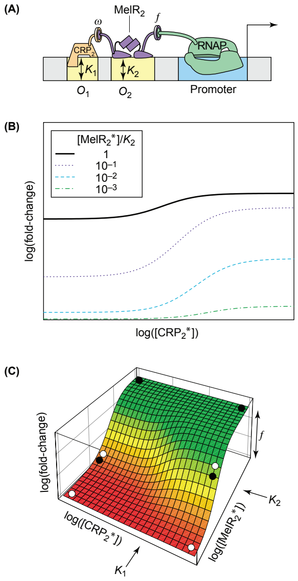


Figure 3.3: Cooperative co-activation. (A) Cis-regulatory construct for co-activation by CRP and MelR. The figure shows the truncated JK15 version of *melAB* promoter studied by Wade et al. [21]. The full *melAB* promoter is more complicated due to the presence of multiple MelR operators. However, the co-activation pattern is similar to that of JK15 discussed here. The yellow boxes denote the operator sites O_1, O_2 and the blue box corresponds to the promoter. The DNA-binding affinity of CRP2 for O_1 and MelR2 for O_2 is described by the dissociation constant K_1 and K_2, respectively. MelR can recruit RNAP (drawn with proteinprotein contacts) and cooperative interaction between MelR$_2$ and CRP$_2$ is indicated by interacting protein subunits. (B) log-log plot of the fold-change in gene expression as a function of activated CRP dimer concentration [CRP$_2^*$] for different activated MelR dimer concentrations [MelR$_2^*$]. Since none of the parameters f, ω, and K values have been determined experimentally, the scales of the plot can only be expressed relative to these parameters. Nevertheless, the plot reveals important qualitative predictions by the thermodynamic model (e.g., the dependence of the maximal CRP-dependent fold-change on the MelR concentration). (C) Three-dimensional log-log plot of the fold-change in gene expression as a function of both CRP$_2$ and MelR$_2$. For different choices of “high” and “low” concentration (the four combinations of “high/low” for these two TFs form a rectangle), the same *melAB* promoter can serve as an OR function (solid circles) or an AND function (open circles).

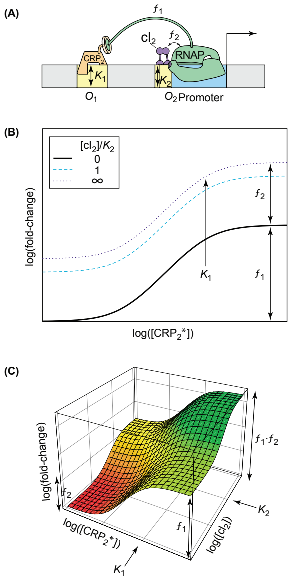


Figure 3.4: Synergistic co-activation. (A) Cis-regulatory architecture for synergistic co-activation in synthetic promoters [29]. The yellow boxes denote the operator sites O_1, O_2 and the blue box corresponds to the promoter. The DNA-binding affinity of CRP$_2$ for O_1 and cI$_2$ for O_2 is described by the dissociation constants K_1 and K_2, respectively. Each activator can independently interact with RNAP and enhance transcription at different strengths f_1, f_2 (as shown with interacting protein-protein subunits). (B) log-log plot of the fold-change in gene expression as a function of [CRP$_2^*$] for different concentrations of [cI$_2$]. (C) Three-dimensional log-log plot of the fold-change in gene expression as a function of both CRP$_2$ and cI$_2$. Note that on log scale, the product appears as an additive shift.

of gene expression on [CRP$_2^*$] is independent of [cI$_2$], except for an overall vertical shift. This is a reflection of the multiplicative nature of independent synergistic activation. An alternative way of visualizing the same result is the three-dimensional plot of figure 3.4(C).

In another experiment by Joung et al. [27], both the proximal site (O_2) and the distal site (O_1) were engineered to bind CRP (see figure 3.5(A), left). An important result of these experiments was that the fold-change with both CRP operators is larger than the product of the fold-changes with one operator alone. This is not consistent with the independent recruitment assumption and suggests additional cooperativity ($\omega > 1$). A possible mechanism proposed by Joung et al. is that DNA bending (see figure 3.5(A), right) induced by the CRP bound to the proximal operator O_2 facilitates the upstream CRP interaction with RNAP, without any direct protein-protein interaction between the two TFs. This cooperative effect can be included in the thermodynamic model as shown in case 8 of table 2.1 (with $[A1] = [A2] = [\mathrm{CRP}_2^*]$, $K_{A1} = K_1$, $K_{A2} = K_2$ and $\omega > 1$) regardless of the specific molecular mechanism. Similar to the case of activation by cI, the expression level is most sensitive when the K values for the two binding sites are equal. In figure 3.5(B), we plot the expected fold-change, with $K_1 = K_2$ and different values of ω. The extra cooperativity increases both the enhancement factor ($\omega \times f_1 \times f_2$) and the sensitivity ($s$) of the transition region.

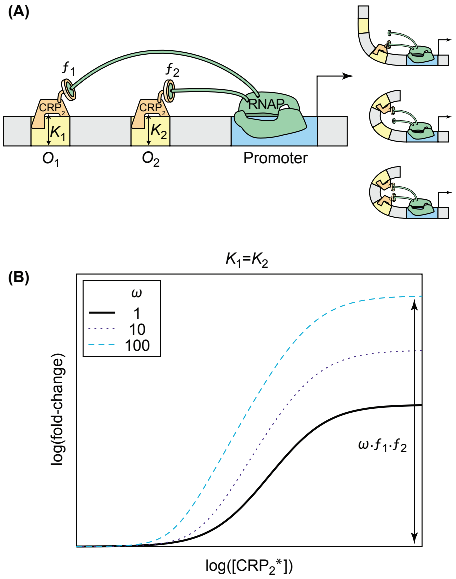


Figure 3.5: Enhancement of sensitivity by synergistic activation. (a) To the left is the cis-regulatory architecture for synergistic activation by the same TF in synthetic promoters [27]. The yellow boxes denote the operator sites O_1, O_2 and the blue box corresponds to the promoter. The DNA-binding affinity of CRP_2 for O_1 and O_2 is described by the dissociation constants K_1 and K_2, respectively. Activators at each operator can recruit RNAP independently at different strengths f_1, f_2 (as shown with interacting proteinprotein subunits). As illustrated to the right, the binding of CRP to proximal O_2 bends DNA and facilitates the bent interaction of RNAP to CRP bound at upstream O_1. (b) log-log plot of the fold-change in gene expression as a function of $[\mathrm{CRP}_2^*]$ for equal dissociation constants ($K_1 = K_2$). We have included the additional cooperativity ω that can occur when the binding of CRP to O_1 promotes the interaction of RNAP to CRP bound at O_2. The additional cooperativity simultaneously increases the maximal fold-change to $\omega \times f_1 \times f_2$ and enhances the transcriptional sensitivity in the transition region.

3.1.5 Simple repression

The simplest example of repression involves the binding of a TF to a single operator site that interferes with the binding of RNAP to the core promoter. This is the case in the truncated lac promoter (e.g., *lacUV5*) which has only the main operator, O_m, of LacI located closely downstream of the core promoter (figure 3.6(A)) [35]. The expected fold-change is given by case 1 of table 2.1, with $[R] = [\mathrm{LacI}_4]$, $K_R = K_m$ and only one unknown parameter, K_m, characterizing the effective dissociation constant of the operator O_m. Here, it is possible to compute K_m [36] directly from the experimental data of Oehler et al. [35], because the cellular concentration of LacI was quantified. In fact, because Oehler et al. characterized gene expression at two distinct LacI concentrations, the two data points can be used to check the consistency of the thermodynamic model.

This analysis was performed for the three lac operator sequences O_1, O_2 and O_3 studied in [35] (results shown in figure 3.6(B)). We note that the K_m values obtained, $K_1 \approx 0.22$ nM, $K_2 \approx 2.7$ nM and $K_3 \approx 110$ nM for the three operators, are significantly different from, for example, the results $K1 \approx 10^{-3}$ nM, $K_2 \approx 10^{-2}$ nM and $K_3 \approx 0.016$ nM to 1 nM obtained from *in vitro* assays [37–39]. These results underscore the fact that the relevant TF-operator binding constant for the thermodynamic model is not given by the *in vitro* measurement — even if the appropriate physiological conditions are used — but must be corrected for by considering the interaction of the TF with the genomic background [2, 40]. Consistent with the theoretical expectation, the ratios of the K values are in reasonable agreement between the *in vivo* and the *in vitro* results. We note also that the expected range of promoter activities is much larger than those for the activator-controlled promoters described above. This follows from the strong excluded volume interaction between the repressor and RNAP, such that more repressor proteins generally lead to stronger repression; whereas in activation more activator protein does not lead to more activation beyond the enhancement factor (f), which is set by the weak activatorRNAP interaction.[1] By contrast, the sensitivity is still limited to $s \leq 1$ with a single operator site.

3.1.6 Repression by DNA looping

For the wild-type *lac* promoter, the degree of repression exceeds 1000-fold with only ~ 10 repressor molecules in a cell [13]. This is substantially larger than the < 100-fold repression achievable by the best of the truncated promoters (figure 3.6) at the same repressor concentration. The additional repression is facilitated by the stabilization of the O_m-bound Lac tetramer, which can simultaneously bind to an auxiliary operator O_a through DNA looping (see figure 3.7(A)). The wild type *lac* promoter has two such auxiliary operators: O2 located 401 bases downstream and O3 located 92 bases upstream. We describe the simpler case studied experimentally by Oehler et al. [35], which involves only repression and looping between the main operator, O_m, and the downstream auxiliary operator, O2. The expected fold-change is given by case 9 of table 2.1,

[1] Not discussed here is a lower plateau of promoter activity for saturating amounts of repressor, sometimes referred to as promoter leakage. Such leakage could result, for example, from the passage of the replication fork through a tightly repressed promoter, leading to basal transcription activity.

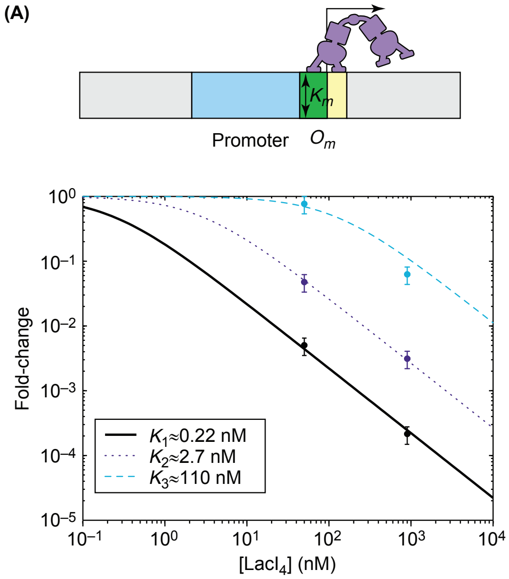


Figure 3.6: Simple repression. (a) Cis-regulatory structure of the truncated *lac* promoter, with the main operator O_m (yellow box) located closely downstream of the core promoter (blue box). Repressor bound at O_m will block RNAP binding to the promoter, as denoted by the overlap (green box). The DNA-binding affinity of LacI$_4$ for O_m is described by the dissociation constant K_m. (b) log-log plot of the fold-change in gene expression as a function of LacI$_4$. Here, the repressor concentration shown on the horizontal axis refers to the cellular LacI tetramers in the absence of inducers. The experiments of Oehler et al. [35] used the operator sequences O1, O2, O3 at position O_m and measured fold-repression at two different LacI concentrations (50 nM and 900 nM); the data are shown as circles. The expected form of the fold-changes are plotted as the solid, dotted and dashed lines as indicated in the legend. The value of K_m for each curve (see legend) is determined by fitting one of the two data points. The fact that the other data point lies closely on the curve supports the applicability of the thermodynamic model to this promoter.

with $[R] = [\mathrm{LacI}_4]$.

Given that the three K values are already determined (see figure 3.6(B)), there is only one unknown parameter in this case in the form for the fold-change (case 9 of table 2.1). This is $[L]$, the effective concentration of repressors that are made available, as a result of DNA-looping, for binding to one of the two operators. This looping is itself caused by the binding of a repressor to the other operator. Oehler et al. [35] did experiments with the main operator, O_m, substituted for one of the three operator sequences (O1, O2 and O3), each for two concentrations of LacI. The results of all six experiments are described consistently by the expected fold-changes according to the thermodynamic model (see figure 3.7(B)), with $[L] \approx 660$ nM [36].

Quantitatively, the strong repression effect (compare figure 3.6(B) and figure 3.7(B)) results directly from the large value of $[L]$ generated by DNA looping, which amplifies the effective concentration of one operator-bound repressor 660-fold. This enhancement of the local repressor concentration is a result of the linkage between O_m and O_a, as already described qualitatively elsewhere [35, 41]. Intuitively, once a LacI tetramer binds to one of the two operators, it is available within a small volume for binding to the other. The actual value of $[L]$ is clearly dependent on the spacing between the two operators, in addition to the energetics of bending the DNA backbone. We have deduced the dependence of $[L]$ on operator spacing (shown in figure 3.7(D)) by analyzing the data of Müller et al. [42], who measured the fold-changes in repression for promoter constructs with different spacing between the main and auxiliary operators (see figure 3.7(C)). In figure 3.7(C), we also show the predicted transcriptional fold-changes for the same constructs of Müller et al. [42], but at different LacI concentrations.

3.1.7 Cooperative repression

Interaction between the TFs can also enhance the sensitivity in transcriptional repression. The $\mathrm{P_R}$ promoter, which controls the expression of cro in phage lambda (illustrated in figure 3.8(A)), is a good example of this mode of repression [3]. When bound to either $\mathrm{O_{R1}}$ or $\mathrm{O_{R2}}$, the lambda repressor, cI, blocks the access of RNAP to the core promoter, thereby repressing transcription. The combined effect of two repressive operators, reinforced by the cooperative interaction between the operator-bound cIs, results in both further repression and enhanced sensitivity. The expected form of fold-change is given by case 6 in table 2.1 ($[R1] = [R2] = [\mathrm{cI2}]$) and plotted in figure 3.8(B). Maximum log-log (i.e., sensitivity) in repression is the largest when K_{R1} and K_{R2} are equal. Similar schemes have been generalized for co-repression by two species of repressors [43–45], and can be used to mimic the logical NAND function [22].

In fact, enhanced sensitivity in repression does not require direct interaction between the repressor molecules. An example is the $\mathrm{P_{LtetO\text{-}1}}$ promoter [46], which contains two operators of TetR; see figure 3.8(C). The expected form of the fold-change is given by case 5 in table 2.1, with $[R_1] = [R_2] = [\mathrm{TetR}_2^*]$, and $K_{R1} = K_1$, $K_{R2} = K_2$. By appropriately decreasing K_1 and K_2, it is possible to make the activity of this promoter (not shown) nearly identical to that represented by the solid line in figure 3.8(B) (i.e.,

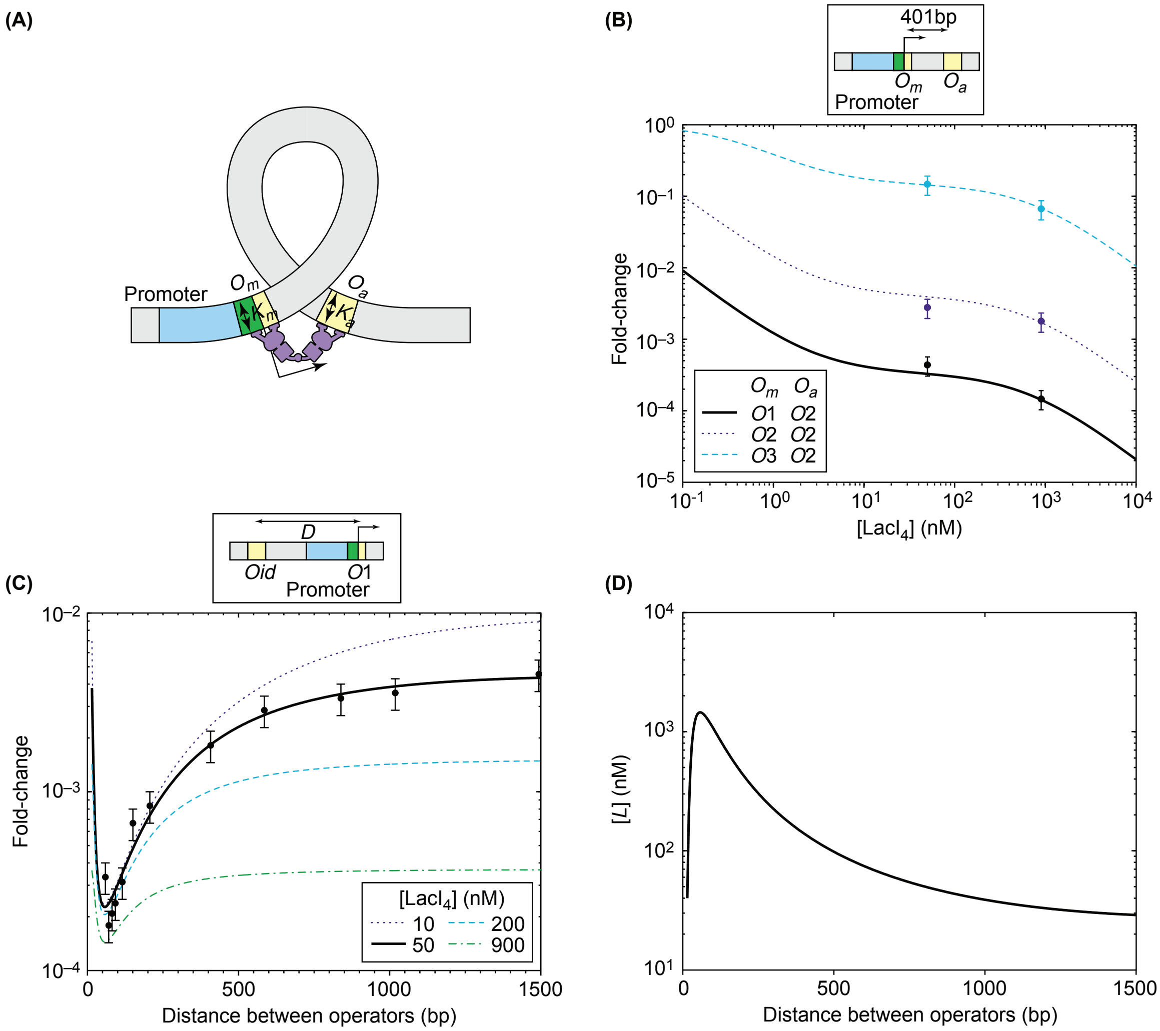


Figure 3.7: Repression by DNA looping. (A) Cis-regulatory layout for looping and repression in the *lac* promoter experiments of Oehler et al. [35]. Yellow boxes are operators and the blue box is the promoter. LacI tetramer bound at the main operator O_m interferes with RNAP binding to the promoter, and this is indicated by the overlap (green box) between the promoter and the operator. This binding is further stabilized if the other two legs of the tetramer bind at O_a through DNA-looping. (B) log-log plot of the fold-change in gene expression as a function of LacI_4 concentration for different constructs where O_m is replaced by O1, O2, or O3 and O_a is O2. The curves are generated by plotting case 9 of table 2.1 using the appropriate dissociation constants shown in figure 3.6 for each pair of operators involved. Note that the six data points (shown with circles) can all be brought into agreement with the expected form (the lines) by the choice of a single parameter, the available LacI_4 concentration $[L]$ due to looping. The best-fit value obtained is $[L] \approx 660$ nM. (C) Loglinear plot of the transcriptional fold-change as a function of distance D between O1 (located at position O_m) and an auxiliary operator Oid located upstream of the promoter, for various repressor concentrations. The data of [42] (filled circles) are fitted to the transcriptional fold-changes expected for looping (solid line) using $[\mathrm{LacI}_4] = 50$ nM and values of $K_1 \approx 0.27$ nM and $Kid \approx 0.05$ nM determined from the data of [35]. The fitting function is the dependence of the available concentration due to looping, $[L]$, on the operator spacing D. We use the form $[L] = \exp(-a/Db\ln(D) + c \times D + e)$ motivated by the worm-like chain model of DNA bending [48]. The other lines correspond to the predicted gene expression of the same constructs at different LacI concentrations as indicated in the legend. (D) Log–linear plot of $[L]$ versus D obtained from the fit described in (C), with $a = 140.6$, $b = 2.52$, $c = 1.4 \times 10^{-3}$, $e = 19.9$.

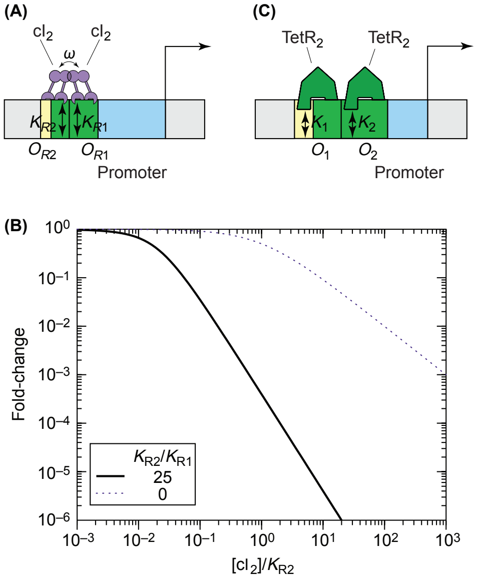


Figure 3.8: Enhanced sensitivity by dual repression. (A) Cis-regulatory architecture for cooperative transcriptional repression in phage lambda P_R promoter. The yellow boxes denote the operator sites O_{R1}, O_{R2} and the blue box corresponds to the promoter. Repression is indicated by the overlap (green box) between the promoter and operator. The cooperative interaction between bound cI_2 at operators O_{R1} and O_{R2} is given by ω (protein-protein contacts). (B) log-log plot of the fold-change in gene expression as a function of cI_2 concentration for two different values of K_{R2}/K_{R1}. At high repressor concentrations, the maximum log-log slope(s) for all the curves is equal to 2 with the exception of $K_{R2}/K_{R1} = 0$ (i.e., deletion of O_{R1}) where the maximum log-log slope is equal to 1. The latter case corresponds to a single repressive site, O_{R2} (see figure 3.6). This plot was generated using $\omega \approx 100$, and $K_{R2}/K_{R1} \approx 25$ extracted from *in vitro* biochemical studies [26]. The absolute *in vivo* values of the K values are unknown, which is why our concentration is expressed in terms of $[cI_2]/K_{R2}$. (c) Cis-regulatory architecture for transcription repression in $\mathrm{P_{LtetO\text{-}1}}$ promoter engineered by Lutz and Bujard [46]. Note that there is no cooperative interaction between the TetR dimers. The log-log plot of fold-change of $\mathrm{P_{LtetO\text{-}1}}$ promoter is similar to that of phage lambda P_R with a maximum log-log slope equal to 2.

with the steepened slope) even though the TetR dimers do not contact each other physically. The enhanced sensitivity is expected here because of the “collaborative” nature of repression — the occupation of either operator is sufficient to block RNAP from the core promoter, leaving the other operator site available for binding for “free” [47]. We expect that a similar construct where the two operators are targets of different, non-interacting TFs would implement co-repression. Comparing the activating and repressive modes of transcription control, we find repressive control to be advantageous because high sensitivity can be generated by TFs without the need of TFTF interaction, and fold changes are not limited by the magnitude of the (typically weak) TFRNAP interaction [48].

3.1.8 Phenomenological model of transcription control

The mathematical description for the different activation and repression mechanisms discussed above can be summarized by very simple forms. For a single TF species with up to two operators in the cis-regulatory region, all of the fold-changes described in table 2.1 can be compactly represented by the general form

$$F_{reg}([TF]) = \frac{1 + a_1[TF] + a_2[TF]^2}{1 + b_1[TF] + b_2[TF]^2}. \tag{3.1}$$

Similarly, for co-regulation by two TFs with cellular concentrations, $[TF_1]$ and $[TF_2]$, and for no more than one operator each in the regulatory region, the foldchange has the form

$$F_{reg}([TF_1], [TF_2]) = \frac{1 + a_{1,0}[TF_1] + a_{0,1}[TF_2] + a_{1,1}[TF_1] \times [TF_2]}{1 + b_{1,0}[TF_1] + b_{0,1}[TF_2] + b_{1,1}[TF_1] \times [TF_2]}. \tag{3.2}$$

The general forms in equation 3.1 and equation 3.2 include many possible mechanisms of activation and repression not discussed above. If 3 binding sites for the TF are involved in the regulatory process, then equation 3.1 or equation 3.2 would be generalized to the ratio of third-degree polynomials of the $[TF]$s.

The above analysis indicates that, by quantitatively measuring the fold-change as a function of the activated TF concentration(s), we can achieve two important goals: (i) by fitting experimental results to an expression such as equation 3.1 or equation 3.2, one would obtain a quantitative characterization of the promoter at all TF concentrations, but with only a few (e.g., four or six) parameters. This can be done regardless of the validity of the thermodynamic model itself. As discussed previously, the compact description will facilitate quantitative higher-level study of gene circuits. (ii) By comparing the values of these parameters to the expected forms according to the thermodynamic model (e.g., table 2.1), one can generate hypotheses on the likely mechanisms of transcriptional control for further experiments. Thus, the form of the fold-change in gene expression itself can be an effective diagnostic tool to distinguish subtle mechanisms of transcriptional control.

3.2 Conclusions

We have illustrated a variety of promoter activities implemented in different cis-regulatory designs. Also illustrated are important functional differences (e.g., in transcriptional cooperativity, and in the nature of combinatorial control) among promoters characterized by different parameters of the same cis-regulatory construct. These differences often cannot be discriminated by the qualitative characterization of promoter activity predominantly practiced in molecular biology today (e.g., fold-change in gene expression caused by deletion of a regulatory protein). Instead, they call for more quantitative characterization, particularly the quantification of the TF concentrations or their relative concentrations, controlling promoter activity. The reward of quantitative characterization includes a compact phenomenological description of promoter activity for higher-level analysis and the elucidation of unknown mechanisms of transcriptional control.

Bibliography

[1] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[2] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[3] M. Ptashne. *A genetic switch: Phage lambda revisited.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 3rd edition, 2004.

[4] C. C. Guet, M. B. Elowitz, W. Hsing, and S. Leibler. Combinatorial synthesis of genetic networks. *Science*, 296(5572):1466–70, 2002.

[5] J. Hasty, D. Mcmillen, F. Isaacs, and J. J. Collins. Computational studies of gene regulatory networks: In numero molecular biology. *Nat Rev Genet*, 2(4):268–79, 2001.

[6] H. H. Mcadams, B. Srinivasan, and A. P. Arkin. The evolution of genetic regulatory systems in bacteria. *Nat Rev Genet*, 5(3):169–78, 2004.

[7] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. Pan, M. J. Schilstra, P. J. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. Mcclay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–78, 2002.

[8] R. Thomas and R. D'ari. *Biological feedback.* CRC Press, Boca Raton, 1990.

[9] D. M. Wolf and F. H. Eeckman. On the relationship between genomic regulatory element organization and gene regulatory dynamics. *J Theor Biol*, 195(2):167–86, 1998.

[10] D. F. Browning and S. J. Busby. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*, 2(1):57–65, 2004.

[11] Y. Wei, J. M. Lee, C. Richmond, F. R. Blattner, J. A. Rafalski, and R. A. Larossa. High-density microarray-mediated gene expression profiling of escherichia coli. *J Bacteriol*, 183(2):545–56, 2001.

[12] R. H. Ebright. Transcription activation at class i cap-dependent promoters. *Mol Microbiol*, 8(5):797–802, 1993.

[13] B. MÜLler-Hill. *The lac operon: A short history of a genetic paradigm.* Walter de Gruyter, Berlin, New York, 1996.

[14] M. E. Wall, W. S. Hlavacek, and M. A. Savageau. Design of gene circuits: Lessons from bacteria. *Nat Rev Genet*, 5(1):34–42, 2004.

[15] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci U S A*, 100(13):7702–7, 2003.

[16] T. Kuhlman, Z. Zhang, M. H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *escherichia coli*. *Proc Natl Acad Sci U S A*, 104(14):6043–8, 2007.

[17] J. Beckwith, T. Grodzicker, and R. Arditti. Evidence for two sites in the lac promoter region. *J Mol Biol*, 69(1):155–60, 1972.

[18] M. M. Garner and A. Revzin. Stoichiometry of catabolite activator protein/adenosine cyclic 3',5'-monophosphate interactions at the lac promoter of escherichia coli. *Biochemistry*, 21(24):6032–6, 1982.

[19] M. G. Fried and D. M. Crothers. Equilibrium studies of the cyclic amp receptor protein-DNA interaction. *J Mol Biol*, 172(3):241–62, 1984.

[20] M. Takahashi, B. Blazy, A. Baudras, and W. Hillen. Ligand-modulated binding of a gene regulatory protein to DNA. Quantitative analysis of cyclic-amp induced binding of crp from escherichia coli to non-specific and specific DNA targets. *J Mol Biol*, 207(4):783–96, 1989.

[21] J. T. Wade, T. A. Belyaeva, E. I. Hyde, and S. J. Busby. A simple mechanism for co-dependence on two activators at an escherichia coli promoter. *EMBO J*, 20(24):7160–7, 2001.

[22] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003.

[23] M. Ptashne and A. Gann. Imposing specificity by localization: Mechanism and evolvability. *Curr Biol*, 8(22):R812–22, 1998.

[24] I. B. Dodd, K. E. Shearwin, A. J. Perkins, T. Burr, A. Hochschild, and J. B. Egan. Cooperativity in long-range gene regulation by the lambda ci repressor. *Genes Dev*, 18(3):344–54, 2004.

[25] D. K. Hawley and W. R. Mcclure. Mechanism of activation of transcription initiation from the lambda prm promoter. *J Mol Biol*, 157(3):493–525, 1982.

[26] K. S. Koblan and G. K. Ackers. Site-specific enthalpic regulation of DNA transcription at bacteriophage lambda or. *Biochemistry*, 31(1):57–65, 1992.

[27] J. K. Joung, L. U. Le, and A. Hochschild. Synergistic activation of transcription by escherichia coli camp receptor protein. *Proc Natl Acad Sci U S A*, 90(7):3083–7, 1993.

[28] S. Busby, D. West, M. Lawes, C. Webster, A. Ishihama, and A. Kolb. Transcription activation by the escherichia coli cyclic amp receptor protein. Receptors bound in tandem at promoters can interact synergistically. *J Mol Biol*, 241(3):341–52, 1994.

[29] J. K. Joung, D. M. Koepp, and A. Hochschild. Synergistic activation of transcription by bacteriophage lambda ci protein and e. Coli camp receptor protein. *Science*, 265(5180):1863–6, 1994.

[30] S. Scott, S. Busby, and I. Beacham. Transcriptional co-activation at the ansb promoters: Involvement of the activating regions of crp and fnr when bound in tandem. *Mol Microbiol*, 18(3):521–31, 1995.

[31] T. A. Belyaeva, V. A. Rhodius, C. L. Webster, and S. J. Busby. Transcription activation at promoters carrying tandem DNA sites for the *escherichia coli* cyclic amp receptor protein: Organisation of the rna polymerase alpha subunits. *J Mol Biol*, 277(4):789–804, 1998.

[32] J. Tebbutt, V. A. Rhodius, C. L. Webster, and S. J. Busby. Architectural requirements for optimal activation by tandem crp molecules at a class i crp-dependent promoter. *FEMS Microbiol Lett*, 210(1):55–60, 2002.

[33] C. M. Beatty, D. F. Browning, S. J. Busby, and A. J. Wolfe. Cyclic amp receptor protein-dependent activation of the escherichia coli acsp2 promoter by a synergistic class iii mechanism. *J Bacteriol*, 185(17):5148–57, 2003.

[34] S. Busby and R. H. Ebright. Transcription activation by catabolite activator protein (cap). *J Mol Biol*, 293(2):199–213, 1999.

[35] S. Oehler, M. Amouyal, P. Kolkhof, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. Quality and position of the three *lac* operators of *e. Coli* define efficiency of repression. *EMBO J*, 13(14):3348–55, 1994.

[36] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[37] S. Lin and A. D. Riggs. The general affinity of lac repressor for e. Coli DNA: Implications for gene regulation in procaryotes and eucaryotes. *Cell*, 4(2):107–11, 1975.

[38] M. Pfahl, V. Gulde, and S. Bourgeois. “second” and “third operator” of the lac operon: An investigation of their role in the regulatory mechanism. *J Mol Biol*, 127(3):339–44, 1979.

[39] R. B. Winter and P. H. Von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. the *escherichia coli* repressor–operator interaction: Equilibrium measurements. *Biochemistry*, 20(24):6948–60, 1981.

[40] U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc Natl Acad Sci U S A*, 99(19):12015–20, 2002.

[41] G. R. Bellomy and M. T. Record Jr. Stable DNA loops in vivo and in vitro: Roles in gene regulation at a distance and in biophysical characterization of DNA. *Prog Nucleic Acid Res Mol Biol*, 39:81–128, 1990.

[42] J. MÜLler, S. Oehler, and B. MÜLler-Hill. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol*, 257(1):21–9, 1996.

[43] M. Dmitrova, G. Younes-Cauet, P. Oertel-Buchheit, D. Porte, M. Schnarr, and M. Granger-Schnarr. A new lexa-based genetic system for monitoring and analyzing protein heterodimerization in escherichia coli. *Mol Gen Genet*, 257(2):205–12, 1998.

[44] J. C. Hu, M. G. Kornacker, and A. Hochschild. Escherichia coli one- and two-hybrid systems for the analysis and identification of protein-protein interactions. *Methods*, 20(1):80–94, 2000.

[45] L. B. Hays, Y. S. Chen, and J. C. Hu. Two-hybrid system for characterization of protein-protein interactions in e. Coli. *Biotechniques*, 29(2):288–90, 292, 294 passim, 2000.

[46] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in *escherichia coli* via the lacr/o, the tetr/o and arac/i1-i2 regulatory elements. *Nucleic Acids Res*, 25(6):1203–10, 1997.

[47] J. A. Miller and J. Widom. Collaborative competition mechanism for gene activation in vivo. *Mol Cell Biol*, 23(5):1623–32, 2003.

[48] M. Ptashne and A. Gann. *Genes and signals*. Cold Spring Harbor Laboratory Press, New York, 2002.

Chapter 4

DNA Looping and Gene Regulation: The Physics of Biological Action at a Distance

This chapter is the reproduction of a manuscript to be submitted to Physical Biology shortly.

DNA looping is a ubiquitous regulatory motif in bacteria where a transcription factor can bind to multiple sites on the DNA that are often hundreds of base pairs apart. Over the past several decades a set of fascinating quantitative experiments have been performed on DNA looping in the context of the *lac* operon. We use thermodynamic models of transcriptional regulation to systematically dissect such experiments leading to the quantification of the looping free energy cost paid in these configurations. We explore the consequences of this parameter both in the context of the *lac* operon and raise questions about the canonical view of DNA looping as a means to reduce fluctuations in gene expression. We also show that the looping free energy is a transferrable parameter that can be extracted from one experiment in order to generate predictions about another. By combining the looping free energy as a measure of the mechanical properties of DNA with simple polymer models we set bounds on the *in vivo* effective mechanical properties at length scales between 100 bp and 1 kbp, showing that it can be thought of in terms of simple excluded-volume polymer models. By taking into account the contribution of non-specific looping to the *in vivo* looping free energy we are able to compare the results of *in vivo* and *in vitro* experiments explicitly. We conclude that our current mechanical understanding of protein-DNA looping cannot account for the much higher DNA flexibility observed *in vivo* with respect to *in vitro* and suggest a new round of experiments to shed light on this issue. This work can serve as the basis of a systematic characterization of more complex regulatory motifs within the framework of thermodynamic models of transcriptional regulation.

4.1 Introduction

Transcriptional regulation is one the most common ways in which cells make decisions about the level of expression of their genes. This regulation is carried out by a variety of transcription factors, proteins that bind to DNA and interact with RNA polymerase (RNAP) by either inhibiting or enhancing its ability to bind the promoter [1]. It is amusing that in the two celebrated examples of transcriptional decision making which led to the formulation of the operon concept, the λ switch and the *lac* operon, DNA looping plays a crucial role [2, 3]. In DNA looping, transcription factors bind to two sites on the DNA simultaneously looping the intervening DNA [4, 5]. The mechanical properties of the DNA can therefore play an active role in the regulation of the level of gene expression.

The role of DNA mechanics in decision making is not limited to bacteria. There is an increasing body of evidence that points to the role of the physical properties of DNA in eukaryotes. On long length scales, elements that are thousands of base pairs away from each other on the DNA communicate to regulate transcription [6, 7]. These long range interactions can be measured directly on a genome-wide scale [8, 9]. The physical state of the DNA has also even been suggested to be involved in the maintenance of epigenetic states [10]. The majority of our knowledge about the mechanical state of the DNA inside the cell comes from techniques that give access to length scales of a kbp and beyond [8, 11]. However, there is limited information for lengths scales shorter than 1 kbp. These length scales are precisely those that are accessed by DNA looping in transcriptional regulation in bacteria and are also relevant in the context of, for example, nucleosomal positioning. There is increasing evidence that nucleosome positioning is encoded by the mechanical properties of the DNA that is wrapped around the histone octamer [12]. However, the subject of strongly bent DNA remains a source of controversy [13, 14].

In this paper we examine the role of DNA looping in transcriptional regulation. We do this by investigating repression by DNA looping using the *lac* operon as a particular case study. One of the regulators of this operon is the Lac repressor (LacI). This repressor has two binding heads, allowing it to bind to two DNA sites several hundreds of base pairs away from each other simultaneously. We use thermodynamic models of transcriptional regulation as the tool to dissect this DNA looping motif. These models have a rich history in quantitatively describing transcriptional regulation (for reviews see [1, 15]). DNA looping has been addressed in the context of these models on multiple occasions allowing for a connection between the level of gene expression and microscopic parameters that are directly related to the *in vivo* mechanical properties of DNA [1, 16–20].

The logic behind our approach is to analyze increasingly complex promoter architectures, starting with the case of simple repression where there is only one repressor binding site. At each stage in the analysis we will use parameters obtained from the previous simpler architecture. As a result, when analyzing complex systems regulated by DNA looping, the only free parameter that remains will be the looping free energy itself.

The remainder of the paper is organized as follows. First, we analyze simple repression by Lac repressor in the absence of DNA looping. This will allow us to build up key concepts such as the energetics of its binding to DNA and the nature of the non-specific reservoir. With these concepts in hand, we explore the theoretical implications of repression by DNA looping in the *lac* operon through the prism of our models. We show that the obtained looping free energy is more than a fitting parameter by using it as a tool to generate predictions. Finally, we analyze experiments on *lac* operon mutants as a way to access the *in vivo* mechanical properties of DNA at length scales spanning from single base pairs to 1 kbp. The paper culminates with a direct comparison of our current best knowledge about the mechanics of DNA looping *in vivo* and *in vitro* that is only possible through the quantification of the contribution of non-specific sites to the *in vivo* looping free energy. This comparison shows significant differences between the two contexts suggesting a new round of quantitative and systematic experimentation.

4.1.1 Lac repressor and the *lac* operon

The Lac repressor is a tetramer built of identical dimeric subunits. Each of these subunits has a DNA binding head [21] which can bind to DNA independently [3, 22]. Its *in vivo* mode of repression is thought to occur by sterically excluding RNAP from the promoter [23]. As a result, a loop is not required to achieve repression. The *lac* operon has three binding sites, or operators, for its repressor (O1, O2 and O3, with their binding affinities decreasing in that order) and one binding site for its activator, CRP. In figure 4.1a we present the architecture of the *lac* promoter (for a review and very interesting history of the *lac* operon see [3]).

Distal, or auxiliary, sites like O2 or O3 do not exert any significant effect on gene expression in the absence of O1 [24]. However, cooperativity between the proximal and distal operators is obtained through the simultaneous binding of Lac repressor to the main site and one of its auxiliary partner sites, which have to be brought together to each of the binding heads by looping the intervening DNA [3]. The contribution of DNA looping to repression in the *lac* operon can be readily observed by measuring changes in gene expression in constructs that delete and mutate different combinations of operators [22, 24], change the concentration of transcription factor [22] and change the spacing between operators [16, 25, 26].

We build our models using experimental data based on simplified constructs like that shown in figure 4.1b. In this case, only the main operator is present. We use the thermodynamic formalism to obtain *in vivo* binding energies of Lac repressor to each one of it operators. These energies will be then used as known parameters when addressing the more complex case of transcriptional regulation by DNA looping.

4.2 Simple repression by Lac repressor

We begin by formally introducing thermodynamic models of transcriptional regulation in bacteria. From there we move on to dissecting the case of simple repression, where only one binding site for the repressor is present. Similar formalism to that presented here have also been applied to simple repression by Lac repressor in [18, 20, 27]. We first consider the simpler case of a repressor with only one binding head. It is convenient that dimeric mutants of Lac repressor with such characteristics do exist [24]. We next consider simple repression by Lac repressor tetramers, which will lead us to consider subtleties such as the role of the extra binding head in simple repression. The analysis in this section will culminate in the determination of

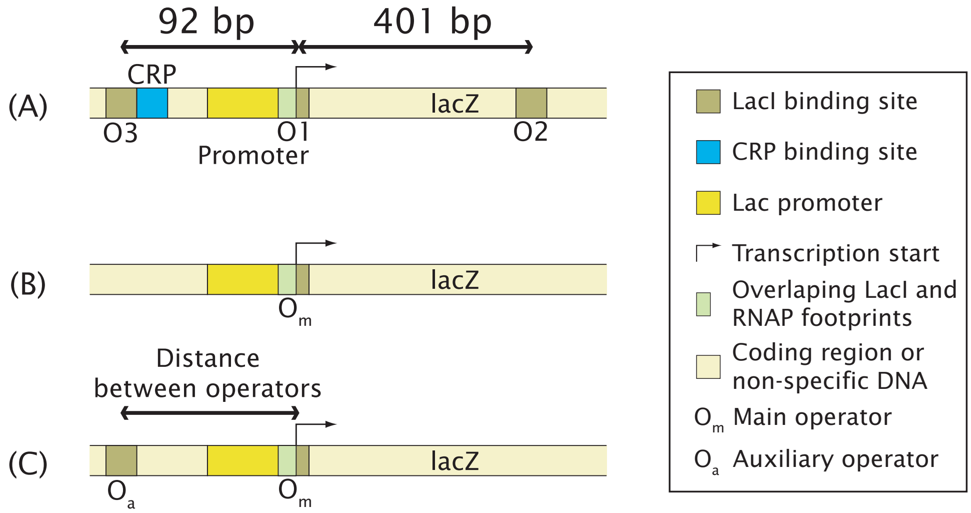


Figure 4.1: Architecture of the wild-type *lac* operon and engineered versions relevant to the models in this paper. (A) Distances between the binding sites for the molecular agents in the wild-type *lac* operon. (B) Simplified construct with only a Lac repressor main operator, O_m, present used by Oehler et al. [22] and Becker et al. [26]. (C) Looping construct with a main operator and an auxiliary operator O_m and O_a, respectively, used by Müller et al. [25] and Becker et al. [26].

the *in vivo* binding energies of Lac repressor to its different binding sites. As a result only the looping free energy will remain as an unknown when addressing repression by DNA looping.

A significant part of this section is a reproduction of the Supplementary Information in [28].

4.2.1 Thermodynamic models of transcriptional regulation

Thermodynamic models of transcriptional regulation are based on computing the probability of finding RNA polymerase (RNAP) bound to the promoter and how the presence of transcription factors (TFs) modulates this probability. These models and their application to bacteria are reviewed in [1, 15]. These models make two key assumptions. First, they assume that the processes leading to transcription initiation by RNAP are in quasi-equilibrium. This means that we can use the tools of statistical mechanics to describe the binding of RNA polymerase and TFs to DNA. Second, they assume that the level of gene expression is proportional to the probability of finding RNAP bound to the corresponding promoter.

We start by analyzing the probability that RNAP will be bound at the promoter of interest in the absence of any transcription factors. We assume that the key molecular players (RNAP and TFs) are bound to the DNA either specifically or non-specifically. In particular, this question has been addressed experimentally in the context of RNAP [29] and the Lac repressor [30, 31] our two main molecules of interest in this paper. Experiments demonstrate that the reservoir for RNAP is the background of non-specific sites. In order to determine the contribution of this reservoir we sum over the Boltzmann weights of all the possible configurations. For P RNAP molecules inside the cell with N_{NS} non-specific DNA sites one finds

$$Z^{NS}(P;N_{NS}) = \frac{N_{NS}!}{P!(N_{NS}-P)!}e^{-\beta\varepsilon_{pd}^{NS}} \simeq \frac{N_{NS}^{P}}{P!}e^{-\beta\varepsilon_{pd}^{NS}}, \tag{4.1}$$

where $\beta = 1/k_BT$. The factor of $N_{NS}!/[P!(N_{NS}-P)!]$ in the previous expression accounts for all the possible configurations of RNAP in the reservoir. This is shown diagrammatically in figure 7.6. The second factor assigns the energy of binding between RNAP and non-specific DNA, ε_{pd}^{NS}, and as a theoretical convenience that may have to be revised in quantitatively dissecting real promoters, is taken to be the same for all non-specific sites. A more sophisticated treatment of this model to account for the differences in the non-specific binding energy has been addressed by [32]. Finally, the last expression corresponds to assuming that $N_{NS} \gg P$, a reasonable assumption given that the *E. coli* genome is approximately 5 Mbp long and that the number of σ^{70} RNAP molecules, the type of RNAP we are interested in for the purposes of this paper, is on the order of a thousand [33].

We calculate the probability of finding one RNAP bound to a promoter of interest in the presence of this non-specific reservoir. Two states are considered: either the promoter is empty and P RNAPs are in the reservoir or the promoter is occupied leaving $P-1$ RNAP molecules in the reservoir. The corresponding total partition function is

$$Z(P;N_{NS}) = \underbrace{Z^{NS}(P;N_{NS})}_{\text{Promoter unoccupied}} + \underbrace{e^{-\beta\varepsilon_{pd}^{S}}Z^{NS}(P-1;N_{NS})}_{\text{Promoter occupied}}, \tag{4.2}$$

where we have now defined ε_{pd}^{S} as the binding energy between RNAP and the promoter. The probability of finding the promoter occupied, p_{bound}, is then

$$p_{bound}(P) = \frac{e^{-\beta\varepsilon_{pd}^{S}}Z^{NS}(P-1;N_{NS})}{Z^{NS}(P;N_{NS}) + e^{-\beta\varepsilon_{pd}^{S}}Z^{NS}(P-1;N_{NS})} = \frac{1}{1+\frac{N_{NS}}{P}e^{\beta\Delta\varepsilon_{pd}}}, \tag{4.3}$$

with $\Delta\varepsilon_{pd} = \varepsilon_{pd}^{S} - \varepsilon_{pd}^{NS}$, the difference in energy between being bound specifically and non-specifically. With this framework in hand we can now turn to addressing regulation by Lac repressor through simple repression.

4.2.2 Simple repression by Lac repressor dimers

As for polymerase, we invoke the assumption that the reservoir for dimeric Lac repressor is the non-specific DNA. This assumption is supported by experimental evidence [30, 31]. Our aim is to examine all of the different configurations available to P RNA polymerase molecules, R LacI dimers and N_{NS} non-specific sites. If the binding energy of RNAP and the LacI head to non-specific DNA are ε_{pd}^{NS} and ε_{rd}^{NS}, respectively, the

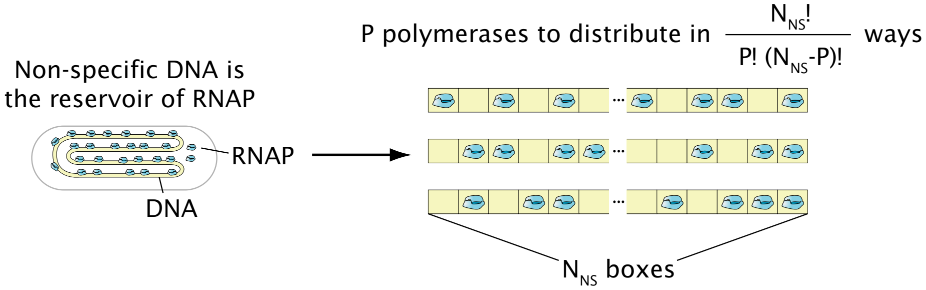

P polymerases to distribute in $\frac{N_{NS}!}{P!\,(N_{NS}-P)!}$ ways
Non-specific DNA is the reservoir of RNAP
RNAP
DNA
N_{NS} boxes

Figure 4.2: Model for the RNA polymerase reservoir. The non-specific sites on the genome are assumed to be the reservoir for RNAP. Different arrangements of RNAP in this reservoir are shown.

non-specific partition function becomes

$$Z^{NS}(P,R_2)=\underbrace{\frac{N_{NS}^{P}}{P!}e^{-P\beta\varepsilon_{pd}^{NS}}}_{Z^{NS}(P)}\underbrace{\frac{(N_{NS})^{R_2}}{R_2!}e^{-R_2\beta\varepsilon_{rd}^{NS}}}_{Z^{NS}(R_2)},\tag{4.4}$$

where we have assumed that both LacI dimers and RNAP are so diluted in the reservoir that they do not interact with each other. We use the notation R_2 with the subscript 2 as a reminder that we are describing Lac repressor dimers.

This model assumes three distinct classes of microscopic state for the promoter: 1) promoter unoccupied, 2) one RNAP taken from the reservoir and placed on the promoter and 3) a LacI dimer taken from the reservoir and placed on the operator. In this scheme, Lac repressor and RNAP cannot be found on the promoter simultaneously [23]. These states and their corresponding normalized weights, which we derive below, are shown in figure 7.7(A). The total partition function is

$$Z_{total}(P,R_2)=\underbrace{Z^{NS}(P,R_2)}_{\text{promoter free}}+\underbrace{Z^{NS}(P-1,R_2)e^{-\beta\varepsilon_{pd}^{S}}}_{\text{RNAP on promoter}}+\underbrace{Z^{NS}(P,R_2-1)e^{-\beta\varepsilon_{rd}^{S}}}_{\text{LacI dimer on Om}},\tag{4.5}$$

where ε_{pd}^{S} and ε_{rd}^{S} are the binding energies of RNAP and a Lac repressor head to their specific sites, respectively. We factor out the term corresponding to all molecules present in the reservoir and define $\Delta\varepsilon_{pd}=\varepsilon_{pd}^{S}-\varepsilon_{pd}^{NS}$ and $\Delta\varepsilon_{rd}=\varepsilon_{rd}^{S}-\varepsilon_{rd}^{NS}$ as the energy gain of RNAP and dimeric LacI when switching from a non-specific site to their specific sites, respectively. The probability of finding RNAP bound to the promoter is given by

$$p_{bound}=\frac{\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_{pd}}}{1+\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_{pd}}+\frac{R_2}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}}}.\tag{4.6}$$

This expression can be rewritten as

$$p_{bound}=\frac{1}{1+\frac{N_{NS}}{P\cdot F_{reg}(R_2)}e^{\beta\Delta\varepsilon_{pd}}},\tag{4.7}$$

where we have defined the regulation factor

$$F_{reg}(R_2) = \frac{1}{1 + \frac{R_2}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}}. \tag{4.8}$$

Notice that in the absence of repressor ($R_2 = 0$), p_{bound} reduces to equation 7.8. The regulation factor can be seen as an effective rescaling of the number of RNAP molecules inside the cell [15].

Though determining the probability of promoter occupancy is interesting, it is not necessarily the most natural quantity to measure experimentally. Relative levels of gene expression are often measured instead by characterizing the concentration or activity of a reporter protein or by measuring changes in its mRNA concentration. The fold-change in gene expression is defined as the ratio of reporter produced in the presence of the transcription factor (repressor in this case) relative to the amount of reporter produced in the absence of the transcription factor. It is important to note here that we are defining the fold-change with respect to the presence and absence of the transcription factor itself and not with respect to the presence and absence of an inducer, a definition that is much more common mainly for its experimental ease.

One of the key assumptions in the thermodynamic class of models is that the level of gene expression is linearly related to p_{bound}. This allows us to equate the fold-change in gene expression to the fold-change in promoter occupancy

$$\text{fold-change}(R_2) = \frac{p_{bound}(R_2 \neq 0)}{p_{bound}(R_2 = 0)}. \tag{4.9}$$

If we substitute p as shorthand for $\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}$ in the expression for p_{bound}, we find

$$\text{fold-change}(R_2) = \frac{p + 1}{p + \frac{1}{F_{reg}(R_2)}}. \tag{4.10}$$

The fold-change becomes independent of the details of the promoter in the case of a weak promoter, where $p \ll 1, \frac{1}{F_{reg}(R_2)}$, which permits us to write the approximate expression

$$\text{fold-change}(R_2) \simeq F_{Reg}(R_2) = \left(1 + \frac{R_2}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}\right)^{-1}. \tag{4.11}$$

In this work we consider two closely related promoters: the wild-type *lac* promoter and a mutant *lacUV5* promoter [25]. In the case of the *lac* promoter if one considers *in vitro* binding energies of RNAP to the promoter, p has the approximate value $\sim 10^{-3}$ [15]. For the *lacUV5* promoter used in some of the experiments analyzed in this work the p is expected to be of order 0.1 or smaller [28].

The expression in equation 7.16 relates the fold-change in gene expression to two inputs that can be controlled experimentally: the binding energy of the operator and the cellular concentration of Lac repressor. Oehler et al. [22] created different constructs where several values of these two "knobs" were characterized. They then measured the fold-change in gene expression as a function of the concentration of LacI dimers inside the cell for constructs bearing Oid, O1, O2 and O3 as the main operator in a simple repression

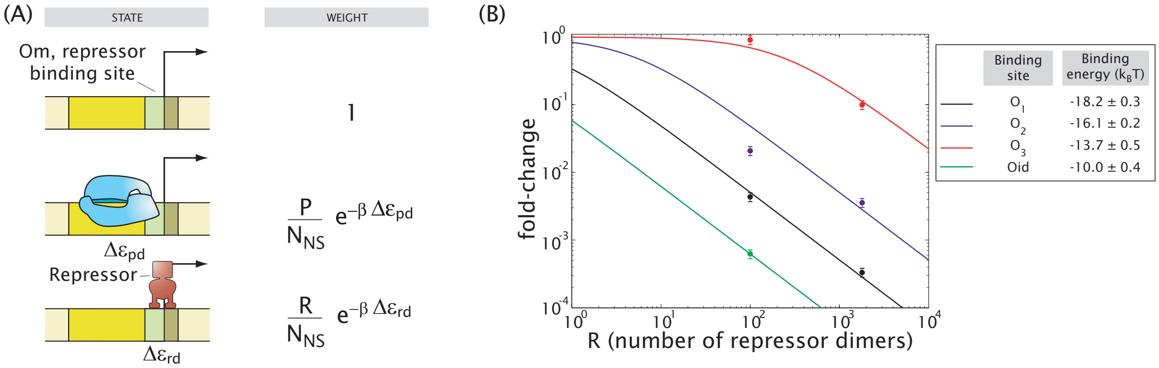


Figure 4.3: Single-site repression by LacI dimers. (A) Schematic listing of the different states and their respective weights when RNAP and the dimeric repressor have overlapping sites. (B) Repression for four different strengths of the main repressor binding site (Om) as a function of the number of dimers inside the cell. The binding energy of dimeric Lac repressor to each operator are calculated by fitting each data set to the repression expression from equation 7.16 and presented in table 7.1.

Table 4.1: Single-site binding energies for repressor dimers and tetramers. The energies are obtained using the data by $^{(1)}$ Oehler et al. [22] and $^{(2)}$ Garcia and Phillips [28] and equations 7.16 and 7.24 for the dimers and tetramers, respectively. The error bars for the Oehler et al. data are calculated assuming an error in the fold-change measurement of 30%. The *in vitro* binding constants are calculated using $K_{rd}^{NS} \approx 40,000 \times 10^{-11}$ M [34], $K_{Oid} \approx 2,4 \times 10^{-12}$ M [35], $K_{O1} \approx 1.35 \times 10^{-11}$ M [34], $K_{O2} \approx 2 \times 10^{-10}$ M [36] and $K_{O3} \approx 15 \times K_{O1}$ [37]. Notice that even though the *in vivo* binding energies do not coincide with the *in vitro* energies the relative difference between the energies for Oid and O1 or O2 are comparable.

Operator	Dimers $(k_BT)^{(1)}$	Tetramers $(k_BT)^{(1)}$	Tetramers $(k_BT)^{(2)}$	*in vitro* (k_BT)
Oid	-18.2 ± 0.3	-17.7 ± 0.3	-16.8 ± 0.2	-12.0
O_1	-16.1 ± 0.2	-16.2 ± 0.1	-15.1 ± 0.2	-10.3
O_2	-13.7 ± 0.5	-13.7 ± 0.1	-13.7 ± 0.2	-7.6
O_3	-10.0 ± 0.4	-10.4 ± 0.4	-9.6 ± 0.1	-7.6

architecture.

In figure 7.7b we present their data as well as a fit of the fold-change in gene expression. Notice that for each construct there is only one unknown: the *in vivo* binding energies, $\Delta\varepsilon_{rd}$. As a result we estimate these *in vivo* binding energies for each one of the operators. The results are summarized in table 7.1. Interestingly, if we are to believe the claims of the thermodynamic models, this data on simple repression can now be inherited for use in the consideration of more complex architectures involving the same operators.

4.2.3 The non-specific reservoir for Lac repressor tetramers

As a next level of transcriptional complexity, we now consider simple repression by Lac repressor tetramers. In principle, when dealing with Lac repressor tetramers only one head has to be bound to the DNA in order to exert any repression. It is not clear, however, what the state of the remaining head is. To derive the consequences of simple repression by tetramers we must address this issue. For example, that extra head could be "hanging" from the DNA without establishing contact with DNA. Another option is that the extra head will also be exploring different non-specific sites. For the purposes of this section we will assume that the second head can also bind to DNA. In section 4.4.2 we explore this model further

We begin by assuming that both Lac repressor binding heads are bound to DNA at all times either specifically or non-specifically. At this point this choice is arbitrary and the final results will not depend on the particular model for the state of the second head. We work with this particular formulation of the problem since it is both concrete and analytically tractable and makes the counting of the accessible states more transparent.

The model for the non-specific reservoir is depicted in figure 7.8. For LacI dimers we assumed that the molecules were exploring all possible non-specific sites. In this model both heads of a tetramer will be exploring all possible non-specific sites as well. As a result we will need to account not only for binding to all sites, but for looping between all these non-specific sites. We start by considering only one LacI molecule. We count the possible ways in which we can arrange the two heads on different non-specific sites on the DNA. We label the site where one of the heads binds i and the other site j. For every choice of sites an energy ε_{rd}^{NS} is gained for each head that is non-specifically bound. These two sites can be joined through four different DNA loops which are depicted in figure 4.5. Each one of these loops correspond to a different orientation between the DNA binding site and the protein binding head. We will label these different tangent orientations with the index σ. A cost in the form of a looping free energy $F_{loop}(i,j,\sigma)$ is also paid for bringing sites i and j together. The sum over all non-specific states can be written as

$$Z^{NS}(R_4=1) = \frac{1}{2} \underbrace{\sum_{i=1}^{N_{NS}} e^{-\beta\varepsilon_{rd}^{NS}}}_{\text{head 1, site } i} \underbrace{\sum_{j=1}^{N_{NS}} e^{-\beta\varepsilon_{rd}^{NS}}}_{\text{head 2,site } j} \qquad \underbrace{\sum_{\sigma} e^{-\beta F_{loop}(i,j,\sigma)}}_{\text{Looping between sites } i \text{ and } j \text{ with configuration } \sigma} \quad . \tag{4.12}$$

We introduce the notation R_4 to specify the number of Lac repressor tetramers, signified by the subscript 4. Note that a factor of $\frac{1}{2}$ has been introduced in order not to over-count loops. This is equivalent to assuming that the two binding heads on a repressor are indistinguishable. Our model assumes that the binding of a tetramer head is independent of the state of the other head. As a result the interaction between a head and DNA are the same in the tetramer and dimer case.

To simplify this expression we chose a particular binding site for the first head, i_0, and sum over all possible positions for the second head. This can now be done for the different N_{NS} positions that can be chosen for i_0, resulting in

$$Z^{NS}(R_4=1) \simeq \frac{1}{2} \underbrace{N_{NS}}_{\text{choices for } i} \; e^{-\beta 2\varepsilon_{rd}^{NS}} \sum_{j=1}^{N_{NS}} \sum_{\sigma} e^{-\beta F_{loop}(i_0,j,\sigma)}. \tag{4.13}$$

Finally, we bury the term $\sum_j \sum_\sigma e^{-\beta F_{loop}(i_0,j,\sigma)}$ into an effective non-specific looping free energy $e^{-\beta F_{loop}^{NS}}$. In section 4.2.3 we will discuss different models for F_{loop}^{NS} and their distinctive predictions.

In order to obtain the partition function for R_4 tetramers we assume that all repressors are independent

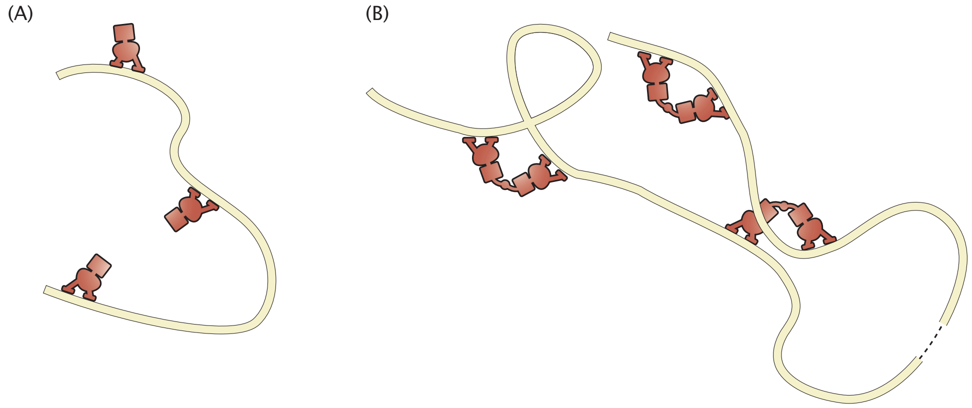


Figure 4.4: Model for the non-specific looping background. Possible states of non-specific DNA bound by Lac repressor. (A) Dimers will explore all available non-specific sites. (B) Tetramers explore all possible loops between non-specific sites.

and indistinguishable. We therefore extend the partition function to the case of R_4 non-interacting tetramers in the reservoir by computing

$$Z^{NS}(R_4) = \frac{\left[Z^{NS}(R_4=1)\right]^{R_4}}{R_4!} = \frac{1}{2^{R_4}}\frac{(N_{NS})^{R_4}}{R_4!}e^{-\beta R_4\, 2\varepsilon_{rd}^{NS}}\, e^{-\beta R_4\, F_{loop}^{NS}}, \tag{4.14}$$

where the binding energy is still defined as in section 4.2.2.

From this point on we will only consider Lac repressor tetramers. As a result, for notational compactness we replace R_4 with R. We obtain the complete non-specific partition function by multiplying the factor corresponding to repressors with a factor corresponding to RNAP being bound non-specifically shown in equation 7.9 resulting in

$$Z^{NS}(P,R) = \frac{(N_{NS})^P}{P!}e^{-\beta P \varepsilon_{pd}^{NS}}\ \frac{1}{2^R}\frac{(N_{NS})^R}{R!}e^{-\beta R\, 2\varepsilon_{rd}^{NS}}\, e^{-\beta R\, F_{loop}^{NS}}, \tag{4.15}$$

which now allows us in the next section to address the case of simple repression by LacI tetramers.

4.2.4 Simple repression by Lac repressor tetramers

We begin by taking one head of one Lac repressor tetramer out of the non-specific reservoir shown in equation 7.19 and binding it specifically to the operator. This can be easily done by going back to equation 7.17. We label the position on the genome corresponding to the specific site i_0. We will choose only those terms in the summation corresponding to the binding site of interest. Since either one of the heads can reach the position labeled by i_0 we obtain the following partition function for a single tetramer bound to a specific site

$$Z_R^{O,NS} = \frac{1}{2}e^{-\beta\varepsilon_{rd}^{S}}e^{-\beta\varepsilon_{rd}^{NS}}\left(\sum_{i=1}^{N_{NS}}\sum_{\sigma}e^{-F_{loop}(i,i_0,\sigma)} + \sum_{j=1}^{N_{NS}}\sum_{\sigma}e^{-F_{loop}(i_0,j,\sigma)}\right). \tag{4.16}$$

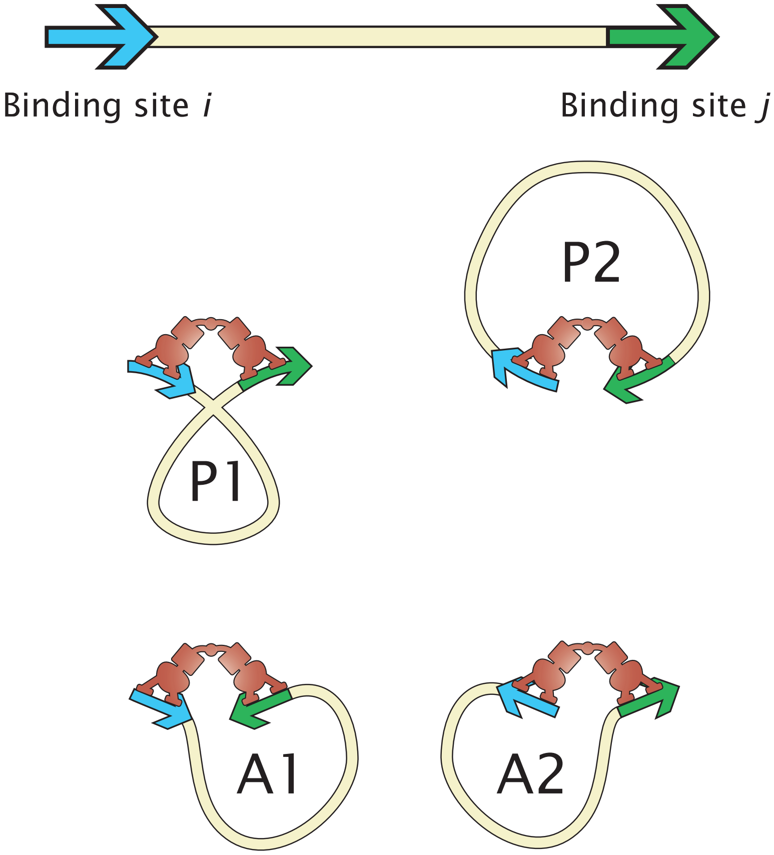


Figure 4.5: Different loop geometries. Different configurations for a DNA loop given a separation between the two binding sites. We label the different configurations using the notation introduced by Geanacopoulos et al. [38]

Because both sums are identical we can reduce this to

$$Z_R^{O,NS} = e^{-\beta\varepsilon_{rd}^S} e^{-\beta\varepsilon_{rd}^{NS}} \sum_{j=1}^{N_{NS}} \sum_{\sigma} e^{-F_{loop}(i_0,j,\sigma)} = e^{-\beta\varepsilon_{rd}^S} e^{-\beta\varepsilon_{rd}^{NS}} e^{-\beta F_{loop}^{NS}}. \tag{4.17}$$

We are now ready to calculate the total partition function. We will consider the three states from figure 7.1. The weights corresponding to the first two states will be the same as in the LacI dimer case. The third state corresponds to the partition function term we just calculated. The total partition function is then

$$Z_{total}(P,R) = Z^{NS}(P,R) + Z^{NS}(P-1,R)e^{-\beta\varepsilon_{pd}^S} + Z^{NS}(P,R-1) \times Z_R^{O,NS}. \tag{4.18}$$

After rewriting these equations using equation 7.22 using the weak promoter approximation we get the following fold-change in gene expression

$$\text{fold-change}(R) \simeq \left(1 + 2\frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{rd}}\right)^{-1}. \tag{4.19}$$

Even though the contribution from the non-specific loops drops out of the expression, we see that there is now a factor of two in front of the number of LacI tetramers. This is different from the fold-change in gene expression for dimers shown in equation 7.14. It can be easily understood if we think about the actual number of binding heads that are now present. In the case of dimers we have R_2 binding heads whereas for tetramers there are $2R_4$ binding heads inside the cell. As a result, no information about the non-specific looping background can be obtained by doing experiments on simple repression. We see that as long as the number of binding heads is the same the fold-change will not vary. Interestingly, this is one of the conclusions from the data by Oehler et al. [22]. They compared repression for Lac repressor dimers and tetramers under the condition $2R_4 = R_2$ and saw a comparable fold-change in gene expression. An alternative way to look at this is by comparing the binding energies obtained for dimers and tetramers. These two set of energies, obtained from equations 7.16 and 7.24, are shown in table 7.1.

In a recent experimental paper, we presented a set of experiments where the consequences of equation 7.24 were explored systematically by tuning the number of repressors and operator strengths over a wide dynamic range [28]. This led to a better estimate of the tetramer binding energies for each of the operators. These binding energies are also shown in table 7.1. We can also compare the *in vivo* binding energies obtained from gene expression measurements to their *in vitro* values. This comparison can be made through the relation [15]

$$\Delta\varepsilon_{rd} = \varepsilon_{rd}^S - \varepsilon_{rd}^{NS} = \frac{K_{rd}^S}{K_{rd}^{NS}}, \tag{4.20}$$

where K_{rd}^S and K_{rd}^{NS} are the specific and non-specific dissociation constants, respectively. In table 7.1 we show the resulting *in vitro* binding energies resulting from various measurements of the dissociation constants discussed in the literature. Note that even though the absolute values of the binding energies are not the

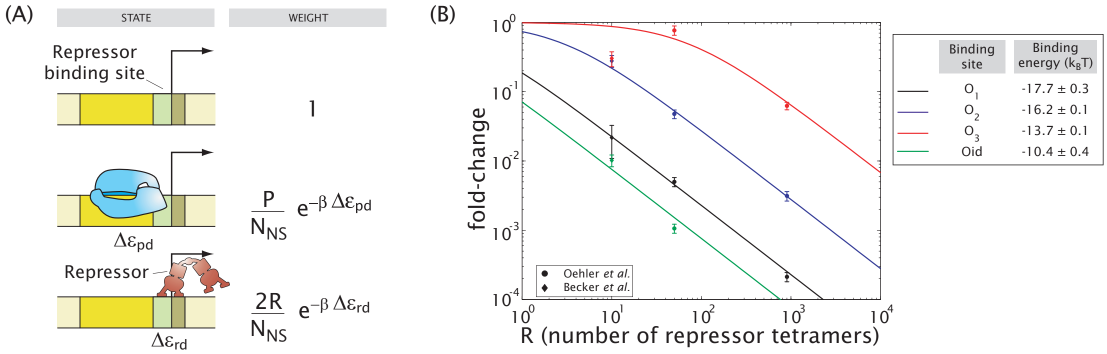


Figure 4.6: Single-site repression by LacI tetramers. (A) States and weights when RNAP and the repressor have overlapping sites. (B) Repression for four different strengths of the main repressor binding site (Om) as a function of the number of repressors. The binding energy of Lac repressor to each operator is calculated by fitting each data set to the repression expression from equation 7.24. The numbers are given in table 7.1.

same *in vivo* as *in vitro*, nevertheless the differences between the binding energies for Oid and O1 and O2 are comparable.

Now that we have obtained the *in vivo* binding energies of Lac repressor to its operators we are ready to address more complex regulatory architectures such as DNA looping. The fact that we have already determined these binding parameters will result in there being only one unknown: the looping free energy, a quantitative measure of the mechanical properties of DNA.

4.3 DNA Looping by Lac Repressor

In the previous section we explored simple repression by Lac repressor as viewed through the prism of thermodynamic models. One of the main outcomes of this approach was the calculation of DNA binding energies of the repressor to its different operator sequences. We now take the next step in this constructionist analysis of DNA looping. We consider the case where there are two specific binding sites for Lac repressor, a main and an auxiliary site as shown in figure 4.1c. Both heads can be bound simultaneously bringing the operators into proximity and forming a DNA loop as shown in figure 4.7. It should be noted that at this point we cannot commit to a particular structural model of the looped DNA because of lack of *in vivo* information about the geometrical and mechanical properties of both the protein and the supercoiled DNA [13]. For now, we choose to think of the looping free energy as a parameter. Later on in the text we will make a series of attempts at connecting this magnitude to DNA mechanics.

In this section we will mathematically describe repression by DNA looping. This will allow us first to develop intuition about the potential usefulness of this regulatory motif. In particular, this section will lead to a dissection of the wild-type *lac* operon in terms of the probabilities of formation of each of its three loops (O3-O1, O1-O2 and O3-O2).

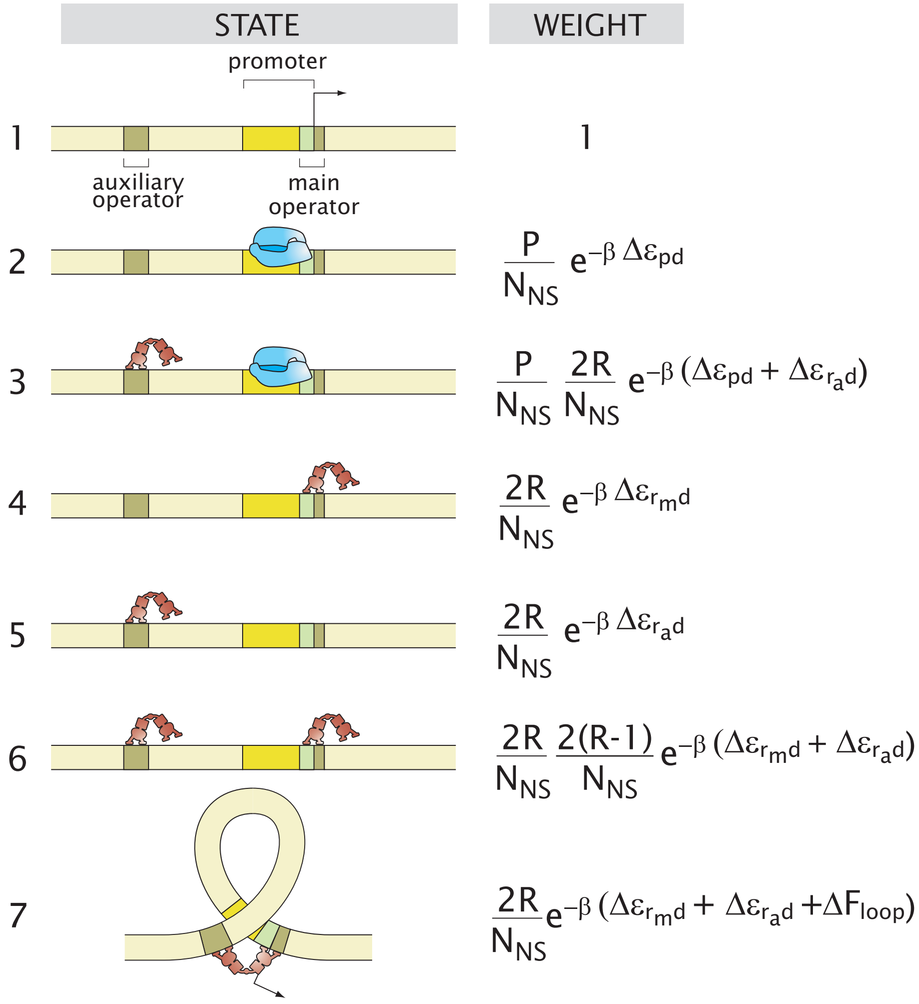

STATE
WEIGHT
promoter
auxiliary operator
main operator
1
2
3
4
5
6
7
1
$\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_{pd}}$
$\frac{P}{N_{NS}} \frac{2R}{N_{NS}} e^{-\beta (\Delta\varepsilon_{pd} + \Delta\varepsilon_{r_a d})}$
$\frac{2R}{N_{NS}} e^{-\beta \Delta\varepsilon_{r_m d}}$
$\frac{2R}{N_{NS}} e^{-\beta \Delta\varepsilon_{r_a d}}$
$\frac{2R}{N_{NS}} \frac{2(R-1)}{N_{NS}} e^{-\beta (\Delta\varepsilon_{r_m d} + \Delta\varepsilon_{r_a d})}$
$\frac{2R}{N_{NS}} e^{-\beta (\Delta\varepsilon_{r_m d} + \Delta\varepsilon_{r_a d} + \Delta F_{loop})}$

Figure 4.7: Statistical mechanics of repression by DNA looping. States and weights used in the looping model. See text for how these weights were determined.

4.3.1 The looping regulation factor

In order to model the construct from figure 4.1c we ask how many different ways we can find LacI bound to any of its sites. In figure 4.7 we show the relevant states and their corresponding weights that will be derived below. Notice that the auxiliary operator cannot exert any repression by itself. This operator can only do so when it is part of the loop since in this case it is helping to stabilize the binding of LacI to the main operator.

To compute the repression, we need the total partition function. Using $Z^{NS}(P,R)$ from equation 7.20 and the result from equation 7.22 regarding the binding of one head to a specific site while the other one is non-specific we can write the total looping partition function

$$\begin{aligned} Z_{total}(P,R) \quad = \quad & \underbrace{Z^{NS}(P,R)}_{\text{empty sites}} + \underbrace{Z^{NS}(P-1,R)e^{-\beta\varepsilon^{S}_{pd}}}_{\text{RNAP on promoter}} + \\ & \underbrace{Z^{NS}(P-1,R-1)e^{-\beta\varepsilon^{S}_{pd}}e^{-\beta(\varepsilon^{S}_{rad}+\varepsilon^{NS}_{rd}+F^{NS}_{loop})}}_{\text{RNAP on promoter, repressor on } O_a} + \\ & \underbrace{Z^{NS}(P,R-1)e^{-\beta(\varepsilon^{S}_{rmd}+\varepsilon^{NS}_{rd}+F^{NS}_{loop})}}_{\text{repressor on } O_m} + \underbrace{Z^{NS}(P,R-1)e^{-\beta(\varepsilon^{S}_{rad}+\varepsilon^{NS}_{rd}+F^{NS}_{loop})}}_{\text{repressor on } O_a} + \\ & \underbrace{Z^{NS}(P,R-2)e^{-\beta(\varepsilon^{S}_{rmd}+\varepsilon^{S}_{rad}+2\varepsilon^{NS}_{rd}+2F^{NS}_{loop})}}_{\text{two repressors: on } O_m \text{ and } O_a} + \underbrace{Z^{NS}(P,R-1)e^{-\beta(\varepsilon^{S}_{rmd}+\varepsilon^{S}_{rad}+F^{S}_{loop})}}_{\text{loop between } O_m \text{ and } O_a}, \end{aligned} \tag{4.21}$$

where we have defined F^{S}_{loop} as the free energy cost of bringing the main and auxiliary sites together by looping the DNA. This free energy cost includes information about the possible loop configurations denoted by σ and shown in figure 4.5. The specific looping free energy can be described in terms of the looping free energy of each configuration as follows

$$e^{-\beta F^{S}_{loop}} = \sum_{\sigma} e^{-\beta F_{loop}(\sigma)}. \tag{4.22}$$

As a result we bury all details of the actual loop geometry into an effective specific looping free energy.

We define $\Delta F_{loop} = F^{S}_{loop} - F^{NS}_{loop}$ as the looping free energy measured with respect to the non-specific looping background. In section 4.4.2 we go back to this issue but, for now, we will focus only on ΔF_{loop}. Once again, we use the weak promoter approximation (equation 7.16) and equate the looping regulation factor to the fold-change in gene expression obtaining

$$\begin{aligned} \text{fold-change} \quad (R) \quad \simeq & \; F_{reg}(R) = \\ & \left(1 + 2\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{rad}}\right) \Big/ \left(1 + 2\frac{R}{N_{NS}}(e^{-\beta\Delta\varepsilon_{rmd}} + e^{-\beta\Delta\varepsilon_{rad}}) + \right. \\ & \left. 4\frac{R(R-1)}{(N_{NS})^2}e^{-\beta(\Delta\varepsilon_{rmd}+\Delta\varepsilon_{rad})} + 2\frac{R}{N_{NS}}e^{-\beta(\Delta\varepsilon_{rmd}+\Delta\varepsilon_{rad}+\Delta F_{loop})}\right). \end{aligned} \tag{4.23}$$

The factor of two arising from the presence of two binding heads could be absorbed into the binding energies, although this would require us to offset ΔF_{loop} by ln(2), as was done in [1]. In the present work, however, we will keep it in an effort to understand the contribution of the non-specific loops to F_{loop}^{NS}.

We now have an expression that relates a magnitude that can be easily determined experimentally such as the fold-change in gene expression to a change in the mechanical state of a protein-DNA complex given by ΔF_{loop}. Having determined the binding energies from simpler experiments previously is key to this as the only unknown in equation 4.23 is the looping free energy. In the remainder of this paper we explore the consequences of equation 4.23. We investigate the relevance of DNA looping in transcriptional regulation and try to connect the looping free energies that can be obtained *in vivo* to both theoretical and *in vitro* experimental expectations.

4.3.2 Why is looping useful for transcriptional regulation?

Why is DNA looping such a persistent regulatory architecture? In this section we will try to develop intuition about the looping motif and suggest some explanations as to why this motif is so common in prokaryotes (see the reviews by Schleif [5] and Matthews [4], for further elaboration).

Before going into the mathematical derivations we wish to pose the problem in a clearer way by reference to figure 4.8. Here we compare the predicted fold-change as a function of the concentration of repressor molecules inside the cell for different choices of single operator constructs (figure 4.1b) and for a particular choice of a looping construct (figure 4.1c). As already noted by Vilar and Leibler [18] and reviewed by Saiz and Vilar [27, 39] we see that DNA looping can give a larger fold-change than simple repression at wild-type concentrations of LacI ($\sim$ 10 repressors/cell). Even if simple repression is implemented using the strongest, unnatural binding site Oid, DNA looping gives a higher fold-change in gene expression at wild-type concentrations of repressor binding to the weaker, natural binding sites.

One interesting characteristic of the looping fold-change as a function of repressor number is the presence of a plateau in the curve. In particular note that there is a regime in which a substantial change in LacI concentration produces a small fold-change in gene expression. This plateau occurs at concentrations that are typical for the wild-type *lac* operon. It has been suggested that the wild-type parameters have been tuned for repression to be "robust" against fluctuations in the concentration of transcription factor [27]. It has also been suggested that DNA looping can have a significant effect on cell-to-cell variability [18]. However, it has been shown recently that this conclusion depends strongly on the particular choice of parameters assumed for the different rates involved in the process [40]. Finally, it is important to note that, though all these ideas have been suggested in the context of repression by DNA looping, the case of the wild-type *lac* operon is more complex. The presence of three different operators leads to multiple loops whose consequences we will explore in section 4.3.3. In the context of the wild-type *lac* operon, DNA looping has been shown to be key for the maintenance of bistability [41]. Here, it was proposed that upon unbinding from O1, the repressor will rebind to it with a different rate if it is already bound to O2 or O3 than if it is coming from

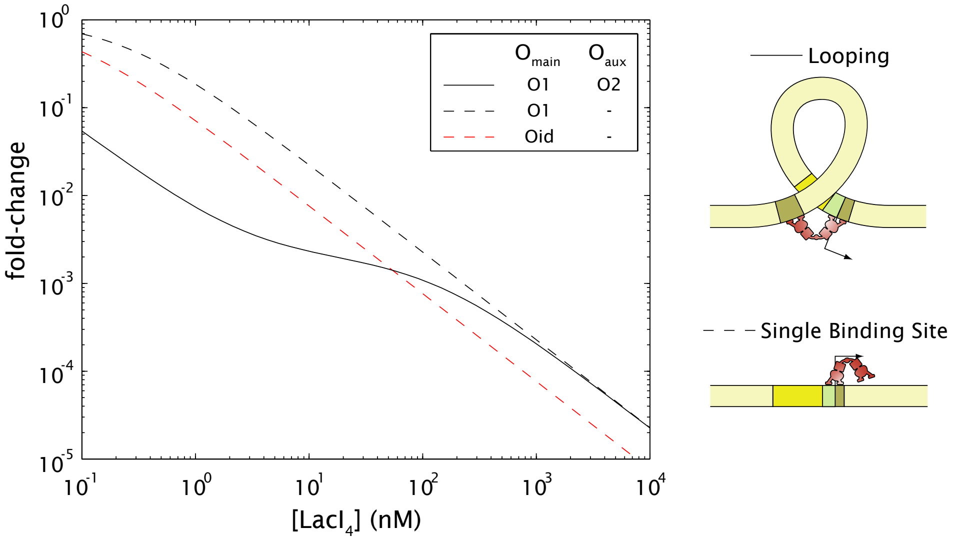


Figure 4.8: Comparison of simple repression and DNA looping. Fold-change from a single $O1$ site (dashed black) and a single Oid site (dashed red) compared to repression by looping from an $O1-O2$ configuration with the difference in looping free energy $\Delta F_{loop} = 10k_BT$ (solid black). Also notice that for high number of repressors the fold-change due to the $O1$-$O2$ loop approaches the fold-change due to $O1$ in simple repression.

the cytoplasm. This difference in rates results in extra stability in the *lac* operon leading presumably to the observed bistability [41–43].

For high repressor concentrations the looping motif reduces to simple repression by the main operator. This can be understood by comparing the last two states of figure 4.7, since in the limit of $R \to \infty$ the term that goes as R^2 becomes dominant over all the others. This term corresponds to having one LacI molecule bound to each site. The concentration is so high that the auxiliary site will be always bound independently of the state of the main site. Under this limit the expression for the fold-change from equation 4.23 becomes

$$\text{fold-change} \overset{R\to+\infty}{\longrightarrow} \frac{2\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{rad}}}{4\frac{R^2}{(N_{NS})^2}e^{-\beta(\Delta\varepsilon_{rmd}+\Delta\varepsilon_{rad})}} = \left[2\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{rmd}}\right]^{-1} \simeq \left[1+2\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{rmd}}\right]^{-1}. \quad (4.24)$$

In conclusion, once the concentration of repressor is high enough the contribution of DNA looping to repression becomes negligible and we are left with only simple repression. An example is shown in figure 4.8, where it can be seen that repression by an O1-O2 loop approaches simple repression by O1 for high number of repressors.

Finally, in figure 4.9 the fold-change in gene expression as a function of LacI concentration is plotted for different choices of main and auxiliary sites and for various values of the looping free energy. From these plots all the intuition developed in this section can be easily confirmed. These plots also give a sense for the range of fold-change values that can be attained by changing all the relevant parameters of the looping motif within reasonable ranges.

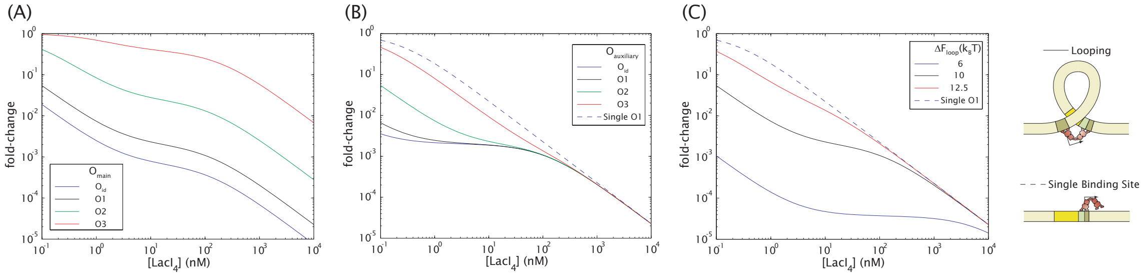


Figure 4.9: Dissecting the looping motif. Repression as a function of Lac repressor concentration for (A) different choices of O_m with O_a=O2, (B) different O_a with O_m=O1 and (C) for different values of the difference in looping free energy ΔF_{loop} with O_m=O1 and O_a=O2.

4.3.3 Reconstructing the *lac* operon

The analysis of the previous section suggests that two of the main potential features of the looping motif are an increase in the level of repression and a robustness with respect to fluctuations in the concentration of repressor. The wild-type architecture of the *lac* operon is different, however, from the two-operator case considered there. Three operators are present in that case allowing for three different loops. In this section we dissect this case in order to determine which features of the simpler looping motif with only one possible loop still apply to the wild-type case.

Having three binding sites translates into three possible loops: O1-O2 (401 bp), O1-O3 (92 bp) and O2-O3 (492 bp). Each loop corresponds to a different interoperator distance and, therefore, to a different looping free energy. It is of importance to point out the presence of a binding site for the activator CRP between O3 and O1. CRP has been reported to interact with LacI or the LacI-mediated O3-O1 loop [44]. One possible explanation for this is the fact that CRP bends the DNA it binds to, potentially facilitating the formation of the DNA loop [44–46]. As a result of this interaction we might expect the looping free energies for looping to be different in the presence or absence of CRP. In order to determine the respective values of F_{loop} we invoke the different *lac* operon deletions characterized by Oehler et al. [22]. These constructs are shown in figure 4.10(A), where it is shown that they measured the fold-change in gene expression for all possible deletions of the wild-type *lac* operon in the context of two different Lac repressor concentrations and in the presence of the CRP binding site.

Since we already have the binding energies corresponding to each operator we can use equation 4.23 in order to determine the looping free energies for the O1-O2 and O3-O1 loops on the basis of the two-operator measurements. However, this strategy does not allow us to determine the looping free energy of the O3-O2 loop from such constructs because this loop does not affect repression directly. To obtain its looping energy we turn to a model of repression in the wild-type *lac* operon. The full *lac* operon involves a significant proliferation of states beyond those shown in figure 4.7 and as a result, we go straight to the expression for the fold-change. In order to do so in a more compact fashion we define the weight corresponding to a single repressor bound to any of the sites as

$$r_i = \frac{2R}{N_{NS}} e^{-\beta \Delta\varepsilon_{rdi}}, \tag{4.25}$$

where the subscript i labels the operator that is bound. When two binding sites are occupied simultaneously we define

$$r_{ij} = \frac{4R(R-1)}{(N_{NS})^2} e^{-\beta(\Delta\varepsilon_{rdi}+\Delta\varepsilon_{rdj})} \tag{4.26}$$

and when all three are bound we use

$$r_{ijk} = \frac{8R(R-1)(R-2)}{(N_{NS})^3} e^{-\beta(\Delta\varepsilon_{rdi}+\Delta\varepsilon_{rdj}+\Delta\varepsilon_{rdk})}. \tag{4.27}$$

The states with a loop are described by

$$r_{loop(ij)} = \frac{2R}{N_{NS}} e^{-\beta(\Delta\varepsilon_{rdi}+\Delta\varepsilon_{rdj}+\Delta F_{loop,ij})}. \tag{4.28}$$

Finally, we can also have a loop and the remaining site occupied by another Lac repressor molecule. We represent this by

$$r_{i,loop(jk)} = \frac{4R(R-1)}{(N_{NS})^2} e^{-\beta(\Delta\varepsilon_{rdi}+\Delta\varepsilon_{rdj}+\Delta\varepsilon_{rdk}+\Delta F_{loop,jk})}. \tag{4.29}$$

In all of these definitions the indices i, j and k correspond to the different operators. Using this notation the fold-change in gene expression for the wild-type *lac* operon is

$$\begin{aligned} \text{fold-change} \;=\; & \left[1 + r_2 + r_3 + r_{23} + r_{23,loop}\right] / & (4.30)\\ & \big[1 + r_1 + r_2 + r_3 + r_{12} + r_{13} + r_{23} + r_{123} + r_{loop(12)} + & (4.31)\\ & r_{loop(13)} + r_{loop(23)} + r_{1,loop(23)} + r_{2,loop(13)} + r_{3,loop(12)}\big]\,. & \end{aligned}$$

Notice that the only unknown in this expression is the looping free energy between O3 and O2. We obtain it by fitting this formula to the data for the wild-type *lac* operon from figure 4.10(A). In table 4.2 we show the various looping energies for the different operator combinations obtained so far. In section 4.4.1 we obtain looping energies for loops of the same lengths as the *lac* operon loops in the absence of CRP. These looping energies are also shown in table 4.2. Notice that for those loops harboring the CRP site within the loop that the difference in looping energy is roughly 2 k_BT. This stabilization of the loops by CRP is in quantitative agreement with previous results both *in vitro* [44] and *in vivo* [46].

Now that we have all the parameters of equation 4.30 we can predict the fold-change in gene expression for the *lac* operon and its mutants at any concentration of Lac repressor. These predictions are shown in figure 4.10(A). It is interesting to note that even though a construct bearing only $O1$ and $O2$ shows the plateau at wild-type concentrations of Lac repressor that was identified in the previous section, the complete wild-type construct does not maintain this feature. As a result, we conclude that in the natural lac operon with all three binding sites the plateau in the repression curve does not exist. Wild-type *E. coli* is not robust against fluctuations in repressor by the mechanism suggested when only the O1-O2 loop is considered. It has been proposed that the absence of CRP can recover this "robustness" in the *lac* operon [39, 47, 48].

Interestingly, all three looping constructs shown in figure 4.10(A) have very similar levels of gene expression at the wild-type concentration of 10 repressors per cell. This becomes more clear in figure 4.10(B) where we plot the probability of the different loops as a function of the number of Lac repressors. It can be seen that the probability of looping between O1 and O2 is approximately equal to that corresponding to O1 and O3 at the wild-type concentration. It is possible that the three-operator system is simply maximizing the amount of repression in the cell over the entire range of repressor concentrations. In figure 4.10(A) it can be seen that over the entire range of repressor concentrations the fold-change resulting from the presence of 3 binding sites is always larger than or equal to the fold-change with 2 binding sites. That is, the fold-change effect is enhanced by the presence of all three operators. Still, this then leaves the question open: if the looping probabilities and the repression levels attained by each loop are the same, what is the functional and evolutionary significance of the full three operator case?

From figure 4.10(B) it is also evident that the predominant loop can be selected by changing the concentration of transcription factor. The fact that such a titration might select for a predominant DNA conformation might be of special interest for eukaryotic domain intercommunication proteins such as SpGCF1 [49].

Finally, a caveat of the model for the *lac* operon proposed here is that it does not account for "deactivation" of CRP when repressor is bound to O3. Oehler et al. [22] observed that there is residual repression in the absence of both O1 and O2 sites. Another possible caveat has to do with an experimental subtlety related to how the operators where deleted. For some of the operator deletions some of the base pairs of the operators were mutated. The choice of bases to mutate corresponded to the ones that had been determined to be most relevant for binding [50]. Still, it was recently proposed that a residual binding energy corresponding to the deleted operators can cause a significant change in the predicted behavior [20, 48]. As a result, a more complex model accounting for both CRP inactivation and residual binding to the deleted operators might be necessary in order to analyze the results by Oehler et al. in more detail.

Table 4.2: Looping energies between the wild-type operators. The energies corresponding to loops in the wild-type *lac* operon are shown in the presence and absence of CRP.

Loop	Distance (bp)	ΔF_{loop} in the presence of CRP[a] (k_BT)	ΔF_{loop} in the absence of CRP[b] (k_BT)
$O3 - O1$	92	6.3	7.9
$O1 - O2$	401	8.4	10.3
$O3 - O2$	493	11[c]	11

[a] Obtained from Oehler et al. [22]

[b] Obtained from Müller et al. [25]

[c] energy is obtained by using the two previous looping energies and solving for the wild-type case with the three operators present

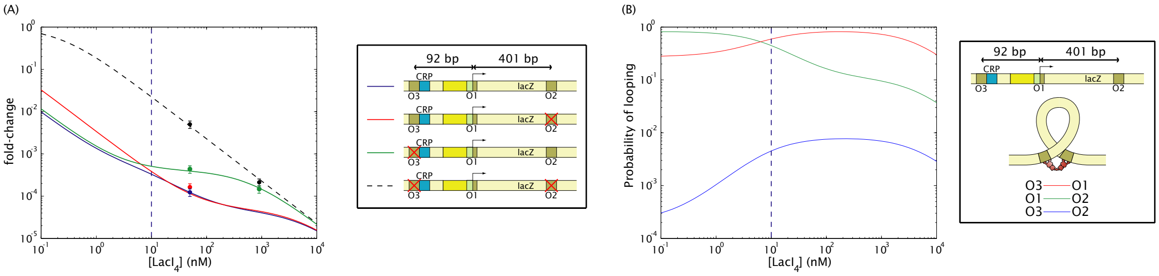


Figure 4.10: Contribution of the various DNA loops in the *lac* operon. (A) Experimental data measured by Oehler et al. [22] and theoretical fits corresponding to the wild-type *lac* operon and various simpler constructs derived from it. (B) Theoretical prediction of the probability of formation of the different loops in the *lac* operon as a function of the repressor concentration. The vertical dashed line in (A) and (B) corresponds to the wild-type concentration of Lac repressor.

4.4 *In vivo* DNA Mechanics

So far we have treated the looping free energy as a parameter that gives us predictive power about *lac* operon mutants, leading to plots like those shown in figure 4.10. However, we have not yet interpreted the mechanical significance of this looping free energy in terms of either the *in vivo* or the *in vitro* mechanical properties of DNA.

The recent development of experimental techniques such as chromosome conformation capture (3C) based methods, DNA FISH and live tagging of nuclear loci using transcription factor-fluorescent protein fusions has opened the door to the analysis of DNA inside the cell at different resolutions spanning kbp to Mbp [8, 9, 11, 51, 52]. However, little is known about *in vivo* DNA mechanics on shorter length scales. It is known, for example, that nucleoid-associated proteins such as HU, IHF and H-NS and that supercoiling affect the effective flexibility of DNA at these short length scales [26, 53]. Despite recent efforts [54] a quantitative understanding of *in vivo* DNA mechanics and their effect on DNA looping based on fundamental properties of DNA is still lacking.

In the remainder of the paper we address DNA looping experiments with Lac repressor as a tool to dissect the *in vivo* physical properties of DNA on length scales up to 1 kbp, in general, and their contribution to transcriptional regulation in particular. We will take a step further from thinking of the looping free energy as a parameter that can be fit to try to understand it in terms of simple mechanical models.

We will show that the looping energy is not just a parameter that is specific to a particular experimental context, but that it can be extracted from one experiment in order to generate predictions about another experiment. Additionally, we will interpret the long-distance behavior of the looping free energy in terms of simple mechanical models of DNA. Finally, we will try to go beyond that by connecting the *in vivo* and *in vitro* description of DNA mechanics. This is key in understanding DNA mechanics inside the cell since our knowledge and intuition about it is built mainly around *in vitro* experiments [13].

4.4.1 Looping free energy vs. distance

Several experiments have been performed where the length of a loop involved in transcriptional regulation was varied systematically while monitoring the resulting change in gene expression both in the *lac* operon [16, 25, 26, 55] and other systems. Length is one of the most easily controllable parameters of the mechanical response of DNA both *in vitro* and *in vivo*. Such experiments have the potential of revealing key aspects of DNA mechanics and how it influences gene expression [13, 56].

We address two different experiments on DNA looping using Lac repressor. The corresponding transcriptional architectures are shown in figure 4.11(A). Here, the distance between the main and auxiliary operators was varied and for each one of these constructs the fold-change in gene expression measured. The data from these two experiments is shown in figure 4.11(B). There are a few subtle differences between these two experimental setups. One of the main differences is the choice of binding sites. While Müller et al. used O_1 and Oid as the main and the auxiliary operators respectively, Becker et al. used O_2 and Oid. The first group worked at a concentration of about 50 repressors per cell, while the latter measured gene expression in the wild-type system that has about 10 repressors per cell. Müller et al. performed their experiments on the chromosome, while Becker et al. had their construct on a single copy F-plasmid of about 180 kbp in length. The sequences introduced between the operators are also different and, in principle, random. It must also be noted that the cells were grown in completely different media and temperature conditions in each experiment. All these differences lead to completely different values for the fold-change in gene expression as can be seen in figure 4.11(B).

The reader is reminded that once we know the binding energies to the operators involved in a DNA looping construct there is only one unknown in the fold-change in gene expression described by equation 4.23, namely the looping free energy. As a result, for each one of the DNA looping experiments described above we can obtain a looping free energy as a function of the distance between the operators, an approach that has previously been applied [1, 19, 27, 57].

In contrast to previous analysis [19] we will show that though we are addressing two completely different experiments on DNA looping with very different characteristics the looping free energy obtained is the same. This suggests that we are dealing indeed with a parameter that is just related to the effective mechanical properties of DNA which are in turn determined by the DNA itself and potentially with various other DNA-binding proteins [26, 53]. We will also be able to recapitulate some of the functional dependence of the looping energy in terms of very simple models of DNA mechanics. As a result we argue that such DNA looping systems can be used as a tool to characterize the *in vivo* mechanical properties of DNA.

4.4.1.1 ΔF_{loop} at short distances

In this section we analyze the length dependence of the looping free energy at short lengths. We address the data from Müller et al. and Becker et al. where the fold-change was measured with single base pair resolution. Since we are dealing with lengths comparable to the persistence length (~ 150 bp), we expect

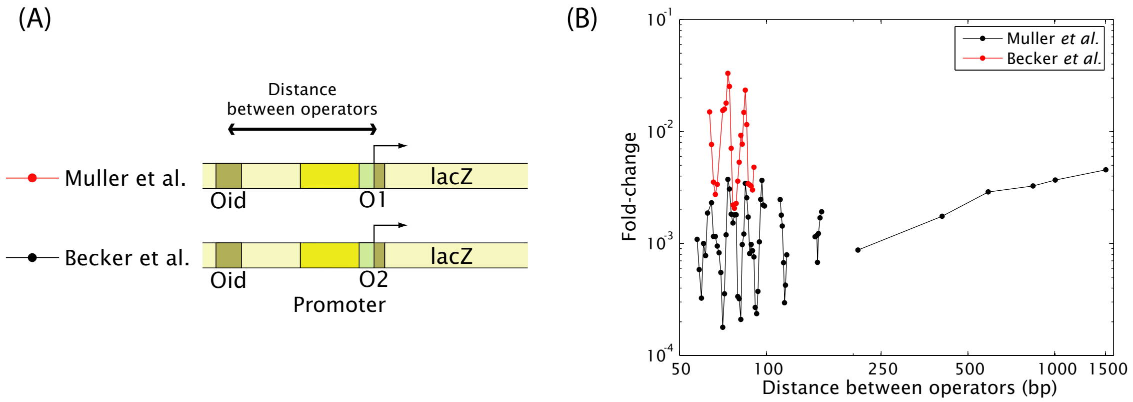


Figure 4.11: Fold-change as a function of interoperator distance. (A) Diagram of the constructs used by Müller et al. and Becker et al. to measure fold-change as a function of distance between operators. (B) Data for fold-change as a function of interoperator distance.

the DNA to be tightly bent in the looped configuration it adopts [13]. In this length regime the size of the loop is comparable to the size of the protein itself. This suggests that the looping free energy will be determined by a combination of DNA bending and the particular geometrical constraints set by the protein. Such geometrical constraints include, for example, the particular angle of alignment of the binding heads with the operators and the separation between the two binding heads [54, 58–61].

In figure 4.12(A) we show the short distance fold-change data from both sets of experiments. If our model is correct, despite the difference in fold-change of the two experiments shown in figure 4.12(A) we should obtain the same looping free energy from them. One strategy is to obtain the looping free energy from the Müller et al. experiment using the looping regulation factor from equation 4.23 and the operator binding energies from table 7.1 and generate a "prediction band". In figure 4.12(B) we overlay the looping free energy obtained from the Becker et al. data with this prediction band. It is clear that, despite certain systematic difference between the prediction and the looping free energy, the amplitude of the modulation in the looping free energy as well as the maximum and minimum value lie within the prediction band. Alternatively, we can obtain the looping free energy for each experiment and compare their values directly. This approach is shown in figure 4.12(C).

Regardless of the approach chosen to compare the looping free energies, the most striking feature is that, despite a difference in phasing, they are comparable. They both oscillate with a comparable amplitude and mean value within experimental error. Saiz et al. [19, 27] performed a similar analysis, but did not obtain overlapping looping free energies probably due to the fact that the fold-change in the Becker et al. work was defined differently than in the Müller et al. experiment. Instead of defining it as the ratio of expression in the presence and absence of repressor, Becker et al. defined it as the ratio of expression in the absence and presence of the inducer IPTG. We obtained their raw gene expression data, which included a construct without any repressor binding sites (Nicole Becker, personal communication). By assuming that this construct has the same level of expression as any other construct in the absence of Lac repressor we

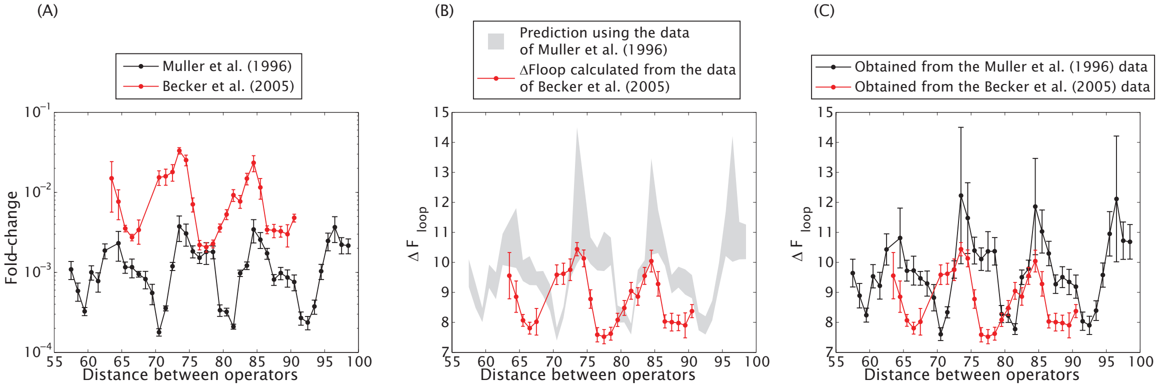


Figure 4.12: From repression to ΔF_{loop} at short distances. (A) Fold-change as a function of distance between operators measured by Müller et al. [25] and Becker et al. [26] for short distances with single base pair resolution. (B) We extract the looping free energy from the Müller et al. data by using the looping regulation factor in equation 4.23 in order to generate a "prediction band" for the looping free energy. This prediction band is contrasted with the looping free energy obtained from the Becker et al. data. Though systematically different, the looping energy from the experiments falls within range of the prediction. (C) Direct comparison between the looping free energies obtained from both experiments.

recover our definition of fold-change in gene expression for their data.

Interestingly, the two looping free energies have maxima that are aligned, but minima that are out of phase. Additionally, the peaks seem to be asymmetric, but in a different way depending on the experiment. The maxima in looping free energy correspond to the lengths at which the twist of DNA makes it the hardest for the intervening DNA to loop. It has been proposed that the phasing and asymmetry of the looping free energy can be explained by the presence of different looping geometries that are favored differently in the two experiments [62]. Even though this is a plausible explanation to the best of our knowledge no mechanistic rationalization of this effect has been reached yet. It is clear, however, that the background of the experiment (for example, the different conditions between the experiments by Müller et al. and Becker et al.) can determine the specific shape of the phasing in DNA looping. A real understanding of such phasing behavior will undoubtedly require much better knowledge of the geometrical details of the Lac repressor-DNA loop complex [54, 59–61].

Despite the obvious differences in the looping free energy, it is important to point out the striking similarities. These two experiments were carried out in different growth media, with different binding sites and repressor concentrations and in different DNA contexts (the chromosome vs. an F-plasmid). Nevertheless, they yield looping free energies which are comparable in many of their features.

As a result this exercise gives us confidence that this looping free energy is more than a fitting parameter. It is a magnitude that captures the mechanical properties and geometrical constraints of DNA and that can be used to predict the outcome of an experiment where all the other parameters (binding energies, repressor concentrations, etc.) are changed.

4.4.1.2 A polymer model for *in vivo* DNA

Over the last few years a significant body of experimental work has addressed the *in vivo* mechanical properties of DNA in various contexts [8, 9, 11, 52, 63–65]. These studies access the properties of the DNA at length scales of several kilobasepairs (kbp) and longer. In contrast, our knowledge of the mechanical properties of DNA *in vivo* at length scales below 1 kbp is rather limited. This is partially due to the fact that most of the current techniques lack the resolving power to query the DNA at these length scales [8]. In this context the experiments by Müller et al. are very suggestive. By measuring repression by DNA looping at lengths between 100 bp and 1 kbp they open the door to a characterization of the DNA in a range not accessible by other techniques.

Our objective is to determine an effective polymer model for DNA at these length scales. A first step in that direction is to estimate the Kuhn length of DNA. This length is a measure of the stiffness of the DNA and gives a sense for the length scale over which DNA behaves like a stiff rod [66] and is double the persistence length, another magnitude often used to express the flexibility of a polymer. Müller et al. [25] and Law et al. [16] performed such an analysis in the context of their DNA looping experiments obtaining a Kuhn length of about 40 bp. In order to infer this value of the persistence length they both used a model in which they examined the free energy penalty to make a DNA circle (i.e cyclization). This observation is confirmed in figure 4.13, where we show the looping free energy obtained from the data of Müller et al. together with different curves describing DNA cyclization for several choices of the Kuhn length. It is important to note that, in contrast to this estimation, recent computational models that take into account many geometrical details of the Lac repressor-DNA loop complex estimated the Kuhn length to be around 200 bp [54].

We expect the behavior of DNA to be different at different length scales. A useful way to quantify these behaviors is to look at the scaling of the average end-to-end distance of the polymer, r_{ee} as a function of its contour length, L. Some of these different regimes are shown diagrammatically in figure 4.14. For lengths below the Kuhn length we expect DNA to behave like a stiff rod. As a result the end-to-end distance scales linearly with the contour length. The looping probability at these length scales will, however, be hard to calculate because of the geometrical and physical details of the repressor [54, 59–61].

As we go higher in scale above the Kuhn length the geometrical details of the Lac repressor-DNA complex will cease to be relevant. This is due to the fact that the DNA implicated in the loop will be much larger than the dimensions of the protein itself. At these length scales the polymer will start feeling its own presence resulting in a self-avoiding random walk with a scaling of the end-to-end distance that goes as $L^{1.76}$ [67]. This self-avoiding random walk defines an average structure that behaves as a "blob" or mesh with a radius r_ξ and a corresponding contour length L_ξ. For longer length scales now the "blobs" start interacting with each other. There is a self-avoiding interaction of each blob with itself and with other chains (or very distant parts of the same chain), the first leading to expansion of the chain and the second to collapse. Interestingly, the Flory theorem states that these two effects will compensate each other. As a result for length scales

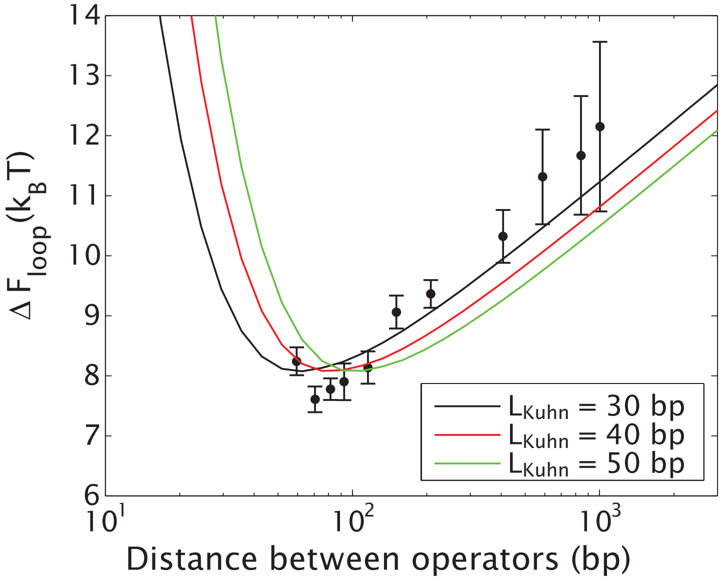


Figure 4.13: Estimation of the *in vivo* Kuhn length at length scales below 1 kbp. We show the looping free energy for cyclization calculated using the wormlike chain model [56] for several choices of the Kuhn length overlaid with the looping free energy calculated from the Müller et al. data. This approach bounds the *in vivo* Kuhn length between 40 and 60 bp.

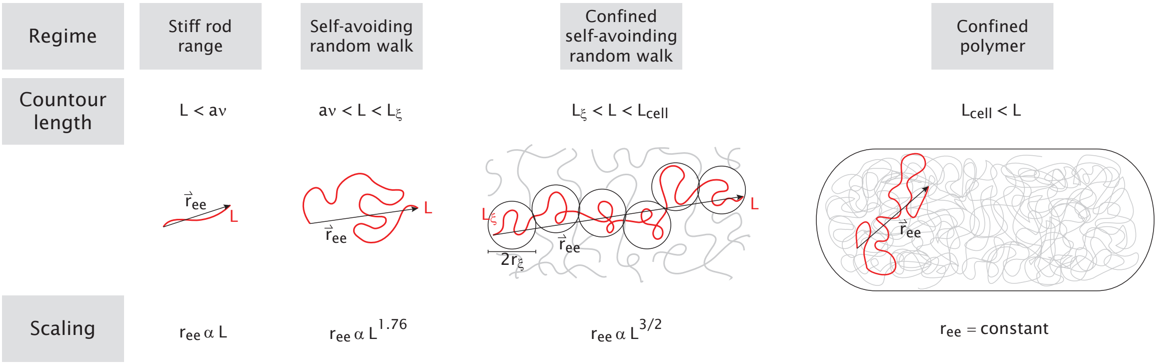
beyond the mesh size the blobs behave like a simple random walk polymer chain resulting in a scaling of the end-to-end distance of $L^{3/2}$ [68]. Finally, when we consider contour lengths much larger than the typical size of the cell the polymer is so tightly packed within the cell that its density becomes constant. Effectively, once we go to such long contour lengths the two ends behave as if they were not connected by the intervening polymer, resulting in uncorrelated behavior. This constant density translates into a constant end-to-end distance regardless of the contour length.

Given our previous estimate of the Kuhn length we have a clear expectation of where DNA will stop behaving like a stiff rod. However, it is not clear *a priori* where the boundary between self-avoiding and confined self-avoiding random walk regimes will lie. In order to determine that boundary we need to calculate the mesh size L_ξ. In order to do so we calculate the free energy corresponding to the polymer within the "blob", F_{blob}. This energy will consist of two terms. First, an entropy term related to the multiple polymer configuration compatible with a given mesh size r_ξ. Second, an energy term which describes the self-avoidance behavior between monomers, $F_{excluded}$ [66].

Fixman calculated the probability distribution for the radius of gyration of a self-avoiding polymer [69]. The entropy associated with this probability distribution is

$$S_{blob}(r) = -\beta^{-1}\left(-2\ln(r) + \frac{3}{2}\frac{r^2\pi^2}{Na^2}\right) + \text{const}, \tag{4.32}$$

where r is the radius of gyration, a is the Kuhn length and N is the number of Kuhn segments. Both r and a are usually expressed in nm. The number of segments is related to the length of the polymer in base pairs



Regime	Stiff rod range	Self-avoiding random walk	Confined self-avoinding random walk	Confined polymer
Countour length	L < av	av < L < L_ξ	L_ξ < L < L_{cell}	L_{cell} < L
Scaling	$r_{ee} \alpha$ L	$r_{ee} \alpha$ $L^{1.76}$	$r_{ee} \alpha$ $L^{3/2}$	r_{ee} = constant

Figure 4.14: Different regimes of scaling for a semiflexible polymer confined to a cell. For contour lengths below the Kuhn length, $a\nu$, DNA behaves like a stiff rod. The end-to-end distance r_{ee} scales linearly with the corresponding contour length L. As we go up in scale the polymer behaves like a self-avoiding random walk up to a length scale defined by the mesh size r_ξ and L_ξ. Because of the constrained volume given by the *E. coli* cell the Flory theorem states that for length scales beyond the mesh size, but smaller than the typical dimensions of the cell the effective polymer will behave like an entropic spring with a scaling $r_{ee} \propto L^{3/2}$. Finally, for contour lengths beyond the typical cell dimension, L_{cell}, the density of polymer monomers is uniform throughout the cell. As a result the end-to-end distance becomes constant and independent of the contour length.

L by $\frac{L}{\nu a}$, with $nu \simeq 3$ bp/nm the conversion between base pairs and nm for DNA. There is also a constant term with no dependence on r. This formula is valid for $\frac{6r^2}{Na^2} \gg 1$.

The term related to the excluded volume for a polymer that is approximated by a gas of hard cylinders is [66]

$$F_{excluded}(r) = \beta^{-1} N^2 \frac{3a^2 d}{8r^2}, \tag{4.33}$$

where d is the diameter of the DNA of about 2 nm.

The total Flory energy is then

$$F_{Flory}(r) = S_{blob}(r) + F_{excluded}(r). \tag{4.34}$$

We wish to find the value r_ξ that minimizes this energy by taking the derivative of F_{Flory} with respect to the radius and equating it to zero. The resulting expression is

$$\frac{dF_{flory}(r)}{dr}\beta = 0 = -\frac{2}{r_\xi} + \frac{3r_\xi \pi^2}{N_\xi a^2} + G_\xi^2 \frac{3a^2 d}{8r_\xi^2}, \tag{4.35}$$

where we have included the mesh contour length L_ξ associated with the mesh size r_ξ. We now make an approximation before solving this polynomial equation. Notice that the ratio between the first and second terms in equation 4.35 corresponding to the entropy is

$$\frac{2}{r_\xi} / \frac{3r_\xi \pi^2}{N_\xi a^2} = \frac{4}{\pi^2} \frac{N_\xi a^2}{6r_\xi^2} \ll 1. \tag{4.36}$$

We are making the assumption that, due to self-avoidance, the size of the blob, r is much larger than the size of the random walk polymer with the same number of Kuhn segments $Na^2/6$. As a result we can neglect the first term in equation 4.35 which leads to

$$r_\xi = \left(\frac{3}{8\pi^2}\right)^{1/5} N_\xi^{3/5} \left(a^4 d\right)^{1/5}. \tag{4.37}$$

We thus obtain a relation between the radius of the mesh and its corresponding contour length.

We now use an extra condition related to the DNA density within the cell. This density is to a first approximation uniform and should be equal to the DNA density within a blob. We can then equate both densities

$$\frac{L_{cell}}{V_{cell}} = \frac{L_\xi}{\frac{4}{3}\pi r_\xi^3}, \tag{4.38}$$

where L_{cell} is the size of the *E. coli* genome and V_{cell} is its volume. We now solve for the contour length L_ξ and obtain

$$L_\xi = \left(\frac{V_{cell}}{L_{cell}}\right)^{5/4} \frac{\nu^{9/4}}{\left(\frac{2}{3^2\pi}\right)^{1/4} (ad)^{3/4}} \simeq 1200 \text{ bp}. \tag{4.39}$$

We conclude that for loops longer than the Kuhn length and shorter than about 1200 bp the expected scaling of the looping probability should be of approximately $L^{1.76}$.

In order to determine if the looping free energy does have the expected scaling we fit it to the following formula

$$\Delta F_{loop}(L) = n \times \ln(L) + B, \tag{4.40}$$

with B being a constant factor that is related to the zero of energy. We fit the minima in looping free energy from the Müller et al. data. These minima correspond to the orientation between operators where the least cost in phasing is paid in order to loop. The resulting fit to equation 4.40 is shown in figure 4.15, where we have only used the data points with a loop length larger than 150 bp. In the same plot we show the best fit when we fix the exponent to be either $n = 1.76$ or $n = 3/2$. The result of the best fit is $n = 1.6 \pm 0.1$, which is in principle consistent with both the self-avoiding random walk and the constrained self-avoiding random walk regimes. Interestingly, our results are different from those obtained by Vilar and Saiz using similar models [57]. We are unaware of the reason for this discrepancy.

In conclusion, gene expression experiments for DNA-looping dependent promoters are able to set constraints on effective polymer models of DNA. They allow access to a length scale that is not usually resolvable using standard techniques [9]. Though our results are not entirely conclusive, we have shown that the behavior of DNA at this scale is consistent with a self-avoiding polymer *in vivo* and with a Kuhn length of 40 bp. This is consistent both with the position of the repression maximum and the scaling of repression as a function of DNA loop length at operator separations between 100 and 1000 bp [68]. More accurate data on DNA looping at these length scales would certainly increase our ability to constrain these polymer models.

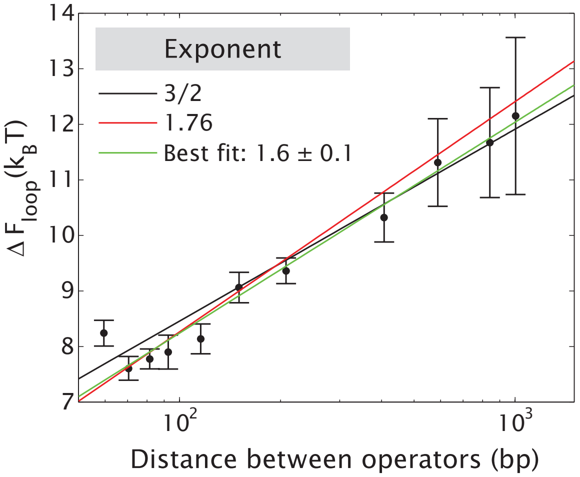


Figure 4.15: Using the looping energy to set constraints on polymer models of DNA. The looping free energy calculated from the data of Müller et al. [25] is fitted to the polymer models shown in figure 4.14. Note that we are only fitting data points for length of 150 bp and higher.

4.4.2 A model for the non-specific looping reservoir

One of the challenges in comparing results from different experiments is treating the arbitrariness of the reference energy. In the *in vivo* case we have worked with a definition in which the zero of energy is determined by the non-specific looping background energy F_{loop}^{NS}. If we were to compare this energy to the looping free energy coming from an *in vitro* assay we would have to shift one of the free energies accordingly. In this section we present an explicit model for the *in vivo* non-specific looping background. We show that we can account for this non-specific background in terms of very simple thermodynamic considerations.

We begin by replacing the looping free energy with an associated looping J-factor. A description in terms of this quantity makes the analysis more transparent. This J-factor is defined as the concentration of operator DNA in the vicinity of a Lac repressor binding head given that the other head is already bound to the other operator. The relation between the looping J-factor J_{loop} and the looping free energy is [70]

$$J_{loop} = c_0 e^{-\beta F_{loop}}, \tag{4.41}$$

where c_0 is the standard biochemical state which is usually taken as 1 M. We will suppress reference to the subscript in J_{loop} but remind the reader that this is not the same J as arises in cyclization.

In section 4.3.1 we defined the *in vivo* change in looping free energy as

$$e^{-\beta \Delta F_{loop}} = e^{-\beta (F_{loop}^{S} - F_{loop}^{NS})} = \frac{J^S}{J^{NS}}, \tag{4.42}$$

where we have used our definition for the looping J-factor from equation 4.41. The *in vivo* change in looping free energy is comprised then of two different looping J-factors: specific and the non-specific. Knowing the value of J^{NS} will then allow us to account for the difference in offset between the *in vivo* and *in vitro* looping free energies.

With reference to figure 7.8(B) we see that our model for the non-specific looping background consists of Lac repressor adopting all possible loop configurations between all available binding sites on the DNA. In equation 4.13 we defined the non-specific looping free energy as

$$e^{-\beta F_{loop}^{N}S} = \sum_{j}\sum_{\sigma} e^{-\beta F_{loop}(i_0,j,\sigma)}. \tag{4.43}$$

Here, i_0 is an arbitrary non-specific site on the DNA where the first binding head is placed. The second head then explores all available non-specific binding sites through the index j. Finally, each loop can be formed with four of the different tangent orientations shown in figure 4.5. The index σ keeps track of these orientations. Since the position i_0 is arbitrary we can just sum over a length L resulting in

$$J^{NS} = e^{-\beta F_{loop}^{N}S} = \sum_{L}\sum_{\sigma} e^{-\beta F_{loop}(L,\sigma)} = \sum_{L}\sum_{\sigma} J(L,\sigma). \tag{4.44}$$

In this last equation we have included the non-specific looping energy and expressed it in terms of a sum over all the possible looping J-factors between non-specific sites. By making a continuum approximation we can rewrite this expression as an integral

$$J^{NS} = \sum_{l,\sigma} J(L,\sigma) = \sum_{\sigma} \frac{1}{l_{bp}} \int_{1}^{N_{NS}} J(L,\sigma)\, dL, \tag{4.45}$$

with l_{bp} the length of a base pair.

To make progress we separate the integral in equation 4.45 into three regimes that are illustrated schematically in figure 4.16(B). First, looping at very short distances can occur, such that the distance between the DNA binding sites is comparable to the dimensions of the protein. As we go up in scale for the length of the loop we reach another regime where the geometrical details of the protein are not very relevant. This corresponds to looping at distances higher than 100 to 150 bp as discussed in section 4.4.1.2. Finally, we can have looping at distances which are so long that the two sites are effectively uncorrelated, as if each site was on separate DNA strands.

Equation 4.45 is basically telling us that the contribution of each of those non-specific looping regimes has to be added in order to determine the energy of the non-specific looping background. In figure 4.16 we present a cartoon of how we expect $J(L,\sigma)$ to behave for the different length regimes. It must be kept in mind that this is just a graphical way of qualitatively separating the behavior into different regimes and that plot is not meant to be interpreted literally with respect to the particular dividing regions. Splitting the looping J-factor into these regimes allows us to split the integral into three different terms, namely,

$$J^{NS} = \sum_{\sigma} \frac{1}{l_{bp}} \left(\int_{1}^{L_0} J_1(L,\sigma)\, dL + \int_{L_0}^{L_1} J_2(L,\sigma)\, dL + \int_{L_1}^{N_{NS}} J_3(L,\sigma)\, dL \right). \tag{4.46}$$

Next we estimate the contribution of each integral in the previous equation.

From the Müller et al. data [25] we know that there will be a maximum in the looping J-factor at around 70 bp and that it will decrease for shorter lengths, at least until 55 bp. Their construct does not allow them to go to much smaller interoperator spacings because the auxiliary operator would have to overlap with their promoter. For shorter distances we expect the J-factor to be highly dependent on the geometric details and flexibility of the LacI-DNA complex [58]. Because of the lack of such information we will assume that in this regime, J_1 is negligible compared to the contribution from the two other regimes.

As we saw in section 4.4.1.2 one description for the DNA that is consistent with the looping free energy is that of DNA in terms of a polymer that behaves as an entropic chain with a Kuhn length of around 40 bp. We will therefore assume that this regime can actually be described by the probability of closure or cyclization of a Gaussian chain [67] given by

$$J_2(L) = \left(\frac{3}{2\pi\, L\, a} \right)^{3/2}, \tag{4.47}$$

where $a = 40$ bp is the Kuhn length.

As we go to even longer lengths we expect the J-factor to converge to a constant. This is an effect of the DNA being constrained to the volume of the cell. *E. coli* has a genome with a length of $\sim 100,000$ Kuhn lengths whereas the typical dimension of the cell is of ~ 60 Kuhn lengths. This tells us that confinement should be a relevant effect. In this regime (also shown in figure 4.14) the concentration of one site in the vicinity of the other one cannot decrease below $V_{cell}^{-1} \simeq 1.67$ nM, since they are uncorrelated.

In order to calculate the contribution from the second regime we integrate $J_2(L)$. We assume that all four possible tangent orientations σ are described by the same expression, which results in a total contribution from this regime of the form

$$\sum_{\sigma} \frac{1}{l_{bp}} \int_{L_0}^{L_1} J_2(L,\sigma)\, dL = 4 \frac{1}{l_{bp}} \int_{L_0}^{L_1} \left(\frac{3}{2\pi\, L\, a} \right)^{3/2} dl < 4 \frac{1}{l_{bp}} \int_{L_0}^{+\infty} \left(\frac{3}{2\pi\, L\, a} \right)^{3/2} dl \simeq 50 \text{ mM}. \tag{4.48}$$

In this last calculation we have overestimated the contribution slightly by taking the limit $L_1 \to +\infty$.

At length scales larger than L_1 the distance between the two ends of a loop is so large that, even though the two ends are on the same polymer chain, they act as if they were uncorrelated or independent of each other. As a result, the concentration of one end in the vicinity of the other one is constant and can be approximated by the concentration of one base pair inside the cell $\frac{1}{V_{cell}} \simeq 1.67$ nM. By taking $J_3(L) \simeq \frac{1}{V_{cell}} \simeq 1.67$ nM the contribution from regime 3 to J^{NS} is then

$$\sum_{\sigma} \frac{1}{L_{bp}} \int_{L_1}^{N_{NS}} J_3(L,\sigma)\, dl \simeq 33 \text{ mM}. \tag{4.49}$$

We conclude the contribution of non-specific looping between uncorrelated, distant sites is comparable to the contribution of non-specific looping between sites that are within a range of 100 kbp. We estimate the non-specific J-factor to be $J^{NS} \simeq 80$ mM or $F_{loop}^{NS} \simeq 1.9\ k_B T$.

With this result in hand we can go back to the long distance looping free energy addressed in section 4.4.1.2. Even though we have a prediction for F_{loop}^{NS} we also fit it using the equation

$$e^{-\beta \Delta F_{loop}} = 4 \left(\frac{3}{2\pi\, L\, a} \right)^{3/2} \frac{1}{J^{NS}}. \tag{4.50}$$

The result of both the prediction and the fit are shown in figure 4.17. Through the fit we obtain $J^{NS} = (24 \pm 2)$ mM or $F_{loop}^{NS} = (-3.7 \pm 0.1)\ k_B T$. This is within a factor of 3.5 of our estimate of J^{NS}. One possible explanation for our overestimation of the non-specific looping background is that we accounted for more base pairs than are actually accessible in regime 3. Using recombination it has been shown that the *E. coli* region is separated into macrodomains [71]. It is possible that Lac repressor can only loop non-specifically between sites in the same macrodomain, reducing the value of J^{NS}. Additionally, it is important to point out that our prediction is based on modeling DNA as an entropic chain given by equation 4.47. Though in section 4.4.1.2 we show that the scaling of the looping free energy is consistent with this model, that does not say that equation 4.47 is a valid expression to describe it. The prefactors could differ, which would result in a different estimation of the contribution of regime 2. Nevertheless, we are satisfied with the relative success of the prediction based on our simple model. Now that we have a value for F_{loop}^{NS} we are ready to compare the *in vivo* looping free energy with different *in vitro* experiments.

4.4.2.1 Predicting the *in vitro* outcome for DNA looping by Lac repressor

To understand the similarities and differences between *in vivo* and *in vitro* DNA looping requires similar experiments to be performed systematically in both settings. The *in vivo* experiments of Müller et al. [25] described in the previous section are a concrete and excellent example of such systematic experimentation. On the *in vitro* front the looping probability of Lac repressor was recently quantified for several loop lengths [70]. However, in order to be able to plot both looping free energies or looping J-factors on the same plot we need to account for the difference in the definition of the zero of energy. This correction to the *in vivo* looping free energy is done by subtracting the contribution from the non-specific looping background F_{loop}^{NS} calculated in the previous section. Operationally, this is done by plugging the *in vivo* looping free energy obtained from the Müller et al. experiment in section 4.4.1 into equation 4.42 relating the looping free energy and the looping J-factor.

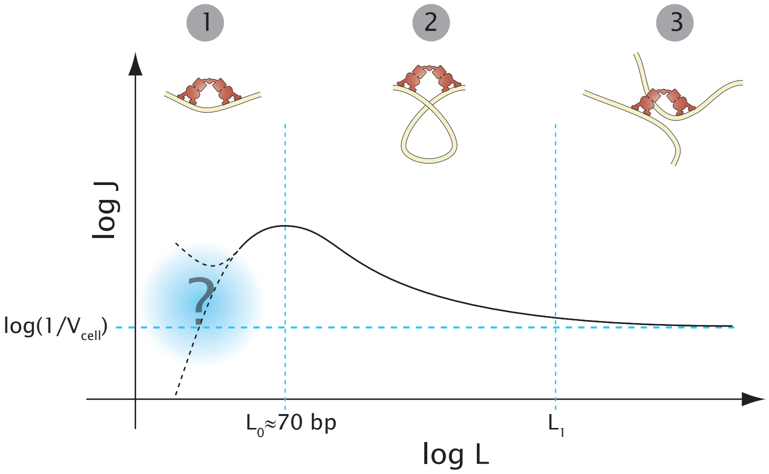


Figure 4.16: J-factor scaling with interoperator distance for Lac repressor. Cartoon of the different non-specific DNA looping regimes we will consider and their proposed corresponding contribution to the looping J-factor.

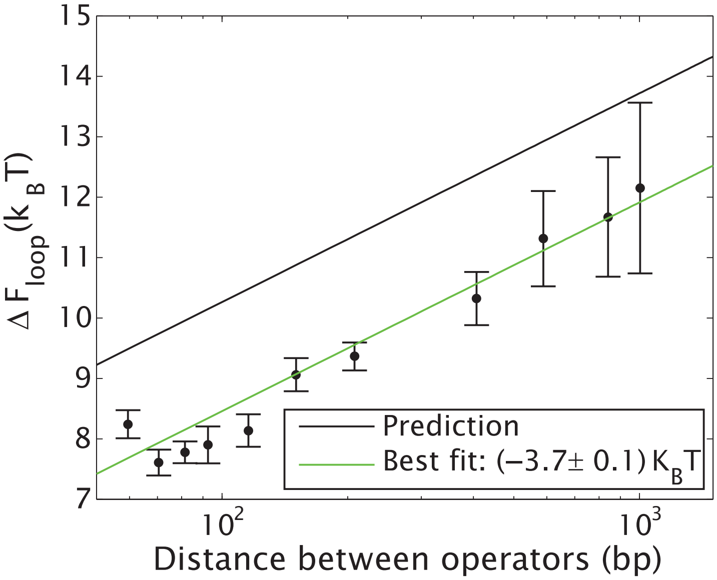


Figure 4.17: Determination of the contribution of the non-specific reservoir to the looping free energy. By fitting equation 4.50 to the looping free energy obtained from the Müller et al. data we estimate the contribution of the non-specific looping to the looping free energy to be $(3.7 \pm 0.1)\ k_BT$ or (24 ± 2) mM. This value is to be compared with our prediction generated using our toy model of 1.9 K_BT or 80 mM.

Figure 4.18 shows the *in vivo* looping J-factor from Müller et al. together with various *in vitro* quantifications of DNA looping in the context of Lac repressor. Additionally, we include recent experiments on DNA mechanics at short distances.

As mentioned above, the two most interesting data sets to compare are those of Müller et al. [25] and Han et al. [70] together with the theoretical expectation for the *in vitro* experiment based on Monte Carlo simulations [61]. First, the most naive use of the wormlike chain model for the *in vitro* looping free energy is not in accord with the *in vitro* data it is trying to describe. This could be attributed to a failure of the wormlike chain model itself. This scenario is reviewed in detail in [14]. However, a more plausible explanation is that the numerical simulations did not account for the geometrical and mechanical properties of the protein adequately [54, 59, 60, 72].

Regardless of our failure to understand the *in vitro* data in terms of first-principle models, it is still interesting to contrast the *in vivo* and *in vitro* looping J-factors. Whereas *in vitro* looping seems to become slightly less costly for lengths of 300 bp, the *in vivo* energy has its optimum at 70 bp. This observation implies that DNA is effectively more flexible *in vivo* than *in vitro*. This higher effective flexibility could come from the DNA itself as a result of a breakdown of the wormlike chain model. In fact, recent experiments on DNA cyclization have sparked significant controversy by suggesting that DNA is much more flexible at short length scales than what the wormlike chain model would predict. The results by Cloutier and Widom [73] shown in figure 4.18 should be contrasted with the theoretical expectation of the wormlike chain model also shown in that figure. Similar conclusions about the breakdown of ellasticity at short length scales have been reached by other works [74]. On the other hand, work by Du et al. [75] has shown a good agreement with the theoretical expectation. It is clear that this issue is far from settled.

The higher *in vivo* flexibility could also be due to the presence of nucleoid-associated proteins such as HU, H-NS and IHF [26, 53] or to the fact that the *in vivo* DNA is supercoiled [76, 77]. By deleting various nucleoid-associated proteins Becker et al. have shown that the fold-change in gene expression due to DNA looping by Lac repressor changes significantly. Developing a quantitative model based on structural information will most likely require the quantification of the effect of these proteins on looping over several length scales [54, 62, 78]. Additionally, being able to control the concentration of these nucleoid-associated proteins would allow the development of rudimentary equilibrium models of their contribution to the effective flexibility of DNA [79–81]. The role of supercoiling in DNA looping remains elusive. Even though Whitson et al. reported a significant increase in DNA looping when using supercoiled DNA [82, 83] the effect seen by Normanno et al. in their single molecule experiments where supercoiling was systematically controlled is much smaller [84]. It is clear that both significant experimentation and theoretical modeling are still necessary to dissect the effect of supercoiling in this system.

Though the comparison of the *in vivo* and *in vitro* looping J-factors leaves us with more questions than answers we view this as progress. Up until now, there was no direct way to compare the absolute values of each experimental outcome. Though our approach is model dependent it allows for an initial contrast

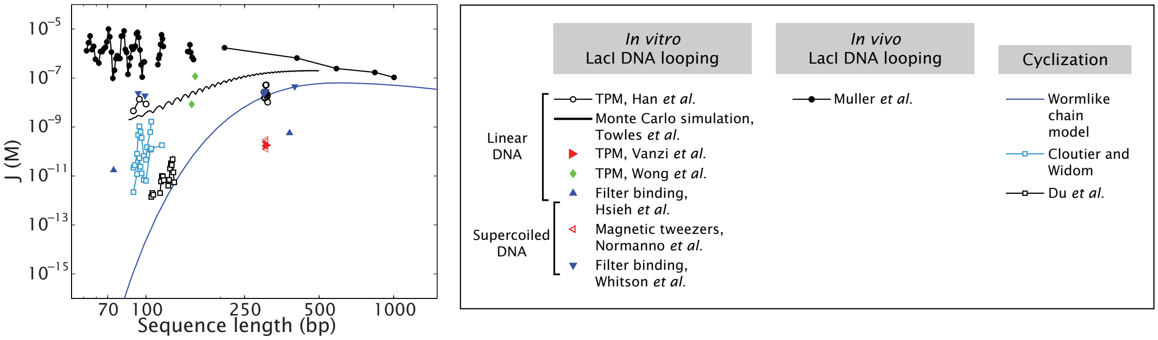


Figure 4.18: *In vivo* and *in vitro* experiments and calculations on DNA looping and DNA mechanics. The *in vivo* looping J-factor based on the Müller et al. [25] data using our model for the non-specific looping reservoir. This prediction is superimposed with looping J-factors derived from a variety of different *in vitro* measurements. Among them are the the looping J-factors derived from the *in vitro* experiments by Han et al., [70] and the corresponding theoretical expectation by Towles et al. [61] showing a significant disagreement between the theoretical, *in vivo* and *in vitro* aspects of the problem. In some of the other experiments shown here the looping J-factor was not reported explicitly and had to be estimated from the data. As such, some of the *in vitro* values should be viewed as approximations. From this plot it is clear that there is still a large spread in experimental data and that the experimental techniques need to be improved upon. Finally, we show results for DNA cyclization, where the propensity of DNA to form circles in the absence of any proteins is measured. Here, the results of Du et al. [75] agreeing with the theoretical expectation based on the wormlike chain model [56] are to be contrasted with the much higher flexibility obtained by Cloutier and Widom [73]. Other data sources are: Hsieh et al. [36], Whitson et al. [83], Vanzi et al. [85], Normanno et al. [84] and Wong et al. [86].

between the two settings based on quantitative arguments rather than qualitative comparisons.

4.5 Conclusion

The increasing availability of systematic and quantitative experiments on transcriptional regulation calls for the development of theoretical models that live up to these experiments by generating quantitative descriptions of the relevant transcriptional motifs and by suggesting new rounds of experimentation. In this paper we have quantitatively and systematically dissected the ubiquitous regulatory looping motif in bacteria.

Our bottom-up approach consisted in analyzing experiments on simpler constructs which allowed us to obtain fundamental parameters of the regulatory motif such as binding energies that could then be used for the more complex DNA looping architecture. As a result we were able to extract the *in vivo* change in looping free energy by Lac repressor. Through this framework we determined that one of the main features of the simple looping motif is a robustness in the level of expression with respect to fluctuations in the number of repressors. However, we also illustrated that this feature is lost in the wild-type *lac* operon. Interestingly, the multiple loops within the *lac* operon have similar probabilities of looping leading to similar levels of repression. Why evolution would lead to such a redundancy is a mystery we can only speculate on and that is beyond the scope of this paper. However, other authors have addressed this issue in detail [20, 39, 48].

Our models also propose a clear microscopic interpretation for the parameters obtained. In this context we analyzed experiments on *lac* operon mutants as a tool for reporting on the *in vivo* properties of DNA. We showed that this looping free energy gives predictive power, namely, that one can use the values deduced from one experiment performed under one set of conditions (binding site energy, number of repressors, etc.) in order to predict the outcome of another experiment under a completely different set of conditions.

DNA looping allows access to information about the *in vivo* properties of DNA on length scales no resolvable by standard techniques [9]. By contrasting the looping free energy with simple polymer models we show that the behavior of DNA at length scales below 1 kbp is consistent with a self-avoiding random walk and with a constrained-volume self-avoiding random walk.

We also showed that though some features of the looping free energy can be understood in terms of simple polymer models of DNA mechanics, there are fundamental differences between *in vivo* and *in vitro* DNA mechanics that can only be accessed though a new round of experimentation. For example, a next generation of *in vitro* DNA looping experiments should explore the effect of supercoiling and the presence of nucleoid associated proteins [84]. These parameters are harder to control *in vivo.* Still, recent experiments show that some level of manipulation of nucleoid associated proteins and supercoiling can be exerted in the *in vivo* context of DNA looping experiments [26, 53]. Rather than quantifying the effect of deletions of proteins such as HU on DNA looping, being able to titrate their intracellular concentration might allow for a better understanding of their role in DNA mechanics from the thermodynamic perspective.

We view such efforts as key in developing a picture of transcriptional regulation in terms of input-output functions. Here by just knowing a set of simple parameters such as the identity of binding sites and concentrations of the relevant transcription factors the output of a regulatory architecture can be predicted. Of course, more and more complex regulatory architectures will have to be dissected experimentally and theoretically in order to determine if thermodynamic models of transcriptional regulation are indeed a good tool for calculating such input-output functions.

Bibliography

[1] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[2] M. Ptashne. *A genetic switch: Phage lambda revisited.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 3rd edition, 2004.

[3] B. MÜLler-Hill. *The lac operon: A short history of a genetic paradigm.* Walter de Gruyter, Berlin, New York, 1996.

[4] K. S. Matthews. DNA looping. *Microbiol Rev*, 56(1):123–36, 1992.

[5] R. Schleif. DNA looping. *Annu Rev Biochem*, 61:199–223, 1992.

[6] Q. Li, S. Harju, and K. R. Peterson. Locus control regions: Coming of age at a decade plus. *Trends Genet*, 15(10):403–8, 1999.

[7] J. P. Noonan and A. S. Mccallion. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet*, 11:1–23, 2010.

[8] N. Naumova and J. Dekker. Integrating one-dimensional and three-dimensional maps of genomes. *J Cell Sci*, 123(Pt 12):1979–88, 2010.

[9] B. Van Steensel and J. Dekker. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol*, 28(10):1089–1095, 2010.

[10] W. Deng and G. A. Blobel. Do chromatin loops provide epigenetic gene expression states? *Curr Opin Genet Dev*, 2010.

[11] P. A. Wiggins, K. C. Cheveralls, J. S. Martin, R. Lintner, and J. Kondev. Strong intranucleoid interactions organize the *escherichia coli* chromosome into a nucleoid filament. *Proc Natl Acad Sci U S A*, 107(11):4991–5, 2010.

[12] E. Segal and J. Widom. What controls nucleosome positions? *Trends Genet*, 25(8):335–43, 2009.

[13] H. G. Garcia, P. Grayson, L. Han, M. Inamdar, J. Kondev, P. C. Nelson, R. Phillips, J. Widom, and P. A. Wiggins. Biological consequences of tightly bent DNA: The other life of a macromolecular celebrity. *Biopolymers*, 85(2):115–30, 2007.

[14] J. P. Peters and L. J. Maher. DNA curvature and flexibility in vitro and in vivo. *Quarterly Reviews of Biophysics*, 43(1):23–63, 2010.

[15] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[16] S. M. Law, G. R. Bellomy, P. J. Schlax, and M. T. J. Record. *in vivo* thermodynamic analysis of repression with and without looping in *lac* constructs. Estimates of free and local *lac* repressor concentrations and of physical properties of a region of supercoiled plasmid DNA *in vivo*. *J Mol Biol*, 230(1):161–73, 1993.

[17] R. R. Seabold and R. F. Schleif. Apo-arac actively seeks to loop. *J Mol Biol*, 278(3):529–38, 1998.

[18] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[19] L. Saiz, J. M. Rubi, and J. M. Vilar. Inferring the in vivo looping properties of DNA. *Proc Natl Acad Sci U S A*, 102(49):17642–5, 2005.

[20] J. M. Vilar. Accurate prediction of gene expression by integration of DNA sequence statistics with detailed modeling of transcription regulation. *Biophys J*, 99(8):2408–13, 2010.

[21] M. Lewis. The lac repressor. *C R Biol*, 328(6):521–48, 2005.

[22] S. Oehler, M. Amouyal, P. Kolkhof, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. Quality and position of the three *lac* operators of *e. Coli* define efficiency of repression. *EMBO J*, 13(14):3348–55, 1994.

[23] P. J. Schlax, M. W. Capp, and M. T. J. Record. Inhibition of transcription initiation by lac repressor. *J Mol Biol*, 245(4):331–50, 1995.

[24] S. Oehler, E. R. Eismann, H. Kramer, and B. MÜLler-Hill. The three operators of the lac operon cooperate in repression. *EMBO J*, 9(4):973–9, 1990.

[25] J. MÜLler, S. Oehler, and B. MÜLler-Hill. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol*, 257(1):21–9, 1996.

[26] N. A. Becker, J. D. Kahn, and L. J. Maher Iii. Bacterial repression loops require enhanced DNA flexibility. *J Mol Biol*, 349(4):716–30, 2005.

[27] L. Saiz and J. M. Vilar. DNA looping: The consequences and its control. *Curr Opin Struct Biol*, 16(3):344–50, 2006.

[28] H. G. Garcia and R. Phillips. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A*, 2011. (*Under review*).

[29] W. Runzi and H. Matzura. *in vivo* distribution of ribonucleic acid polymerase between cytoplasm and nucleoid in escherichia coli. *J Bacteriol*, 125(3):1237–9, 1976.

[30] P. H. Von Hippel, A. Revzin, C. A. Gross, and A. C. Wang. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: i. the *lac* operon: Equilibrium aspects. *Proc Natl Acad Sci U S A*, 71(12):4808–12, 1974.

[31] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'conner, D. W. Noble, and P. H. Von Hippel. Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound *escherichia coli lac* repressor *in vivo*. *Proc Natl Acad Sci U S A*, 74(10):4228–32, 1977.

[32] U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc Natl Acad Sci U S A*, 99(19):12015–20, 2002.

[33] M. Jishage and A. Ishihama. Regulation of rna polymerase sigma subunit synthesis in *escherichia coli*: Intracellular levels of sigma 70 and sigma 38. *J Bacteriol*, 177(23):6832–5, 1995.

[34] X. Zhang and P. A. Gottlieb. Thermodynamic and alkylation interference analysis of the *lac* repressor-operator substituted with the analogue 7-deazaguanine. *Biochemistry*, 32(42):11374–84, 1993.

[35] D. E. Frank, R. M. Saecker, J. P. Bond, M. W. Capp, O. V. Tsodikov, S. E. Melcher, M. M. Levandoski, and M. T. J. Record. Thermodynamics of the interactions of *lac* repressor with variants of the symmetric *lac* operator: Effects of converting a consensus site to a non-specific site. *J Mol Biol*, 267(5):1186–206, 1997.

[36] W. T. Hsieh, P. A. Whitson, K. S. Matthews, and R. D. Wells. Influence of sequence and distance between two operators on interaction with the *lac* repressor. *J Biol Chem*, 262(30):14583–91, 1987.

[37] M. Fried and D. M. Crothers. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 9(23):6505–25, 1981.

[38] M. Geanacopoulos, G. Vasmatzis, V. B. Zhurkin, and S. Adhya. Gal repressosome contains an antiparallel DNA loop. *Nat Struct Biol*, 8(5):432–6, 2001.

[39] L. Saiz and J. M. Vilar. Protein-protein/DNA interaction networks: Versatile macromolecular structures for the control of gene expression. *IET Syst Biol*, 2(5):247–55, 2008.

[40] A. Sanchez, H. G. Garcia, D. Jones, R. Phillips, and J. Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput Biol*, 2011. (*Under review*).

[41] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–6, 2008.

[42] A. Novick and M. Weiner. Enzyme induction as an all-or-none phenomenon. *Proc Natl Acad Sci U S A*, 43(7):553–566, 1957.

[43] E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. Van Oudenaarden. Multistability in the lactose utilization network of escherichia coli. *Nature*, 427(6976):737–40, 2004.

[44] J. M. Hudson and M. G. Fried. Co-operative interactions between the catabolite gene activator protein and the lac repressor at the lactose promoter. *J Mol Biol*, 214(2):381–96, 1990.

[45] A. Balaeff, L. Mahadevan, and K. Schulten. Structural basis for cooperative DNA binding by cap and lac repressor. *Structure*, 12(1):123–32, 2004.

[46] T. Kuhlman, Z. Zhang, M. H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *escherichia coli*. *Proc Natl Acad Sci U S A*, 104(14):6043–8, 2007.

[47] L. Saiz and J. M. Vilar. Efficiency and versatility of distal multisite transcription regulation. *arXiv:0704.3264v1 [q-bio.SC]*, 2007.

[48] L. Saiz and J. M. Vilar. Ab initio thermodynamic modeling of distal multisite transcription regulation. *Nucleic Acids Res*, 36(3):726–31, 2008.

[49] R. W. Zeller, J. D. Griffith, J. G. Moore, C. V. Kirchhamer, R. J. Britten, and E. H. Davidson. A multimerizing transcription factor of sea urchin embryos capable of looping DNA. *Proc Natl Acad Sci U S A*, 92(7):2989–93, 1995.

[50] N. Lehming, J. Sartorius, M. Niemoller, G. Genenger, B. V Wilcken-Bergmann, and B. Muller-Hill. The interaction of the recognition helix of lac repressor with lac operator. *EMBO J*, 6(10):3145–53, 1987.

[51] P. H. Viollier, M. Thanbichler, P. T. Mcgrath, L. West, M. Meewan, H. H. Mcadams, and L. Shapiro. Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc Natl Acad Sci U S A*, 101(25):9257–62, 2004.

[52] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009.

[53] N. A. Becker, J. D. Kahn, and L. J. Maher 3rd. Effects of nucleoid proteins on DNA repression loop formation in *escherichia coli*. *Nucleic Acids Res*, 35(12):3988–4000, 2007.

[54] Y. Zhang, A. E. Mcewen, D. M. Crothers, and S. D. Levene. Analysis of in-vivo lacr-mediated gene repression based on the mechanics of DNA looping. *PLoS ONE*, 1:e136, 2006.

[55] D. H. Lee and R. F. Schleif. In vivo DNA loops in aracbad: Size limits and helical repeat. *Proc Natl Acad Sci U S A*, 86(2):476–80, 1989.

[56] H. Yamakawa. *Helical wormlike chains in polymer solutions*. Springer, Berlin, New York, 1997.

[57] J. M. Vilar and L. Saiz. DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. *Curr Opin Genet Dev*, 15(2):136–44, 2005.

[58] K. Rippe. Making contacts on a nucleic acid polymer. *Trends Biochem Sci*, 26(12):733–40, 2001.

[59] D. Swigon, B. D. Coleman, and W. K. Olson. Modeling the lac repressor-operator assembly: The influence of DNA looping on lac repressor conformation. *Proc Natl Acad Sci U S A*, 103(26):9879–84, 2006.

[60] Y. Zhang, A. E. Mcewen, D. M. Crothers, and S. D. Levene. Statistical-mechanical theory of DNA looping. *Biophys J*, 90(6):1903–12, 2006.

[61] K. B. Towles, J. F. Beausang, H. G. Garcia, R. Phillips, and P. C. Nelson. First-principles calculation of DNA looping in tethered particle experiments. *Phys Biol*, 6(2):25001, 2009.

[62] L. Saiz and J. M. Vilar. Multilevel deconstruction of the in vivo behavior of looped DNA-protein complexes. *PLoS ONE*, 2(4):e355, 2007.

[63] L. Ringrose, S. Chabanis, P. O. Angrand, C. Woodroofe, and A. F. Stewart. Quantitative comparison of DNA looping in vitro and in vivo: Chromatin increases effective DNA flexibility at short distances. *Embo J*, 18(23):6630–41, 1999.

[64] J. Dekker. Mapping in vivo chromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction. *J Biol Chem*, 283(50):34532–40, 2008.

[65] A. Miele and J. Dekker. Long-range chromosomal interactions and gene regulation. *Mol Biosyst*, 4(11):1046–57, 2008.

[66] R. Phillips, J. Kondev, and J. Theriot. *Physical biology of the cell*. Garland Science, New York, 2009. (Illustrated by N. Orme; with problems, solutions, and editorial assistance of H. G. Garcia.).

[67] M. Doi and S. F. Edwards. *The theory of polymer dynamics*. Clarendon Press, Oxford University Press, New York, 1986.

[68] P. G. De Gennes. *Scaling concepts in polymer physics.* Cornell University Press, Ithaca, N.Y., 1979.

[69] M. Fixman. Radius of gyration of polymer chains. *Journal of Chemical Physics*, 36(2):306, 1962.

[70] L. Han, H. G. Garcia, S. Blumberg, K. B. Towles, J. F. Beausang, P. C. Nelson, and R. Phillips. Concentration and length dependence of DNA looping in transcriptional regulation. *PLoS One*, 4(5):e5621, 2009.

[71] M. Valens, S. Penaud, M. Rossignol, F. Cornet, and F. Boccard. Macrodomain organization of the escherichia coli chromosome. *Embo J*, 23(21):4330–41, 2004.

[72] J. F. Allemand, S. Cocco, N. Douarche, and G. Lia. Loops in DNA: An overview of experimental and theoretical approaches. *European Physical Journal E*, 19(3):293–302, 2006.

[73] T. E. Cloutier and J. Widom. DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc Natl Acad Sci U S A*, 102(10):3645–50, 2005.

[74] P. A. Wiggins, T. Van Der Heijden, F. Moreno-Herrero, A. Spakowitz, R. Phillips, J. Widom, C. Dekker, and P. C. Nelson. High flexibility of DNA on short length scales probed by atomic force microscopy. *Nat Nanotechnol*, 1(2):137–41, 2006.

[75] Q. Du, C. Smith, N. Shiffeldrim, M. Vologodskaia, and A. Vologodskii. Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proc Natl Acad Sci U S A*, 102(15):5397–402, 2005.

[76] A. Vologodskii and N. R. Cozzarelli. Effect of supercoiling on the juxtaposition and relative orientation of DNA sites. *Biophys J*, 70(6):2548–56, 1996.

[77] P. K. Purohit and P. C. Nelson. Effect of supercoiling on formation of protein-mediated DNA loops. *Phys Rev E Stat Nonlin Soft Matter Phys*, 74(6 Pt 1):061907, 2006.

[78] L. Czapla, D. Swigon, and W. K. Olson. Effects of the nucleoid protein hu on the structure, flexibility, and ring-closure properties of DNA deduced from monte carlo simulations. *J Mol Biol*, 382(2):353–70, 2008.

[79] R. T. Dame and N. Goosen. Hu: Promoting or counteracting DNA compaction? *FEBS Lett*, 529(2–3):151–6, 2002.

[80] J. Van Noort, S. Verbrugge, N. Goosen, C. Dekker, and R. T. Dame. Dual architectural roles of hu: Formation of flexible hinges and rigid filaments. *Proc Natl Acad Sci U S A*, 101(18):6969–74, 2004.

[81] P. A. Wiggins, R. T. Dame, M. C. Noom, and G. J. Wuite. Protein-mediated molecular bridging: A key mechanism in biopolymer organization. *Biophys J*, 97(7):1997–2003, 2009.

[82] P. A. Whitson, W. T. Hsieh, R. D. Wells, and K. S. Matthews. Supercoiling facilitates lac operator-repressor-pseudooperator interactions. *J Biol Chem*, 262(11):4943–6, 1987.

[83] P. A. Whitson, W. T. Hsieh, R. D. Wells, and K. S. Matthews. Influence of supercoiling and sequence context on operator DNA binding with *lac* repressor. *J Biol Chem*, 262(30):14592–9, 1987.

[84] D. Normanno, F. Vanzi, and F. S. Pavone. Single-molecule manipulation reveals supercoiling-dependent modulation of *lac* repressor-mediated DNA looping. *Nucleic Acids Res*, 36(8):2505–13, 2008.

[85] F. Vanzi, C. Broggio, L. Sacconi, and F. S. Pavone. Lac repressor hinge flexibility and DNA looping: Single molecule kinetics by tethered particle motion. *Nucleic Acids Res*, 34(12):3409–20, 2006.

[86] O. K. Wong, M. Guthold, D. A. Erie, and J. Gelles. Interconvertible lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biol*, 6(9):e232, 2008.

Chapter 5

Effect of Promoter Architecture on the Cell-to-Cell Variability in Gene Expression

This chapter is a reproduction of reference [1].

According to recent experimental evidence, promoter architecture, defined as the number, strength and regulatory role of the operators that control transcription from a promoter, plays a major role in determining the level of cell-to-cell variability in gene expression. These quantitative experiments call for a corresponding modeling effort that addresses the question of how changes in promoter architecture affect noise in gene expression, in a systematic rather than case-by-case fashion. In this article, we make such a systematic investigation, based on a microscopic model of gene regulation that incorporates stochastic effects. In particular, we show how operator strength and operator multiplicity affect this variability. We examine different modes of transcription factor binding to complex promoters (cooperative, independent, simultaneous) and how each of these affects the level of variability in transcriptional output from cell-to-cell. We propose that direct comparison between *in vivo* single-cell experiments and theoretical predictions for the moments of the probability distribution of mRNA number per cell can be used to test kinetic models of gene regulation. The emphasis of the discussion is on prokaryotic gene regulation, but our analysis can be extended to eukaryotic cells as well.

5.1 Introduction

A fundamental property of all living organisms is their ability to gather information about their environment and adjust their internal, physiological state in response to environmental conditions. This property, shared by all organisms, includes the ability of single-cells to respond to changes in their environment by regulating their patterns of gene expression. By regulating the genes they express, cells are able to survive, for example, changes in the extracellular pH or osmotic pressure, switch the mode of sugar utilization when the sugar content in their medium changes, or respond to shortages in key metabolites by adapting their metabolic pathways. Perhaps more interesting is the organization of patterns of gene expression in space and time resulting in the differentiation of cells into different types, which is one of the defining features of multicellular organisms. Much of this regulation occurs at the level of transcription initiation, and is mediated by simple interactions between transcription factor proteins and DNA, leading to genes being turned on or off. Understanding how genes are turned on or off (as well as the more nuanced expression patterns in which the level of expression takes intermediate levels) at a mechanistic level has been one of the great challenges of molecular biology and has attracted intense attention over the past 50 years.

The current view of transcription and transcriptional regulation has been strongly influenced by recent experiments with single-cell and single-molecule resolution [2–12]. These experiments have confirmed the long-suspected idea that gene expression is stochastic [13, 14], meaning that different steps on the path from gene to protein occur at random. This stochasticity also causes variability in the number of messenger RNAs (mRNA) and proteins produced from cell-to-cell in a colony of isogenic cells [12, 15–18]. The question of how transcriptional regulatory networks function reliably in spite of the noisy character of the inputs and outputs has attracted much experimental and theoretical interest [19, 20]. A different, but also very relevant, question is whether cells actually exploit this stochasticity to fulfill any physiologically important task. This issue has been investigated in many different cell types and it has been found that stochastic gene expression is a key player in processes as diverse as cell fate determination in the retina of *D. melanogaster* [21], entrance to the competent state of *B. subtilis* [8], resistance of yeast colonies to antibiotic challenge [18], maintenance of HIV latency [22], promoting host infection by pathogens [23] or the induction of the lactose operon in *E. coli* [24]. Other examples have been found, and reviewed elsewhere [25, 26]. The overall conclusion of all of these studies is that noise in gene expression can have important physiological consequences in natural and synthetic systems and that the overall architecture of the gene regulatory network can greatly affect the level of stochasticity.

A number of theoretical and experimental studies have revealed multiple ways in which the architecture of the gene regulatory network affects cell-to-cell variability in gene expression. Examples of mechanisms for the control of stochasticity have been proposed and tested, including the regulation of translational efficiency [9], the presence of negative feedback loops [27–29], or the propagation of fluctuations from upstream regulatory components [30]. Another important source of stochasticity in gene expression is fluctuations in promoter activity, caused by stochastic association and dissociation of transcription factors, chromatin remodeling events, and formation of stable pre-initiation complexes [6, 16, 17, 24, 31]. In particular, it has been reported that perturbations to the architecture of yeast and bacterial promoters, such as varying the strength of transcription factor binding sites [18], the number and location of such binding sites [12, 32], the presence of auxiliary operators that mediate DNA looping [24], or the competition of activators and repressors for binding to the same stretch of DNA associated with the promoter [33], may strongly affect the level of variability.

Our goal is to examine all of these different promoter architectures from a unifying perspective provided by stochastic models of transcription leading to mRNA production. The logic here is the same as in earlier work where we examined a host of different promoter architectures using thermodynamic models of transcriptional regulation [34, 35]. We generalize those systematic efforts to examine the same architectures, but now from the point of view of stochastic models. Stochastic models allow us to assess the unique signature provided by a particular regulatory architecture in terms of the cell-to-cell variability it produces.

First, we investigate in general theoretical terms how the architecture of a promoter affects the level of cell-to-cell variability. The architecture of a promoter is defined by the collection of transcription factor binding sites (also known as operators), their number, position within the promoter, their strength, as well as what kind of transcription factors bind them (repressors, activators or both), and how those transcription factors bind to the operators (independently, cooperatively, simultaneously). We apply the master-equation model of stochastic gene expression [36–39] to increasingly complex promoter architectures [31], and compute the moments of the mRNA and protein distributions expected for these promoters. Our results provide an expectation for how different architectural elements affect cell-to-cell variability in gene expression.

The second point of this paper is to make use of stochastic kinetic models of gene regulation to put forth *in vivo* tests of the molecular mechanisms of gene regulation by transcription factors that have been proposed as a result of *in vitro* biochemical experiments. The idea of using spontaneous fluctuations in gene expression to infer properties of gene regulatory circuits is an area of growing interest, given its non-invasive nature and its potential to reveal regulatory mechanisms *in vivo*. Different theoretical methods have recently been proposed, which could be employed to distinguish between different modes (e.g., AND/OR) of combinatorial gene regulation, and to rule out candidate regulatory circuits [28, 40, 41] based solely on properties of noise in gene expression, such as the autocorrelation function of the fluctuations [28] or the three-point steady-state correlations between multiple inputs and outputs [40, 41].

Here, we make experimentally testable predictions about the level of cell-to-cell variability in gene expression expected for different bacterial promoters, based on the physical kinetic models of gene regulation that are believed to describe these promoters *in vivo*. In particular, we focus on how varying the different parameters (i.e., mutating operators to make them stronger or weaker, varying the intracellular concentration of transcription factors, etc.) should affect the level of variability. This way, cell-to-cell variability in gene expression is used as a tool for testing kinetic models of transcription factor mediated regulation of gene expression *in vivo*.

The remainder of the paper is organized as follows: First we describe the theoretical formalism we use to determine analytic expressions for the moments of the probability distribution for both mRNA and protein abundances per cell. Next, we examine how the architecture of the promoter affects cell-to-cell variability in gene expression. We focus on simple and cooperative repression, simple and cooperative activation, and transcriptional regulation by distal operators mediated by DNA looping. We investigate how noise in gene expression caused by promoter activation differs from repression, how operator multiplicity affects noise in gene expression, the effect of cooperative binding of transcription factors, as well as DNA looping. For each one of these architectures we present a prediction of cell-to-cell variability in gene expression for a bacterial promoter that has been well characterized experimentally in terms of their mean expression values. These predictions suggest a new round of experiments to test the current mechanistic models of gene regulation at these promoters.

5.2 Methods

In order to investigate how promoter architecture affects cell-to-cell variability in gene expression, we use a model based on classical chemical kinetics (illustrated in figure 5.1(A)), in which a promoter containing multiple operators may exist in as many biochemical states as allowed by the combinatorial binding of transcription factors to its operators. The promoter transitions stochastically between the different states as transcription factors bind and fall off. Synthesis of mRNA is assumed to occur stochastically at a constant rate that is different for each promoter state. Further, transcripts are assumed to be degraded at a constant rate per molecule.

This kind of model is the kinetic counterpart of the so-called "thermodynamic model" of transcriptional regulation [42], and it is the standard framework for interpreting the kinetics of gene regulation in biochemical experiments, both *in vivo* [3, 24] and *in vitro* [42, 43]. This class of kinetic models can easily accommodate stochastic effects, and it leads to a master equation from which the probability distribution of mRNA and protein copy number per cell can be computed. It is often referred to as the standard model of stochastic gene expression [39, 44, 45]. The degree of cell-to-cell variability in gene expression can be quantified by the stationary variance, defined as the ratio of the standard deviation and the mean of the probability distribution of mRNA or protein copy number per cell [36], or else by the Fano factor, the ratio between the variance and the mean. These two are the two most common metrics of noise in gene expression, and the relation between them will be discussed later.

In order to compute the noise strength from this class of models, we follow the same approach as in a previous article [31], which extends a master equation derived elsewhere [37, 38, 46] to promoters with arbitrary combinatorial complexity. The complexity refers to the existence of a number of discrete promoter states corresponding to different arrangements of transcription factors on the promoter DNA. Promoter dynamics are described by trajectories involving stochastic transitions between promoter states which are induced by the binding and unbinding of transcription factors. A detailed derivation of the equations which describe promoter dynamics can be found in the Appendix, but the essentials are described below.

There are only two stochastic variables in the model: the number of mRNA transcripts per cell, which is represented by the unitless state variable m, and the state of the promoter, which is defined by the pattern of transcription factors bound to their operator sites. The promoter state is described by a discrete and finite stochastic variable (s) (for an example, see figure 5.1(A)). The example in figure 5.1(A) illustrates the simplest model of transcriptional activation by a transcription factor. When the activator is not bound (state 1), mRNA is synthesized at rate r_1. When the activator is bound to the promoter (state 2), mRNA is synthesized at the higher rate r_2. The promoter switches stochastically from state 1 to state 2 with rate k_A^{on}, and from state 2 to state 1 with rate k_A^{off}. Each mRNA molecule is degraded with rate γ.

The time evolution for the joint probability of having the promoter in states 1 or 2, with m mRNAs in the cell (which we write as $p(1,m)$ and $p(2,m)$, respectively), is given by a master equation, which we can build by listing all possible reactions that lead to a change in cellular state, either by changing m or by changing s (figure 5.1(B)). The master equation takes the form:

$$\begin{aligned}
\frac{d}{dt}p(1,m) &= -k_A^{on}p(1,m) + k_A^{off}p(2,m) - r_1 p(1,m) - \gamma m p(1,m) + r_1 p(1,m-1) + \gamma(m+1)p(1,m+1), \qquad (5.1)\\
\frac{d}{dt}p(2,m) &= k_A^{on}p(1,m) - k_A^{off}p(2,m) - r_2 p(2,m) - \gamma m p(2,m) + r_2 p(2,m-1) + \gamma(m+1)p(2,m+1).
\end{aligned}$$

Inspecting this system of equations, we notice that by defining the vector:

$$\vec{p}(m) = \begin{pmatrix} p(1,m) \\ p(2,m) \end{pmatrix}, \tag{5.2}$$

and the matrices

$$\hat{K} = \begin{bmatrix} -k_A^{on} & k_A^{off} \\ k_A^{on} & k_A^{off} \end{bmatrix}; \hat{R} = \begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix}; \hat{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.3}$$

we can rewrite the system of equations 5.1 in matrix form

$$\frac{d}{dt}\vec{p}(m) = \left[\hat{K} - \hat{R} - m\gamma\hat{I}\right]\vec{p}(m) + \hat{R}\vec{p}(m-1) + (m+1)\gamma\hat{I}\vec{p}(m+1). \tag{5.4}$$

This has several advantages, but the most important one is that the matrix approach reduces the task of obtaining analytical expressions for the moments of the steady-state mRNA distribution for an arbitrarily complex promoter to solving two simple linear matrix equations (more details are given in the Appendix).

The matrices appearing in equation 5.4 all have simple and intuitive interpretations. The matrix $\hat{K}$ describes the stochastic transitions between promoter states: The off-diagonal elements of the matrix $\hat{K}_{ij}$ are the rates of making transitions from promoter state (j) to promoter state (i). The diagonal elements of the matrix $\hat{K}_{jj}$ are negative, and they represent the net probability flux out of state (j): $\hat{K}_{jj} = -\sum_{i \neq j} \hat{K}_{ij}$. The matrix $\hat{R}$ is a diagonal matrix whose element $\hat{R}_{jj}$ gives the rate of transcription initiation when the promoter is in state (j). Finally, the matrix $\hat{I}$ is the identity matrix.

An example of matrices $\hat{K}$ and $\hat{R}$ is presented pictorially in figure 5.2. It is straightforward to see that even though equation 5.4 has been derived for a two-state promoter, it also applies to any other promoter architecture. What will change for different architectures are the dimensions of the matrices and vectors (these are given by the number of promoter states) as well as the values of the rate constants that make up the matrix elements of the various matrices.

An important limit of the master equation, which is often attained experimentally, is the steady-state limit, where the probability distribution for mRNA number per cell does not change with time. Although the time dependence of the moments of the mRNA distribution can be easily computed from our model, for the sake of simplicity and because most experimental studies have been performed on cells in steady-state, we focus on this limit. As shown in the Appendix, analytic expressions for the first two moments of the steady-state mRNA probability distribution are found by multiplying both sides of equation 5.4 by m and m^2, respectively, and then summing m from 0 to infinity. After some algebra (elaborated in an earlier paper and in the SI), we find that the first two moments can be written as:

$$\langle m \rangle = \frac{\vec{r} \cdot \vec{m}_{(0)}}{\gamma}, \tag{5.5}$$

$$\langle m^2 \rangle = \langle m \rangle + \frac{\vec{r} \cdot \vec{m}_{(1)}}{\gamma}. \tag{5.6}$$

The vector $\vec{r}$ contains the ordered list of rates of transcription initiation for each promoter state. For the two-state promoter shown in figure 5.1, $\vec{r} = (r_1, r_2)$. The vector $\vec{m}_{(0)}$ contains the steady-state probabilities for finding the promoter in each one of the possible promoter states, while $\vec{m}_{(1)}$ is the steady-state mean mRNA number in each promoter state. The vector $\vec{m}_{(0)}$ is the solution to the matrix equation

$$\hat{K}\vec{m}_{(0)} = 0, \tag{5.7}$$

while the vector is obtained from

$$\left(\hat{K} - \gamma\hat{I}\rangle\right)\vec{m}_{(1)} + \hat{R}\vec{m}_{(0)} = 0. \tag{5.8}$$

Figure 5.1 illustrates the following algorithm for computing the intrinsic variability of mRNA number for promoters of arbitrarily complex architecture:

1. Make a list of all possible promoter states and their kinetic transitions (figure 5.1(B)).
2. Construct the matrices $\hat{K}$ and $\hat{R}$, and the vector $\vec{r}$, (figure 5.2).
3. Solve equations 5.7 and 5.8 to obtain $\vec{m}_{(0)}$ and $\vec{m}_{(1)}$.
4. Plug solutions of equations 5.7 and 5.8 into equations 5.5 and 5.6 to obtain the moments.

The normalized variance of the mRNA distribution in steady-state is then computed from the equation:

$$\eta^2 = \frac{Var(m)}{\langle m \rangle^2} = \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle^2} = \frac{1}{\langle m \rangle} + 1over\langle m \rangle \left(\frac{\vec{r} \cdot \vec{m}_{(1)}}{\gamma} - \langle m \rangle^2\right). \tag{5.9}$$

Equation 5.9 reveals that, regardless of the specific details characterizing promoter architecture, the intrinsic noise is always the sum of two components, and it can be written as

$$\eta^2 = \frac{1}{\langle m \rangle} + \eta^2_{promoter}. \tag{5.10}$$

The first component is due to spontaneous stochastic production and degradation of single mRNA molecules, it is always equal to the Poissonian expectation of $1/\langle m \rangle$, and is independent of the architecture of the promoter. For an unregulated promoter that is always active and does not switch between multiple states (or does so very fast compared to the rates of transcription and mRNA degradation), the mRNA distribution is well described by a Poisson distribution [45, 47], and the normalized variance is equal to $1/\langle m \rangle$. The second component ("promoter noise") results from promoter state fluctuations, and captures the effect of the promoter's architecture on the cell-to-cell variability in mRNA:

$$\eta^2_{promoter} = \frac{1}{\langle m \rangle^2} \left(\frac{\vec{r} \cdot \vec{m}_{(1)}}{\gamma} - \langle m \rangle^2 \right). \tag{5.11}$$

In order to quantify the effect of the promoter architecture in the level of cell-to-cell variability in mRNA expression, we define the deviation in the normalized variance caused by gene regulation relative to the baseline Poisson noise for the same mean (see figure 5.3):

$$\text{Fold-change in mRNA noise} = \frac{\eta^2}{\eta^2_{Poisson}} = \frac{Var(m)/\langle m \rangle^2}{1/\langle m \rangle} = \frac{Var(m)}{\langle m \rangle}. \tag{5.12}$$

Therefore, the deviation in the normalized variance caused by gene regulation is equal to the ratio between the variance and the mean. This parameter is also known as the Fano factor. Thus, for any given promoter architecture, the Fano factor quantitatively characterizes how large the mRNA noise is relative to that of a Poisson distribution of the same mean (i.e., how much the noise for the regulated promoter elevates with respect to the Poisson noise). This is the parameter that we will use throughout the paper as the metric of cell-to-cell variability in gene expression.

5.2.1 Promoter noise and variability of mRNA and protein numbers

For proteins, the picture is only slightly more complicated. As shown in the Appendix, in the limit where the lifetime of mRNA is much shorter than that of the protein it encodes for (a limit that is often fulfilled [31]), the noise strength of the probability distribution of proteins per cell takes the following form (where we define n as a state variable that represents the copy number of proteins per cell):

$$\frac{Var(n)}{\langle n \rangle^2} = \frac{\langle n^2 \rangle - \langle n \rangle^2}{\langle n \rangle^2} = \frac{1+b}{\langle n \rangle} + \frac{1}{\langle n \rangle^2} \left(b \frac{\vec{r} \cdot \vec{n}_{(1)}}{\gamma_{protein}} - \langle n \rangle^2 \right), \tag{5.13}$$

where $\gamma_{protein}$ stands for the protein degradation rate, and the constant b is equal to the protein burst size (the average number of proteins produced by one mRNA molecule). The mean protein per cell is given by

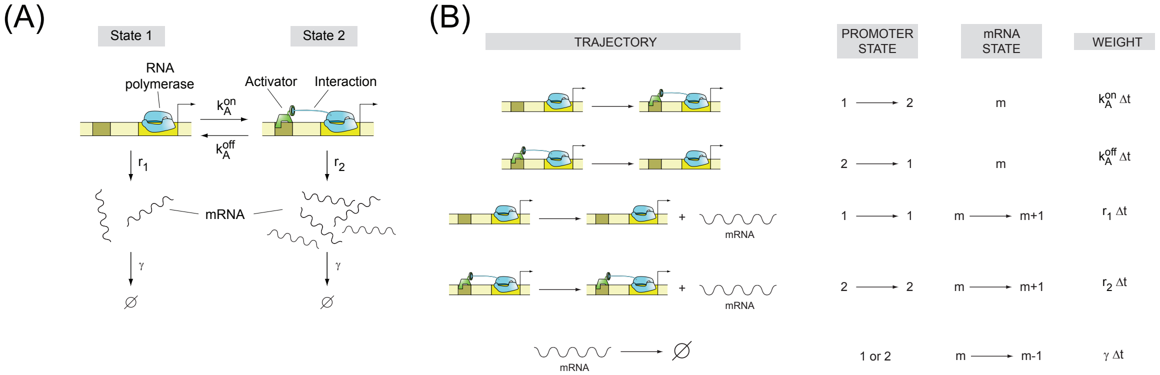


Figure 5.1: Two-state promoter. (A) Simple two-state bacterial promoter undergoing stochastic activation by a transcriptional activator binding to a single operator site. The rates of activator association and dissociation are given by k_A^{on} and k_A^{off}, respectively, and the rates of mRNA production for the basal and active states are r_1 and r_2, respectively. The mRNA degradation rate is assumed to be constant for each molecule, and is given by the parameter γ. (B) List of all possible stochastic transitions affecting either the copy number of mRNA (m) or the state of the promoter (s) and their respective statistical weight. State 1 has the operator free. State 2 is the activator bound state. The weights represent the probability that each change of state will occur during a time increment Δt . The master equation is constructed based on these rules.

$\langle n \rangle = b\frac{\vec{r}\cdot\vec{m}_{(0)}}{\gamma_{protein}}$, and the vector $\vec{n}_{(1)}$ is the solution of the algebraic equation:

$$\left(\hat{K} - \gamma_{protein}\hat{I}\right)\vec{n}_{(1)} + b\hat{R}\vec{m}_{(0)} = 0. \tag{5.14}$$

The reader is referred to the Appendix for a detailed derivation and interpretation of these equations. In the previous section we have shown that the noise for proteins and mRNA take very similar analytical forms. Indeed, if we define $\vec{r_n} = b\vec{r}$ and $\hat{R}_n = b\hat{R}$, as the vector and matrix containing the average rates of protein synthesis for each promoter state, it is straightforward to see that equations 5.8 and 5.14 are mathematically equivalent, with the only difference being that in equation 5.14 the matrix $\hat{R}_n$ represents the rates of protein synthesis, so all the rates of transcription are multiplied by the translation burst size b. Therefore, the vectors $\vec{m}_{(1)}$ and $\vec{n}_{(1)}$ are only going to differ in the prefactor b multiplying all the different transcription rates. We conclude that the promoter contribution to the noise takes the exact same analytical form both for proteins and for mRNA, with the only other quantitative difference being the different rates of degradation for proteins and mRNA. Therefore, promoter architecture has the same qualitative effect on cell-to-cell variability in mRNA and protein numbers. All the conclusions about the effect of promoter architecture on cell-to-cell variability in mRNA expression are also valid for proteins, even though quantitative differences do generally exist. For the sake of simplicity we focus on mRNA noise for the remainder of the paper.

5.2.2 Parameters and assumptions

In order to evaluate the equations in our model, we use parameters that are consistent with experimental measurements of rates and equilibrium constants *in vivo* and *in vitro*, which we summarize in table 5.1.

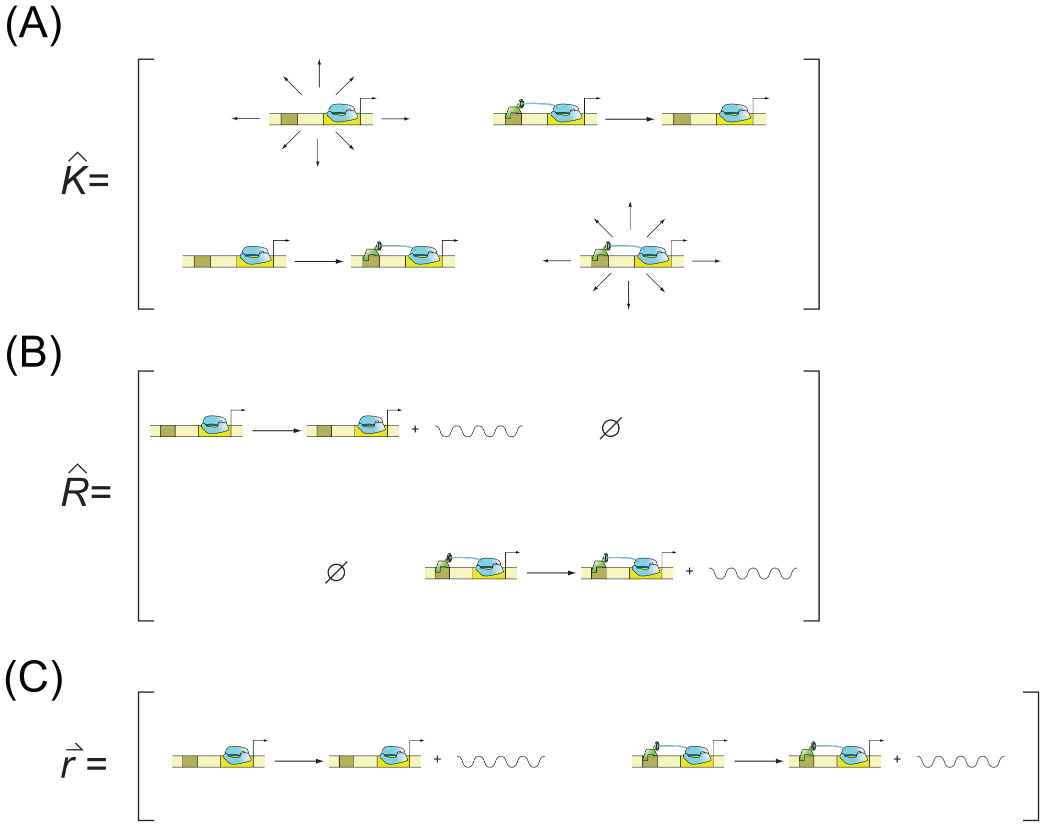


Figure 5.2: Cartoon depiction of the construction of kinetic rate matrices and vectors. (A) Cartoon representation of the kinetic rate matrix $\hat{K}$. The diagonal elements represent the net rate at which the promoter abandons each state. For instance, element $\{\hat{K}\}_{11}$ is the rate at which the promoter abandons state 1 due to stochastic association of the activator with the promoter: $\{\hat{K}\}_{11} = -k_A^{on}$, and element $\{\hat{K}\}_{22} = -k_A^{off}$ is the rate of dissociation of the activator from the promoter, abandoning state 2. The non-diagonal element $\{\hat{K}\}_{21} = k_A^{on}$ is the rate at which the promoter makes a transition from state 1 to state 2 (by dissociation association of one activator to the promoter), and the non-diagonal element $\{\hat{K}\}_{12} = -k_A^{off}$ is the rate at which the promoter makes a transition from state 2 to state 1 (by dissociation of the activator). (B) The transcription rate matrix contains, in its diagonal elements, the net rate of transcription at each promoter state. Element $\{\hat{R}\}_{11} = r_1$ is the rate of transcription in promoter state 1 and $\{\hat{R}\}_{22} = r_2$ is the rate of transcription in promoter state 2. (C) The vector $\vec{r} = (r_1, r_2)$ contains the rates of transcription at states 1 and 2, and is identical to the diagonal of matrix $\hat{R}$.

Kinetic rate	Symbol	Value	Reference
Unregulated promoter transcription rate	r	0.33 s^{-1}	[52]
Repressor and activator association rates	k_R^0, k_A^0	0.0027 (s nM)$^{-1}$	[3]
Repressor and activator dissociation rates	k_R^{off}, k_A^{off}	0.0023 s^{-1}	[42]
mRNA decay rate	γ	0.011 s^{-1}	[11]
Ratio between transcription rates due to activation	$f = r_1/r_2$	11	[50]
Cooperativity in repression	$\Omega_{repression}$	0.013	[50]
Cooperativity in activation	$\Omega_{activation}$	0.1	[35]
Looping J-factor	$[J]$	660 nM	[35]
Protein translation burst size	b	31.2 proteins/mRNA	[6]
Protein decay rate	$\gamma_{protein}$	0.00083 s^{-1}	[53]

Table 5.1: Kinetic parameters used to make the quantitative estimates in the text and plots in the figures. These parameters are all measured for model systems such as the $\mathrm{P_{lac}}$ or $\mathrm{P_{RM}}$ promoters in *E. coli*, and are here considered representative for promoter-transcription factor interactions.

Although these values correspond to specific examples of *E. coli* promoters, like the $\mathrm{P_{lac}}$ or the $\mathrm{P_{RM}}$ promoter, we extend their reach by using them as “typical” parameters characteristic of bacterial promoters, with the idea being that we are trying to demonstrate the classes of effects that can be expected, rather than dissecting in detail any particular promoter. The rate of association for transcription factors to operators *in vivo* is assumed to be the same as the recently measured value for the Lac repressor, which is close to the diffusion limited rate [48].

Operator strength reflects how tightly operators bind their transcription factors, and it is quantitatively characterized by the equilibrium dissociation constant K_{O-TF}. The dissociation constant has units of concentration and is equal to the concentration of free transcription factor at which the probability for the operator to be occupied is 1/2. K_{O-TF} is related to the association and dissociation rates by $K_{O-TF} = k_{off}/k_{on}^0$, where k_{off} is the rate (i.e., the probability per unit time) at which a transcription factor dissociates from the promoter, and k_{on}^0 is a second-order rate constant, which represents the association rate per unit of concentration of transcription factors (i.e., $k_{on} = k_{on}^0[N_{TF}]$). For simplicity, we assume that the binding reaction is diffusion limited, namely, k_{on}^0 is already close to its maximum possible value, so the only parameter that can differ from operator to operator is the dissociation rate: strong operators have slow dissociation rates, and weak operators have large dissociation rates.

Throughout this paper, we also make the assumption that the mean expression level is controlled by varying the intracellular concentration of transcription factors, a scenario that is very common experimentally [49–51]. We also assume that changing the intracellular concentration of transcription factors only affects the association rate of transcription factors to the operators, but the dissociation rate and the rates of transcription at each promoter state are not affected. In other words, k_{ofF} is a constant parameter for each operator, and it is not changed when we change the mean by titrating the intracellular repressor level. All of these general assumptions need to be revisited when studying a specific gene-regulatory system. Here our focus is on illustrating the general principles associated with different promoter architectures typical of those found in prokaryotes.

5.2.3 Simulations

To generate mRNA time traces, we applied the Gillespie algorithm [54] to the master equation described in the text. A single time step of the simulation is performed as follows: one of the set of possible trajectories is chosen according to its relative weight, and the state of the system is updated appropriately. At the same time, the time elapsed since the last step is chosen from an exponential distribution, whose rate parameter equals the sum of rate parameters of all possible trajectories. This process is repeated iteratively to generate trajectories that exactly reflect dynamics of the underlying master equation. For the figures, simulation lengths were set long enough for the system to reach steady-state and for a few promoter state transitions to occur.

To generate the probability distributions, it is convenient to reformulate the entire system of mRNA master equations in terms of a single matrix equation. To do this, we first define a vector

$$\vec{P} = \begin{pmatrix} p(1,0) \\ p(2,0) \\ \vdots \\ p(N,0) \\ p(1,1) \\ \vdots \\ p(N,1) \\ p(1,2) \\ \vdots \\ p(N,2) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vec{p}(0) \\ \vec{p}(1) \\ \vec{p}(2) \\ \vdots \end{pmatrix}, \tag{5.15}$$

where is the joint probability of having mRNAs while in the ith promoter state. Then the master equation for time evolution of this probability vector is

$$\frac{d\vec{P}}{dt} = \begin{pmatrix} \hat{K} - \hat{R} & \gamma\hat{I} & 0 & \dots \\ \hat{R} & \hat{K} - (\hat{R} + \gamma\hat{I}) & 2\gamma\hat{I} & \dots \\ 0 & \hat{R} & \hat{K} - (\hat{R} + 2\gamma\hat{I}) & \dots \\ 0 & 0 & \hat{R} & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix} \begin{pmatrix} \vec{p}(0) \\ \vec{p}(1) \\ \vec{p}(2) \\ \vdots \end{pmatrix}, \tag{5.16}$$

where each element of the matrix is itself an N by N matrix as described in the text. Then finding the steady-state distribution $\vec{P}_{ss}$ is equivalent to finding the eigenvector of the above matrix associated with eigenvalue 0. To perform this calculation numerically, one must first choose an upper bound on mRNA copy number in order to work with finite matrices. In this work, we chose an upper bound six standard deviations above mean mRNA copy number as an initial guess, and then modified this bound if necessary. Computations were performed using the SciPy (Scientific Python) software package.

5.3 Results

5.3.1 Promoters with a single repressor binding site

We first investigate a promoter architecture consisting of a single repressor binding site, and examine how operator strength affects intrinsic variability in gene expression. Although this particular mode of gene regulation has been well studied theoretically before [2, 17, 37, 38, 45], it is a useful starting point for illustrating the utility of this class of models. Within this class of models, when the repressor is bound to the operator, it interferes with transcription initiation and transcription does not occur. When the repressor dissociates and the operator is free, RNAP can bind and initiate transcription at a constant rate r. The probability per unit time that a bound repressor dissociates is k_R^{off}, and the probability per unit time that a free repressor binds the empty operator is $k_R^{on} = k_{on}^0[N_R]$, where k_{on}^0 is the second-order association constant and $[N_R]$ is the intracellular repressor concentration. The rate of mRNA degradation per molecule is γ. This mechanism is illustrated in figure 5.3(A).

We compute the mean and the Fano factor for this architecture following the algorithm described in the Mathematical Methods section. The kinetic rate and transcription rate matrices $\hat{K}$ and $\hat{R}$ are shown in table 5.2. For this simple architecture, the mean of the mRNA probability distribution and the normalized variance take simple analytical forms:

$$\langle m \rangle = \frac{r}{\gamma} \frac{k_R^{off}}{k_R^{off} + k_R^{on}} = \frac{r}{\gamma} \frac{1}{1 + k_R^{on}/k_R^{off}}, \tag{5.17}$$

$$\eta^2 = \frac{1}{\langle m \rangle} + \frac{k_R^{on}}{k_R^{off}} \frac{\gamma}{\gamma + k_R^{off} + k_R^{ob}}. \tag{5.18}$$

Using the relationship between k_R^{on} and the intracellular concentration of repressor, we can write the mean as:

$$\langle m \rangle = \frac{r}{\gamma} \frac{1}{1 + k_{on}^0[N_R]/k_R^{off}} = \langle m \rangle_{max} \frac{1}{1 + [N_R]/K_{OR}}. \tag{5.19}$$

Here we have defined the equilibrium dissociation constant between the repressor and the operator as: $K_{OR} = k_R^{off}/k_{on}^0$. It is interesting to note that equation 5.19 could have been derived using the thermodynamic model approach [34, 35, 42, 55]. In particular we see that this expression is equal to the product of the maximal activity in the absence of repressor $\langle m \rangle_{max} = r/\gamma$, and the so-called fold-change in gene expression: $\left(1 + k_R^{on}/k_R^{off}\right)^{-1} = (1 + [N_R]/K_{OR})^{-1}$ [35]. The fold-change is defined as the ratio of the level of expression in the presence of the transcription factor of interest, and the level of expression in the absence of the transcription factor.

The Fano factor for the mRNA distribution can be computed from equation 5.12 and we obtain:

$$Fano = 1 + \frac{k_R^{on}}{k_R^{off} + k_R^{on}} \frac{r}{\gamma + k_R^{off} + k_R^{on}}, \tag{5.20}$$

which is also shown as the first entry of table 5.3. In many experiments [5, 16, 32, 50], the concentration of repressor inside the cell $[N_R]$ (and therefore the association rate $k_R^{on} = k_{on}^0[N_R]$) can be varied by either expressing the repressor from an inducible promoter, or by adding an inducer that binds directly to the repressor rendering it incapable of binding specifically to the operators in the promoter region. When such an operation is performed, the only parameter that is varied is typically k_R^{on}, and all other kinetic rates are constant. The Fano factor can thus be re-written as a function of the mean mRNA, and we find:

$$Fano = 1 + \langle m \rangle \frac{1 - \langle m \rangle / \langle m \rangle_{max}}{k_R^{off}/\gamma + \langle m \rangle / \langle m \rangle_{max}}. \tag{5.21}$$

Therefore, for any given value of the mean, the Fano factor depends only on two parameters: the maximal mRNA or protein expression per cell, and a parameter that reflects the strength of binding between the repressor and the operator: k_R^{off}. Equations 5.19 and 5.20 reveal that changes in the mean due to repressor titration affect the noise as well as the mean. Since neither the repressor dissociation rate k_R^{off} nor the mRNA degradation rates are affected by the concentration of repressors, k_R^{off}/γ is a constant parameter that will determine how large the cell-to-cell variability is: The Fano factor is maximal for promoters with very strong operators, ($k_R^{off} \ll \gamma$), and it goes to 1 (i.e., the distribution tends to a Poisson distribution) when the operator is very weak and the rate of dissociation extremely fast ($k_R^{off} \gg \gamma$). In the latter limit of fast promoter kinetics, the fast fluctuations in promoter occupancy are filtered by the long lifetime of mRNA. Effectively, mRNA degradation acts as a low-pass frequency filter [56, 57], and fast fluctuations in promoter occupancy are not propagated into mRNA fluctuations. Therefore, promoters with strong operators are expected to be noisier than promoters with weak operators [58]. From this discussion it should also be clear that the mRNA degradation rate critically affects cell-to-cell variability. Any processes that tend to accelerate degradation will tend to increase noise, and mRNA stabilization (i.e., protection of the transcript by RNA binding proteins) leads to reduction of variability. However, the focus of this article is on promoter architecture and transcriptional regulation. Therefore, we do not consider regulation of transcription by mRNA degradation, and assume that all the promoters transcribe the same mRNA as is often the case in experimental studies.

The effect of operator strength on the mRNA distribution is illustrated in figures 5.3(B) and 5.3(C), where we show the normalized variance and the Fano factor, as a function of the fold-change in the mean mRNA concentration for a single strong operator whose dissociation rate is $k_R^{off} = 0.0027$ s^{-1} (a value that is representative of well characterized repressor-operator interactions such as the Lac repressor-O1, or the cI$_2$-O$_{\text{R1}}$), and for a single operator whose dissociation rate k_R^{off} is 10 times faster. The Poisson noise is shown for reference. The level of variability is always smaller for the weak operator than for the strong operator, due to faster promoter switching leading to smaller mRNA fluctuations and a more Poisson-like mRNA distribution (figure 5.3(E)), in which most cells are close to the mean. Slow dissociation, on the other hand, causes slower promoter fluctuations and highly non-Poissonian mRNA distributions, with few cells near the mean expression level (see figure 5.3(E), strong promoter). In figure 5.3(D) we plot the fold-change in protein noise due to gene regulation for the simple repression architecture. As expected, we find that the effect of operator strength in protein noise is qualitatively identical to what we found for mRNA. Since the same can be said of all the rest of architectures studied, we will limit the discussion to mRNA noise for the rest of the paper, with the understanding that for the class of models considered here, all the conclusions about the effect of promoter architecture in cell-to-cell variability that are valid for mRNA, are true for intrinsic protein noise as well.

An example of the single repressor-binding site architecture is a simplified version of the $\mathrm{P_{lacUV5}}$ promoter, which consists of a single operator overlapping with the promoter. Based on a simple kinetic model of repression, in which the Lac repressor competes with RNAP for binding at the promoter, we can write down the $\hat{K}$ and $\hat{R}$ matrices and compute the cell-to-cell variability in mRNA copy number. The matrices are presented in table 5.2. Based on our previous analysis, we know that stronger operators are expected to cause larger noise and higher values of the Fano factor than weaker operators. Therefore, we expect that if we replace the wild-type O1 operator by the 10 times weaker O2 operator, or by the ~ 500 times weaker operator O3, the fold-change in noise should go down. Using our best estimates and available measurements for the kinetic parameters involved, we find that noise is indeed much larger for O1 than for O2, and it is negligible for O3. This prediction is presented as an inset in figure 5.3(C).

5.3.2 Promoters with two repressor-binding operators

Dual repression occurs when promoters contain two or more repressor binding sites. Here, we consider three different scenarios for architectures with two operators: 1) repressors bind independently to the two operators, 2) repressors bind cooperatively to the two operators and 3) one single repressor may be bound to the two operators simultaneously thereby looping the intervening DNA. At the molecular level, cooperative repression is achieved by two weak operators that form long-lived repressor-bound complexes when both operators are simultaneously occupied. Transcription factors may stabilize each other either through direct protein-protein interactions [55], or through indirect mechanisms mediated by alteration of DNA conformation [59].

5.3.2.1 Cooperative and independent repression

The kinetic mechanisms of gene repression for both the cooperative and independent repressor architectures are reproduced in figure 5.4(A). For simplicity, we assume that both sites are of equal strength, so the rates of association and dissociation to both sites are equal. Cooperative binding is reflected in the fact that the rate of dissociation from the state where the two operators are occupied is slower (by a factor $\Omega \ll 1$) than the dissociation from a single operator. This parameter is related to the cooperativity factor ω often found

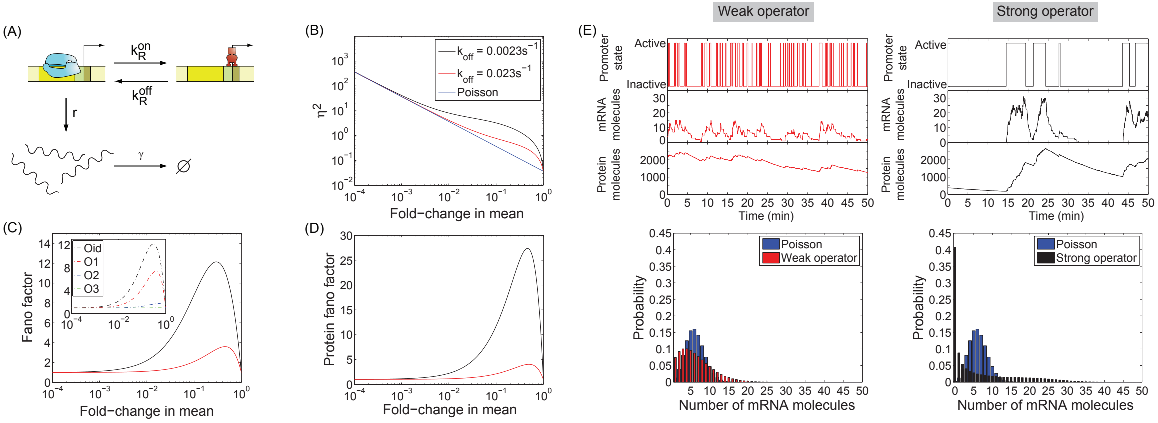


Figure 5.3: Simple repression architecture. (A) Kinetic mechanism of repression for an architecture involving a single repressor binding site. The repressor turns off the gene when it binds to the promoter (with rate k_R^{on}), and transcription occurs at a constant rate r when the repressor falls off (with rate k_R^{off}). (B) Normalized variance as a function of the fold-change in mean mRNA copy number. The parameters used are drawn from table 5.1. The value of $k_R^{off} = 0.0023$ s^{-1} from table 5.1 corresponds to the *in vitro* dissociation constant of the Lac repressor from the Oid operator (black). The results for an off-rate 10-times higher are also plotted (red). As a reference for the size of the fluctuations, we show the normalized variance for a Poisson promoter. (C) Fano factor for two promoters bearing the same off-rates as in (B). Inset. Prediction for the Fano factor for the $\Delta_{O3}\Delta_{O2}P_{lacUV5}$ promoter, a variant of the P_{lacUV5} promoter for which the two auxiliary operators have been deleted. The fold-change in mRNA noise is plotted as a function of the fold-change in mean mRNA copy number for mutants of the promoter that replace O1 for Oid, O2 or O3. The parameters are taken from table 5.1 and [35]. Lifetimes of the operator-repressor complex are 7 min for Oid, 2.4 min for O1, 11 s for O2 and 0.47 s for O3. (D) Fold-change in protein noise as a function of the fold-change in mean expression. As expected, the effect of operator strength is the same as observed for mRNA noise. (E) Time traces for promoter activity, mRNA and protein copy number are shown for both the weak operator and the strong operator. The mRNA histograms are also shown. The weaker operator with a faster repressor dissociation rate leads to small promoter noise, and an mRNA probability distribution resembling a Poisson distribution (shown by the blue-bar histogram), in which most cells express mRNA near the population average. In contrast, the stronger operator with a slower repressor dissociation rate, leads to larger promoter noise and strongly non-Poissonian mRNA statistics.

in thermodynamic models [56] by $\Omega = 1/\omega$. A typical value of Ω for cooperative binding is on the order of $10^{-3} - 10^{-2}$ [50, 55]. By way of contrast, independent binding is characterized by a value of $\Omega = 1$, which reflects the fact that the rate of dissociation from each operator is not affected by the presence of the other operator.

The $\hat{K}$ and $\hat{R}$ matrices for these two architectures are defined in table 5.2. Using these matrices, we can compute the mean gene expression and the Fano factor for these two architectures as a function of the concentrations of repressor. The resulting expression for the fold-change in noise is shown as entry number 3 of table 5.3. As shown in figure 5.4(B), the noise for cooperative repression is substantially larger than for the independent repression architecture. The high levels of intrinsic noise associated with cooperative repression can be understood intuitively in terms of the kinetics of repressor-operator interactions. At low repressor concentration, the lifetime of the states where only one repressor is bound to either one of the two operators can be shorter than the time it takes for a second repressor to bind. This makes simultaneous binding of two repressors to the two operators a rare event. However, when it occurs, the two repressors stabilize each other, forming a very long-lived complex with the operator DNA. This mode of repression, with rare but long-lived repression events, is intrinsically very noisy, since the promoter switches slowly between active (unrepressed) and inactive (repressed) states, generating wide bimodal distributions of mRNA (see figure 5.4(C)). On the other hand, independent binding to two operators causes more frequent transitions between repressed and unrepressed states, leading to lower levels of intrinsic noise and long-tailed mRNA distributions (see figure 5.4(C)).

As an example of the two repressor-binding sites architecture, we consider a simplified version of the lytic phage-λ $\mathrm{P_R}$ promoter, which is controlled by the lysogenic repressor cI. The wild-type PR promoter consists of three proximal repressor binding sites, $\mathrm{O_{R1}}$, $\mathrm{O_{R2}}$ and $\mathrm{O_{R3}}$, with different affinities for the repressor ($\mathrm{O_{R2}}$ is ~ 25 times weaker than $\mathrm{O_{R1}}$) [60], and three distal operators $\mathrm{O_{L1}}$, $\mathrm{O_{L2}}$ and $\mathrm{O_{R3}}$. For simplicity, we consider a simpler version of $\mathrm{P_R}$, harboring a deletion of the three distal operators. In the absence of these operators, the $\mathrm{O_{R3}}$ operator plays only a very minor role in the repression of this promoter, and it can be ignored [50, 61]. We are then left with only $\mathrm{O_{R1}}$ and $\mathrm{O_{R2}}$. The cI repressor binds cooperatively to $\mathrm{O_{R1}}$ and $\mathrm{O_{R2}}$, and that cooperativity is mediated by direct protein-protein interactions between cI bound at each operator [61]. Mutant forms of cI that are cooperativity deficient (i.e., not able to bind cooperatively to the promoter) have been designed [62]. In the inset in figure 5.4(B), we compare the normalized variance of the mRNA distribution, both for wild-type cI repressor, and for a cooperativity deficient mutant such as Y210H [62]. The cooperative repressor is predicted to have significantly larger promoter noise than the cooperativity deficient mutant.

5.3.2.2 Simultaneous binding of one repressor to two operators: DNA looping

Repression may also be enhanced by the presence of distant operators, which stabilize the repressed state by allowing certain repressors to simultaneously bind to both distant and proximal operators, forming a DNA

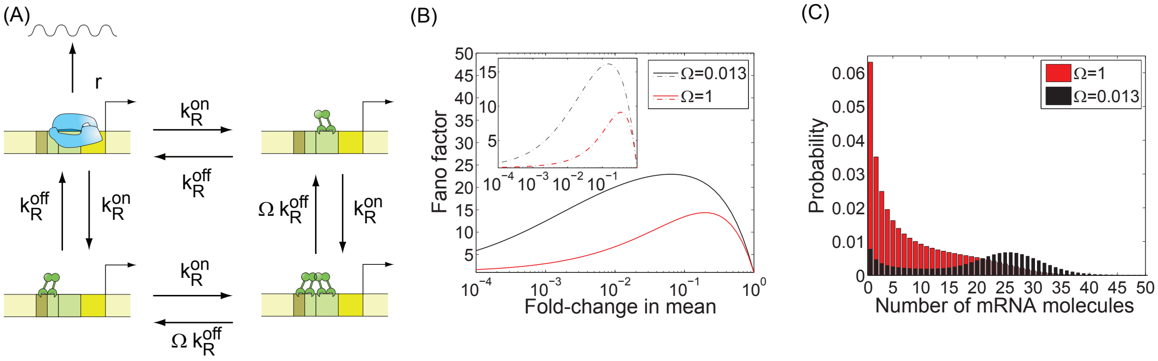


Figure 5.4: Dual repression architecture. (A) Kinetic mechanism of repression for a dual-repression architecture. The parameters k_R^{off} and k_R^{on} are the rates of repressor dissociation and association to the operators, and Ω is a parameter reflecting the effect of cooperative binding on the dissociation rate. For independent binding, $\Omega = 1$ and for cooperative binding $\Omega = 0.013$ (see table 5.1). (B) Fold-change in the mRNA noise caused by gene regulation for independent (red) and cooperative (black) repression as a function of the mean mRNA copy number. Inset: Prediction for a variant of the λ $\mathrm{P_R}$ promoter where the upstream operators $\mathrm{O_{L1}}$, $\mathrm{O_{L2}}$ and $\mathrm{O_{L3}}$ are deleted. The promoter mRNA noise is plotted as a function of the mean mRNA number for both wild-type cI repressor (blue line) and a repressor mutant (Y210H) that abolishes cooperativity (red line). Parameters taken from [43, 63]. The lifetime of the $\mathrm{O_{R1}}$-cI complex is 4 min. Lifetime of $\mathrm{O_{R2}}$-cI complex is 9.5 s. (C) mRNA distribution for the same parameters used in (B).

loop [64, 65]. The $\mathrm{P_{lac}}$ promoter is a prominent example of this architecture. The kinetic mechanism of repression characterizing this promoter architecture is presented in figure 5.5(A). The repressor only prevents transcription when it is bound to the main operator Om, but not when it is only bound to the auxiliary operator Oa. DNA loop formation is characterized by a kinetic rate $k_{loop} = k_{on}^0[J]$ where $[J]$, the looping J-factor, can be thought of as the local concentration of repressor in the vicinity of one operator when the repressor is bound to the other operator [34, 35]. The rate of dissociation of the operator-repressor complex in the looped conformation is given by $k_{unloop} = ck_R^{off}$. The parameters $[J]$ and c have both been measured *in vitro* for the particular case of the Lac repressor [42, 66], and also estimated from *in vivo* data [34, 67]. The $\hat{K}$ and $\hat{R}$ matrices for this architecture are defined in table 5.2. We use these matrices to compute the mean and the noise strength, according to equations 5.5–5.12 resulting in the fifth entry of table 5.3.

We first examine how the presence of the auxiliary operator affects the level of cell-to-cell variability in mRNA expression. In figure 5.5(B) we compare the Fano factor in the absence of the auxiliary operator with the Fano factor in the presence of the auxiliary operator, which is assumed to be of the same strength as the main operator. We use parameters in table 5.1, and we first assume that the dissociation rate of the operator-repressor complex in the looped state is the same as the dissociation rate in the unlooped state, so $c = 1$ and $k_{unloop} = k_R^{off}$. This assumption is supported by single-molecule experiments in which the two operators are on the same side of the DNA double-helix, separated by multiples of the helical period of DNA [42, 66]. Under these conditions we find that the presence of an auxiliary operator results in a larger Fano factor, in spite of the fact that the auxiliary operator Oa does not stabilize the binding of the repressor to the main operator Om. Interestingly, we find that the Fano factor is maximal at intermediate concentrations of repressor for which only one repressor is bound to the promoter, making the simultaneous occupancy of the auxiliary and main operators mediated by DNA looping possible. In contrast, the Fano factor is identical to that of the simple repression case if the concentration of repressor is so large that it saturates both operators and looping never occurs. It had been previously hypothesized that DNA looping might be a means to reduce noise in gene expression, due to rapid re-association kinetics between Om and a repressor that is still bound to Oa, which may cause short and frequent bursts of transcription [67, 68]. Here, by applying a simple stochastic model of gene regulation, we show that the presence of the auxiliary operator does not, by itself, decrease cell-to-cell variability. On the contrary, it is expected to increase it. The reason for this increase is that the rate of dissociation from the main operator is not made faster by DNA looping; instead the presence of the auxiliary operator causes the repressor to rapidly rebind the main operator, extending the effective period of time when the promoter is repressed.

Indeed, we find that only if the dissociation rate for a repressor in the looped state is faster than in the unlooped state, the presence of the auxiliary operator might reduce the cell-to-cell variability. To illustrate this limit, we have assumed a value of $c = 100$, so that $k_{unloop} = 100k_R^{off}$, and find that the Fano factor goes down, below the expectation for the simple repression architecture. A modest increase in the dissociation rate in the looped conformation has been reported in recent single-molecule experiments for promoter architectures in which the two operators are out of phase (located on different faces of the DNA) [42].

An example of this type of architecture is a simplified variant of the $\mathrm{P_{lacUV5}}$ promoter, which consists of one main operator and one auxiliary operator upstream from the promoter. The kinetic mechanism of repression is believed to be identical to the one depicted in figure 5.5(A) [24, 42, 66, 67]. We can use the stochastic model of gene regulation described in the theory section to make precise predictions that will test this kinetic model of gene regulation by DNA looping. We find that the kinetic model predicts that, if we move the center of the auxiliary operator further upstream from its wild-type location, in increments of distance given by the helical period of the DNA, such that both operators stay in phase, the fold-change in noise should behave as represented in figure 5.5(C). In order to model the effect of DNA looping, we assume that the dependence of the rate of DNA looping on the inter-operator distance D (in units of base-pairs) is given by [34], $k_{loop} = k_R^{on} \times \exp\left[-\frac{u}{D} - \nu\ln(D) + wD + z\right]$, where $u = 140.6$, $v = 2.52$, $w = 0.0014$, $z = 19.9$, [34], and we assume the same concentration of repressors (and therefore the same value for k_R^{on}) for all of the different loop lengths. Note that in figure 5.5(C), the Fano factor is not plotted as a function of the mean, but as a function of the inter-operator distance D. That is, we keep the number of repressors constant, and instead we alter the distance between the two operators. This results in mRNA distributions that differ both in the mean and the variance.

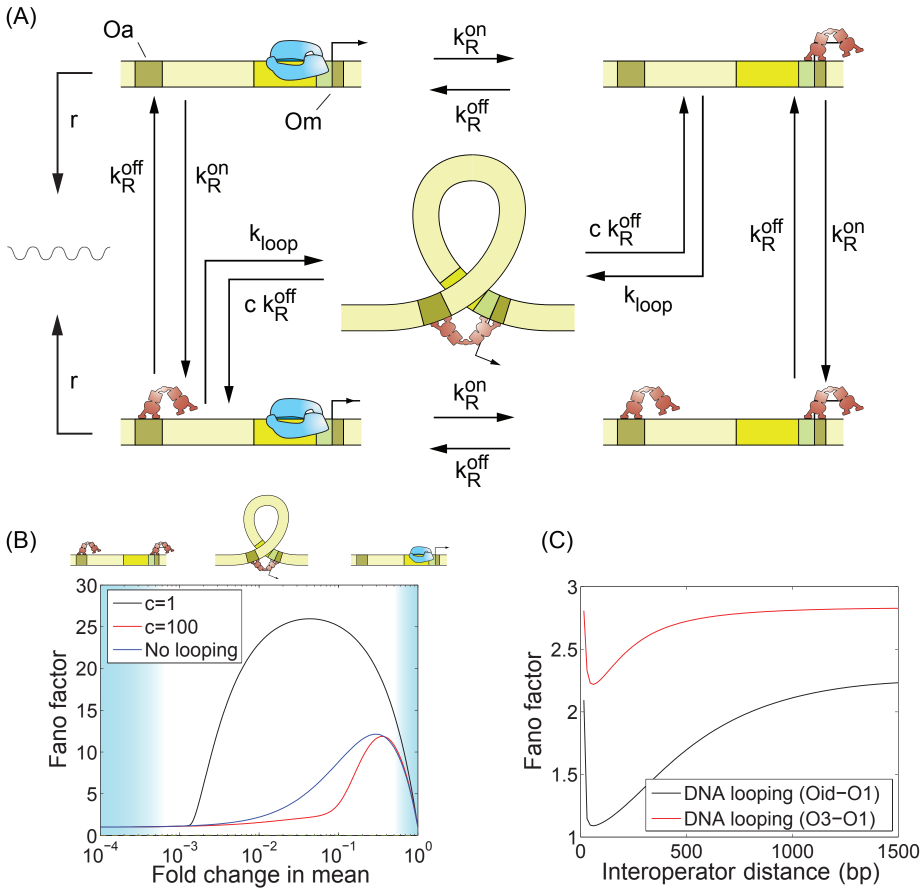


Figure 5.5: Repression by DNA looping. (A) Kinetic mechanism of repression. k_R^{off} and k_R^{on} are the rates of repressor dissociation and association. The rate of loop formation is $k_{loop} = [J]k_R^0$, where $[J]$ can be thought of as the local concentration of repressor in the vicinity of one operator when it is bound to the other operator. The rate of dissociation of the operator-repressor complex in the looped conformation is given by $k_{unloop} = ck_R^{off}$. The parameter c captures the rate of repressor dissociation in the looped state relative to the rate of dissociation in a non-looped state. (B) Effect of DNA looping on cell-to-cell variability. The Fano factor is plotted as a function of the fold-change in the mean expression level, in the absence (blue) and presence (black) of the auxiliary operator, and assuming that dissociation of the operator from Om is the same in the looped and the unlooped state ($c = 1$). The presence of the auxiliary operator, which enables repression by DNA looping, increases the cell-to-cell variability. The regions over which the state with two repressors bound, the state with one repressor bound or the looped DNA state are dominant are indicated by the shading in the background. The noise is larger at intermediate repression levels, where only one repressor is found bound to the promoter region, simultaneously occupying the auxiliary and main operators through DNA looping. The rate of DNA loop formation is $k_{loop} = 660$ nMk_R^0 [35]. We also show the effect of DNA looping in the case where the kinetics of dissociation from the looped state are 100 times faster than the kinetics of dissociation from the unlooped state: $c = k_{unloop}/k_r^{off}$ (red). In this limit, the presence of the auxiliary operator leads to less gene expression noise. (C) Prediction for a library of P$_{\text{lacUV5}}$ promoter variants, harboring an O2 deletion, and with the position of O3 moved upstream by multiples of 11 bp while keeping its identity (red), or replaced by the operator by Oid (black). Parameters are taken from the analysis in [35] of the data in [69]. We assume a concentration of 50 Lac repressor tetramers per cell. The association rate of the tetrameric repressor to the operators is taken from table 5.1. The lifetimes of the operator-repressor complex are given in the caption to figure 5.3. The dependence of the rate of DNA looping on the inter-operator distance is taken from [35], and equal to: $k_{loop} = k_R^{on} \times \exp\left[-\frac{u}{D} - v\ln(D) + wD + z\right]$, where $u = 140.6$, $v = 2.52$, $w = 0.0014$, $z = 19.9$. Note that the Fano factor is not plotted as a function of the mean, but as a function of the inter-operator distance D. In this case, as we change D, we vary both the mean and the Fano factor.

5.3.3 Simple activation

Transcriptional activators bind to specific sites at the promoter from which they increase the rate of transcription initiation by either direct contact with one or more RNAP subunits or indirectly by modifying the conformation of DNA around the promoter [59]. The simplest example of an activating promoter architecture consists of a single binding site for an activator in the vicinity of the RNAP binding site. When the activator is not bound, transcription occurs at a low basal rate. When the activator is bound, transcription occurs at a higher, activated rate. Stochastic association and dissociation of the activator causes fluctuations in transcription rate which in turn cause fluctuations in mRNA copy number.

This simple activation architecture is illustrated in figure 5.1(A). The $\hat{K}$ and $\hat{R}$ matrices for this architecture are given in table 5.2. Solving equations 5.5–5.8 for this particular case, we find that the mean mRNA per cell for this simple mechanism takes the form:

$$\langle m \rangle = \frac{r_2}{\gamma} \frac{k_A^{on}}{k_A^{on} + k_A^{off}} + \frac{r_1}{\gamma} \frac{k_A^{off}}{k_A^{on} + k_A^{off}}. \tag{5.22}$$

The mean mRNA can be changed by adjusting the intracellular concentration of the activator. The rate at which one of the activators binds to the promoter is proportional to the activator concentration: $k_A^{on} = k_{on}^0 [N_A]$. Following the same argument as we used in the simple repression case, the equilibrium dissociation constant for the activator-promoter interaction is given by $K_{OA} = k_A^{off}/k_{on}^0$. Finally, it is convenient to define the enhancement factor: the ratio between the rate of transcription in the active and the basal states $f = r_2/r_1$. The mean mRNA can be written in terms of these parameters as:

$$\langle m \rangle = \frac{r_1}{\gamma} \left(\frac{K_{OA}}{[N_A] + K_{OA}} + f \frac{[N_A]}{[N_A] + K_{OA}} \right). \tag{5.23}$$

The Fano factor can be computed using equations 5.5–5.12 and it is shown as entry 2 of table 5.3. We can rewrite the equation appearing in table 5.3 by writing k_A^{on} as a function of the mean:

$$Fano = 1 + \langle m \rangle \left(\frac{f - \langle m \rangle / \langle m \rangle_{basal}}{\langle m \rangle / \langle m \rangle_{basal}} \right)^2 \frac{\langle m \rangle / \langle m \rangle_{basal} - 1}{(f - \langle m \rangle / \langle m \rangle_{basal}) + \frac{k_A^{off}}{\gamma}(f-1)}. \tag{5.24}$$

With these equations in hand, we explore how operator strength affects noise in gene expression in the case of activation. Stronger operators bind to the activator more tightly than weak operators, leading to longer residence times of the promoter in the active state.

In figure 5.6(A) we plot the Fano factor as a function of the fold-change in mean expression for a strong operator as well as a 10 times weaker operator. We have used the parameters in table 5.1. Just as we saw for the simple repression architecture, it is also true for the simple activation architecture that stronger operators cause larger levels of noise for activators than weaker operators.

To get a sense of the differences between these two standard regulatory mechanisms, we compare simple repression with simple activation. In figure 5.6(B), we plot the Fano factor as a function of the mean for a repressor and an activator with identical dissociation rates. We assume that the promoter switches between a transcription rate $r = 0$ in its inactive state (which happens when the repressor is bound in the simple repression case, or the activator is not bound in the simple activation case), and a rate equal to $r = 0.33\ \mathrm{s}^{-1}$ (see table 5.1) in the active state (repressor not bound in the simple repression case, activator bound in the simple activation case). As shown in figure 5.6(B), at low expression levels the simple activation is considerably (> 20 times) noisier than the simple repression promoter. At high expression levels both architectures yield very similar noise levels, with the simple repression architecture being slightly noisier. A low level of gene expression may be achieved either by low concentrations of an activator, or by high concentrations of a repressor. Low concentrations of an activator will lead to rare activation events. High concentrations of a repressor will lead to frequent but short-lasting windows of time for which the promoter is available for transcription. As a result, and as we illustrate in figure 5.6(C), the activation mechanism leads to bursty mRNA expression whereas the repressor leads to Poissonian mRNA production. This result suggests that in order to maintain a homogeneously low expression level, a repressive strategy in which a high concentration of repressor ensures low expression levels may be more adequate than a low activation strategy.

An example of simple activation is the wild-type $\mathrm{P_{lac}}$ promoter, which is activated by CRP when complexed with cyclic AMP (cAMP). CRP is a ubiquitous transcription factor, and is involved in the regulation of dozens of promoters, which contain CRP binding sites of different strengths [70]. In the inset of figure 5.6(A) we include CRP as an example of simple activation, and make predictions for how changing the wild-type CRP binding site in the $\mathrm{P_{lac}}$ promoter by the CRP binding site of the $\mathrm{P_{gal}}$ promoter (which is ~ 8 times weaker [71]) should affect the Fano factor. As expected from our analysis of this class of promoters, the noise goes down.

5.3.4 Dual activation: independent and cooperative activation

Dual activation architectures have two operator binding sites. Simultaneous binding of two activators to the two operators may lead to a larger promoter activity in different ways. For instance, in some promoters each of the activators may independently contact the polymerase, recruiting it to the promoter. As a result, the probability to find RNAP bound at the promoter increases and so does the rate of transcription [34, 72]. In other instances, there is no increase in enhancement factor when the two activators are bound. However, the first activator recruits the second one through protein-protein or protein-DNA interactions, stabilizing the active state and increasing the fraction of time that the promoter is active [61]. These two modes are not mutually exclusive, and some promoters exhibit a combination of both mechanisms [73].

We first investigate the effect of dual activation in the limit where binding of the two transcription factors is not cooperative. Assuming that activators bound at the two operators independently recruit the polymerase, we compare this architecture with the simple activation architecture. The mechanism of

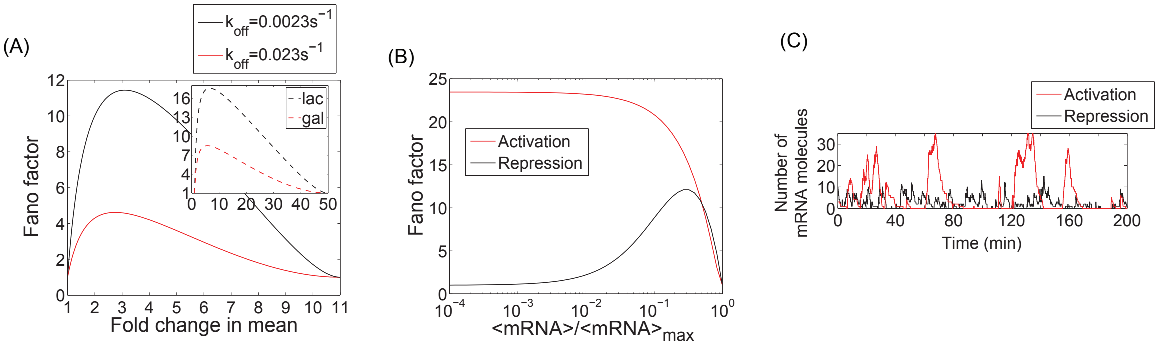


Figure 5.6: Simple activation architecture. (A) The Fano factor is plotted as a function of the fold-change gene expression (blue line). In red, we show the effect of reducing operator strength (i.e., reducing the lifetime of the operator-activator complex) by a factor of 10. Just as we observed with single repression, weak activator binding operators generate less promoter noise than strong activating operators. The parameters used are shown in table 5.1 with the exception of $r_1 = 0.33\ \mathrm{s}^{-1}/f$, where f is the enhancement factor. Inset: Prediction for the activation of the $\mathrm{P}_{\mathrm{lac}}$ promoter. The fold-change in noise is plotted as a function of the fold-change in mean mRNA expression for both the wild-type $\mathrm{P}_{\mathrm{lac}}$ (CRP dissociation time = 8 min), represented by a blue line, and a $\mathrm{P}_{\mathrm{lac}}$ promoter variant where the *lac* CRP binding site has been replaced by the weaker gal CRP binding site (dissociation time = 1 min). The enhancement factor was set to $f = 50$ [35]. These parameters are taken from [71] and [35]. The remaining parameters are taken from table 5.1. (B) Fano factor as a function of $\langle \mathrm{mRNA} \rangle / \langle \mathrm{mRNA} \rangle_{max}$ for a repressor (black) and an activator (red) with the same transcription factor affinity. The transcription rate in the absence of activator is assumed to be zero. The transcription rate in the fully activated case is equal to the transcription rate of the repression construct in the absence of repressor and is $r = 0.33\ \mathrm{s}^{-1}$ as specified by table 5.1. For low expression levels $\langle m \rangle / \langle m \rangle_{max} < 0.5$ simple activation is considerably noisier than simple repression. (C) The results of a stochastic simulation for the simple activation and simple repression architectures. We assume identical dissociation rates for the activator and repressor, and identical rates of transcription in their respective active states. As shown in (B), low concentrations of an activator result in few, but very productive transcription events, whereas high concentrations of a repressor lead to the frequent but short-lived excursions into the active state.

activation is depicted in figure 5.7(A), and matrices $\hat{K}$ and $\hat{R}$ are presented in table 5.2. For simplicity, we assume that both operators have the same strength, and both have the same enhancement factor $f = r_2/r_1 = r_3/r_1$ .When the two activators are bound, the total enhancement factor is given by the product of the individual enhancement factors, which in this case is $f \times f = r_4/r_1$ [34]. All of the other relevant kinetic parameters are given in table 5.1. The Fano factor is plotted in figure 5.7(B). We find that compared to the single operator architecture, the second operator increases the level of variability, even when binding to the operators is non-cooperative.

We then ask whether this is also true when the binding of activators is cooperative. We assume a small cooperativity factor $\Omega = 0.1$. Just as we found for repressors, cooperative binding of activators generates larger cell-to-cell variability than independent binding, which in turn generates larger cell-to-cell variability than simple activation. This is illustrated in the stochastic simulation in figure 5.7(C). As expected the dual activation architectures are noisier than the simple activation, characterized by rare but long-lived activation events that lead to large fluctuations in mRNA levels. In contrast, the simple activation architecture leads to more frequent but less intense activation events.

Together with the results from the dual repressor mechanism, these results indicate that multiplicity in operator number may introduce significant intrinsic noise in gene expression. Multiple repeats of operators commonly appear in eukaryotic promoters [2, 74, 75], but are often found in prokaryotic promoters as well [61, 72, 76]. It is interesting to note that this prediction of the model is in qualitative agreement with the findings by Raj et al. [3] who report an increase in cell-to-cell variability in mRNA when the number of activator binding sites was changed from one to seven.

An example of cooperative activation is the lysogenic phage-λ P_{RM} promoter [61]. This promoter contains three operators (O_{R1}, O_{R2} and O_{R3}) for the cI protein, which acts as an activator. When O_{R2} is occupied, cI activates transcription. O_{R1} has no direct effect on the transcription rate, but it helps recruit cI to O_{R2}, since cI binds cooperatively to the two operators. Finally, O_{R3} binds cI very weakly, but when it is occupied, P_{RM} becomes repressed. There are variants of this promoter [50] that harbor mutations in O_{R3} that make it unable to bind cI. In figure 5.7(D), we include one of these variants, r1-P_{RM} [51] as an example of dual activation, and we present a theoretical prediction for the promoter noise as a function of the mean mRNA. We examine the role of cooperativity by comparing the wild-type cI, with a cooperativity deficient mutant. We find that the cooperative activator causes substantially larger cell-to-cell variability than the mutant, emphasizing our expectation that cooperativity may cause substantial noise in gene expression in bacterial promoters such as P_{RM}.

5.4 Discussion

The DNA sequence of a promoter encodes the binding sites for transcriptional regulators. In turn, the collection of these regulatory sites, known as the architecture of the promoter, determines the mechanism of

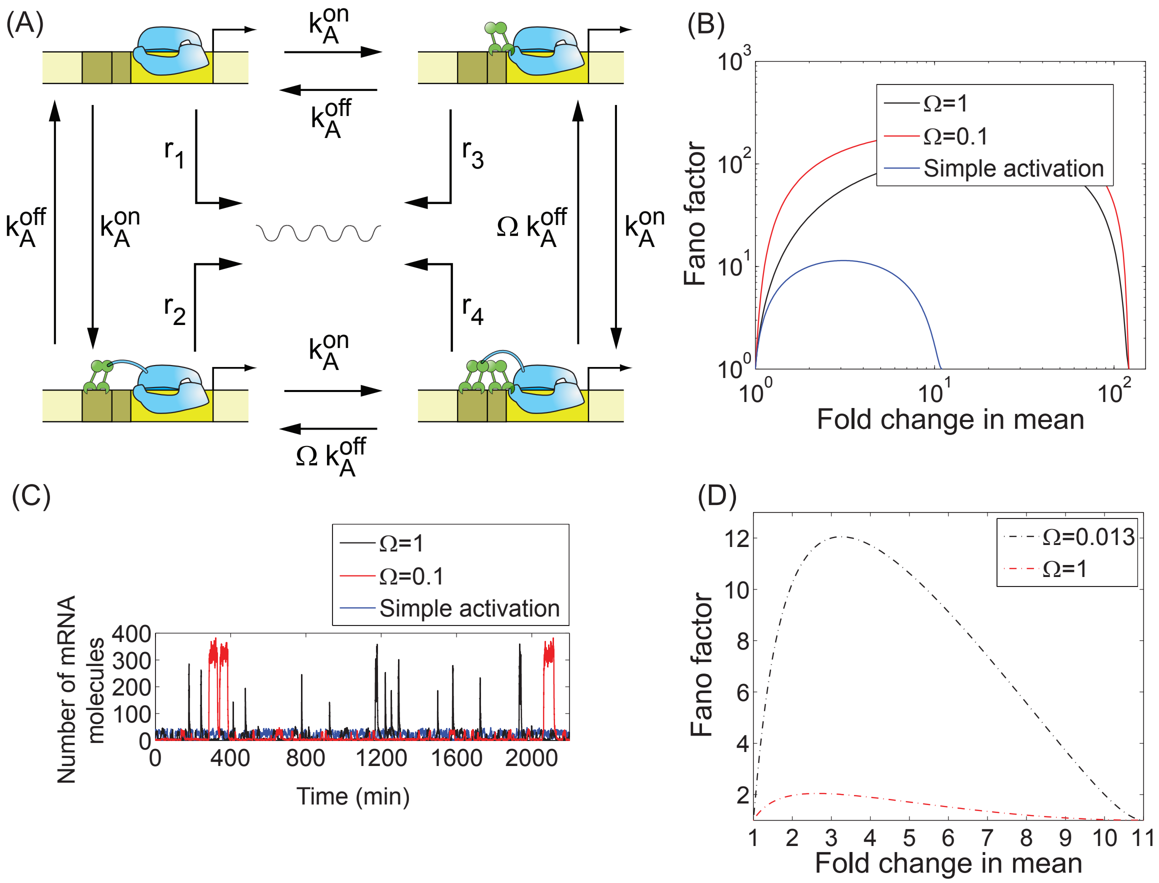


Figure 5.7: Dual activation architecture. (A) Kinetic mechanism of dual activation. The parameters k_A^{off} and k_A^{on} are the rates of activator dissociation and association to the operators, and Ω is a parameter reflecting the effect of cooperative binding on the dissociation rate. (B) Fano factor as a function of the mean mRNA for independent ($\Omega = 1$, black), cooperative ($\Omega = 0.1$, red), and for simple activation (blue). The parameters are taken from table 5.1 and $r_1 = 0.33\ \mathrm{s}^{-1}/f$, $r_2 = f \times r_1$, $r_3 = f \times r_1$, and $r_4 = f^2 \times r_1$; f is the enhancement factor. (C) A stochastic simulation shows the effect of independent and cooperative binding in creating a sustained state of high promoter activity, resulting in high levels of mRNA in the active state and large cell-to-cell variability. (D) Prediction for the r1-P$_{mboxRM}$ promoter (a P$_{mboxRM}$ promoter variant that does not exhibit O$_{mboxR3}$ mediated repression [51]). This promoter is activated by cI, which binds cooperatively to O$_{mboxR1}$ and O$_{mboxR2}$. The prediction is shown for wild-type cI ($\Omega = 0.013$) and for a cooperativity deficient mutant (Y210H, $\Omega = 1$). Parameters are taken from [35, 43, 60, 63]. The lifetime of O$_{mboxR1}$-cI complex is 4 min. Lifetime of O$_{mboxR2}$-cI complex is 9.5 s.

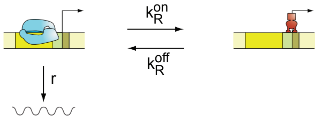
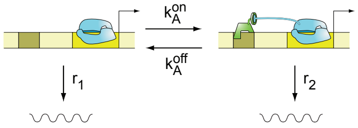
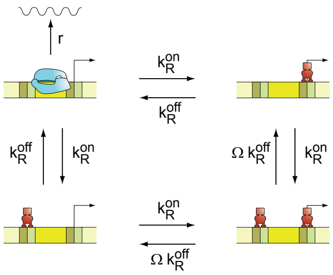
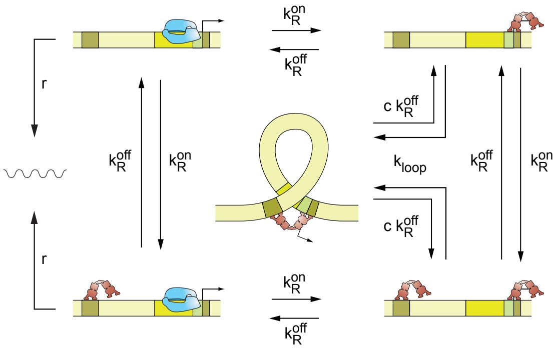
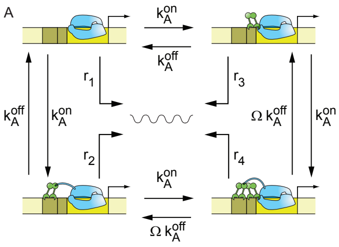
Mechanism	Matrices and vectors	
	$\hat{K}$	$\vec{u}\cdot\hat{R}$
	$\begin{pmatrix} -k_R^{on} & k_R^{off} \\ k_R^{on} & -k_R^{off} \end{pmatrix}$	$\begin{pmatrix} r \\ 0 \end{pmatrix}$
	$\begin{pmatrix} -k_A^{on} & k_A^{off} \\ k_A^{on} & -k_A^{off} \end{pmatrix}$	$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$
 r k_R^{on} k_R^{off} k_R^{off} k_R^{on} $\Omega\ k_R^{off}$ k_R^{on} k_R^{on} $\Omega\ k_R^{off}$	$\begin{pmatrix} -2k_R^{on} & k_R^{off} & k_R^{off} & 0 \\ k_R^{on} & -(k_R^{off}+k_R^{on}) & 0 & \Omega\, k_R^{off} \\ k_R^{on} & 0 & -(k_R^{off}+\ k_R^{on}) & \Omega\, k_R^{off} \\ 0 & k_R^{on} & k_R^{on} & -2\Omega\, k_R^{off} \end{pmatrix}$	$\begin{pmatrix} r \\ 0 \\ 0 \\ 0 \end{pmatrix}$
 k_R^{on} k_R^{off} r $c\ k_R^{off}$ k_R^{off} k_R^{on} k_{loop} k_R^{off} k_R^{on} $c\ k_R^{off}$ r k_R^{on} k_R^{off}	$\begin{pmatrix} -2k_R^{on} & k_R^{off} & 0 & 0 & 0 \\ k_R^{on} & -(k_R^{off}+k_R^{on}+k_l) & 0 & k_R^{off} & c\, k_R^{off} \\ k_R^{on} & 0 & -(k_R^{off}+k_R^{on}+k_l) & k_R^{off} & c\, k_R^{off} \\ 0 & k_R^{on} & k_R^{on} & -2k_R^{off} & 0 \\ 0 & k_l & k_l & 0 & -2c\, k_R^{off} \end{pmatrix}$	$\begin{pmatrix} r \\ r \\ 0 \\ 0 \\ 0 \end{pmatrix}$
 A k_A^{on} k_A^{off} r_1 r_3 k_A^{off} k_A^{on} $\Omega\ k_A^{off}$ k_A^{on} r_2 r_4 k_A^{on} $\Omega\ k_A^{off}$	$\begin{pmatrix} -2k_A^{on} & k_A^{off} & k_A^{off} & 0 \\ k_A^{on} & -(k_A^{off}+k_A^{on}) & 0 & \Omega\, k_A^{off} \\ k_A^{on} & 0 & -(k_A^{off}+k_A^{on}) & \Omega\, k_A^{off} \\ 0 & k_A^{on} & k_A^{on} & -2\Omega\, k_A^{off} \end{pmatrix}$	$\begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{pmatrix}$

Table 5.2: Kinetic rate matrices for all mechanisms in the text. In the first column, we represent the kinetic mechanisms of gene regulation for all of the architectures considered in the text. In the second and third columns, we show the corresponding promoter kinetic transition rate matrices $\hat{K}$ and the vector $\vec{r} = \vec{u}\hat{R}$ for all of the mechanisms.

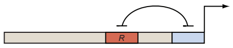
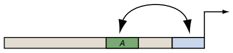
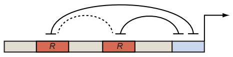
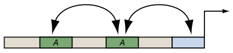
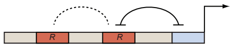
Promoter architecture	Fold-change in noise
1. Simple repression  	$1+\frac{r\,k_R^{on}}{\left(k_R^{\mathrm{off}}+k_R^{\mathrm{on}}\right)\left(\gamma\ +k_R^{\mathrm{off}}+k_R^{\mathrm{on}}\right)}$
2. Simple activation  	$1+\left(\frac{\left(\frac{r_2}{r_1}-1\right)^2 k_A^{off} k_A^{on} r_2}{\left(k_A^{off}+k_A^{on}\right)\left(\gamma\ +k_A^{off}+k_A^{on}\right)\left(k_A^{off}+\frac{r_2}{r_1}\ k_A^{on}\right)}\right)$
3. Dual repression  	$1+\frac{\left(r\,k_R^{on}\left(k_R^{2\,on}+2\,\Omega\,k_R^{off}\left(\gamma\ +2\,\Omega\,k_R^{off}\right)+k_R^{on}\left(\gamma\ +k_R^{off}+4\,\Omega\,k_R^{off}\right)\right)\right)}{\left(\left(2\,(k_R^{on})^2\ \ +\left(\gamma\ +k_R^{off}\right)\left(\gamma\ +2\,\Omega\,k_R^{off}\right)+k_R^{on}\left(3\,\gamma\ +4\,\Omega\,k_R^{off}\right)\right)\left((k_R^{on})^2+\Omega\,k_R^{off}\left(k_R^{off}+2\,k_R^{on}\right)\right)\right)}$
4. Cooperative activation  	$1+\frac{(r_2/r_1-1)^2 r_2\ \Omega\,k_A^{off} k_A^{on}}{\Omega\,k_A^{off}\left(k_A^{off}+k_A^{on}\right)+r_2/r_1\,k_A^{on}\left(\Omega\,k_A^{off}+k_A^{on}\right)}\left(\frac{1}{2\left(\gamma\ +k_A^{off}+k_A^{on}\right)}+\frac{2\,(k_A^{on})^3+(k_A^{on})^2\left(\gamma\ +6\,k_A^{off}\right)+\Omega\left(k_A^{off}\right)^2\left(\gamma\ +2\,\Omega\,k_A^{off}\right)+2\,k_A^{off} k_A^{on}\left(\gamma\ +k_A^{off}+2\,\Omega\,k_A^{off}\right)}{2\left(2\,(k_A^{on})^2+\left(\gamma\ +k_A^{off}\right)\left(\gamma\ +2\,\Omega\,k_A^{off}\right)+k_A^{on}\left(3\,\gamma\ +4\,\Omega\,k_A^{off}\right)\right)\left((k_A^{on})^2+\Omega\,k_A^{off}\left(k_A^{off}+2\,k_A^{on}\right)\right)}\right)$
5. Repression by DNA looping  	$1+r\,k_R^{on} k_l\frac{\left(k_R^{off}+k_R^{on}\right)\left(\gamma\ +k_R^{off}+k_R^{on}\right)\left(\gamma\ +2\left(k_R^{off}+k_R^{on}\right)\right)}{\left(\gamma\ +k_l+k_R^{off}+k_R^{on}\right)\left(k_l\,k_R^{on}+\left(k_R^{off}+k_R^{on}\right)^2\right)}$ $\frac{1+\left(k_l k_R^{off}+2\,(k_R^{on})^2+2\,k_R^{off}\left(\gamma\ +2\,k_R^{off}\right)+k_R^{on}\left(\gamma\ +5\,k_R^{off}\right)\right)\left(\left(k_R^{off}+k_R^{on}\right)+\left(\ \gamma^2+4\left(k_R^{off}+k_R^{on}\right)^2+\gamma\left(5\,k_R^{off}+4\,k_R^{on}\right)\right)\right)}{k_l\left(\gamma\ +2\,k_R^{on}\right)+\left(\gamma\ +k_R^{off}+k_R^{on}\right)\left(\gamma\ +2\left(k_R^{off}+k_R^{on}\right)\right)}$

Table 5.3: Fold-change in noise for different promoter architectures. The fold-change in promoter noise is shown as a function of the different kinetic parameters corresponding to each promoter architecture considered throughout the text. Refer to table 5.2 for the definition and value of each rate.

gene regulation. The mechanism of gene regulation determines the transcriptional response of a promoter to a specific input, in the form of the concentration of one or more transcription factors or inducer molecules. In recent years we have witnessed an increasing call for quantitative models of gene regulation that can serve as a conceptual framework for reflecting on the explosion of recent quantitative data, testing hypotheses, and proposing new rounds of experiments [35, 77, 78]. Much of this data has come from bulk transcription experiments with large numbers of cells, in which the average transcriptional response from a population of cells (typically in the form of the level of expression of a reporter protein) was measured as a function of the concentration of a transcription factor or inducer molecule [50, 79]. Thermodynamic models [35, 42, 55] of gene regulation are a general framework for modeling gene regulation and dealing with this kind of bulk transcriptional regulation experiments. This class of models has proven to be very successful at predicting gene expression patterns from the promoter architecture encoded in the DNA sequence [49, 77–81]. However, a new generation of experiments now provides information about gene expression at the level of single cells, with single-molecule resolution [3, 5–7, 10, 11, 24, 32, 47, 51]. These experiments provide much richer information than just how the mean expression changes as a function of an input signal: they tell us how that response is spread among the population of cells, distinguishing homogeneous responses, in which all cells express the same amount of proteins or mRNA for the same input, from heterogeneous responses in which some cells achieve very high expression levels while others maintain low expression. Thermodynamic models are unable to explain the single-cell statistics of gene expression, and therefore are an incomplete framework for modeling gene regulation at the single-cell level.

A class of stochastic kinetic models have been formulated that make it possible to calculate either the probability distribution of mRNA or proteins per cell or its moments, for simple models of gene regulation involving one active and one inactive promoter state [37, 38, 45, 82]. Recently, we have extended that formalism to account for any number of promoter states [31], allowing us to model any promoter architecture within the same mathematical framework. Armed with this model, we can now ask how promoter architecture affects not only the response function, but also how that response is distributed among different cells.

In this paper we have explored the feasibility of this stochastic analog of thermodynamic models as a general framework to understand gene regulation at the single-cell level. Using this approach we have examined a series of common promoter architectures of increasing complexity, and established how they affect the level of cell-to-cell variability of the number of mRNA molecules, and proteins, in steady-state. We have found that, given the known kinetic rates of transcription factor association and dissociation from operators, the level of variability in gene expression for many well-studied bacterial promoters is expected to be larger than the simple Poissonian expectation, particularly for mRNA and short-lived proteins. We have investigated how the level of variability generated by a simple promoter consisting of one single operator differs from more complex promoters containing more than one operator, and found that the presence of multiple operators increases the level of cell-to-cell variability even in the absence of cooperative binding. Cooperative binding makes the effect of operator multiplicity even larger. We also found that operator strength is one of the major determinants of cell-to-cell variability. Strong operators cause larger levels of cell-to-cell variability than weak operators. We have also examined the case where one single repressor may bind simultaneously to two operators by looping the DNA inbetween. We have found that the stability of the DNA loop is the key parameter in determining whether DNA looping increases or decreases the level of variability, suggesting a potential role of DNA mechanics in regulating cell-to-cell variability.

We have examined the difference between activators and repressors, and found that repressors tend to generate less cell-to-cell variability than activators at low expression levels, whereas at high expression levels repressors and activators generate similar levels of cell-to-cell variability. We conclude that induction of gene expression by increasing the concentration of an activator leads to a more heterogeneous response at low and moderate expression levels than induction of gene expression by degradation, sequestration or dilution of a repressor. In addition, we have used this model to make quantitative predictions for a few well-characterized bacterial promoters, connecting the kinetic mechanism of gene regulation that we believe applies for these promoters *in vivo* with single-cell gene expression data. Direct comparison between the model and experimental data offers an opportunity to validate these kinetic mechanisms of gene regulation.

5.4.1 Intrinsic and extrinsic noise

There are two different classes of sources of cell-to-cell variability in gene expression. The first class has its origins in the intrinsically stochastic nature of the chemical reactions leading to the production and degradation of mRNAs and proteins, including the binding and unbinding of transcription factors, transcription initiation, mRNA degradation, translation and protein degradation. The noise coming from these sources is known as intrinsic noise [83]. A different source of variability originates in cell-to-cell differences in cell size, metabolic state, copy number of transcription factors, RNA polymerases, ribosomes, nucleotides, etc. This second kind of noise is termed extrinsic noise [83]. The contributions from intrinsic and extrinsic sources can be separated experimentally, and the total noise can be written as the sum of intrinsic and extrinsic components [4]. In this paper we focus exclusively on intrinsic noise, and the emphasis is on bacterial promoters. This double focus requires us to discuss to what extent intrinsic noise is relevant in bacteria.

The experimental evidence gathered so far indicates that intrinsic noise is the dominant source of cell-to-cell variability in bacteria of the mRNA copy number. In a recent single-molecule study, transcription was monitored in real time for two different *E. coli* promoters, $\mathrm{P_{RM}}$ and $\mathrm{P_{lac/ara}}$ [5]. The authors measured the rates of mRNA synthesis and dilution, as well as the rates of promoter activation and inactivation in single cells. The intrinsic noise contribution was calculated from all of these rates. It was found to be responsible for the majority of the total cell-to-cell variability, accounting for over 75% of the total variance. Another recent experiment in *B. subtilis* [8] found that mRNA expressed from the ComK promoter is also dominated by intrinsic noise. Furthermore, this study indicated that intrinsic mRNA noise is responsible for activation of a phenotypic switch that drives a fraction of the cells to competence for the uptake of DNA [8]. A third recent report investigated the activation of the genetic switch in *E. coli* which drives the entrance of a fraction of cells into a lactose metabolizing phenotype [24]. The authors of the study found evidence that stochastic binding and unbinding of the Lac repressor to the main operator was responsible for the observed cell-to-cell variability in gene expression and, consequently the choice of phenotype. Furthermore, the authors discovered that the deletion of an auxiliary operator that permits transcriptional repression by DNA looping leads to a strong increase in the level of cell-to-cell variability in the expression of the lactose genes, indicating that promoter architecture plays a big role in determining the level of noise and variability in this system. Taken all together, these experiments suggest that intrinsic mRNA noise is dominant and may have important consequences for cell fate determination. In addition, at least in one case, promoter architecture has been shown to be of considerable importance.

At the protein level, the contribution of extrinsic and intrinsic noise to the total cell-to-cell variability has also been determined experimentally for a variety of promoters and different kinds of bacteria. The first reports examined intrinsic and extrinsic protein noise in *E. coli* and found that extrinsic noise was the dominant source of cell-to-cell variability in protein expressed from a variant of the $\mathrm{P_L}$ promoter in a variety of different strains [4]. However, the intrinsic component was non-negligible and for some strains, dominant [4]. A second team of researchers examined a different set of *E. coli* promoters involved in the biosynthetic pathway of lysine [84]. The authors found that the intrinsic noise contribution was significant for some promoters (i.e., lysA), but not for others. In a third study the total protein noise was measured for a Lac repressor-controlled promoter in *B. subtilis*, and it was reported that the data could be well explained by a model consisting only of intrinsic noise [9]. The authors found that the rates of transcription and translation could be determined by directly comparing the total cell-to-cell variability to the predictions of a simple stochastic model that considered only intrinsic sources of noise. They also found that the model had predictive power, and that mutations that enhanced the rate of translation or transcription produced expected effects in the total noise.

In summary, all studies that have measured mRNA noise in bacteria so far report that intrinsic noise contributes substantially to the total cell-to-cell variability. This is further supported by observations that most of the mRNA variability comes from intrinsic sources in yeast [32] and mammalian cells [2]. The issue is less clear for protein noise. Some reports indicate that it is mostly extrinsic [4], but others suggest that intrinsic noise may also be important [9, 24, 84]. It seems likely that the relative importance of intrinsic and extrinsic noise depends on the context, and that for some promoters and genes extrinsic noise will be larger, whereas for others the intrinsic component may dominate. In any case, it is clear that both contributions are important, and both need to be understood.

5.4.2 Comparison with experimental results

The aim of this paper is to formulate a set of predictions that reflect the class of kinetic models of gene regulation in bacteria that one routinely finds in the literature [42, 66, 67, 85–87]. Our analysis indicate that if these models are correct, and if the kinetic and thermodynamic parameters that have been measured over the years are also reasonably close to their real values in live cells [88], the effect of promoter architecture in cell-to-cell variability in bacteria should be rather large and easily observable. In this sense, our intention is more to motivate new experiments than to explain or fit any currently available data. We only know of one published report in which the effect of perturbing the architecture of a bacterial promoter on the cell-to-cell variability in gene expression has been determined [24]. Given that there are several examples of promoters in bacteria for which a molecular kinetic mechanism of gene regulation has been formulated [42, 66, 67, 85–87, 89], we hope that the computational analysis in this paper may serve as an encouragement for researchers to do for bacteria the same kind of experiments that have been already performed in eukaryotes [2, 12, 16, 18, 32]. Indeed, several different studies have examined the effect of promoter architectural elements in cell-to-cell variability in protein and mRNA in eukaryotic cells. Although our efforts in this paper have focused on bacterial promoters rather than eukaryotic promoters, it is worthwhile to discuss the findings of these studies and compare them (if only qualitatively) with the predictions made in this paper.

Two recent studies measured intrinsic mRNA noise in yeast [32] and mammalian cells [2]. Both papers concluded that stochastic promoter activation and inactivation was the leading source of intrinsic noise. While stochastic chromatin remodeling is suspected to be the origin of those activation events, neither one of these studies was conclusive about the precise molecular mechanism responsible for promoter activation. However, both studies found that promoter architecture had an important role and strongly affected the level of total mRNA noise. In both studies, the authors found that when the number of binding sites for a transcriptional activator was raised from one to seven, the normalized variance increased several-fold. This qualitative behavior is in agreement with our prediction that dual activation causes larger intrinsic mRNA noise than simple activation. It is possible that this agreement is coincidental, since the actual mechanism of gene regulation at these promoters could be much more complicated than the simple description of gene activation at a bacterial promoter adopted here.

Other studies [12, 16, 18] have measured the total protein noise from variants of the GAL1 promoter in yeast, and found that their data could be well explained by a model that considered only intrinsic noise sources. These studies also concluded that the main sources of intrinsic noise were stochastic activation and inactivation of the promoter due to chromatin remodeling. However, it was also found that the stable formation of pre-initiation complex at the TATA box and the stochastic binding and unbinding of transcriptional repressors contributed to the total noise [12, 16, 18]. The authors of these studies found that for point mutations in the TATA box of the GAL1 promoter in yeast, which made the box weaker, the level of cell-to-cell variability went down significantly. This is also in good agreement with our prediction that the stronger the binding site of a transcriptional activator, the larger the intrinsic noise should be. However, since this study measured the total noise strength, and did not isolate the intrinsic noise, the observed decrease in noise strength as a result of making the TATA box weaker may have other origins. These experiments were conducted under induction conditions that minimize repression by nucleosomes and activation by chromatin remodeling. A more recent report by the same lab [12] found that the copy number and location of a transcriptional repressor binding site greatly affects the total protein noise. The authors found that when they increased the number of repressor binding sites, the noise went up. This is also in qualitative agreement with our prediction that operator number positively correlates with intrinsic noise in the case of dual repression. However, the same caveat applies here as in the previous case studies, which is that only the total noise was measured. Although the authors of this study attributed all of the noise to intrinsic sources, it is still possible that extrinsic noise was responsible for the observed dependence of noise strength on operator number.

Finally, it is worth going back to bacteria, and discussing the only study that has yet examined the effect of a promoter-architecture motif on cell-to-cell variability in gene expression. In this paper, the authors investigated the effect of DNA looping on the total cell-to-cell variability for the P_{lacUV5} promoter in *E. coli* [24]. Using a novel single-protein counting technique, Choi and co-workers measured protein distributions for promoters whose auxiliary operator had been deleted (leaving them with a simple repression architecture), and compared them to promoters with the auxiliary operator O3 present, which allows for DNA looping. They report a reduction in protein noise due to the presence of O3, which according to our analysis, may indicate that the dissociation of the repressor from the looped state is faster than the normal dissociation rate. The authors attributed this looping-dependent decrease in noise to intrinsic origins, related to the different kinetics of repressor binding and rebinding to the main operator in the presence of the auxiliary operator, and in its absence. However, their measurements also reflect the total noise, and not only the intrinsic part, so the explanation may lie elsewhere. These results emphasize the need for more experiments in which the intrinsic noise is isolated and measured directly.

More recently, several impressive experimental studies have measured the noise in mRNA in bacteria for a host of different promoters ([90], and Ido Golding, private communication). In both of these cases, simplified low-dimensional models which do not consider the details of the promoter architecture have been exploited to provide a theoretical framework for thinking about the data. Our own studies indicate that the differences between a generic two-state model and specific models that attempt to capture the details of a given architecture are sometimes subtle and that the acid test of ideas like those presented in this paper can only come from experiments which systematically tune parameters, such as the repressor concentration, for a given transcriptional architecture.

5.4.3 Future directions

Some recent theoretical work has analyzed the effect of cooperative binding of activators in the context of particular examples of eukaryotic promoters [91, 92]. The main focus of this study is bacterial promoters. The simplicity of the microscopic mechanisms of transcriptional regulation for bacterial promoters makes them a better starting point for a systematic study like the one we propose. However, many examples of eukaryotic promoters have been found whose architecture affects the cell-to-cell variability [2, 12, 18, 32, 33]. Although the molecular mechanisms of gene regulation in these promoters are much more complex, with many intervening global and specific regulators [53], the stochastic model employed in this paper can be applied to any number of promoter states, and thus can be applied to these more complex promoters. Recent experimental work is starting to reveal the dynamics of nucleosomes and transcription factors with single-molecule sensitivity [93, 94], allowing the formulation of quantitative kinetic and thermodynamic mechanistic models of transcriptional regulation at the molecular level [77, 81]. The framework for analyzing gene expression at the single-cell level developed in this paper will be helpful in investigating the kinetic mechanisms of gene regulation in eukaryotic promoters, as the experimental studies switch from ensemble, to single cell.

5.4.4 Shortcomings of the approach

Although the model of transcriptional regulation used in this paper is standard in the field, it is important to remark that it is a very simplified model of what really happens during transcription initiation. There are many ways in which this kind of model can fail to describe real situations. For instance, mRNA degradation requires the action of RNases. These may become saturated if the global transcriptional activity is very large, and the degradation becomes non-linear [57]. Transcription initiation and elongation are assumed to be jointly captured in a single constant rate of mRNA synthesis for each promoter state. This is an oversimplification also. When considered explicitly, and in certain parameter ranges, the kinetics of RNAP-promoter interaction may cause noticeable effects in the overall variability [46]. Similarly, as pointed out elsewhere [95–97], translational pausing, backtracking or road-blocking may also cause significant deviations in mRNA variability from the predictions of the model used in this paper. How serious these deviations are depends on the specifics of each promoter-gene system. The model explored in this paper also assumes that the cell is a well-mixed environment. Deviations from that approximation can significantly affect cell-to-cell variability [58, 98]. Another simplification refers to cell growth and division, which are not treated explicitly by the model used in this paper: cell division and DNA replication cause doubling of gene and promoter copy number every cell cycle, as well as binomial partitioning of mRNAs between mother and daughter cells [4]. In eukaryotes, mRNA often needs to be further processed by the splicing apparatus before it becomes transcriptionally active. It also needs to be exported out of the nucleus, where it can be translated by ribosomes.

To study the effect of transcription factor dynamics on mRNA noise we assume that the unregulated promoter produces mRNA in a Poisson manner, at a constant rate. This assumption can turn out to be wrong if there is another process, independent of transcription factors, that independently turns the promoter on and off. In eukaryotes examples of such processes are nucleosome positioning and chromatin remodeling, while in prokaryotes analogous processes are not as established, but could include the action of non-specifically bound nucleoid proteins such as HU and H-NS, or DNA supercoiling. Experiments that measure cell-to-cell distributions of mRNA copy number in the absence of transcription factors (say without Lac repressor for the *lac* operon case) can settle this question. In case the Fano factor for this distribution is not one (as expected for a Poisson distribution) this can signal a possible transcription factor-independent source of variability. The stochastic models studied here can be extended to account for this situation. For example, the promoter can be made to switch between an on and an off state, where the transcription factors are allowed to interact with promoter DNA only while it is in the on state. In this case the mRNA fluctuations produced by an unregulated promoter will not be Poissonian. One can still investigate the affect of transcription factors by measuring how they change the nature of mRNA fluctuations from this new base-line. Comparison of this extended model with single-cell transcription experiments would then have the exciting potential for uncovering novel modes of transcriptional regulation in prokaryotes.

For the purpose of isolating the effect of individual promoter architectural elements on cell-to-cell variability in gene expression, we have artificially changed the value of one of those parameters, while keeping the other parameters constant. For instance, we have investigated the effect of altering the strength of an operator on the total cell-to-cell variability. In order to do this, we ask how changes in the dissociation rate of the transcription factor alter the cell-to-cell variability, given that all other rates (say the rate of transcription, or mRNA degradation) remain constant. This assumption is not necessarily always correct, since very often the operator sequence overlaps the promoter, and therefore changes in the sequence that alter operator strength also affect the sequence from which RNAP initiates transcription, which can potentially affect the overall rates of transcription. As is usually the case, biology presents us with a great diversity of forms, shapes and functions, and promoters are no exception. One needs to examine each promoter independently on the basis of the assumptions made in this paper, as many of these assumptions may apply for some promoters, but not for others.

For the same reason of isolating the effect of promoter architecture and cis-transcriptional regulation on cell-to-cell variability in gene expression, when we compare different architectures we make the simplifying assumption that they are transcribing the same gene, and therefore that the mRNA transcript has the same degradation rate. Care must be taken to take this into account when promoters transcribing different genes are investigated, since the mRNA degradation rate has a large effect on the level of cell-to-cell variability. We have also assumed that when transcription factors dissociate from the operator, they dissociate into an averaged out, well-mixed, mean-field concentration of transcription factors inside the cell. The possibility of transcription factors being recaptured by the same or another operator in the promoter right after they fall off the operator is not captured by the class of models considered here. Recent *in vivo* experiments suggest that this scenario may be important in yeast promoters containing arrays of operators [32].

In spite of all of the simplifications inherent in the class of models analyzed in this paper, we believe they are an adequate jumping off point for developing an intuition about how promoter architecture contributes to variability in gene expression. Our approach is to take a highly simplified model of stochastic gene expression, based on a kinetic model for the processes of the central dogma of molecular biology, and add promoter dynamics explicitly to see how different architectural features affect variability. This allows us to isolate the effect of promoter dynamics, and develop an intuitive understanding of how they affect the statistics of gene expression.

It must be emphasized, however, that the predictions made by the model may be wrong if any of the complications mentioned above are significant. This is not necessarily a bad outcome. If the comparison between experimental data and the predictions made by the theory for any particular system reveals inconsistencies, then the model will need to be refined and new experiments are required to identify which of the sources of variability that are not accounted for by the model are in play. In other words, experiments that test the quantitative predictions outlined stand a chance of gaining new insights about the physical mechanisms that underlie prokaryotic transcriptional regulation.

Appendix: The moments of the distributions

The moments of the mRNA probability distribution

We start by considering the same mechanism as in the text (see figure 5.1), in which the promoter switches between one active and one inactive state. There are only two stochastic variables in the model: the number of mRNA transcripts per cell (m), and the state of the promoter which reflects which transcription factors are bound where. The promoter state is always a discrete and finite stochastic variable (s) (for an example, see figure 5.1(A)). The example in figure 5.1(A) illustrates the simplest model of transcriptional activation by a transcription factor.

When the activator is bound to the promoter (state 1) mRNA is synthesized at rate r_1. When the activator is not bound (state 2) mRNA is synthesized at a lower rate r_2. The promoter switches stochastically from state 1 to state 2 with rate k_A^{off}, and from state 2 to state 1 with rate k_A^{on}. Each mRNA molecule is degraded with rate γ.

The time evolution for the joint probability of having the promoter in states 1 or 2, with m mRNAs in the cell (which we write as $p(1,m)$ and $p(2,m)$, respectively), is given by a master equation, which we can build by listing all possible reactions that lead to a change in cellular state, either by changing m or by changing s (figure 5.1(B)). The master equation takes the form:

$$\begin{aligned}
&\tfrac{d}{dt}p(1,m) = -k_A^{off}p(1,m) + k_A^{on}p(2,m) - r_1p(1,m) - \gamma mp(1,m) + r_1p(1,m-1) + \gamma(m+1)p(1,m+1),\\
&\tfrac{d}{dt}p(2,m) = k_A^{off}p(1,m) - k_A^{on}p(2,m) - r_2p(2,m) - \gamma mp(2,m) + r_2p(2,m-1) + \gamma(m+1)p(2,m+1).
\end{aligned}$$

Inspecting this system of equations, we notice that by defining the vector:

$$\vec{p}(m) = \begin{pmatrix} p(1,m) \\ p(2,m) \end{pmatrix},$$

and the matrices

$$\hat{K} = \begin{bmatrix} -k_A^{off} & k_A^{on} \\ k_A^{off} & -k_A^{on} \end{bmatrix};\hat{R} = \begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix};\hat{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

we can rewrite the system of equations 5.4.4 in matrix form

$$\frac{d}{dt}\vec{p}(m) = \left[\hat{K} - \hat{R} - m\gamma\hat{I}\right]\vec{p}(m) + \hat{R}\vec{p}(m-1) + (m+1)\gamma\hat{I}\vec{p}(m+1).$$

In steady-state, the left-hand side of equation 5.4.4 is equal to 0:

$$0 = \left[\hat{K} - \hat{R} - m\gamma\hat{I}\right]\vec{p}(m) + \hat{R}\vec{p}(m-1) + (m+1)\gamma\hat{I}\vec{p}(m+1).$$

In order to find the first two moments of the steady-state mRNA probability distribution, we follow the same strategy as in references [31, 38]: we multiply both sides of equation 5.4.4 by m and m^2 respectively, and then sum over all values of m, from 0 to ∞. We start from the first moment of the mRNA distribution, which requires us to multiply equation 5.4.4 by m and then sum:

$$\sum_{m=0}^{\infty} m\left(\left[\hat{K} - \hat{R} - m\gamma\hat{I}\right]\vec{p}(m) + \hat{R}\vec{p}(m-1) + (m+1)\gamma\hat{I}\vec{p}(m+1)\right) = \sum_{m=0}^{\infty} m\hat{K}\vec{p}(m) - \sum_{m=0}^{\infty} m^2\gamma\hat{I}\vec{p}(m)$$
$$- \sum_{m=0}^{\infty} m\hat{R}\vec{p}(m) + \sum_{m=0}^{\infty} m\hat{R}\vec{p}(m-1) + \sum_{m=0}^{\infty} m(m+1)\gamma\hat{I}\vec{p}(m+1).$$

Since none of the three matrices $\hat{K}$, $\hat{R}$ and $\hat{I}$ are functions of m, they can be taken out of the sums, and we find:

$$0 = \hat{K}\sum_{m=0}^{\infty} m\vec{p}(m) - \gamma\hat{I}\sum_{m=0}^{\infty} m^2\vec{p}(m) - \hat{R}\sum_{m=0}^{\infty} m\vec{p}(m) + \hat{R}\sum_{m=0}^{\infty} m\vec{p}(m-1) + \gamma\hat{I}\sum_{m=0}^{\infty} m(m+1)\vec{p}(m+1).$$

It will be convenient in what follows to define the following vectors of partial moments of the mRNA probability distribution:

$$\vec{m}_{(0)} = \sum_{m=0}^{\infty} m^0\vec{p}(m) = \begin{pmatrix} \sum_{m=0}^{\infty} m^0 p(1,m) \\ \sum_{m=0}^{\infty} m^0 p(2,m) \end{pmatrix} = \begin{pmatrix} p(1) \\ p(2) \end{pmatrix},$$
$$\vec{m}_{(1)} = \sum_{m=0}^{\infty} m\vec{p}(m) = \begin{pmatrix} \sum_{m=0}^{\infty} m p(1,m) \\ \sum_{m=0}^{\infty} m p(2,m) \end{pmatrix},$$
$$\vec{m}_{(2)} = \sum_{m=0}^{\infty} m^2\vec{p}(m) = \begin{pmatrix} \sum_{m=0}^{\infty} m^2 p(1,m) \\ \sum_{m=0}^{\infty} m^2 p(2,m) \end{pmatrix}.$$

The usefulness of these vectors of partial moments of the mRNA distribution lies in the fact that they are related to the moments of the probability distribution. For instance, the mean mRNA is given by

$$\langle m\rangle = \sum_{s=1}^{2}\sum_{m=0}^{\infty} mp(s,m) = \sum_{m=0}^{\infty} mp(1,m) + \sum_{m=0}^{\infty} mp(2,m).$$

If we define, again for convenience, the vector $\vec{u} = (1, 1)$, we find that the mean of the mRNA distribution is related to the vectors of partial moments by $\langle m \rangle = \vec{u} \cdot \vec{m}_{(1)}$. Following this example, it is also straightforward to prove that the second moment of the mRNA distribution is given by: $\langle m^2 \rangle = \vec{u}\vec{m}_{(2)}$. Given these definitions, we return to equation 5.4.4 which we can now write as:

$$\hat{K}\vec{m}_{(1)} - \gamma\hat{I}\vec{m}_{(2)} - \hat{R}\vec{m}_{(1)} + \hat{R}\sum_{m=0}^{\infty} m\vec{p}(m-1) + \gamma\hat{I}\sum_{m=0}^{\infty} m(m+1)\vec{p}(m+1) = 0.$$

We can re-arrange terms in the last two sums so that we write them as operations on the vectors of partial moments of the probability distributions. For instance, by making the change of variables: $m \rightarrow m+1$, and taking into account the fact that the number of mRNA molecules inside the cell can never fall below 0 (so that $\vec{p}(-1) = 0$), we find:

$$\sum_{m=0}^{\infty} m\vec{p}(m-1) = \sum_{m=0}^{\infty} (m+1)\vec{p}(m) = \vec{m}_{(1)} + \vec{m}_{(0)}.$$

Similarly, by making the change of variables $m+1 \rightarrow m$, the last sum takes the simpler form:

$$\sum_{m=0}^{\infty} m(m+1)\vec{p}(m+1) = \sum_{m=0}^{\infty} m(m-1)\vec{p}(m) = \vec{m}_{(2)} - \vec{m}_{(1)}.$$

Entering these results into equation 5.4.4, we finally find:

$$\hat{K}\vec{m}_{(1)} - \gamma\hat{I}\vec{m}_{(2)} - \hat{R}\vec{m}_{(1)} + \hat{R}\left(\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \gamma\hat{I}\left(\vec{m}_{(2)} - \vec{m}_{(1)}\right) = \hat{K}\vec{m}_{(1)} - \gamma\hat{I}\vec{m}_{(1)} + \hat{R}\vec{m}_{(0)}.$$

The vector of partial moments $\vec{m}_{(1)}$ is therefore the solution to the matrix equation:

$$\left(\hat{K} - \gamma\hat{I}\right)\vec{m}_{(1)} + \hat{R}\vec{m}_{(0)} = 0.$$

The final step is to multiply both sides of equation 5.4.4 by the vector $\vec{u} = (1, 1)$. Because of how it was constructed (i.e., $p(1, m)s$ loss is $p(2, m)s$ gain during transitions between promoter states), the matrix $\hat{K}$ has the property that the sum of the elements of any one of its columns is always 0. Therefore, we find that $\vec{u}\hat{K} = 0$. The matrix $\hat{R}$ is diagonal, so if we multiply matrix $\hat{R}$ on the left by vector $\vec{u}$, we get a vector that is equal to the list of diagonal elements of matrix $\hat{R}$. Thus, we define the vector $\vec{r} = (\hat{R}_{11}, \hat{R}_{22})$, as the vector for which it is true that $\vec{u}\hat{R} = \vec{r}$. Finally, the identity matrix is $\hat{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Therefore, multiplying $\hat{I}$ on the left by the vector $\vec{u}$ leads us to: $\vec{u}\hat{I} = \vec{u}$. Therefore, when we multiply equation 5.4.4 by the vector we find:

$$0 = \vec{u}\hat{K}\vec{m}_{(1)} - \vec{u}\gamma\hat{I}\vec{m}_{(1)} + \vec{u}\hat{R}\vec{m}_{(0)} = -\gamma\vec{u}\vec{m}_{(1)} + \vec{r}\vec{m}_{(0)}.$$

Knowing that the mean of the mRNA distribution is related to the vector of partial moments by: $\langle m \rangle =$

$\vec{u}\vec{m}_{(1)}$, we find that:

$$\langle m \rangle = \frac{\vec{r}\vec{m}_{(0)}}{\gamma}.$$

Note that, by definition,

$$\vec{m}_{(0)} = \sum_{m=0}^{\infty} m^0 \vec{p}(m) = \left(\begin{array}{c} \sum\limits_{m=0}^{\infty} m^0 p(1,m) \\ \sum\limits_{m=0}^{\infty} m^0 p(2,m) \end{array} \right) = \left(\begin{array}{c} p(1) \\ p(2) \end{array} \right).$$

In other words, the first element of vector $\vec{m}_{(0)}$ is the steady-state probability to find the promoter in state 1, and the second element is the steady-state probability to find the promoter in state 2. This vector is straightforward to obtain by summing equation 5.4.4 over all m, and it is the solution of $\hat{K}\vec{m}_{(0)} = 0$, normalized so that $p(1) + p(2) = 1$. In order to find the second moment, we just multiply equation 5.4.4 by m^2 and sum over all m from 0 to ∞. As a result of this manipulation, we find:

$$\begin{array}{l} \sum\limits_{m=0}^{\infty} m^2 \left(\left[\hat{K} - \hat{R} - m\gamma\hat{I} \right] \vec{p}(m) + \hat{R}\vec{p}(m-1) + (m+1)\gamma\hat{I}\vec{p}(m+1) \right) = \sum\limits_{m=0}^{\infty} m^2 \hat{K}\vec{p}(m) - \sum\limits_{m=0}^{\infty} m^3 \gamma\hat{I}\vec{p}(m) \\ - \sum\limits_{m=0}^{\infty} m^2 \hat{R}\vec{p}(m) + \sum\limits_{m=0}^{\infty} m^2 \hat{R}\vec{p}(m-1) + \sum\limits_{m=0}^{\infty} m^2 (m+1)\gamma\hat{I}\vec{p}(m+1) = \\ \hat{K}\vec{m}_{(2)} - \gamma\hat{I}\vec{m}_{(3)} - \hat{R}\vec{m}_{(2)} + \sum\limits_{m=0}^{\infty} m^2 \hat{R}\vec{p}(m-1) + \sum\limits_{m=0}^{\infty} m^2 (m+1)\gamma\hat{I}\vec{p}(m+1). \end{array}$$

The last two terms of the right-hand side of equation 5.4.4 can be simplified by writing the two sums in terms of the vectors of partial moments. In order to do that, we must make the same changes of variables that we invoked above when dealing with the mean. First, the change of variables $m \to m+1$ allows us to rewrite the first sum as:

$$\sum_{m=0}^{\infty} m^2 \vec{p}(m-1) = \sum_{m=0}^{\infty} (m+1)^2 \vec{p}(m) = \vec{m}_{(2)} + 2\vec{m}_{(1)} + \vec{m}_{(0)}.$$

Finally, the change of variables $m+1 \to m$, allows us to re-write the last sum as:

$$\sum_{m=0}^{\infty} m^2 (m+1)\vec{p}(m+1) = \sum_{m=0}^{\infty} m(m-1)^2 \vec{p}(m) = \vec{m}_{(3)} - 2\vec{m}_{(2)} + \vec{m}_{(1)}.$$

Entering these last two sums in equation 5.4.4, we find:

$$\begin{array}{l} \hat{K}\vec{m}_{(2)} - \gamma\hat{I}\vec{m}_{(3)} - \hat{R}\vec{m}_{(2)} + \hat{R}\left(\vec{m}_{(2)} + 2\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \gamma\hat{I}\left(\vec{m}_{(3)} - 2\vec{m}_{(2)} + \vec{m}_{(1)}\right) = \\ \hat{K}\vec{m}_{(2)} + \hat{R}\left(2\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \gamma\hat{I}\left(-2\vec{m}_{(2)} + \vec{m}_{(1)}\right) = 0. \end{array}$$

As we did before, we can transform this equation into an equation for the moments of the mRNA distribution by multiplying both sides of this equation on the left by the vector $\vec{u}$. Performing these operations, we find:

$$\vec{u}\hat{K}\vec{m}_{(2)} + \vec{u}\hat{R}\left(2\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \vec{u}\gamma\hat{I}\left(-2\vec{m}_{(2)} + \vec{m}_{(1)}\right) = \vec{r}\left(2\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \gamma\vec{u}\left(-2\vec{m}_{(2)} + \vec{m}_{(1)}\right) =$$
$$2\vec{r}\vec{m}_{(1)} + \vec{r}\vec{m}_{(0)} - 2\gamma\left\langle m^2\right\rangle + \gamma\left\langle m\right\rangle = 0.$$

Therefore, the second moment of the mRNA distribution in steady-state is given by:

$$\left\langle m^2\right\rangle = \frac{\vec{r}\vec{m}_{(1)}}{\gamma} + \frac{\vec{r}\vec{m}_{(0)} + \gamma\left\langle m\right\rangle}{2\gamma}.$$

Using the fact that the first moment is given by:

$$\left\langle m\right\rangle = \frac{\vec{r}\vec{m}_{(0)}}{\gamma}.$$

, we can further simplify the second moment as:

$$\left\langle m^2\right\rangle = \left\langle m\right\rangle + \frac{\vec{r}\vec{m}_{(1)}}{\gamma}.$$

Therefore, the normalized variance can be written as:

$$\eta^2 = \frac{\left\langle m^2\right\rangle - \left\langle m\right\rangle^2}{\left\langle m\right\rangle^2} = \frac{1}{\left\langle m\right\rangle} + \frac{1}{\left\langle m\right\rangle^2}\left(\frac{\vec{r}\vec{m}_{(1)}}{\gamma} - \left\langle m\right\rangle^2\right).$$

The moments of the protein probability distribution

We can use the same method to compute the normalized variance of the protein distribution. We will start from a promoter that is constitutively active, and then extend our analysis to a promoter that switches between two or more active and inactive states. We assume that each transcription event leads to the production of multiple proteins (a "burst"). The number of proteins produced per mRNA (which we denote as β) obeys a geometric distribution [6, 11, 14] with an average burst size . Therefore, the probability for β is given by: $h\left(\beta\right) = \frac{b^\beta}{(1+b)^{\beta+1}}$. We assume that proteins are also degraded with a constant rate per molecule of δ . In order to write down the master equation for this process, we have to consider all the possible ways in which the cell can enter or leave a state with n proteins during a small increment of time dt. If we assume that mRNA lifetime is much shorter than protein lifetime (an approximation that is realistic in many experimental systems — see refs [6, 11, 52]), then all of the proteins may be assumed to be made simultaneously. Therefore, we need to consider the possibility that the cell will jump from a state with n proteins to a state with $n+\beta$, for all possible values of β. The probability that the cell will leave a state with n proteins, by making a transition to a state with $n+\beta$ proteins is equal to the product of the probability that the cell is in a state with n proteins ($p(n)$), the probability that the cell will make a transcript during dt (rdt), and the probability that the transcript makes β proteins before it is degraded ($h(\beta)$). Thus, the total probability per unit time to abandon a state with n proteins is given by $rh(\beta)p(n)$. Since β can in principle take any integer value, the total probability to abandon the state with n proteins by the occurrence of a protein burst is given by the sum of $rh(\beta)p(n)$ over all possible values of β. This term will be given by: $\sum_{\beta=1}^{\infty} rh\left(\beta\right)p(n) = rp(n) \sum_{\beta=1}^{\infty} h\left(\beta\right)$. Also, we need to consider that the cell may enter a state with n proteins from any state with less than n proteins. The probability per unit time that the cell enters a state with n proteins, from a state with $n-\beta$ proteins is given by: $rh(\beta)p(n-\beta)$. Therefore, following the same logic as we did before, the net probability per unit time that the cell enters a state with n proteins is $\sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta)$. With these considerations, the master equation for a constitutive promoter is given by:

$$\frac{d}{dt}p(n) = -\sum_{\beta=1}^{\infty} rh\left(\beta\right)p(n) + \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta) - \delta np(n) + \delta(n+1)p(n+1).$$

As discussed above, the first sum can be further simplified to:

$$\sum_{\beta=1}^{\infty} rh\left(\beta\right)p(n) = rp(n)\sum_{\beta=1}^{\infty} h\left(\beta\right) = rp(n)\sum_{\beta=1}^{\infty} \frac{b^{\beta}}{\left(1+b\right)^{\beta+1}} = r\left(\frac{b}{1+b}\right)p(n).$$

As a result, the master equation takes the form:

$$\frac{d}{dt}p(n) = -r\left(\frac{b}{1+b}\right)p(n) + \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta) - \delta np(n) + \delta(n+1)p(n+1).$$

In steady-state, the right-hand side of equation 5.4.4 is equal to 0, and we have:

$$0 = -r\left(\frac{b}{1+b}\right)p(n) + \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta) - \delta np(n) + \delta(n+1)p(n+1).$$

The first two moments of the steady-state protein distribution $p(n)$ can be obtained, in exactly the same way we used to find out the moments of the mRNA distribution in the previous section: by multiplying both sides of equation 5.4.4 by n and n^2 respectively, and then summing over all n. Before we do that, it is useful to evaluate the sums $\sum_{n=0}^{\infty} n^2 \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta)$ and $\sum_{n=0}^{\infty} n \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta)$. We can find the general term of the first sum by expanding the series:

$$\sum_{n=0}^{\infty} n^2 \sum_{\beta=1}^{n} h\left(\beta\right)p(n-\beta) = 1^2\left(h(1)p(0)\right) + 2^2\left(h(1)p(1) + h(2)p(0)\right) + 3^2\left(h(1)p(2) + h(2)p(1) + h(3)p(0)\right) + ... =$$
$$\left(1^2h(1) + 2^2h(2) + 3^2h(3)...\right)p(0) + \left(2^2h(1) + 3^2h(2) + 4^2h(3)...\right)p(1) + \left(3^2h(1) + 4^2h(2) + 5^2h(3)...\right)p(2) + ... =$$
$$\sum_{n=0}^{\infty} p(n)\left(\sum_{\beta=1}^{\infty} h\left(\beta\right)\left(n+\beta\right)^2\right) = \sum_{n=0}^{\infty}\left(b + 2b^2 + 2bn + \tfrac{b}{1+b}n^2\right)p(n).$$

We can do the same for the second sum, and we find:

$$\sum_{n=0}^{\infty} n \sum_{\beta=1}^{n} h\left(\beta\right) p(n-\beta) = 1\left(h(1)p(0)\right) + 2\left(h(1)p(1) + h(2)p(0)\right) + 3\left(h(1)p(2) + h(2)p(1) + h(3)p(0)\right) + ... =$$
$$\left(h(1) + 2h(2) + 3h(3)...\right)p(0) + \left(2h(1) + 3h(2) + 4h(3)...\right)p(1) + \left(3h(1) + 4h(2) + 5h(3)...\right)p(2) + ... =$$
$$\sum_{n=0}^{\infty} p(n)\left(\sum_{\beta=1}^{\infty} h\left(\beta\right)\left(n+\beta\right)\right) = \sum_{n=0}^{\infty}\left(b + \tfrac{b}{1+b}n\right)p(n).$$

Likewise, it will be necessary to recall from the first section of this Appendix, that the sum $\sum_{n=0}^{\infty} n(n+1)p(n+1)$ can be computed by using the change of variables: $n+1 \rightarrow n$, and we find:

$$\sum_{n=0}^{\infty} n(n+1)p(n+1) = \sum_{n=0}^{\infty} n(n-1)p(n).$$

With these results in hand, we can finally solve the first two moments of the protein distribution $p(n)$. As explained above, we can find the first moment by multiplying both sides of equation 5.4.4 by n and then summing over all n. In order to find the second moment, we multiply both sides of equation 5.4.4 by n^2 and then sum over all n. For the first moment, we find:

$$0 = -r\left(\tfrac{b}{1+b}\right)\sum_{n=0}^{\infty} np(n) + r\sum_{n=0}^{\infty} n\sum_{\beta=1}^{n} h\left(\beta\right)p(n-\beta) - \delta\sum_{n=0}^{\infty} n^2p(n) + \delta\sum_{n=0}^{\infty} n(n+1)p(n+1) =$$
$$= -r\left(\tfrac{b}{1+b}\right)\left\langle n\right\rangle + r\left(b + \tfrac{b}{1+b}\left\langle n\right\rangle\right) - \delta\left\langle n^2\right\rangle + \delta\left\langle n^2\right\rangle - \delta\left\langle n\right\rangle = rb - \delta\left\langle n\right\rangle.$$

Solving this equation, we find that the mean protein per cell is equal to:

$$\left\langle n\right\rangle = \frac{rb}{\delta}.$$

For the second moment, we find:

$$0 = -r\left(\tfrac{b}{1+b}\right)\sum_{n=0}^{\infty} n^2p(n) + r\sum_{n=0}^{\infty} n^2\sum_{\beta=1}^{n} h\left(\beta\right)p(n-\beta) - \delta\sum_{n=0}^{\infty} n^3p(n) + \delta\sum_{n=0}^{\infty} n^2(n+1)p(n+1) =$$
$$= -r\left(\tfrac{b}{1+b}\right)\left\langle n\right\rangle + r\left(b + 2b^2 + 2b\left\langle n\right\rangle + \tfrac{b}{1+b}\left\langle n^2\right\rangle\right) - \delta\left\langle n^3\right\rangle + \delta\left\langle n^3\right\rangle - 2\delta\left\langle n^2\right\rangle + \delta\left\langle n\right\rangle =$$
$$= r\left(b + 2b^2 + 2b\left\langle n\right\rangle\right) - 2\delta\left\langle n^2\right\rangle + \delta\left\langle n\right\rangle.$$

Solving this last equation, we find that the second moment of the protein distribution is equal to:

$$\left\langle n^2\right\rangle = \frac{r}{2\delta}b + \frac{r}{2\delta}2b^2 + \frac{r}{2\delta}2b\left\langle n\right\rangle + \frac{\left\langle n\right\rangle}{2} = (1+b)\left\langle n\right\rangle + \left\langle n\right\rangle^2.$$

Therefore, the normalized variance of the protein distribution for a constitutive promoter takes the form:

$$\frac{Var(n)}{\left\langle n\right\rangle^2} = \frac{\left\langle n^2\right\rangle - \left\langle n\right\rangle^2}{\left\langle n\right\rangle^2} = \frac{(1+b)\left\langle n\right\rangle + \left\langle n\right\rangle^2 - \left\langle n\right\rangle^2}{\left\langle n\right\rangle^2} = \frac{(1+b)}{\left\langle n\right\rangle}.$$

If now we consider that the promoter can exist in two states, characterized by having different rates of transcription, then the cell's state is characterized not only by the number of proteins present, but also by the state of the promoter. Therefore, the master equation must consider two variables: one characterizing the state of the promoter (s), and one representing the number of proteins per cell (n). By analogy with the mRNA master equation, and the master equation for the protein distribution of a constitutive promoter, the two-state master equation for the protein distribution can be written as:

$$\begin{aligned}
\tfrac{d}{dt}p(1,n) &= -k_A^{on}p(1,n) + k_A^{off}p(2,n) - \sum_{\beta=1}^{\infty} r_1 h\left(\beta\right)p(1,n) + \sum_{\beta=1}^{n} r_1 h\left(\beta\right)p(1,n-\beta) - \delta n p(1,n) + \delta(n+1)p(1,n+1),\\
\tfrac{d}{dt}p(2,n) &= k_A^{on}p(1,n) - k_A^{off}p(2,n) - \sum_{\beta=1}^{\infty} r_2 h\left(\beta\right)p(2,n) + \sum_{\beta=1}^{n} r_2 h\left(\beta\right)p(2,n-\beta) - \delta n p(2,n) + \delta(n+1)p(2,n+1).
\end{aligned}$$

Just as we did in order to compute the moments of the mRNA distribution, we can define the vector $\vec{p}(n) = (p(1,n), p(2,n))$. By doing so, we will be able to re-write the master equation 5.4.4 as a matrix equation, that will be applicable to any promoter with any number of states. This matrix equation can be written in terms of exactly the same matrices we used for the mRNA probability distribution. We find:

$$\frac{d}{dt}\vec{p}(n) = \left[\hat{K} - \frac{b}{1+b}\hat{R} - n\delta\hat{I}\right]\vec{p}(n) + \hat{R}\sum_{\beta=1}^{n} h(\beta)\vec{p}(n-\beta) + (n+1)\delta\hat{I}\vec{p}(n+1).$$

In steady-state, the left side of equation 5.4.4 is equal to 0, and the master equation has the form:

$$0 = \left[\hat{K} - \frac{b}{1+b}\hat{R} - n\delta\hat{I}\right]\vec{p}(n) + \hat{R}\sum_{\beta=1}^{n} h(\beta)\vec{p}(n-\beta) + (n+1)\delta\hat{I}\vec{p}(n+1).$$

Just as we did in order to calculate the moments of the mRNA distribution, it will be convenient to define the vectors of partial moments:

$$\begin{aligned}
\vec{n}_{(0)} &= \sum_{n=0}^{\infty} n^0\vec{p}(n) = \begin{pmatrix} \sum_{n=0}^{\infty} n^0 p(1,n) \\ \sum_{n=0}^{\infty} n^0 p(2,n) \end{pmatrix} = \begin{pmatrix} p(1) \\ p(2) \end{pmatrix},\\
\vec{n}_{(1)} &= \sum_{n=0}^{\infty} n\vec{p}(n) = \begin{pmatrix} \sum_{n=0}^{\infty} n p(1,n) \\ \sum_{n=0}^{\infty} n p(2,n) \end{pmatrix},\\
\vec{n}_{(2)} &= \sum_{n=0}^{\infty} n^2\vec{p}(n) = \begin{pmatrix} \sum_{n=0}^{\infty} n^2 p(1,n) \\ \sum_{n=0}^{\infty} n^2 p(2,n) \end{pmatrix}.
\end{aligned}$$

It is straightforward to see that the vector is exactly identical to the vector $\vec{m}_{(0)}$. The next two vectors $\vec{n}_{(1)}$ and $\vec{n}_{(2)}$ can be obtained by multiplying equation 5.4.4 by n and n^2 respectively, and then summing over all

n. We end up with the following two equations:

$$\begin{aligned}
0 &= \sum_{n=0}^{\infty} n\left[\hat{K} - \tfrac{b}{1+b}\hat{R} - n\delta\hat{I}\right]\vec{p}(n) + \hat{R}\sum_{n=0}^{\infty} n \sum_{\beta=1}^{n} h(\beta)\vec{p}(n{-}\beta) + \sum_{n=0}^{\infty} n(n+1)\delta\hat{I}\cdot\vec{p}(n+1) = \\
&= \hat{K}\vec{n}_{(1)} - \tfrac{b}{1+b}\hat{R}\vec{n}_{(1)} - \delta\hat{I}\vec{n}_{(2)} + \delta\hat{I}\left(\vec{n}_{(2)} - \vec{n}_{(1)}\right) + \hat{R}\left(\tfrac{b}{1+b}\vec{n}_{(1)} + b\vec{n}_{(0)}\right) \\
&= \left(\hat{K} - \delta\hat{I}\right)\vec{n}_{(1)} + b\hat{R}\vec{n}_{(0)},
\end{aligned}$$

and

$$\begin{aligned}
0 &= \sum_{n=0}^{\infty} n^2\left[\hat{K} - \tfrac{b}{1+b}\hat{R} - n\delta\hat{I}\right]\vec{p}(n) + \hat{R}\sum_{n=0}^{\infty} n^2 \sum_{\beta=1}^{n} h(\beta)\vec{p}(n{-}\beta) + \sum_{n=0}^{\infty} n^2(n+1)\delta\hat{I}\cdot\vec{p}(n+1) = \\
&= \hat{K}\vec{n}_{(2)} - \tfrac{b}{1+b}\hat{R}\vec{n}_{(2)} - \delta\hat{I}\vec{n}_{(3)} + \delta\hat{I}\left(\vec{n}_{(3)} - 2\vec{n}_{(2)} + \vec{n}_{(1)}\right) + \hat{R}\left(\tfrac{b}{1+b}\vec{n}_{(2)} + 2b\vec{n}_{(1)} + b(1+2b)\vec{n}_{(0)}\right) = \\
&= \hat{K}\vec{n}_{(2)} + \delta\hat{I}\left(-2\vec{n}_{(2)} + \vec{n}_{(1)}\right) + \hat{R}\left(2b\vec{n}_{(1)} + b(1+2b)\vec{n}_{(0)}\right) = \\
&= \left(\hat{K} - 2\delta\hat{I}\right)\vec{n}_{(2)} + \left(\delta\hat{I} + 2b\hat{R}\right)\vec{n}_{(1)} + b(1+2b)\hat{R}\vec{n}_{(0)}.
\end{aligned} .$$

Now by multiplying the vector $\vec{u} = (1,1)$ on the left of equations 5.4.4 and 5.4.4, we find

$$0 = -\delta\left\langle n\right\rangle + b\vec{r}\vec{n}_{(0)},$$

and

$$0 = -2\delta\left\langle n^2\right\rangle + \delta\left\langle n\right\rangle + 2b\vec{r}\vec{n}_{(1)} + b(1+2b)\vec{r}\vec{n}_{(0)}.$$

Thus, we find analytical equations for the first two moments of the protein distribution:

$$\left\langle n\right\rangle = \frac{b\vec{r}\vec{n}_{(0)}}{\delta},$$

$$\left\langle n^2\right\rangle = (1+b)\left\langle n\right\rangle + \frac{b\vec{r}\vec{n}_{(1)}}{\delta}.$$

Where $\vec{n}_{(1)}$ is the solution of equation 5.4.4:

$$0 = \left(\hat{K} - \delta\hat{I}\right)\vec{n}_{(1)} + b\hat{R}\vec{n}_{(0)}.$$

Armed with these equations, we can finally compute the stationary variance of the protein distribution:

$$\frac{Var(n)}{\left\langle n\right\rangle^2} = \frac{(1+b)\left\langle n\right\rangle + \frac{b\vec{r}\cdot\vec{n}_{(1)}}{\delta} - \left\langle n\right\rangle^2}{\left\langle n\right\rangle^2} = \frac{(1+b)}{\left\langle n\right\rangle} + \frac{1}{\left\langle n\right\rangle^2}\left(\frac{b\vec{r}\vec{n}_{(1)}}{\delta} - \left\langle n\right\rangle^2\right).$$

Bibliography

[1] A. Sanchez, H. G. Garcia, D. Jones, R. Phillips, and J. Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput Biol*, 2011. (*Under review*).

[2] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biol*, 4(10):e309, 2006.

[3] J. Elf, G. W. Li, and X. S. Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–4, 2007.

[4] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002.

[5] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.

[6] L. Cai, N. Friedman, and X. S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62, 2006.

[7] J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer. Transcriptional pulsing of a developmental gene. *Curr Biol*, 16(10):1018–25, 2006.

[8] H. Maamar, A. Raj, and D. Dubnau. Noise in gene expression determines cell fate in bacillus subtilis. *Science*, 317(5837):526–9, 2007.

[9] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. Van Oudenaarden. Regulation of noise in the expression of a single gene. *Nat Genet*, 31(1):69–73, 2002.

[10] A. Raj and A. Van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys*, 38:255–70, 2009.

[11] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–3, 2006.

[12] K. F. Murphy, G. Balazsi, and J. J. Collins. Combinatorial promoter design for engineering noisy gene expression. *Proc Natl Acad Sci U S A*, 104(31):12726–31, 2007.

[13] D. R. Rigney and W. C. Schieve. Stochastic model of linear, continuous protein synthesis in bacterial populations. *J Theor Biol*, 69(4):761–6, 1977.

[14] O. G. Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *J Theor Biol*, 71(4):587–603, 1978.

[15] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'shea, Y. Pilpel, and N. Barkai. Noise in protein expression scales with natural protein abundance. *Nat Genet*, 38(6):636–43, 2006.

[16] W. J. Blake, K. A. M, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–7, 2003.

[17] J. M. Raser and E. K. O'shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–4, 2004.

[18] W. J. Blake, G. Balazsi, M. A. Kohanski, F. J. Isaacs, K. F. Murphy, Y. Kuang, C. R. Cantor, D. R. Walt, and J. J. Collins. Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell*, 24(6):853–65, 2006.

[19] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: From theories to phenotypes. *Nat Rev Genet*, 2005.

[20] N. Maheshri and E. K. O'shea. Living with noisy genes: How cells function reliably with inherent variability in gene expression. *Annu Rev Biophys Biomol Struct*, 36:413–34, 2007.

[21] M. F. Wernet, E. O. Mazzoni, A. Celik, D. M. Duncan, I. Duncan, and C. Desplan. Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature*, 440(7081):174–80, 2006.

[22] L. S. Weinberger, J. C. Burnett, J. E. Toettcher, A. P. Arkin, and D. V. Schaffer. Stochastic gene expression in a lentiviral positive-feedback loop: Hiv-1 tat fluctuations drive phenotypic diversity. *Cell*, 122(2):169–82, 2005.

[23] M. Ackermann, B. Stecher, N. E. Freed, P. Songhet, W. D. Hardt, and M. Doebeli. Self-destructive cooperation mediated by phenotypic noise. *Nature*, 454(7207):987–90, 2008.

[24] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–6, 2008.

[25] R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65–8, 2008.

[26] A. Singh and L. S. Weinberger. Stochastic gene expression as a molecular switch for viral latency. *Curr Opin Microbiol*, 12(4):460–6, 2009.

[27] D. W. Austin, M. S. Allen, J. M. Mccollum, R. D. Dar, J. R. Wilgus, G. S. Sayler, N. F. Samatova, C. D. Cox, and M. L. Simpson. Gene network shaping of inherent noise spectra. *Nature*, 439(7076):608–11, 2006.

[28] C. D. Cox, J. M. Mccollum, M. S. Allen, R. D. Dar, and M. L. Simpson. Using noise to probe and characterize gene circuits. *Proc Natl Acad Sci U S A*, 105(31):10809–14, 2008.

[29] D. Nevozhay, R. M. Adams, K. F. Murphy, K. Josic, and G. Balazsi. Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of gene expression. *Proc Natl Acad Sci U S A*, 106(13):5123–8, 2009.

[30] J. M. Pedraza and A. Van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717):1965–9, 2005.

[31] A. Sanchez and J. Kondev. Transcriptional control of noise in gene expression. *Proc Natl Acad Sci U S A*, 105(13):5081–6, 2008.

[32] T. L. To and N. Maheshri. Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science*, 327(5969):1142–5, 2010.

[33] F. M. Rossi, A. M. Kringstein, A. Spicher, O. M. Guicherit, and H. M. Blau. Transcriptional control: Rheostat converted to on/off switch. *Mol Cell*, 6(3):723–8, 2000.

[34] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[35] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[36] J. Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–8, 2004.

[37] J. Peccoud and B. Ycart. Markovian modelig of gene product synthesis. *Theor Popul Biol*, 48:222–234, 1995.

[38] T. B. Kepler and T. C. Elston. Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys J.*, 81:3116–36., 2001.

[39] P. J. Ingram, M. P. Stumpf, and J. Stark. Nonidentifiability of the source of intrinsic noise in gene expression from single-burst data. *PLoS Comput Biol*, 4(10):e1000192, 2008.

[40] A. Warmflash and A. R. Dinner. Signatures of combinatorial regulation in intrinsic biological noise. *Proc Natl Acad Sci U S A*, 105(45):17262–7, 2008.

[41] M. J. Dunlop, R. S. Cox Iii, J. H. Levine, R. M. Murray, and M. B. Elowitz. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics*, 40:1493–1498, 2007.

[42] O. K. Wong, M. Guthold, D. A. Erie, and J. Gelles. Interconvertible lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biol*, 6(9):e232, 2008.

[43] Y. Wang, L. Guo, I. Golding, E. C. Cox, and N. P. Ong. Quantitative transcription factor binding kinetics at the single-molecule level. *Biophys J*, 96(2):609–20, 2009.

[44] M. Thattai and A. Van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A*, 98(15):8614–9, 2001.

[45] J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175, 2005.

[46] T. Hofer and M. J. Rasch. On the kinetic design of transcription. *Genome Informatics*, 16:73–82., 2005.

[47] D. Zenklusen, D. R. Larson, and R. H. Singer. Single-rna counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*, 15(12):1263–71, 2008.

[48] S. E. Halford. An end to 40 years of mistakes in DNA-protein association kinetics? *Biochem Soc Trans*, 37(Pt 2):343–8, 2009.

[49] H. D. Kim and E. K. O'shea. A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol*, 15:1192–1198, 2008.

[50] I. B. Dodd, K. E. Shearwin, A. J. Perkins, T. Burr, A. Hochschild, and J. B. Egan. Cooperativity in long-range gene regulation by the lambda ci repressor. *Genes Dev*, 18(3):344–54, 2004.

[51] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–5, 2005.

[52] D. Kennell and H. Riezman. Transcription and translation initiation frequencies of the *escherichia coli lac* operon. *J Mol Biol*, 114(1):1–21, 1977.

[53] H. Boeger, J. Griesenbeck, and R. D. Kornberg. Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. *Cell*, 133:716–726, 2008.

[54] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*, 81:2340–2361, 1977.

[55] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003.

[56] C. D. Cox, J. M. Mccollum, D. W. Austin, M. S. Allen, R. D. Dar, and M. L. Simpson. Frequency domain analysis of noise in simple gene circuits. *Chaos*, 16(2):026102, 2006.

[57] J. M. Pedraza and J. Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861):339–43, 2008.

[58] J. S. Van Zon, M. J. Morelli, S. TanaseNicola, and P. R. Ten Wolde. Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophys J*, 91:43504367, 2006.

[59] D. F. Browning and S. J. Busby. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*, 2(1):57–65, 2004.

[60] K. S. Koblan and G. K. Ackers. Site-specific enthalpic regulation of DNA transcription at bacteriophage lambda or. *Biochemistry*, 31(1):57–65, 1992.

[61] M. Ptashne. *A genetic switch: Phage lambda revisited.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 3rd edition, 2004.

[62] A. C. Babic and J. W. Little. Cooperative DNA binding by ci repressor is dispensable in a phage lambda variant. *Proc Natl Acad Sci U S A*, 104(45):17741–6, 2007.

[63] C. Zurla, C. Manzo, D. Dunlap, D. E. A. Lewis, S. Adhya, and L. Finzi. Direct demonstration and quantification of long-range DNA looping by the lambda-bacteriophage repressor. *Nucleic Acids Res.*, 37:2789–2795, 2009.

[64] S. Semsey, M. Geanacopoulos, D. E. Lewis, and S. Adhya. Operator-bound galr dimers close DNA loops by direct interaction: Tetramerization and inducer binding. *Embo J*, 21(16):4349–56, 2002.

[65] B. MÜLler-Hill. *The lac operon: A short history of a genetic paradigm.* Walter de Gruyter, Berlin, New York, 1996.

[66] F. Vanzi, C. Broggio, L. Sacconi, and F. S. Pavone. Lac repressor hinge flexibility and DNA looping: Single molecule kinetics by tethered particle motion. *Nucleic Acids Res*, 34(12):3409–20, 2006.

[67] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[68] J. M. Vilar and L. Saiz. DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. *Curr Opin Genet Dev*, 15(2):136–44, 2005.

[69] J. MÜLler, S. Oehler, and B. MÜLler-Hill. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol*, 257(1):21–9, 1996.

[70] A. S. Cameron and R. J. Redfield. Crp binding and transcription activation at crp-s sites . *J Mol Biol*, 383:313–23, 2008.

[71] K. Gaston, A. Kolb, and S. Busby. Binding of the escherichia coli cyclic amp receptor protein todna fragments containing consensus nucleotide sequences. *Biochem J*, 261:649–53, 1989.

[72] J. K. Joung, D. M. Koepp, and A. Hochschild. Synergistic activation of transcription by bacteriophage lambda ci protein and e. Coli camp receptor protein. *Science*, 265(5180):1863–6, 1994.

[73] J. K. Joung, L. U. Le, and A. Hochschild. Synergistic activation of transcription by escherichia coli camp receptor protein. *Proc Natl Acad Sci U S A*, 90(7):3083–7, 1993.

[74] B. S. Burz, R. Rivera-Pomar, H. Jackle, and S. D. Hanes. Cooperative DNA-binding by bicoid provides a mechanism for threshold-dependent gene activation in the drosophila embryo. *EMBO J*, 17:5998–6009, 1998.

[75] T. S. Karpova, M. J. Kim, C. Spriet, K. Nalley, T. J. Stasevich, Z. Kherrouche, L. Heliot, and J. G. Mcnally. Concurrent fast and slow cycling of a transcriptional activator at an endogenous promoter. *Science*, 5862:466–469, 2008.

[76] M. Shin, S. Kang, S.-J. Hyun, N. Fujita, A. Ishishama, P. Valentin-Hansen, and H. E. Choy. Repression of deop2 in escherichia coli by cytr: Conversion of a transcription activator into a repressor. *EMBO J*, 19:5392–5399, 2001.

[77] E. Segal and J. Widom. From DNA sequence to transcriptional behaviour: A quantitative approach. *Nat Rev Genet*, 10(7):443–56, 2009.

[78] H. D. Kim, T. Shay, E. K. O'shea, and A. Regev. Transcriptional regulatory circuits: Predicting numbers from alphabets. *Science*, 325:429–432, 2009.

[79] T. Kuhlman, Z. Zhang, M. H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *escherichia coli*. *Proc Natl Acad Sci U S A*, 104(14):6043–8, 2007.

[80] J. Gertz, E. D. Siggia, and B. A. Cohen. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–8, 2009.

[81] T. Raveh-Sadka, M. Levo, and E. Segal. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res*, 19:1480–1496, 2009.

[82] V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci U S A*, 105(45):17256–61, 2008.

[83] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A*, 99:12795., 2002.

[84] J. Ou, C. Furusawa, T. Yomo, and H. Shimizu. Analysis of stochasticity in promoter activation by using a dual-fluorescence reporter system. *Biosystems*, 97:160–164, 2009.

[85] P. J. Schlax, M. W. Capp, and M. T. J. Record. Inhibition of transcription initiation by lac repressor. *J Mol Biol*, 245(4):331–50, 1995.

[86] L. Saiz and J. M. Vilar. Stochastic dynamics of macromolecular-assembly networks. *Mol Syst Biol*, 2:2006 0024, 2006.

[87] T. P. Malan, A. Kolb, H. Buc, and W. R. Mcclure. Mechanism of crp-camp activation of lac operon transcription initiation activation of the p1 promoter. *J Mol Biol*, 180(4):881–909, 1984.

[88] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer. Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res*, 38:D750–3, 2010.

[89] S. B. Straney and D. M. Crothers. Lac repressor is a transient gene-activating protein. *Cell*, 51(5):699–707, 1987.

[90] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying *e. Coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329:533, 2010.

[91] P. S. Gutierrez, D. Monteoliva, and L. Diambra. Role of cooperative binding on noise expression. *Phys Rev E*, 80:011914, 2009.

[92] D. Muller and J. Stelling. Precise regulation of gene expression dynamics favors complex promoter architectures. *PLoS Comput Biol*, 5:e1000279, 2009.

[93] G. Li, M. Levitus, C. Bustamante, and J. Widom. Rapid spontaneous accessibility of nucleosomal DNA. *Nat Struct Mol Biol*, 12(1):46–53, 2005.

[94] A. Gansen, A. Valeri, F. Hauger, S. Felekyan, S. Kalinin, K. Toth, J. Langowski, and C. A. Seidel. Nucleosome disassembly intermediates characterized by single-molecule fret. *Proc Natl Acad Sci U S A*, 106:15308–13, 2009.

[95] S. Klumpp and T. Hwa. Stochasticity and traffic jams in the transcription of ribosomal rna: Intriguing role of termination and antitermination. *Proc Natl Acad Sci U S A*, 105:18159–64, 2008.

[96] M. Voliotis, N. Cohen, C. Molina-Paris, and T. B. Liverpool. Fluctuations, pauses and backtracking in DNA transcription. *Biophys J*, 94:334–48, 2008.

[97] M. Dobrzynski and F. Bruggeman. Elongation dynamics shape bursty transcription and translation. *Proc Natl Acad Sci U S A*, 106:2583–8, 2009.

[98] G. Tkacik, T. Gregor, and W. Bialek. The role of input noise in transcriptional regulation. *PLoS One*, 3(7):e2774, 2008.

Part II

Experimental Dissection of Gene Regulatory Motifs

Chapter 6

Comparison and Calibration of Different Reporters for Quantitative Analysis of Gene Expression

This chapter is a reproduction of reference [1].

Absolute levels of gene expression in bacteria are observed to vary over as much as six orders of magnitude. Thermodynamic models have been proposed as a tool to describe these expression levels as a function of the number and interaction energies of the transcription factors involved in regulating a given transcriptional circuit. In this context, it is essential to understand both the limitations and linear range of the different methods for measuring gene expression and to determine to what extent measurements from different reporters can be directly compared, with one aim being the stringent testing of theoretical descriptions of gene expression. In this paper we compare two protein reporters by measuring both the absolute level of expression and fold-change in expression using the fluorescent protein EYFP and the enzymatic reporter β-galactosidase. We determine their dynamic and linear range and show that they are interchangeable over four orders of magnitude. By calibrating these reporters such that they can be interpreted in terms of absolute molecular counts, we establish limits for their applicability: autofluorescence on the lower end of expression for EYFP (at about 10 molecules per cell) and interference with cellular growth on the high end for β-galactosidase (at about 20,000 molecules per cell). These qualities make the reporters complementary and necessary when trying to experimentally verify the predictions from the theoretical models.

6.1 Introduction

In recent years our understanding of transcriptional regulation has increased dramatically. This is true both in terms of the number of regulatory circuits that have been dissected and of the precision with which they have been characterized [2–8]. As illustrated in figure 6.1, quantitative measurements of gene expression have determined that the mean absolute level of expression of different promoters range over more than six orders of magnitude. The majority of gene products regulated under bacterial and viral promoters are

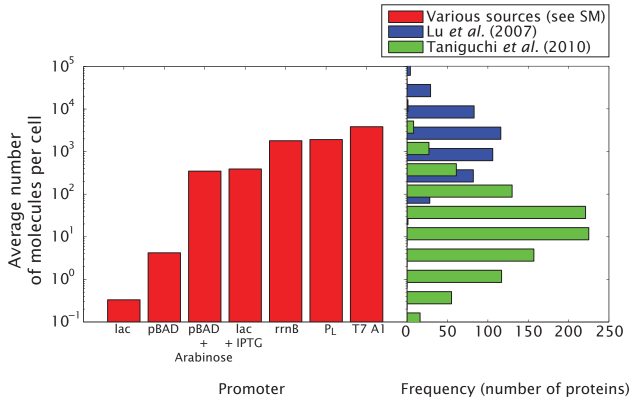


Figure 6.1: Gene expression levels in *E. coli*. The estimated absolute expression level of several bacterial promoters and a strong viral promoter obtained from the literature are shown in red (see Supporting Materials and table 6.4 for the corresponding references and assumptions made in order to determine the level of expression). For comparison the results from two recent cell censuses of *E. coli* are also shown as histograms of the number of proteins [8, 9]. Note that the range of expression spans over about six orders of magnitude for a given set of measurements illustrating the wide dynamical range associated with bacterial promoters. The discrepancy between the two cell census of *E. coli* are further explored in figure 6.15.

present at levels from 0.1 to 10^5 molecules per cell.

In addition to being concerned with these absolute levels of expression, it is also of interest to know the range over which these promoters can be regulated. As shown in figure 6.2(C,D) the level of expression of a given promoter can in turn be regulated to vary over several more orders of magnitude. These results make it clear that a quantitative genome-wide characterization of transcriptional regulation requires techniques with a broad dynamic range and for which the experimental uncertainties have been precisely characterized.

Quantitative experiments like those described above are making it possible to directly compare measurements of regulatory response with the predictions of an increasingly sophisticated host of theoretical ideas for describing regulatory circuits [10, 12–20]. This poses an experimental challenge: is there a technique or techniques that can reliably span this range of expression?

There are a wide variety of different methods for carrying out measurements of gene expression like those described above [21–24]. One classic scheme for measuring the level of gene expression is based upon the enzyme action of β-galactosidase (LacZ) as a reporter in which a substrate for this enzyme can be detected colorimetrically upon cleavage [21]. However, the use of fluorescent reporters is increasingly becoming the method of choice, especially with the construction of a variety of libraries in which nearly each and every gene in a model organism can be read out fluorescently [2, 4, 8]. In certain cases, this idea has been pushed all the way to the single-molecule limit where individual molecules in regulatory circuits are detected through their fluorescence [25]. mRNA counting, both in bulk through quantitative PCR and at the single molecule level, is also becoming widespread as a means for quantifying levels of gene expression [8, 26]. In the cases where antibodies against the protein of interest are available, Western or immunoblotting can provide a quantitative measure of the protein contents of a cell [27]. Finally, another popular enzymatic technique is

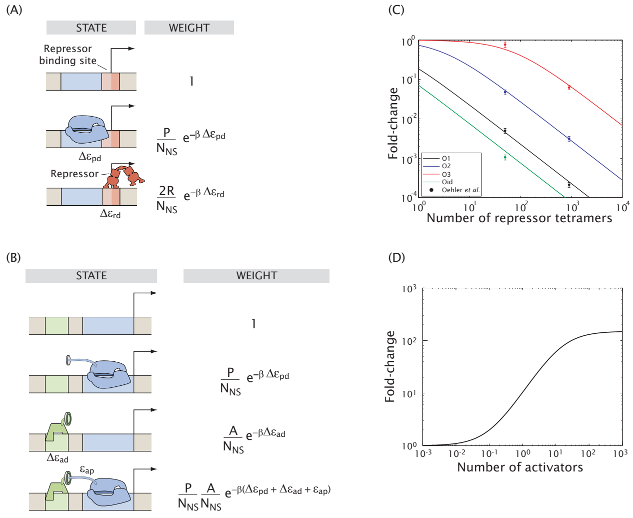


Figure 6.2: Fold-change of different regulatory motifs. The states and weights from the thermodynamic models are shown for the case of (A) simple repression by Lac repressor and (B) simple activation by CRP. The corresponding fold-change in gene expression as a function of transcription factor concentration predicted by the model is shown in (C) for repression and (D) for activation. The fold-change values span over several orders of magnitude. Refer to [10] for a derivation of the respective formulas and their parameters which are characteristic of bacterial promoters. The data for Lac repressor in (A) has been taken from [11].

based on reporting gene expression levels through bioluminescence [22, 28, 29].

The quest to quantitatively dissect regulatory networks of all types [5, 6, 26, 30] raises questions about the relative merits of these different measurement techniques. When trying to compare the significance of these different measurements and to use them as the basis for the development of theoretical models, it is important to have some calibration which reveals their respective dynamic ranges and how they are related. For example, one important question is whether they are linearly related, thus rendering them able to report reliably on the level of expression. Additionally, it is important to determine whether the use of reporters affects cellular processes in any observable way. To that end, in this work, we use systematic experimentation in the context of a well-characterized regulatory network to compare enzymatic and fluorescent reporters as a measure of level of gene expression. Similarly, recent measurements have begun to systematically explore the relation between the amount of expressed protein and the level of mRNA [8, 26, 31] and the quantification of protein levels through Western or immunoblots [27, 32]. Luminescence has an advantage related to its low background levels [33]. Being an enzymatic reporter, it requires the addition of a substrate to the medium or the encoding of genes that can produce the substrate within the cell. To our knowledge, even though it has been established that a constant luminescence level per cell can be used to quantify the number of cells in culture with a very high dynamic range [28], only a very limited comparison of luminescence as a reporter for gene expression has been done with respect to another reporter [34]. The necessity for providing the cells with a substrate has certainly diminished its use with respect to the widespread fluorescent protein reporters. Additionally, there have been reports of the bioluminescence genes potentially affecting their own expression [35]. As such we did not address this technique in this work, though studies similar in spirit to those performed here would be useful.

Our aim was to compare enzymatic and fluorescence reporters for the same promoters in a way that spans the large absolute dynamic range found in bacterial and viral promoters. In analogy to previous work [8, 36] the main strategy consists in engineering a promoter regulated by Lac repressor into *E. coli* which we subsequently induce to a variety of different levels with constructs harboring either a fluorescent or an enzymatic reporter.

The theoretical models mentioned previously predict the fold-change in gene expression, namely, the change in gene expression due to the presence of a transcription factor measured with respect to the level of gene expression in the absence of the same transcription factor as a function of the concentrations and interaction energies of the different molecular players [10, 13, 15]. Contrasting such relative predictions with experimental data relies heavily on the linearity of the reporter used. As such, we require not only that reporters span a high dynamic range, but that they also be linear over the fold-change range of the theoretical predictions.

To implement this alternative strategy we constructed a variety of different realizations of the network in which the binding affinities for Lac repressor are varied in a way that leads to different fold-change levels that differ over several orders of magnitude. Using these schemes, we can explore the presumed linearity of response of the enzymatic assays and their fluorescent reporter protein counterparts.

In the following sections we show a comparison and absolute calibration of the two reporters both in terms of their absolute levels and the fold-change in gene expression. We show that they are interchangeable over several orders of magnitude of expression, but each method has a limited dynamic range either due to limitations of the reporter or to how the reporter acts on the cells. We conclude that they are both complementary and necessary if a systematic characterization of the predictions generated by thermodynamic models spanning over multiple orders of magnitude is to be achieved.

6.2 Materials and Methods

6.2.1 DNA constructs and strains

The construction of all plasmids and strains are described in detail in the Supporting Materials.

In short, plasmid pZS25O1+11, pZS25O2+11, pZS25O3+11 and pZS25Oid+11 have a *lacUV5* promoter controlling the expression of either a EYFP or LacZ reporter. Care was taken to delete the O2 binding site present in the wild-type *lacZ* coding region [37]. These plasmids are shown schematically in figure 6.13.

A construct bearing the same antibiotic resistance, but no reporter, was created by deleting YFP from one of our previous constructs. This construct serves for determining the cell autofluorescence (for fluorescence measurements) and spontaneous hydrolysis (for enzymatic measurements).

Plasmid pZS3*1-lacI expresses Lac repressor off of a $p_{\mathrm{LtetO\text{-}1}}$ promoter [38]. The ribosomal binding sequence of this plasmid was weakened by mutating it (Alon Zaslaver, personal communication) resulting in pZS3*1BRS1-lacI.

Plasmid pLAU53-NoLacI-TetR-YFP is a derivative from plasmid pLAU53 ([39], kindly provided by Paul Wiggins). It expresses a fusion of TetR to the YFP gene used in this work under the control off the arabinose inducible promoter p_{BAD}.

The *E. coli* strains used in this experiment are shown in table 7.3. Chromosomal deletions were generated using the protocol developed by Datsenko and Wanner [40].

Chromosomal integrations were performed using recombineering [41]. Primers used for these integrations are shown in table 7.2. The reporter constructs were integrated into the *galK* region [42] of strain HG105 (lacI-) using primers HG6.1 and HG6.3. Strain HG205 (lacI++) was created by integrating pZS3*1RBS1-lacI into the phage-associated protein *ybcN* [43] using primers HG11.1 and HG11.3.

Integrations of the reporters were moved from strain HG105 (lacI-) to strains HG104 (lacI+) and HG205 (lacI++) using P1 transduction (openwetware.org/wiki/Sauer:P1vir_phage_transduction). All integrations and transductions were confirmed by PCR amplification of the replaced chromosomal region and by sequencing.

For YFP measurements of the fold-change in gene expression, strains MG1655 and TK140 [5] were used. MG1655 is wild-type *E. coli* encoding the *lac* operon and wild-type levels of Lac repressor. TK140 has a deletion of the *lacI* gene. Unlike strains HG104, HG105 and HG205, these two strains have the *lac* operon, which will result in significant levels of β-galactosidase for strain TK140. As a result, strains MG1655 and TK140 can only be used for fluorescence measurements. Constructs bearing LacZ as a reporter were integrated into strains HG104, HG105 and HG205.

The region of plasmid pLAU53-NoLacI-TetR-YFP covering the p_{BAD} promoter, TetR-YFP and the ampicilin resistance gene was integrated into the *galK* locus of strain H104 using primers HG22.04 and HG22.05 and transduced to strain 563 (kindly provided by Paul Wiggins) to create strain 563::TetR-YFP. This strain has both Tet and Lac repressors binding arrays located at the *lac* operon and near *oriC*, respectively.

All sequences, plasmids and strains are available upon request.

6.2.2 Growth conditions and gene expression measurements

Strains to be assayed for gene expression were grown overnight in 5 ml LB plus 30 μg/ml of kanamycin at 37 °C and 300 RPM shaking. The cells were then diluted 1000- to 4000-fold into 4 ml of M9 minimal medium + 0.5% glucose in triplicate culture tubes. Kanamycin was only added at this step for the strains bearing plasmids. The inducer IPTG was also added at this stage if necessary. These cells were grown for 6 to 9 hours until an OD600 of approximately 0.3 was reached after which they were once again diluted 1:10 and grown for 3 more hours to an OD600 of 0.3 for a total of more than 10 cell divisions. At this point, cells were harvested and their level of gene expression measured.

Induction and single cell microscopy was performed on the YFP samples as described in the Supporting Materials. Our protocol for measuring LacZ activity is basically a slightly modified version of the one described in [21, 44]. Details are given in the Supporting Materials.

6.2.3 *In vivo* YFP calibration

Strain 563::TetR-YFP was grown as described in [45], but in the absence of any inducers. In order to reduce the autofluorescence coming from the background buffer/media, we "sandwiched" a small volume of the the cells between two cover glasses corresponding to approximately 25 mm × 25 mm × 1 μm. We used low autofluorescence (Corning D263) coverglass and imaged using a 473 nm laser in epifluorescence and a EM-CCD Andro iXon camera. The glass was cleaned using an RCA wash [46]. Such a reduction in the background was necessary to get an acceptable signal-to-noise ratio. The fluorescence of bright spots attributed the EYFP-Tet repressor fusion bound to the DNA was tracked over multiple frames using a customized version of the the Matlab code "PolyParticleTracker" [47]. The data was analyzed using custom Matlab code. Representative traces and images are shown in figure 6.3(A–D). The resulting distribution is shown in figure 6.3(E).

6.3 Results

In the following sections we show a strategy for obtaining an absolute calibration of our two protein reporters. We then compare these reporters side-by-side and determine their ranges of applicability. Finally, we take these experiments one step further by characterizing the fold-change in gene expression measured with both reporters for a simple transcriptional network. This final analysis allows us to determine a range over which thermodynamic models of gene regulation can be tested using this approach.

6.3.1 Absolute calibration of the reporters

Absolute measurements of gene expression are often reported in arbitrary units, especially for fluorescence measurements where the signal depends on the particular details of the microscope used. Such a quantification of fluorescence makes it hard, if not impossible, to compare results between setups and establish unified standards. On the other hand, having a simple way to turn these arbitrary units into an absolute number of molecules would be helpful both in the context of taking the census of cellular proteins [2, 8, 48] and also in the context of characterizing the limits of each reporter.

In the following sections we obtain an absolute calibration for both the enzymatic and fluorescent reporters characterized throughout this work. In turn, this calibration will allow us to set absolute bounds on the interchangeability of these reporters as well as their effectiveness as reporters of the level of gene expression.

6.3.1.1 Calibration of the absolute number of EYFP molecules

Several previous experiments have performed absolute calibrations of fluorescence levels by looking at a bulk solution of purified fluorophore in buffer [7, 8, 49, 50] or in cell extract [51]. These approaches require a known volume of illumination which can be achieved, for example, by performing either confocal microscopy [49, 50] or two-photon microscopy [7].

These methods should be considered in light of at least two caveats. First, they rely on the extinction coefficient of the fluorophore to determine its concentration. However, the solution will be comprised of active and bleached fluorophores. Therefore the effective extinction coefficient of the solution will be an unknown combination of the extinction coefficients for the active and bleached fluorophores. Second, they are performed outside the cell. Even in the case of cell extract the local environment the protein sees might be different than that of the unperturbed cellular interior.

Counting fluorescent proteins inside the cell is, however, not straightforward. Because of the fast diffusion time of free fluorescent proteins in the cytoplasm [52] the signal of individual fluorescent molecules gets blurred over the cell on the time scale of tens of milliseconds. As a result the fluorescence per unit area of a single fluorophore in the cytoplasm becomes comparable to the cell autofluorescence and, hence, not detectable under common continuous illumination conditions. A way to circumvent this is by immobilizing the fluorophore. For example, membrane proteins fused to fluorescent reporters present a much slower diffusion on the membrane than that of proteins in the cytoplasm. Single fluorophores can be then imaged in this way [8, 53, 54].

Our main approach for calibrating the fluorescence of a single EYFP molecule consists in immobilizing EYFP molecules *in vivo* by fusing them to a transcription factor which is in turn strongly bound to the genomic DNA of *E. coli.* Though this method has the advantage of being *in vivo*, one caveat is that in this case we are not imaging free cytoplasmic EYFP like in the gene expression measurements in this work. The fact that EYFP is fused to another protein that is in turn bound to the DNA could result in a difference of fluorescence.

Puncta of EYFP fused to Tet repressor could be observed inside the cells despite the poor signal-to-noise ratio of about 1.5. In some cases these puncta could be observed to disappear in a single step as shown in the trace in figure 6.3(A,B). We associate this with the photobleaching step of a single EYFP molecule. More often though the puncta would correspond to multiple EYFP molecules. These fluorescence traces manifest multiple discrete levels as shown in figure 6.3(C,D). By integrating the fluorescence of the steps over a small area we can obtain the distributions of steps shown in figure 6.3(E). Please, refer to the Supporting Materials for a detailed discussion of the data analysis process.

We compared the fluorescence per EYFP molecule to the total fluorescence coming from a particular strain, HG105::galK<>25O2+11-YFP, under the same conditions. This strain expresses cytoplasmic EYFP. As a result we estimate the number of EYFP molecules in this strain to be $2,600 \pm 600$. The reasoning behind choosing a strain where we directly measure the number of EYFP molecules is that all further gene

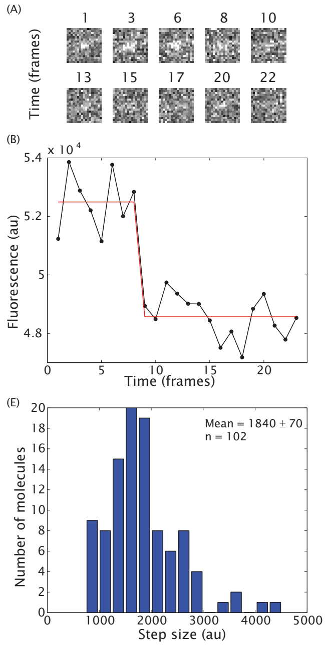


Figure 6.3: Absolute *in vivo* fluorescence calibration. Representative fluorescence snapshots (A) and their corresponding fluorescence traces (B) for a single bleaching event of the EYFP-Tet repressor fusion-bound to the genomic DNA. (C,D) Snapshots and fluorescent traces for multiple bleaching events of the EYFP-Tet repressor fusion. The red lines correspond to a least-squares fit to a single- or multiple-step function. (E) Distribution of fluorescence of bleaching steps for the *in vivo* sample.

expression measurements with EYFP as a reporter will be measured with respect to this "reference" strain. In this way we can easily estimate the number of EYFP molecules in any other strain we measure. Finally, we also quantified the fluorescence of single purified EYFP molecules and obtained a consistent result within 15 %. Please, refer to the Supporting Materials for details of this single molecule fluorescence quantification.

As a sanity check on these results, we estimate the expected number of EYFP molecules. The average number of EYFP molecules in steady state can be approximated by

$$\langle \text{EYFP} \rangle = \frac{\alpha \times b}{\beta}, \tag{6.1}$$

where α is the mRNA production rate, b is the number of proteins made per mRNA molecule, and β is the protein decay rate [55]. Because of the long lifetime of EYFP, the "decay rate" is actually nothing more than the cell doubling time since each cell division effectively halves the number of proteins. For the experiments considered here, we have a cell division time of around 1 hour. The number of proteins per mRNA has been measured for the *lac* operon to be about 20 protein molecules per mRNA molecule [56]. Therefore, we take $b = 20$ proteins/mRNA. This number is within the range of the various protein/mRNA measurements performed by [8]. Finally, the transcription rate for the fully induced *lac* operon has been reported to be between 1 min^{-1} and 20 min^{-1} [57, 58]. However, the lacUV5 promoter is about 30% weaker than the wild type *lac* promoter [59] resulting in a range of $\alpha = 0.7 - 14$ min^{-1}. When combining the decay rate β, the translation rate b and the transcription rate α we obtain an expected number of EYFP molecules of 1200 to 20000 per cell, a range comparable to our measurement.

6.3.1.2 Calibration of the absolute number of LacZ molecules

A simplified version of the reaction describing the breakdown of ONPG into ONP by β-galactosidase is given by

$$\mathrm{ONPG} + \mathrm{LacZ} \xrightarrow{k} \mathrm{ONP} + \mathrm{LacZ}. \tag{6.2}$$

From this reaction scheme we can derive the rate equation for the production of the yellow compound ONP which is given in turn by

$$\frac{d[\mathrm{ONP}]}{dt} = k[\mathrm{ONPG}][\mathrm{LacZ}] \tag{6.3}$$

and a rate equation for the decay in ONPG concentration due to its hydrolysis

$$\frac{d[\mathrm{ONPG}]}{dt} = -k[\mathrm{ONPG}][\mathrm{LacZ}]. \tag{6.4}$$

We wish to obtain the concentration of β-galactosidase, [LacZ], in our reaction in order to calculate the number of LacZ molecules per cell in the culture that was used to perform it.

If we assume that we have an excess concentration of ONPG and that the time of the reaction is short compared to $1/(k[\mathrm{LacZ}])$ we can neglect its depletion during the reaction. As a result we take [ONPG] as a constant in equation 6.3. The reaction described by equation 6.3 is the one we perform in the β-galactosidase assay to measure the amount of LacZ molecules per cell in Miller units (MU). In this assay we monitor the production of ONP over time given by the increase in absorbance at 420 nm of the solution. The standard definition of the Miller units [21] is

$$\mathrm{MU} = 1000\frac{\mathrm{OD}_{420} - 1.75 \times \mathrm{OD}_{550}}{t \times v \times \mathrm{OD}_{600}}, \tag{6.5}$$

where v is the volume of cells used in ml at a cell density given by OD_{600} and t is the reaction time in minutes. These Miller units were defined such that the fully induced wild-type *lac* operon has an activity of 1000 MU and such that its non-induced level would yield approximately 1 MU. We seek to relate these arbitrarily defined Miller units defined in equation 6.5 to equation 6.3 in order to obtain an actual number of LacZ molecules inside the cell.

First, the term $\mathrm{OD}_{420} - 1.75 \times \mathrm{OD}_{550}$ in equation 6.5 is a measure of the amount of ONP, the product of the breakdown of ONPG by β-galactosidase, in the reaction corrected for the cell debris (see Materials and Methods). We relate the absolute concentration of ONP in the reaction to the absorption reading through this term such that $\gamma[\mathrm{ONP}] = \mathrm{OD}_{420} - 1.75 \times \mathrm{OD}_{550}$. From an experimental point of view, the key assumption is that of a linear increase in the amount of ONP over time. Given that at the moment the experiment starts, there is as yet no ONP, we can obtain $d[\mathrm{ONP}]/dt$ simply by taking the accumulated ONP at time t and dividing by this elapsed time, that is,

$$\frac{d[\text{ONP}]}{dt} \approx \frac{[\text{ONP}]}{t}. \tag{6.6}$$

We also invoke a relation between the OD_{600} reading and the density of cells such that $\text{OD}_{600} \times v \times \delta = N_{cells}$, where N_{cells} is the number of cells. Finally, we wish to obtain the number of LacZ tetramers present in the reaction N_{LacZ} from this previous equation. This can be done by rewriting the concentration as $[\text{LacZ}] = N_{LacZ}/V$, where V is now the reaction volume of the standard Miller LacZ assay. If we insert this in our definition of MU we get

$$\text{MU} = 1000\,\gamma\, k[\text{ONPG}]\delta \frac{1}{V}\frac{N_{LacZ}}{N_{cells}}. \tag{6.7}$$

We determined δ for our strains to be $(8.9 \pm 0.8) \times 10^8$ /ml. The relation between ONP absorption at 420 nm and concentration is $0.0045/\mu\text{M}_{\text{ONP}}$ approximately [21, 60]. The volume of the reaction in the standard Miller assay is $V = 1.2$ ml. However, before the ONP reading the sample gets diluted to around 1.7 ml by the addition of Na_2CO_3. Therefore we define γ=$0.0045/\mu\text{M}_{\text{ONP}}$ × (1.7 ml/1.4 ml). Finally, we need to obtain the turnover rate of LacZ given by k. Wallenfels and Weil [61] report a turnover rate of $138 \times 10^6 \frac{\text{M}_{\text{ONP}}}{\text{min}\times\text{M}_{\text{LacZ}}\times\text{M}_{\text{ONPG}}}$, where we are referring to LacZ tetramers. Similar values have been reported by other authors [58, 62]. Since the initial concentration of ONPG in the reaction is 1.86 mM we get

$$\text{MU} \times \text{ml} \times \text{min} \times 0.5 \approx \frac{N_Z}{N_{cells}}. \tag{6.8}$$

Although this is only a rough estimate because of a lack of error bars associated with the reported values for the specific activity of LacZ, this gives us a direct connection between Miller units and number of LacZ molecules per cell.

This LacZ calibration that we have just calculated is consistent with previous experimental results on the *lac* operon. For example, the expression level of the repressed operon is about 0.6 MU [5]. Our calibration suggests that this corresponds to 0.3 LacZ tetramers/cell. Using single molecule techniques, the average number of LacZ tetramers under repressed conditions was estimated to be 1.2 tetramers/cell [56]. The internal consistency of these different estimates is encouraging.

6.3.2 Limits of LacZ and YFP as absolute reporters of gene expression

Recall that our aim was to compare enzymatic and fluorescence reporters for the same promoters in a way that spans the large dynamic range found in natural bacterial and viral promoters.

As we have already noted, the level of expression in such bacterial promoters span over six orders of magnitude as shown in figure 6.1 and table 6.4. Our interest was to design an experiment that would permit us to capture a similar dynamical range in a way that would result in a systematic comparison between the enzymatic and fluorescent reporters. To that end, we use an approach based on induction.

We use an inducible *lacUV5* promoter with a single binding site for Lac repressor (Oid) located directly downstream from its transcriptional start (see figure 6.13). Two versions of this construct regulating the expression of either the *lacZ* or EYFP genes were created. These constructs were either located on the bacterial chromosome or a low copy plasmid in strains that bear wild-type levels of Lac repressor (lacI+), high levels of Lac repressor (lacI++) or no Lac repressor (lacI-). By growing the different combinations of resulting strains at different concentrations of the inducer IPTG we were able to compare the total EYFP fluorescence and LacZ enzymatic activity per cell. These induction curves are shown in figure 6.14.

In figure 6.4 we present the corresponding expression levels measured using the two reporters over four orders of magnitude. For comparison, these results are juxtaposed with the literature expression levels of some naturally occurring promoters such as those presented in figure 6.1 and table 6.4. The blue line corresponds to a fit to a linear model showing that the data is consistent with a linear relation between the two reporters. This observation is consistent with recently published results [8]. The slope or conversion factor is $(9.6 \pm 0.7) \times 10^{-5}$ arbitrary fluorescence units/Miller units (MU). Even if we fit the relation between the two reporters with a more general functional form such as a power law, we find a linear dependence as shown in figure 6.16.

Although β-galactosidase activity is measured in absolute units, the fluorescence intensity depends strongly on details of the experimental apparatus used for the measurement such as the illumination intensity and transmission of the optical elements. The calibrations mentioned above that convert YFP arbitrary fluorescent units and LacZ Miller units into a number of molecules allows for our expression levels to be converted into an approximate absolute number of molecules of each reporter as shown by the labeling on the alternative axes in figure 6.4. We estimate the EYFP-LacZ relation to be around roughly 0.1 EYFP molecules/LacZ monomer. This value seems to be at odds with the fact that they are being expressed from the same promoter. If the transcription rate is the same because of this then that leaves some difference at the translation initiation and translation levels. However, we lack sufficient information to estimate those differences. Alternatively, an underestimation of the number of EYFP molecules inside the cells could be due to issues related to the fluorescence of the molecule itself such as quenching and misfolding [63].

Fluorescence measurements are fundamentally limited for low levels of gene expression. When the fluorescence signal becomes comparable to the autofluorescence level (below 10 molecules/cell) the determination of the level of gene expression has a high associated error. In contrast to free cytoplasmic fluorescent proteins, this limitation is less stringent in the case of fluorescent proteins that are immobilized on the cell membrane [53, 63] or DNA by a fusion (this work and [64]). The high error in the determination of low expression levels is reflected in the fluorescence distributions shown in figure 6.5. For the lowest expression levels, the dominant error comes from variations in the autofluorescence. For example, we observe a slight systematic bias towards overestimating the level of autofluorescence. Given the size of the autofluorescence variation, we do not regard the mean value of fluorescence as statistically significant. This limitation is indicated as a grey shaded area in figure 6.4. To give a sense of the scale, the expression level of the repressed wild-type *lac*

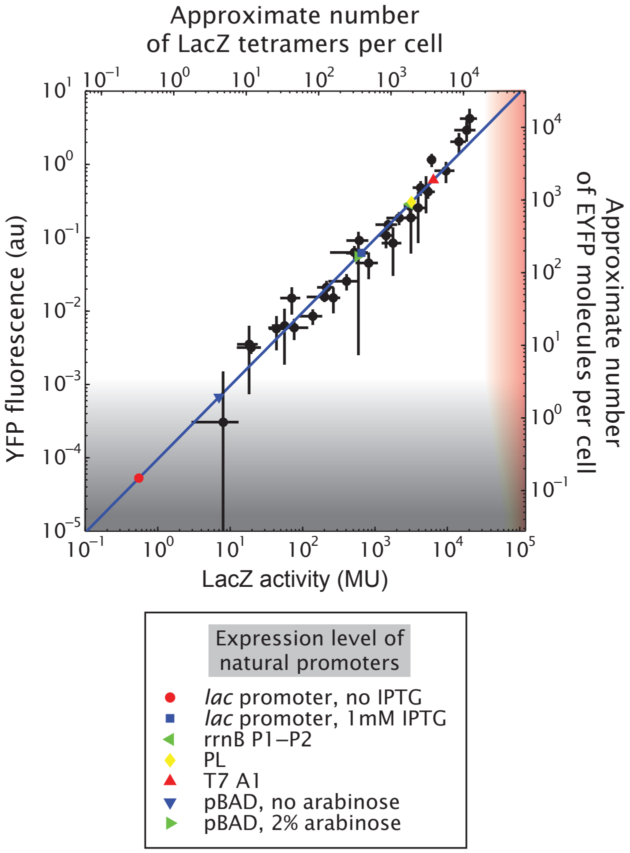


Figure 6.4: Relation between the mean cell fluorescence and β-galactosidase activity. The fluorescence per cell is plotted against the $\beta-$galactosidase activity. Each point corresponds to the same construct bearing either EYFP or *lacZ* as a reporter in the same strain background and at the same concentration of IPTG. The blue line is a linear fit fixing the intercept to zero with a slope of $(9.6 \pm 0.7) \times 10^{-5}$ fluorescence units/MU or an estimated 0.1 YFP molecules/LacZ monomer. The grey shaded area represents the range of YFP where the fluorescence signal is comparable to the cell autofluorescence (see discussion in the main text and figure 6.5). The red shaded area corresponds to the range where our assay can detect LacZ expression affecting cell growth (refer to the main text and to table 6.1). The expression values of several natural promoters, some of which are also shown in figure 6.1, are plotted on the blue line.

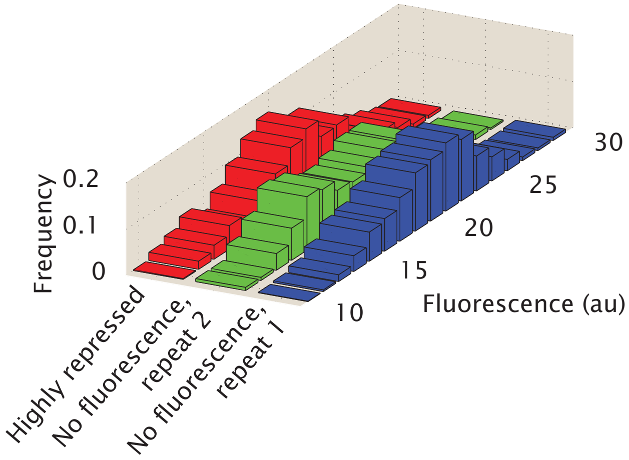


Figure 6.5: Reproducibility of low fluorescence levels. Histograms of the mean fluorescence per area in single cells corresponding to highly repressed samples and two repeats of the same non-fluorescent control are shown. The variation observed in these samples is comparable to the separation between non-fluorescent and low fluorescent distributions resulting in a considerable error in the estimation of the fluorescence of the sample.

promoter could not be measured with fluorescence unless a more sophisticated technique to visualize single fluorescent proteins is invoked [53]. By way of contrast, no significant analogous background was observed in any of our LacZ measurements, showing that this method is more reliable for quantifying very low levels of gene expression. In fact, linearity of the LacZ activity has been reported down to 0.03 MU [5].

When performing a measurement of gene expression using reporters it is important to demonstrate that the presence of the reporter itself is not affecting the state of the cell. We choose the growth rate as an indicator of the cellular state. For all the expression levels shown in figure 6.4 the doubling rate is approximately one hour regardless of the reporter. However, strain lacI- bearing a plasmid with LacZ as a reporter showed a longer doubling time of (74 ± 1) minutes, which was not the case for the corresponding EYFP strain. These growth rates are shown in table 6.1 and the corresponding growth curves are shown in figure 6.17. We confirm previous observations that expression levels above 20,000 LacZ tetramers/cell start affecting the cell significantly [65]. Unlike the low end of expression, where EYFP was limited by the autofluorescence, we find that for the high end of expression LacZ becomes limiting not because of signal issues, but because the cell is affected by the fact that high levels of LacZ are being expressed. Interestingly, even some of the stronger promoters such as rrnB and the T7 A1 promoter have levels below this threshold.

Though our measurements primarily focused on the use of microscopy to quantify EYFP fluorescence it is by no means the only option. An alternative is, for example, to use a plate reader. Though this method does not provide single cell information, it is able to produce data in much higher throughput than microscopy. On the other hand, plate readers will be more limited in terms of the minimum level of fluorescence they can quantify reliably. We perform a comparison between fluorescence measurements on the same strains using microscopy and a plate reader in the Supporting Materials leading to figure 6.12. We reach the conclusion

Strain	**Location**	**Reporter**	IPTG (μM)	Doubling time (min)
lacI++	Chromosome	No reporter	0	59 ± 1
lacI++	Plasmid	No reporter	0	57 ± 1
lacI-	Plasmid	EYFP	0	59 ± 2
lacI+	Plasmid	LacZ	1000	62 ± 1
lacI+	Plasmid	LacZ	0	59 ± 1
lacI-	Plasmid	LacZ	0	74 ± 1

Table 6.1: Effect of expression level on growth rate. Cells expressing a high level of EYFP (lacI-/Plasmid) have effectively the same doubling time as a strain without any reporter (lacI++/Chromosome/No reporter and lacI++/Plasmid/No reporter). However, the same is not true for high LacZ levels (lacI-/Plasmid), where the level of expression affects the doubling time in a measurable way.

that they are completely interchangeable, but that the lower limit of detection is now on the order of 50 molecules/cell, roughly 5 times more than with microscopy.

6.3.3 Limits of LacZ and YFP as reporters of the fold-change in gene expression

The fold-change in gene expression due to regulation by a transcription factor is defined as the level of gene expression in the presence of that molecule divided by the level of gene expression in its absence. In particular, it is the key magnitude predicted by thermodynamic models of transcriptional regulation [10, 15]. These models can predict fold-changes in gene expression that span over multiple orders of magnitude for both repression (fold-change< 1) and activation (fold-change> 1).

In order to test these models it is then necessary to be able to decide which reporter will be the best to assay a particular type of regulatory architecture. For example, in the previous section we found that we can reliably measure EYFP fluorescence down to 10 molecules/cell. If we are dealing with a promoter with a basal expression level of approximately $3,000$ YFP molecules/cell like the *lacUV5* promoter integrated on the chromosome used in this work, this means that the lowest fold-change we can measure with YFP is $10/3000 \approx 10^{-3}$. On the other hand, the maximum LacZ activity attainable before cell growth starts being compromised is around 20,000 LacZ tetramers/cell. This means that we can only *increase* the number of LacZ tetramers beyond the basal level up to this level before the cell senses the presence of these molecules as measured by its growth rate. Since the basal level of our promoter corresponds to $4,000$ LacZ tetramers/cell this translates into a maximum measurable fold-change of $20000/4000 \approx 10^1$ using LacZ as a reporter.

In order to test part of this assertion about the maximum fold-change in repression we performed fold-change measurements on constructs bearing the operators O1, O2 or Oid and the reporters LacZ or EYFP. Figure 6.6 shows the fold-change measured using EYFP as a function of the fold-change measured using LacZ for the different single binding site constructs (O1, O2 and Oid) in two different Lac repressor backgrounds: lacI+ and lacI++. We see that the fold-change levels measured with both reporters are in good correspondence. As expected, when the fold-change in gene expression reaches 10^{-3} the EYFP readings start becoming too noisy to determine the fold-change in gene expression reliably, setting a limit on the range of fold-change that can be measured using EYFP as a reporter.

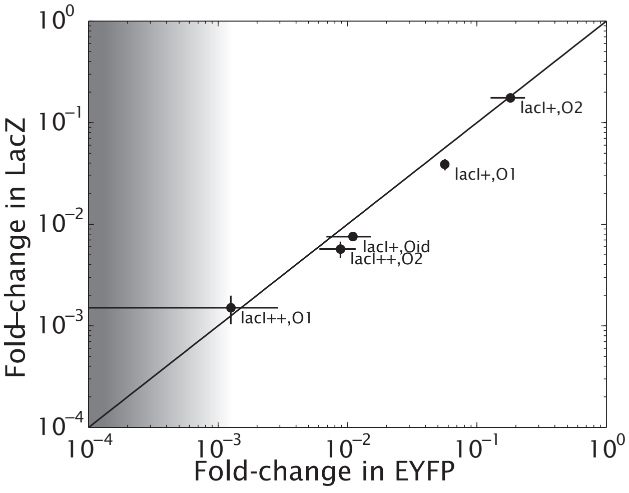


Figure 6.6: Fold-change in gene expression measured by LacZ and EYFP. The fold-change of a construct bearing a single Lac repressor binding site (Oid, O1 and O2) in the lacI+ and lacI++ backgrounds is compared when *lacZ* and EYFP are used as a reporter. The line has a slope of one. The point in the plot displaying the lowest fold-change corresponds to fluorescence levels that are near the detection limit. This results in the very large error bar shown.

6.4 Discussion

In this work we explored the feasibility of testing theoretical models of gene regulation using two reporters of protein expression: EYFP and LacZ. The calibration between EYFP and LacZ levels shown in this work is an important methodological prerequisite for testing quantitative models of gene expression. One important outcome is that it makes it possible to compare previously available data, generally taken using LacZ as a reporter, with single-cell expression data obtained using EYFP over most of the range of expression of bacterial promoters.

Fluorescent molecules have generally been the method of choice recently because they allow for live imaging of single cells. Our work establishes a clear absolute boundary for these measurements: the autofluorescence level. The intuitive expectation that autofluorescence will contaminate fluorescent gene expression measurements is converted into a concrete and precise numerical criterion. We expect this absolute boundary to be dependent on the particular fluorescent protein used, as they can vary widely in their spectral properties and as the autofluorescence is also measurably different at different wavelengths [66]. Interestingly, the enzymatic activity of LacZ shows no such limitation. However, for high levels of expression the presence of LacZ affects cell growth in a detectable way before any similar effect from EYFP can be detected. The experimental capacity to use both methods and to switch between one reporter and the other presented in this work makes it possible to obtain the best of both worlds: very low expression levels can be measured accurately in bulk using LacZ in absolute units whereas slightly higher levels of expression can be measured at the single cell level using fluorescence. Because of fundamental limitations associated with each reporter we conclude that both techniques need to be used together if the full range of absolute gene expression is to be measured. The outcome of this work has direct consequences on the fold-change in gene expression detectable with each reporter and, in turn, on the range of predictions that these measurements can be contrasted against.

6.5 Supporting Materials

6.5.1 Promoter activities

The promoter activities in figure 6.1 are also shown in table 6.4. These correspond to various measurements found in the literature and were obtained as follows. Some promoters such as *lac* [5] and *rrnB* [67] had expression levels directly reported in Miller units corresponding to single-copy chromosomal integrations. However, the data for the P_L and the T7 P_{A1} promoters was available in pBLA units rather than Miller units. This is measured by comparing the rate of mRNA synthesis of the promoter of interest with the rate of mRNA synthesis of the β-lactamase promoter [59]. Lanzer and Bujard [68] estimated the relation between pBLA and Miller units, to be around 5000 MU/pBLA units.

The measurements for the P_L, T7 P_{A1} [59] and pBAD [69] promoters were performed on plasmids bearing a ColE1 origin of replication. Earlier measurements have estimated this origin of replication to result in a plasmid copy number of approximately 60 per cell [38]. As a result, we estimate the level of expression of a single copy of each of the promoter on plasmids by dividing their expression by the plasmid copy number.

Finally, the levels of expression calculated in Miller units were converted to an absolute number of molecules using the absolute LacZ calibration described in the text of $0.6 \frac{\text{LacZ tetramers/cell}}{\text{MU}}$.

6.5.2 Supplementary materials and methods

6.5.2.1 Plasmids

Plasmid pZS22-YFP was kindly provided by Michael Elowitz. The EYFP gene comes from plasmid pDH5 (University of Washington Yeast Resource Center [3]). The main features of the pZ plasmids are located between unique restriction sites [38]. The sequence corresponding to the *lacUV5* promoter [70] between positions -36 and +21 was synthesized from DNA oligos and placed between the EcoRI and XhoI sites of pZS22-YFP in order to create pZS25O1+11-YFP. Note that we follow the notation of Lutz and Bujard [38] and assign the promoter number 5 to the *lacUV5* promoter. The O1 binding site in pZS25O1+11-YFP was changed to O2, O3 and to Oid using site-directed mutagenesis (Quikchange II, Stratagene), resulting in pZS25O2+11-YFP, pZS25O3+11-YFP and pZS25Oid+11-YFP. These plasmids are shown diagrammatically together with the promoter sequence in figure 6.13.

The *lacZ* gene was cloned from *E. coli* between the KpnI and HindIII sites of all the single-site constructs mentioned in the previous paragraph. The O2 binding site inside the *lacZ* coding region was deleted without changing the LacZ protein [71] using site-directed mutagenesis. Successful mutagenesis was confirmed by sequencing the new constructs around the mutagenized area.

After we had generated these constructs and integrated them on the *E. coli* chromosome (as described below) we determined that the different LacZ constructs had acquired some mutations. On average there were three different point mutations in each construct, though pZS25O3+11-lacZ had lost both the KpnI and HindIII sites. All these constructs still expressed functional LacZ. This problem did not present itself in the case of the EYFP constructs. We attribute this higher number of mutations in part to possible problems in the PCR amplification of the *lacZ* coding region. Another possible explanation is related to having a longer plasmid with the *lacZ* gene as opposed to the EYFP gene (3213 bp versus 714 bp). However, it must be noted that none of their EYFP counterparts had any mutations in the coding region, giving less strength to this argument since a simple estimate assuming the same proportion of mutations would have resulted in roughly 1/4 the mutations seen in the LacZ case.

Plasmid pZS21-lacI was kindly provided by Michael Elowitz. This plasmid has kanamycin resistance. The chloramphenicol resistance gene flanked by FLIP recombinase sites was obtained by PCR from plasmid pKD3 [40]. The insert was placed between the SacI and AatII sites of pZS21-lacI to generate pZS3*1-lacI. The ribosomal binding sequence of pZS3*1-lacI was weakened by performing the mutation AGAGGAGAAAGG → AGA**TTT**GAAAGG (Alon Zaslaver, personal communication) resulting in pZS3*1RBS1-lacI. Higher levels of Lac repressor with respect to wild-type can be confirmed by comparing the expression of a construct such as pZS25O1+11 in the two different backgrounds.

Plasmid pET11a-His-YFP was used for the EYFP over-expression and purification described below. His-YFP was created by PCR amplifying the EYFP gene from pZS25O1+11-YFP adding a His-tag at the N-terminus of EYFP and the restriction sites for NheI and BamHI (see table 7.2 for primer sequences). This insert was ligated in pET11a (New England Biolabs). The resulting plasmid was transformed into strain BL21(DE3).

Plasmid pLAU53-NoLacI-TetR-YFP was constructed from plasmid pLAU53 ([39], kindly provided by Paul Wiggins). Its Lac repressor-CFP fusion was deleted by making use of the EcoRI restriction sites flanking the coding sequence. The plasmid was digested and the relevant fragment gel purified and re-ligated to obtain pLAU53-NoLacI. Though this plasmid already has TetR-YFP fusion, the EYFP it contains differs from the EYFP used in this work in some key amino acids [72]. As a result we swapped the EYFP in the plasmid for the one used in our reporter constructs. The EYFP gene was amplified from plasmid pZS25O1+11-YFP using primers HG22.03 and HG25.11 (see table 7.2). These primers added flanking restriction sites for HindIII and XhoI. Plasmid pLAU530-NoLacI was digested with the same restriction enzymes and the relevant fragment gel purified and ligated with the PCR product to generate pLAU53-NoLacI-TetR-YFP.

6.5.2.2 EYFP purification

His-tagged YFP was expressed in *E. coli* BL21(DE3) cells harboring the pET11a-His-YFP expression plasmid and purified using Ni-NTA affinity chromatography (Qiagen) according to the manufacturer's protocol.

6.5.2.3 *In vitro* EYFP calibration

The purified protein was diluted by 10^5 to 10^6 in PBS. Single YFP molecules bound nonspecifically to low autofluorescence (Corning D263) coverglass were imaged using a 473 nm laser in epifluorescence. The rest of the imaging was done as described for the *in vivo* calibration in the main text.

6.5.2.4 Gene expression measurements

Three replicates of each strain were grown in different tubes in order to be able to obtain a mean expression level and its standard deviation. However, the day-to-day variation tended to be more significant than the variation from sample to sample on a given day. Therefore all data points shown in this experiment are the result of averaging over mean values obtained on at least three different days. The error bars are the standard deviation over these days.

The level of gene expression as a function of different IPTG concentrations was measured for strains HG104::galK<>25Oid+11, HG105::galK<>25Oid+11, HG205::galK<>25Oid+11, HG104 + pZS25Oid+11, HG105 + pZS25Oid+11 and HG205 + pZS25Oid+11 for EYFP and LacZ. For the repression measurements in EYFP strains HG104 was replaced by MG1655 and strain HG105 by TK140 as described in the main text.

The results of the induction measurements are shown in figure 6.14, where the levels of gene expression are normalized by their maximum. The levels of expression for both reporters were then combined to generate figures 6.4 and 6.16. The level of fluorescence was normalized by the mean fluorescence per cell of strain HG105::galK<>25O2+11-YFP, which became our fluorescence standard.

Single-cell microscopy

Cells bearing EYFP were imaged at 100x magnification. In order to check for uniformity of the epi-illumination field we first imaged 0.5 μm fluorescent beads (TetraSpeck, Invitrogen) resulting in a typical inhomogeneity throughout the field of view of less than 5%. The cells were immobilized between a number 1.5 coverglass and a pad of 1.5% low-melting-temperature agarose in PBS. Images of the cells in phase contrast and fluorescence were taken. The time between the initial placement of cells on the pad and the last picture taken was about five hours. No detectable difference in the level of gene expression that could be attributed to these five hours on the pad was observed.

We used automated microscopes to take 20 snapshots per strain. Fluorescence was quantified using either a Hamamatsu Orca-285 or a Roper Scientific CoolSnap camera. With every data acquisition we quantified the reference strain HG105::galK<>25O2+11-YFP. This allowed us to directly compare the result from the different microscopy setups.

The phase contrast images were used for automatic segmentation of the cells using custom Matlab code, based on code kindly provided by Michael Elowitz. The total fluorescence per cell was calculated by integrating the fluorescence per pixel over the cell area. The average fluorescence per cell was calculated by averaging the cell's intensity over the cell's area, as determined by the segmentation. Cell and pad autofluorescence were determined by looking at strains bearing the no reporter construct.

β-galactosidase assay

LacZ activity was measured by the classic colorimetric assay [21, 44] with some slight modifications as follows. A volume of the cells between 2.5 μl and 200 μl was added to Z-buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol, pH 7.0) for a total volume of 1 ml. The volume of cells was chosen such that the yellow color would develop in no less than 15 minutes. For the case of the no-reporter constructs 200 μl of cell culture was used. Additionally, we included a blank sample with 1 ml of Z-buffer. The whole assay was performed in 1.5 ml Eppendorf tubes.

In order to lyse the cells, 25 μl of 0.1% SDS and 50 μl of chloroform was added and the mixture was vortexed for 10 s. Finally, 200 μl of 4 mg/ml 2-Nitrophenyl β-D-galactopyranoside (ONPG) in Z-buffer were added to the solution and its color, related to the concentration of the product ONP, monitored visually. Once enough yellow developed in a tube the reaction was stopped by adding 200 μl of 2.5 M Na_2CO_3 instead of adding 500 μl of a 1 M solution as done in other protocols. At this point the tubes were spun down at $> 13,000$ g for three minutes in order to reduce the contribution of cell debris to the measurement.

200 μl of solution was read for OD420 and OD550 on a Tecan Safire2 and blanked using the Z-buffer sample. The OD600 of 200 μl of each culture was read with the same instrument. The absolute activity of LacZ was measured in Miller units using the formula

$$\mathrm{MU} = 1000\frac{\mathrm{OD}_{420} - 1.75 \times \mathrm{OD}_{550}}{t \times v \times \mathrm{OD}_{600}}0.826, \tag{6.9}$$

where t is the reaction time in minutes and v is the volume of cells used in ml. The factor of 0.826 is not present in the usual formula used to calculate Miller units. It is related to using 200 μl Na_2CO_3 as opposed to 500 μl. When using 500 μl, the final volume of the reaction is 1.725 ml (1ml Z-buffer, 25 μl 0.01% SDS, 200 l ONPG, 500 μl Na_2CO_3). However, when using only 200 μl of Na_2CO_3 the total volume is 1.425 ml. The factor of 0.826 adjusts for the difference in concentration of ONP.

All reactions were performed at room temperature. No significant difference in activity was observed with respect to performing the assay at 25C in an incubator.

Plate reader measurements

Cells were grown using the protocol described above in section "Gene expression measurements". When the culture reached an OD600 higher than 0.3 we loaded 200 μL of each culture onto a 96-well plate with a flat, clear bottom. The plate was measured in a Tecan Safire II. Fluorescence was measured from the top (height set manually to 5250 μm) with an excitation wavelength of 505 nm, and an emission wavelength of 535 nm, both with a bandwidth of 12 nm. The gain was automatically adjusted from the brightest well.

Absorbance at 600 nm was also measured. After subtracting the readings from a blank sample (media without cells) we calculated the fluorescence per absorbance unit. The cell autofluorescence was obtained by performing this measurement on a strain without a fluorescent reporter. Its fluorescence per absorbance unit was subtracted from all other samples. Finally, all resulting fluorescence values were normalized by the fluorescence of strain HG105::galK<>25O2+11-YFP.

6.5.3 *In vivo* and *in vitro* single molecule measurements

Fluorescent molecules were tracked using the Matlab code "PolyParticleTracker" [47]. This code finds local maxima in an image and follows these maxima over successive frames of the same field of view. When the molecule bleaches, however, there is nothing to track and the code stops quantifying the fluorescence from that region. Additionally, for some frames, the code's selection criteria (involving area, skewness, etc.) fail to continue tracking particles which were clearly there.

In order to circumvent these two issues we modified the code such that whenever the tracking of a particle was lost it would report its last known position. Additionally, we only kept traces where the particles had been found successfully in at least the first three initial frames. This particle finding scheme resulted in a significant amount of false positives per field of view because random fluctuations in the field of view would be transiently recognized as real particles. On average, we would track on the order of 100 puncta per field of view of which no more than 10 would correspond to real molecules. In order to distinguish the molecules from the vast number of false positives we manually screened all molecules using custom Matlab code. We confirmed that indeed there was a molecule at the position as shown by the snapshots in figure 6.3(A,C). We also checked that the position of the particle did not change by more than two to four pixels over the course of the analyzed trace.

The fluorescence of the selected molecules was then integrated over a box of a certain area. We present a detailed discussion regarding the choice of this parameter further below. For now, the following examples will be given using an area of 9 × 9 pixels, but the general conclusions are independent of this choice. Our traces consisted of 200 continuous frames with an exposure of 250 ms. However, in most cases we would only keep the fragment of the trace around the photobleaching steps as shown in figure 6.7(A). In figure 6.7(B) we present the tracking of the centroid and in figure 6.7(C) we show a set of snapshots over the selected time window.

The steps in the traces such as the one shown in figure 6.7(A) were fitted to a step or multiple step function using least-squares minimization. We manually called the number of steps within a trace and the position of the transitions as starting parameters for the fit. We compared this scheme to using an automated Hidden Markov Model (HMM) approach [73]. We analyze the complete traces corresponding to molecules we had manually selected using both our manual fit scheme and and HMM. For the HMM analysis we used vbFRET, a Matlab package [73]. A sample trace with both fitting schemes is shown in figure 6.8(A). Notice that in the case of HMM we do not constrain the time window for each trace. By doing this over our whole data set we can compare the distributions of steps obtained by the manual fit to the distribution of discrete levels found by HMM. This is shown in figure 6.8(B), where it is clear that both distributions are comparable. This gives us confidence that our manual fitting approach is not significantly different from other automated schemes.

One of the main problems in measuring low fluorescent signals *in vivo* is the contribution from the cell autofluorescence [66]. When taking a time lapse this cell autofluorescence will bleach, resulting in a time-varying background. An example of this effect is shown in figure 6.9(A), where we present the fluorescence per pixel as a function of time of a small box located inside and outside one of the cells. We see that there is a change of about 40 counts per pixel of the cell autofluorescence over the time trace. For a box with an area of $9 \times 9 = 81$ this corresponds to a total fluorescent signal of approximately 3,000 counts, which is comparable to the magnitude of the steps we are trying to detect. A way to reduce the contribution of this effect is to "pre-bleach" the field of view before actually taking data. Operationally, this can be done by keeping only photobleaching steps that occurred after a certain time. In figures 6.9(B,C) we show the effect of doing such filtering on the step size distribution. Notice that the distributions seem to converge once we only keep steps that occurred after 40 frames or later. As a result we apply this filter with a threshold of 40 frames for all our *in vivo* data.

In order to quantify both the *in vivo* and *in vitro* fluorescence of the single molecules and obtain traces such as those shown in figure 6.3(A,B) we integrated the signal over a box centered around their centroids. If the integration box has an area given by A and this area is bigger than the size of the diffraction limited spot corresponding to the molecule we are observing then the fluorescence before the photobleaching step is given by

$$Fluo_0 = YFP + Fluo_{back,0} \times A, \tag{6.10}$$

where YFP is the fluorescence of the YFP molecule and $Fluo_{back,0}$ is the fluorescence per unit area coming from the background. This last magnitude will be a combination of the camera offset and the autofluorescence coming from the glass, buffer and, in the *in vivo* case, of the cell autofluorescence. Notice that we gave the background fluorescence the subscript "0" to denote that there might be a time dependence. After the photobleaching step the fluorescence detected in that same area is

$$Fluo_f = Fluo_{back,f} \times A. \tag{6.11}$$

In this last equation we have defined the fluorescence of the background after the photobleaching event. Of course, if the background does not change over the time course of the experiment then the difference $Fluo_f - Fluo_0$ corresponds to YFP, the fluorescence coming from a single molecule. If the level of background of baseline fluorescence is slightly different between the initial and final time points we get

$$Fluo_0 - Fluo_f = YFP + (Fluo_{back,f} - Fluo_{back,0})A. \tag{6.12}$$

We worry that this effect could be present leading to a systematic error in the estimation of the fluorescence steps. This could be due to changes in background autofluorescence over time or to non-linearities in the detection of low fluorescence levels. This effect would manifest itself as a linear increase in step size with the area of integration. There are obvious limits to this equation. If the area is too small then it will not capture the total fluorescence of the diffraction limited spot. If the area is too large then it will go beyond the cell itself and the background will be significantly different. In figure 6.10 we show the mean step fluorescence as a function of the size of the integration box. As expected, a small area (of 3×3 pixels^2) results in a small step size with respect to the other areas. The remaining areas correspond to 5×5, 7×7 and 9×9 pixels^2. Given that a cell under our magnification conditions is about 8 to 9 pixels wide we view the area of 9×9 pixels^2 as the biggest box that can be fit within the cell. In the figure we show a fit to equation 6.12 and the mean resulting from averaging the over the three larger area boxes. Both the values calculated from the intercept of the linear fit and from averaging over the different data points give comparable results. We conclude that we cannot detect a difference between the background fluorescence before and after the bleaching step within our experimental error.

Interestingly, the *in vitro* values did not show a flat response of the step size as a function of the integration area in the same way than its *in vivo* counterpart did. In figure 6.11(A) we show the scaling of the *in vitro* step fluorescence with the area. Interestingly, we see a significant slope of (11.5 ± 0.4) au/pixel^2. If we are to believe the model in equation 6.12, such a slope would correspond to an underestimation of the value of the background before the bleaching step, $Fluo_{back,0}$. This shift can be clearly observed in the distributions as shown in figure 6.11(B). Additionally, it can be seen in the trace corresponding to a single molecule. This is shown in figure 6.11(C), where we shifted all the fluorescence levels after the bleaching step in order to show all the traces on the same plot. The fact that we see this effect when directly integrating the signal corresponding to a single trace without going through any automated analysis script suggests that this is not an artifact of our data analysis, but a true feature of the data.

We are unable to determine where this systematic error is coming from. This effect was also present when imaging the single molecules under Total Internal Reflection Fluorescence microscopy (TIRF), where there is a significant reduction of the background fluorescence coming from the buffer. Our main hypothesis is that it is due to a non-linearity in the acquisition process. However, we were unable to determine this unequivocally. Regardless of the origin of this systematic error if we assume that the measured step fluorescence as a function of the integration area follows a form such as the one shown in equation 6.12 we can use the intercept of the linear fit to account for the systematic shift and determine the real fluorescence per molecule. The intercept in the fit of figure 6.11(A) is 1470 ± 30 au/pixel^2. This implies that the fluorescence detected *in vitro* is 15% lower than its *in vivo* counterpart. If we were to take a box of a size comparable to the cell size of 9×9 pixel^2 we would estimate the *in vitro* fluorescence to be higher than the *in vivo* fluorescence by about 40%. Interestingly, this result is consistent with recent measurements comparing the *in vivo* single molecule and *in vitro* bulk fluorescence of the fluorescent protein Venus [8].

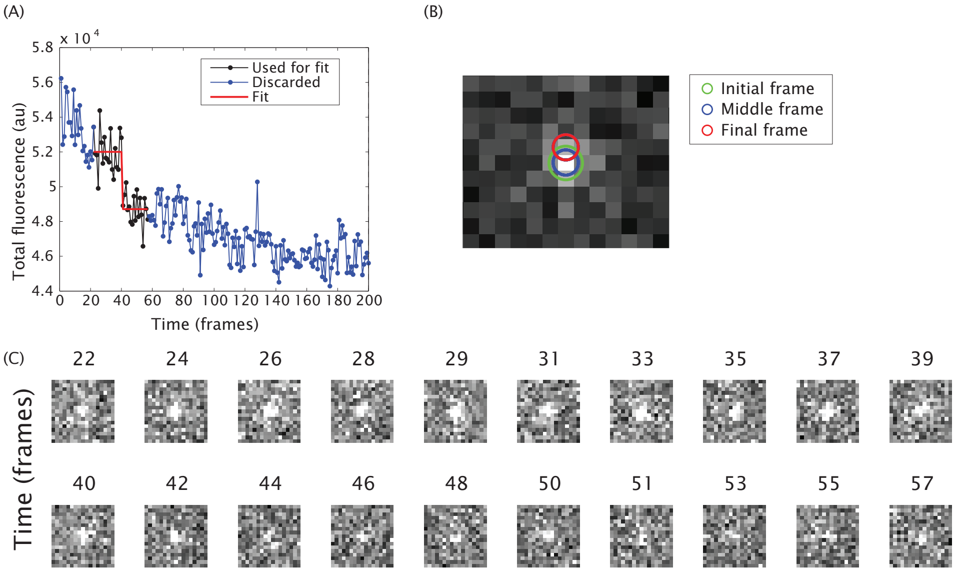


Figure 6.7: Selection scheme for single molecule traces. (A) A region of the trace presenting a discrete step or photobleaching is selected and fitted to a step function. (B) We confirm that the tracking of the molecule was successful by corroborating that there wasn't any significant movement of the particle over the selected time period. The image shown corresponds to the first selected frame and the circles show where the centroid of the particle was found as a function of time. (C) A window around the centroid at different time points is monitored to make sure there aren't any extraneous objects and that we do indeed have a diffraction limited spot within the integration area. The size of the window monitored is twice that of the integration window. In this example the window has a side of 17 pixels.

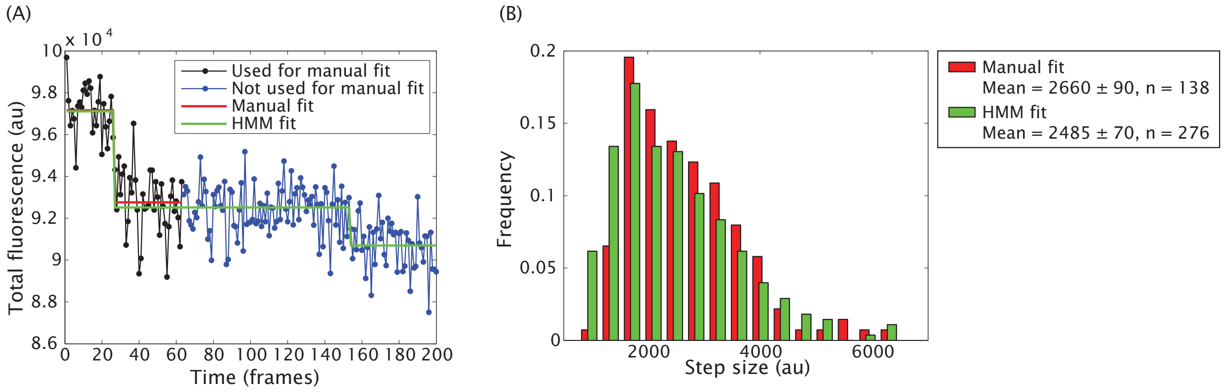


Figure 6.8: Comparison of different analysis schemes to obtain the fluorescence steps. (A) A single trace is shown where we fitted the step manually to a step function using least-squares and using a Hidden Markov Model approach (HMM). For the latter we don't constrain the fit to a particular time window. (B) Histogram of steps obtained manually compared to a histogram of the different levels found by HMM over our whole data set.

6.5.4 Plate reader vs. microscopy for determining fluorescent levels

One of the advantages of using a fluorescent reporter to measure gene expression is that, unlike using LacZ, no further reactions are needed. Once the cells have reached the desired point in their growth all that remains is to quantify their fluorescence level. However, even though it can be highly automated, microscopy remains

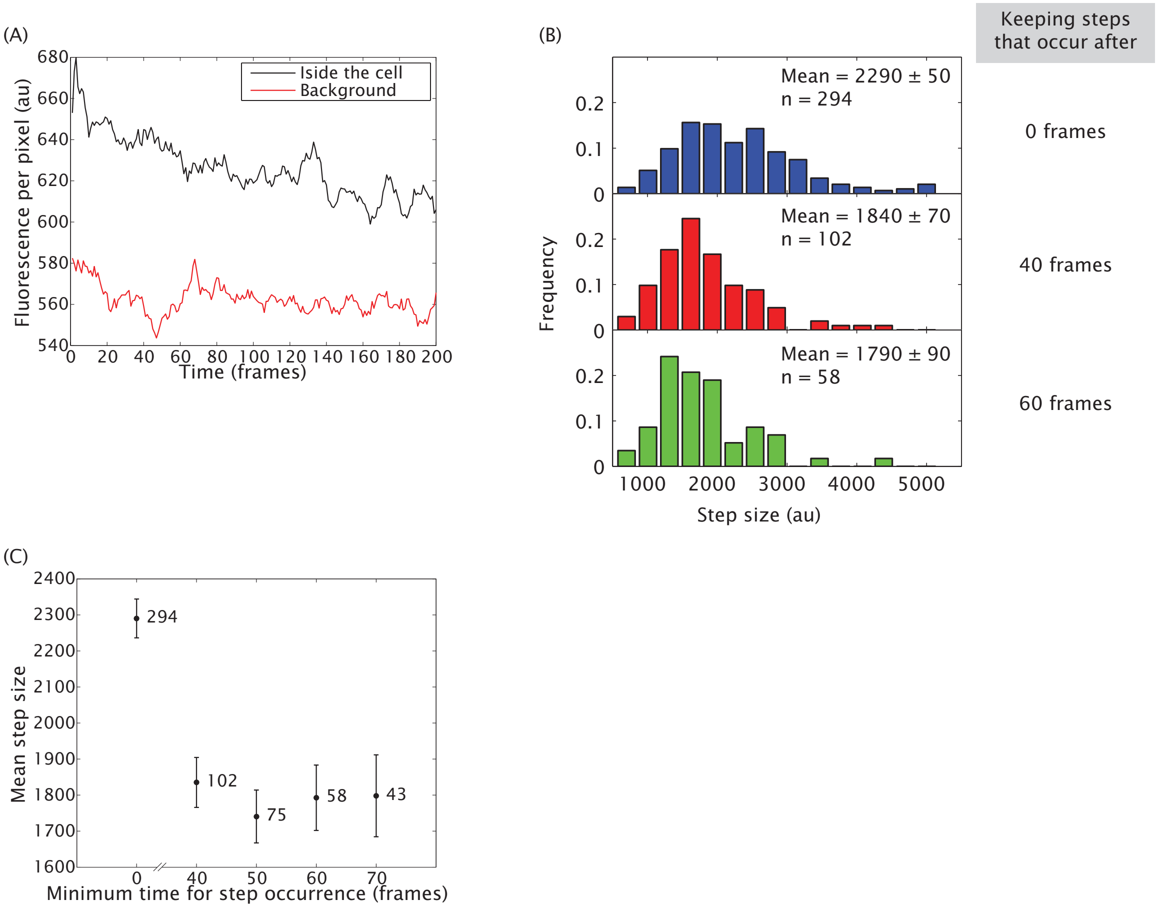


Figure 6.9: Effect of photobleaching on the *in vivo* calibration. (A) Fluorescence per pixel of a small region inside a cell without any fluorescent puncta and outside the cell. A moving average has been applied to smooth the traces. (B) Effect of keeping only steps that occurred after a certain time point on the step size distributions. (C) Mean step size as a function the minimum time of occurrence of steps. The error bars are the standard error of the means. The numbers next to the data points correspond to the number of steps analyzed.

a slow technique when many strains are to be assayed. A compromise is to quantify fluorescence using bulk methods such as a plate reader. Such a device can query the level of fluorescence of multiple strains much faster. However, there is a price to be paid in the form of dynamic range. Whereas with microscopy we could detect down to 10 molecules/cell, with the plate reader used for this work (Tecan Safire II, see Supplementary Methods) the minimum level of fluorescence that could be detected reliably corresponded to about 50 molecules/cell. In figure 6.12 we show a direct comparison of the two techniques. Here, it is clear that both of them give the same result as long as the signal is not close to the detection limit.

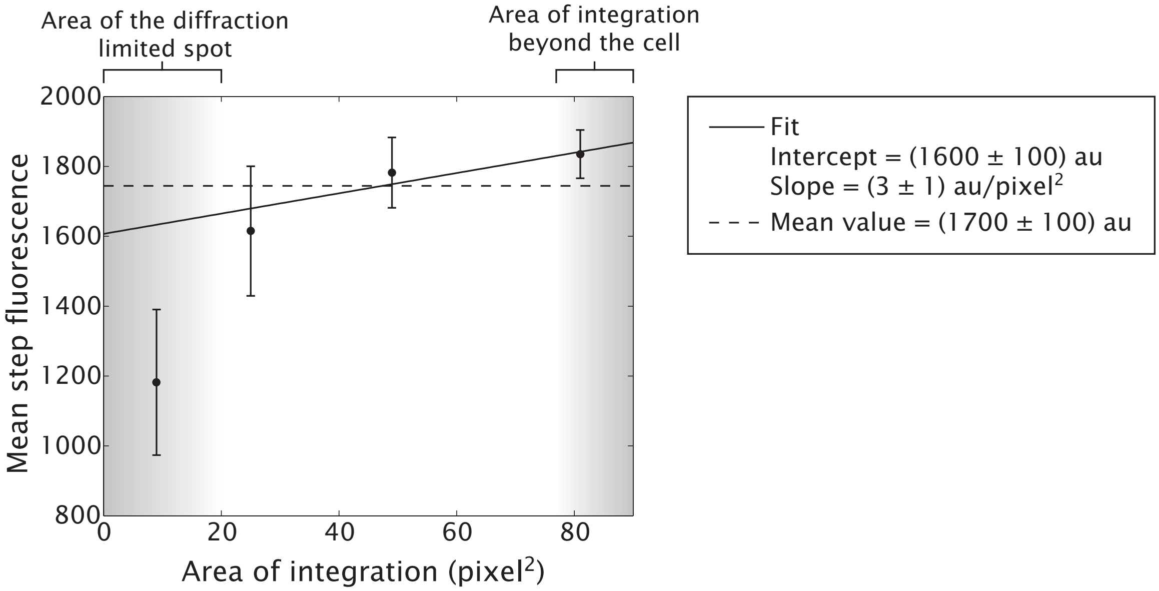


Figure 6.10: Scaling of the step size with the integration area for the *in vivo* calibration. The mean step size is shown as a function of the area of the integration box around the centroid of the molecule. The shaded regions mark the area sizes that are either too small to capture the fluorescence of the diffraction limited spot or too big such that the integration area would span beyond the cell. The linear fit corresponds to a fit to equation 6.12. This approach yields comparable values to just taking the average fluorescence step for any of the different area values.

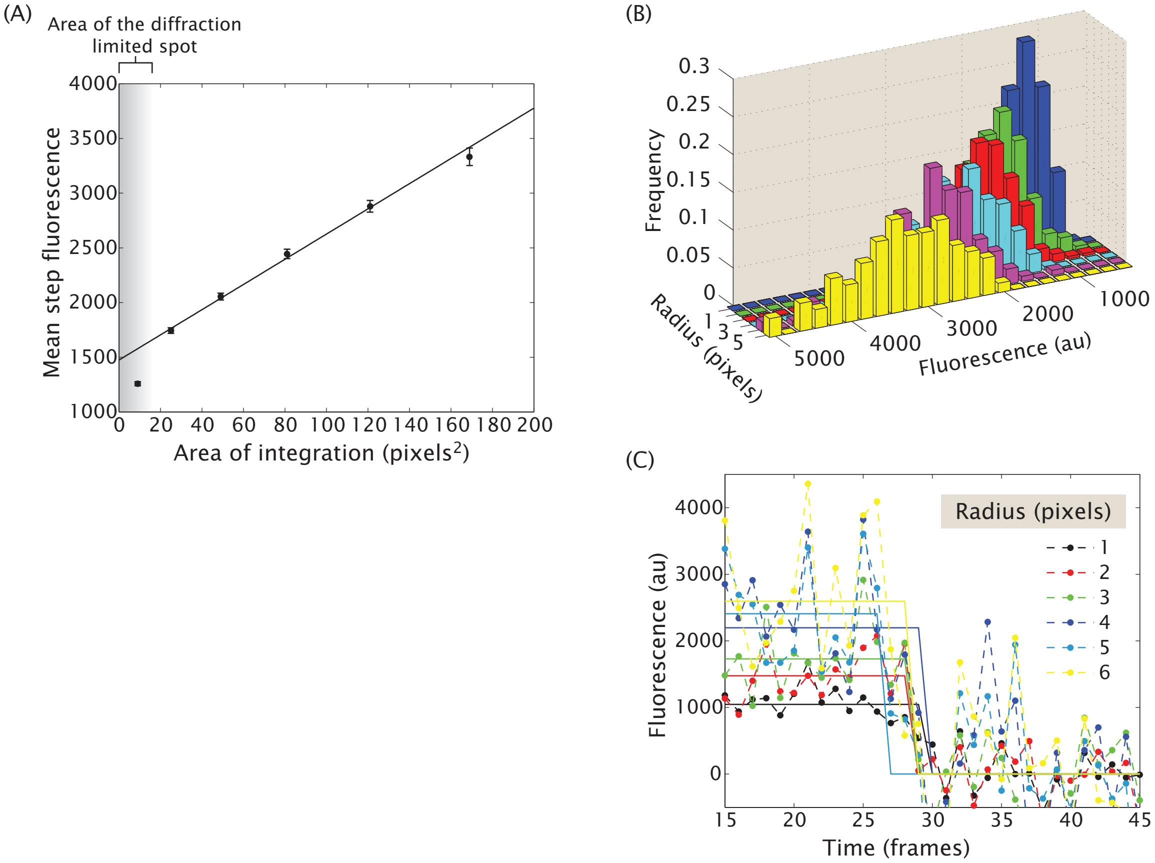


Figure 6.11: Scaling of the step size with the integration area for the *in vitro* calibration. (A) The mean step size is shown as a function of the area of the integration box. The fit corresponds to equation 6.12 with a slope of 11.5 ± 0.4 au/pixel^2 and an intercept of 1470 ± 30 au. (B) This difference in the mean step can be clearly observed at the distribution level. Here $\text{Area} = (2 \times \text{Radius} + 1)^2$. (C) Trace of fluorescence vs. time obtained for a single molecule for different choices of the size of the integration window. We have shifted the fluorescent level of all traces after photobleaching so that they would coincide for easier comparison.

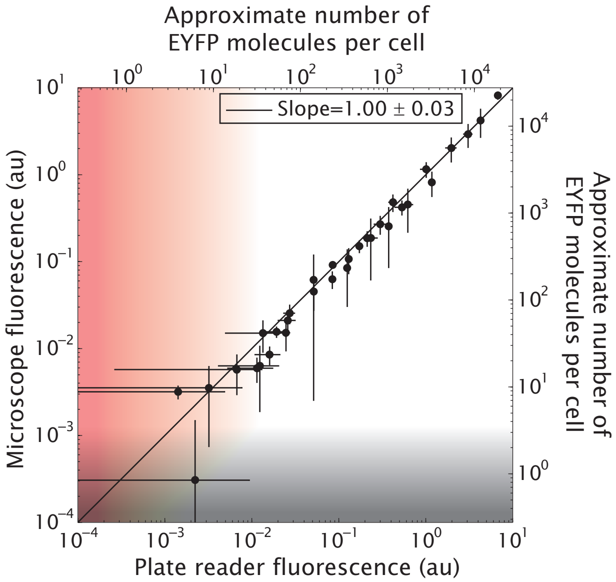


Figure 6.12: Comparison of microscopy and plate reader as methods to quantify gene expression. The same strains were quantified both using microscopy and using a plate reader (Tecan Safire II). The results show a 1:1 correlation between plate reader and microscopy, although the plate reader has a lower limit of detection which is greater than that of microscopy. Whereas using microscopy we can detect as few as roughly 10 EYFP molecules/cell, the plate reader can detect molecules only in excess of concentrations of 50 molecules/cell approximately. These detection limits are marked by the shaded regions.

6.5.5 Supplementary figures and tables

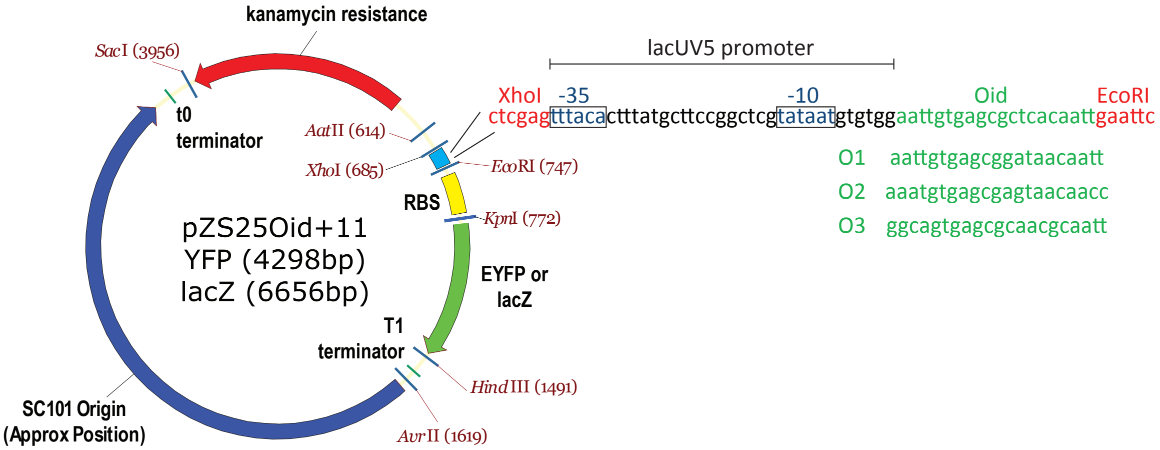


Figure 6.13: Plasmid diagram and promoter sequence. The main features of the plasmids pZS25O1+11-YFP and pZS25O1+11-lacZ are shown flanked by unique restriction sites. The particular promoter sequence based on the *lacUV5* promoter is shown together with the sequences of the different Lac repressor binding sites used.

Table 6.2: List of *E. coli* strains used throughout this experiment. Chromosomal positions correspond to the sequence in GenBank accession no. U00096.

Strain	Alternative name	Genotype	Derived from	Comment
HG104	lacI+	Δ*lacZYA*	MG1655	Deletion from 360,483 to 365,579
HG105	lacI-	Δ*lacZYA*, Δ*lacI*	MG1655	Deletion from 360,483 to 366,637
HG205	lacI++	Δ*lacZYA*, Δ*lacI*, ybcN<>3*1RBS1-lacI	HG105	
TK140		Δ*lacI*	MG1655	[5]

Table 6.3: Primers used throughout this work. For integration primers, lowercase indicates the portion of the primer that is homologous to the *E. coli* gene where the integration is made and uppercase indicates primer homology to the plasmid where PCR was carried out.

Primer	Sequence	Comment
HG6.1 HG6.3	gtttgcgcgcagtcagcgatatccattttcgcgaatccggagtg taagaaACTAGCAACACCAGAACAGCC ttcatattgttcagcgacagcttgctgtacggcaggcaccagct cttccgGGCTAATGCACCCAGTAAGG	Integration of the EYFP and *lacZ* reporter constructs into the *galK* gene.
HG11.1 HG11.3	acctctgcggaggggaagcgtgaacctctcacaagacggcatca aattacACTAGCAACACCAGAACAGCC ctgtagatgtgtccgttcatgacacgaataagcggtgtagccat tacgccGGCTAATGCACCCAGTAAGG	Integration of pZS3*1RBS1-lacI into the *ybcN* gene.
HG22.03 HG22.03R	attatagctagcatgggtcatcaccatcaccatcacggtcgtaa aggagaagaacttttcactgg tattaatggatccttatttgtatagttcatccatgccatgt	Make His-EYFP and insert into pET11a.
HG25.02 HG22.11	atattaaagcttatttgtatagttcatccatgccatg attatctcgagttggtgcgtaaaggagaagaacttttcactgg	Fuse TetR for EYFP in pLAU53-NoLacI
HG22.04 HG22.05	gtttgcgcgcagtcagcgatatccattttcgcgaatccggagt gtaagaaTTAATGCGCCGCTACAGGG ttcatattgttcagcgacagcttgctgtacggcaggcaccagc tcttccgTACTTTTCATACTCCCGCCATTCA	Integration of pLAU53-NoLacI-TetR-YFP into the *galK* gene.

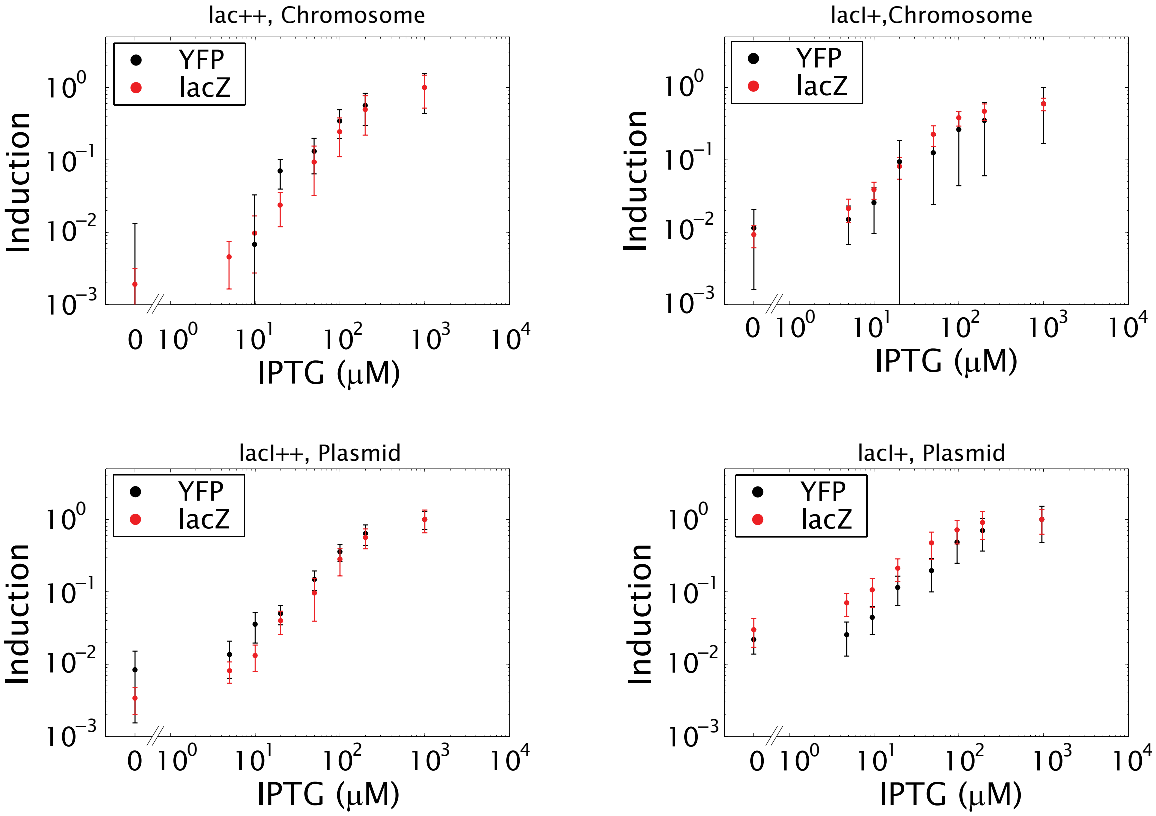


Figure 6.14: Induction curves used in this work. The level of gene expression of the EYFP and *lacZ* constructs is shown as a function of the IPTG concentration for the different construct locations (chromosome or low copy plasmid) and strain background. The level of expression is normalized by the corresponding maximum levels of activities. Error bars correspond to the standard deviation of measurements performed over at least four different days. Refer to the Materials and Methods for a description of the different strains.

Table 6.4: Promoter activities from the literature measured using the LacZ assay. These activities from the literature have been measured for a range of different promoters and conditions. The promoter strengths quoted here are often approximate and should therefore not be considered as accurate. Refer to "Promoter activities" in these Supplementary Materials for details of how these activities were calculated. (1) Measured in M9 + 0.5% glucose. (2) For a cell doubling of about 1.25/h. (3) Calibrating P_{bla} activity and LacZ units [68]. (4) Assuming a plasmid copy number of 60 copies/cell [38].

Promoter	**Strength (MU)**	**Reference**	**Comment**
lac promoter, no IPTG	0.5-0.6	[5]	(1)
lac promoter, 1 mM IPTG	600-700	[5]	(1)
rrnB P1-P2	3,000	[67]	(2)
P_L, no cI	3,200	[59]	(3,4)
T7 P_{A1}	6,400	[59]	(3,4)
pBAD, no arabinose	7	[69]	(4)
pBAD, 2% arabinose	580	[69]	(4)

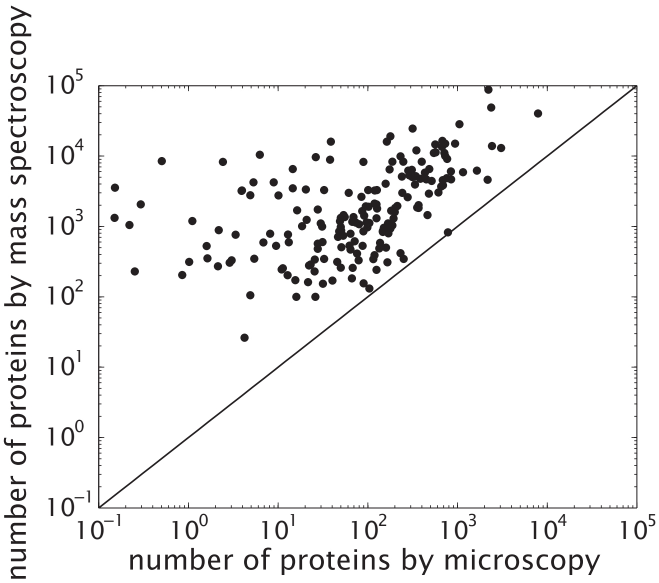


Figure 6.15: Comparison of different *E. coli* cell censuses. The number of proteins for a particular protein measured by mass spectroscopy [9] is compared to the same magnitude measured by fluorescence in single cells [8]. As a reference, a black line with a slope of one is plotted in order to emphasize the systematic disagreement between the two techniques.

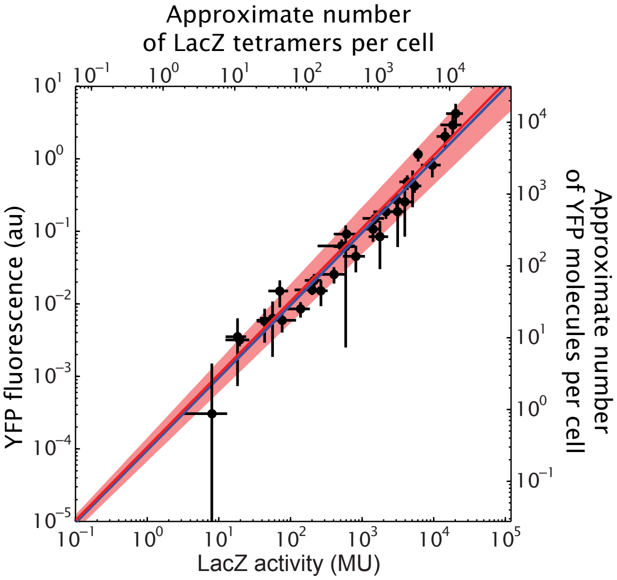
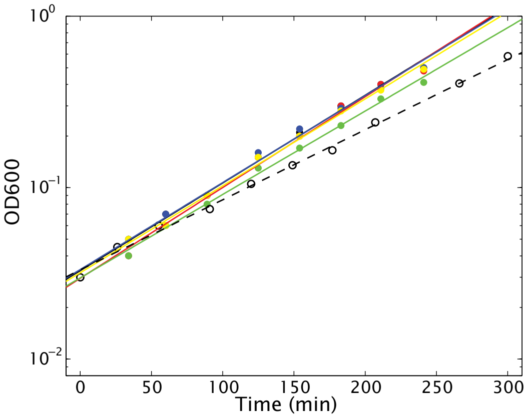


Figure 6.16: Relation between the mean cell fluorescence and the β-galactosidase activity. The total fluorescence per cell is plotted against the β-galactosidase activity. Each point corresponds to the same promoter bearing either EYFP or *lacZ* as a reporter in the same strain background and at the same concentration of IPTG. The blue line is a linear fit fixing the intercept to zero (see figure 6.4). The red line is a fit to a power law with a resulting exponent of 1.01 ± 0.05, consistent with a linear relation between the two reporters. The shaded area is defined by the standard error for the power law fit.



	Strain background	Location of construct	Reporter	IPTG concentration (μM)	Growth rate (min)
⚫	lacI++	Chromosome	None	0	59 ± 1
🔴	lacI++	Plasmid	None	0	57 ± 1
🟢	lacI+	Plasmid	LacZ	1000	62 ± 1
🔵	lacI–	Plasmid	EYFP	0	59 ± 2
🟡	lacI+	Plasmid	LacZ	1000	59 ± 1
○	lacI–	Plasmid	LacZ	0	74 ± 1

Figure 6.17: Effect of reporter proteins on growth rate. The different levels of EYFP assayed in this work do not affect the growth of the cells significantly. However, high β-galactosidase levels slow down growth in a detectable fashion. This level of LacZ is reached when our plasmid reporter is present in strain lacI-.

Bibliography

[1] H. G. Garcia, H. J. Lee, J. Q. Boedicker, and R. Phillips. The limits and validity of methods of measuring gene expression for the testing of quantitative models. *Biophys J*, 2011. (*Under review*).

[2] S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–41, 2003.

[3] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–5, 2005.

[4] A. Zaslaver, A. Bren, M. Ronen, S. Itzkovitz, I. Kikoin, S. Shavit, W. Liebermeister, M. G. Surette, and U. Alon. A comprehensive library of fluorescent transcriptional reporters for *escherichia coli*. *Nat Methods*, 3(8):623–8, 2006.

[5] T. Kuhlman, Z. Zhang, M. H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *escherichia coli*. *Proc Natl Acad Sci U S A*, 104(14):6043–8, 2007.

[6] S. Ben-Tabou De-Leon and E. H. Davidson. Gene regulation: Gene control network in development. *Annu Rev Biophys Biomol Struct*, 36:191, 2007.

[7] T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek. Probing the limits to positional information. *Cell*, 130(1):153–64, 2007.

[8] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying *e. Coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329:533, 2010.

[9] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1):117–24, 2007.

[10] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[11] S. Oehler, M. Amouyal, P. Kolkhof, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. Quality and position of the three *lac* operators of *e. Coli* define efficiency of repression. *EMBO J*, 13(14):3348–55, 1994.

[12] U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc Natl Acad Sci U S A*, 99(19):12015–20, 2002.

[13] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003.

[14] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[15] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[16] S. Klumpp, Z. Zhang, and T. Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1366–75, 2009.

[17] R. Hermsen, S. Tans, and P. R. Ten Wolde. Transcriptional regulation by competing transcription factor modules. *PLoS Comput Biol*, 2(12):e164, 2006.

[18] M. J. Morelli, P. R. Ten Wolde, and R. J. Allen. DNA looping provides stability and robustness to the bacteriophage lambda switch. *Proc Natl Acad Sci U S A*, 106(20):8101–6, 2009.

[19] R. Hermsen, B. Ursem, and P. R. Ten Wolde. Combinatorial gene regulation using auto-regulation. *PLoS Comput Biol*, 6(6):e1000813, 2010.

[20] K. Sneppen, S. Krishna, and S. Semsey. Simplified models of biological networks. *Annu Rev Biophys*, 39:43–59, 2010.

[21] J. H. Miller. *Experiments in molecular genetics.* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1972.

[22] T. K. Van Dyk, E. J. Derose, and G. E. Gonye. Luxarray, a high-density, genomewide transcription analysis of escherichia coli using bioluminescent reporter strains. *J Bacteriol*, 183(19):5496–505, 2001.

[23] R. J. Bongaerts, I. Hautefort, J. M. Sidebotham, and J. C. Hinton. Green fluorescent protein as a marker for conditional gene expression in bacterial cells. *Methods Enzymol*, 358:43–66, 2002.

[24] D. R. Larson, R. H. Singer, and D. Zenklusen. A single molecule view of gene expression. *Trends Cell Biol*, 19(11):630–7, 2009.

[25] A. Raj and A. Van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys*, 38:255–70, 2009.

[26] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.

[27] J. Q. Wu, C. D. Mccormick, and T. D. Pollard. Counting proteins in living cells by quantitative fluorescence microscopy with internal standards. *Methods Cell Biol*, 89:253–73, 2008.

[28] R. Kishony and S. Leibler. Environmental stresses can alleviate the average deleterious effect of mutations. *J Biol*, 2(2):14, 2003.

[29] R. S. Cox Iii, M. G. Surette, and M. B. Elowitz. Programming gene expression with combinatorial promoters. *Mol Syst Biol*, 3:145, 2007.

[30] A. C. Oates, N. Gorfinkiel, M. Gonzalez-Gaitan, and C. P. Heisenberg. Quantitative approaches in developmental biology. *Nat Rev Genet*, 10(8):517–30, 2009.

[31] C. C. Guet, L. Bruneaux, T. L. Min, D. Siegal-Gaskins, I. Figueroa, T. Emonet, and P. Cluzel. Minimally invasive determination of mrna concentration in single living bacteria. *Nucleic Acids Res*, 36(12):e73, 2008.

[32] T. Gregor, E. F. Wieschaus, A. P. Mcgregor, W. Bialek, and D. W. Tank. Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*, 130(1):141–52, 2007.

[33] I. Hautefort and J. C. Hinton. Measurement of bacterial gene expression in vivo. *Philos Trans R Soc Lond B Biol Sci*, 355(1397):601–11, 2000.

[34] C. G. Pfeifer and B. B. Finlay. Monitoring gene expression of salmonella inside mammalian cells: Comparison of luciferase and beta-galactosidase fusion systems. *J of Microbio Meth*, 24(2):155–164, 1995.

[35] A. J. Forsberg, G. D. Pavitt, and C. F. Higgins. Use of transcriptional fusions to monitor gene expression: A cautionary tale. *J Bacteriol*, 176(7):2128–32, 1994.

[36] O. Scholz, A. Thiel, W. Hillen, and M. Niederweis. Quantitative analysis of gene expression with an improved green fluorescent protein. P6. *Eur J Biochem*, 267(6):1565–70, 2000.

[37] J. MÜLler, S. Oehler, and B. MÜLler-Hill. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol*, 257(1):21–9, 1996.

[38] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in *escherichia coli* via the lacr/o, the tetr/o and arac/i1-i2 regulatory elements. *Nucleic Acids Res*, 25(6):1203–10, 1997.

[39] I. F. Lau, S. R. Filipe, B. Soballe, O. A. Okstad, F. X. Barre, and D. J. Sherratt. Spatial and temporal organization of replicating *escherichia coli* chromosomes. *Mol Microbiol*, 49(3):731–43, 2003.

[40] K. A. Datsenko and B. L. Wanner. One-step inactivation of chromosomal genes in *escherichia coli* k-12 using pcr products. *Proc Natl Acad Sci U S A*, 97(12):6640–5, 2000.

[41] S. K. Sharan, L. C. Thomason, S. G. Kuznetsov, and D. L. Court. Recombineering: A homologous recombination-based method of genetic engineering. *Nat Protoc*, 4(2):206–23, 2009.

[42] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002.

[43] G. Posfai, G. Plunkett 3rd, T. Feher, D. Frisch, G. M. Keil, K. Umenhoffer, V. Kolisnychenko, B. Stahl, S. S. Sharma, M. De Arruda, V. Burland, S. W. Harcum, and F. R. Blattner. Emergent properties of reduced-genome *escherichia coli*. *Science*, 312(5776):1044–6, 2006.

[44] N. A. Becker, J. D. Kahn, and L. J. Maher Iii. Bacterial repression loops require enhanced DNA flexibility. *J Mol Biol*, 349(4):716–30, 2005.

[45] P. A. Wiggins, K. C. Cheveralls, J. S. Martin, R. Lintner, and J. Kondev. Strong intranucleoid interactions organize the *escherichia coli* chromosome into a nucleoid filament. *Proc Natl Acad Sci U S A*, 107(11):4991–5, 2010.

[46] M. Unger, E. Kartalov, C. S. Chiu, H. A. Lester, and S. R. Quake. Single-molecule fluorescence observed with mercury lamp illumination. *Biotechniques*, 27(5):1008–14, 1999.

[47] S. S. Rogers, T. A. Waigh, X. Zhao, and J. R. Lu. Precise particle tracking against a complicated background: Polynomial fitting with gaussian weight. *Phys Biol*, 4(3):220–7, 2007.

[48] J. Q. Wu and T. D. Pollard. Counting cytokinesis proteins globally and locally in fission yeast. *Science*, 310(5746):310–4, 2005.

[49] K. Hirschberg, C. M. Miller, J. Ellenberg, J. F. Presley, E. D. Siggia, R. D. Phair, and J. Lippincott-Schwartz. Kinetic analysis of secretory protein traffic and characterization of golgi to plasma membrane transport intermediates in living cells. *J Cell Biol*, 143(6):1485–503, 1998.

[50] D. W. Piston, G. H. Patterson, and S. M. Knobel. Quantitative imaging of the green fluorescent protein (gfp). *Methods Cell Biol*, 58:31–48, 1999.

[51] V. Sourjik and H. C. Berg. Binding of the *escherichia coli* response regulator chey to its target measured *in vivo* by fluorescence resonance energy transfer. *Proc Natl Acad Sci U S A*, 99(20):12669–74, 2002.

[52] M. B. Elowitz, M. G. Surette, P. E. Wolf, J. B. Stock, and S. Leibler. Protein mobility in the cytoplasm of escherichia coli. *J Bacteriol*, 181(1):197–203, 1999.

[53] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–3, 2006.

[54] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–6, 2008.

[55] N. Friedman, L. Cai, and X. S. Xie. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys Rev Lett*, 97(16):–, 2006.

[56] L. Cai, N. Friedman, and X. S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62, 2006.

[57] R. Young and H. Bremer. Analysis of enzyme induction in bacteria. *Biochem J*, 152(2):243–54, 1975.

[58] D. Kennell and H. Riezman. Transcription and translation initiation frequencies of the *escherichia coli lac* operon. *J Mol Biol*, 114(1):1–21, 1977.

[59] U. Deuschle, W. Kammerer, R. Gentz, and H. Bujard. Promoters of *escherichia coli*: A hierarchy of *in vivo* strength indicates alternate structures. *EMBO J*, 5(11):2987–94, 1986.

[60] J. Lederberg. The beta-d-galactosidase of *escherichia coli*, strain k-12. *J Bacteriol*, 60(4):381–92, 1950.

[61] K. Wallenfels and R. Weil. Beta-galactosidase. *The Enzyme*, 7:617–663, 1972.

[62] G. R. Craven, Jr. E. Steers, and C. B. Anfinsen. Purification, composition, and molecular weight of the beta-galactosidase of *escherichia coli* k12. *J Biol Chem*, 240:2468–77, 1965.

[63] M. H. Ulbrich and E. Y. Isacoff. Subunit counting in membrane-bound proteins. *Nat Methods*, 4(4):319–21, 2007.

[64] J. Elf, G. W. Li, and X. S. Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–4, 2007.

[65] H. Dong, L. Nilsson, and C. G. Kurland. Gratuitous overexpression of genes in *escherichia coli* leads to growth inhibition and ribosome destruction. *J Bacteriol*, 177(6):1497–504, 1995.

[66] G. S. Harms, L. Cognet, P. H. Lommerse, G. A. Blab, and T. Schmidt. Autofluorescent proteins in single-molecule research: Applications to live cell imaging microscopy. *Biophys J*, 80(5):2396–408, 2001.

[67] S. Liang, M. Bipatnath, Y. Xu, S. Chen, P. Dennis, M. Ehrenberg, and H. Bremer. Activities of constitutive promoters in *escherichia coli*. *J Mol Biol*, 292(1):19–37, 1999.

[68] M. Lanzer and H. Bujard. Promoters largely determine the efficiency of repressor action. *Proc Natl Acad Sci U S A*, 85(23):8973–7, 1988.

[69] R. R. Seabold and R. F. Schleif. Apo-arac actively seeks to loop. *J Mol Biol*, 278(3):529–38, 1998.

[70] B. MÜLler-Hill. *The lac operon: A short history of a genetic paradigm*. Walter de Gruyter, Berlin, New York, 1996.

[71] S. Oehler, E. R. Eismann, H. Kramer, and B. MÜLler-Hill. The three operators of the lac operon cooperate in repression. *EMBO J*, 9(4):973–9, 1990.

[72] R. Y. Tsien. The green fluorescent protein. *Annu Rev Biochem*, 67:509–44, 1998.

[73] J. E. Bronson, J. Fei, J. M. Hofman, R. L. G. Jr., and C. H. Wiggins. Learning rates and states from biophysical time series: A bayesian approach to model selection and single-molecule fret data. *arXiv:0907.3156v1*, 2009.

Chapter 7

Quantitative Dissection of the Simple Repression Input-Output Function

This chapter is a reproduction of reference [1].

We present a quantitative case study in transcriptional regulation in which we carry out a systematic dialogue between theory and measurement for an important and ubiquitous regulatory motif in bacteria, namely, that of simple repression. This architecture is realized by a single repressor binding site overlapping the promoter. From the theory point of view, this motif is described by a single gene regulation function based upon only a few parameters that are convenient theoretically and accessible experimentally. The usual approach is turned on its side by using the mathematical description of these regulatory motifs as a predictive tool to determine the number of repressors in a collection of strains with a large variation in repressor copy number. The predictions and corresponding measurements are carried out over a large dynamic range in both expression fold-change (spanning nearly four orders of magnitude) and in repressor copy number (spanning about two orders of magnitude). The predictions are tested by measuring the resulting level of gene expression and are then validated by using quantitative immunoblots. The key outcomes of this study include: a systematic quantitative analysis of the limits and validity of the input-output relation for simple repression, a precise determination of the *in vivo* binding energies for DNA-repressor interactions for several distinct repressor binding sites, and a repressor census for Lac repressor in *E. coli.*

7.1 Introduction

It is now possible not only to make quantitative, precise and reproducible measurements on the response of a variety of different genetic regulatory architectures, but even to synthesize novel architectures *de novo.* These successes have engendered hopeful analogies between the circuits found in cells and those that are the basis of many familiar electronic devices [2, 3]. However, in many cases, unlike the situation with the electronic circuit analogy, our understanding of these circuits is based upon enlightened empiricism rather than systematic, quantitative knowledge of the input-output relations of the underlying genetic circuits.

Regulatory biology has shed light on the space-time response of a wide variety of these genetic circuits.

Examples range from the complex regulatory networks that govern processes such as embryonic development [4, 5] to the synthetic biology setting of building completely new regulatory circuits in living cells [6]. In particular, the dissection of genetic regulatory networks is resulting in the elucidation of ever more complex wiring diagrams (see, as an example [7]). With these advances it is becoming increasingly difficult to develop intuition for the behavior of these networks in space and time. In addition, often, the diagrams used to depict these regulatory architectures make no reference to the census of the various molecular actors (the intracellular number of polymerases, activators, repressors, inducers, etc.) or to the quantitative details of their interactions that dictate their response. As a result, there is a growing need to put the description of these networks on a firm quantitative footing.

Often, the default description of regulatory response is offered by phenomenological Hill functions [8–13] which in the case of repression have the form

$$\text{gene expression level} = \frac{\alpha}{1 + ([R]/K_d)^n} + \beta, \tag{7.1}$$

where n is the Hill coefficient which determines the sensitivity of the gene regulatory function, K_d is a dissociation constant, and α and β are constants that determine the maximum and basal levels of expression, respectively. Though such descriptions might provide a satisfactory fit of the data, they can deprive us of insights into the mechanistic underpinnings of a given regulatory response, or worse, can force us into thinking about the behavior of a given circuit in a way that is not faithful to the known architecture.

Alternatively, using thermodynamic models, it has been shown for a wide class of regulatory architectures that for each and every circuit, one can derive a corresponding "governing equation" that provides the fold-change in gene expression as a function of the relevant regulatory tuning variables [14–16]. The goal of our work is to carry out a detailed experimental characterization of the predictions posed by one such governing equation for the regulatory motif describing simple repression (see figure 7.1(A)) in which a repressor can bind to a site overlapping the promoter resulting in the shutting down of expression of the associated gene. This is a particularly fundamental case study since in *E. coli* alone, there are over 400 circuits that are regulated by different transcription factors that repress by binding to a single site in the vicinity of the promoter [17]. Indeed, simple repression and activation are often thought of as the elementary ingredients of a much more diverse range of real regulatory circuits [18, 19].

As seen in figure 7.1, the level of expression in circuits governed by simple repression can be tuned by several different parameters. One of the key tuning variables in nearly all regulatory and signaling networks is the concentrations (or numbers) of the relevant molecular players in the process of interest. We use the repressor concentration as one of the main tunable parameters in the experiments described below, with a 100-fold range of different repressor concentrations considered. In order to explore our understanding of how this parameter dictates regulatory response, we need to know how many repressors our strains of interest harbor. A series of beautiful recent experiments has made important progress in carrying out the molecular census using a variety of clever methods. These molecular counts include the census of all actinrelated proteins in *S. pombe* cells [20], a count of essentially *all* the proteins in *S. cerevisiae* cells [21], a determination of the distribution of both lipids and proteins in synaptic vesicles [22], several counts of the proteins in *E. coli* [23, 24] and other cell types as well [25]. Most relevant to the current work is a recent experiment using a fluctuation-based counting method to determine the number of transcription factors in *E. coli* that control a synthetic circuit of interest [11]. Our work adds a new twist to protein census taking by using thermodynamic models as a way to count the number of repressors in a simple regulatory motif.

A quantitative control of the absolute number of transcription factors is seldom employed in experiments that aim to dissect regulatory architectures even though it is one of the main strategies to to verify the predictions from thermodynamic models [14–16]. Previous work has usually relied on the control of an external inducer to vary the regulatory response of a genetic circuit [6, 10, 12, 13, 26, 27]. However, the use of inducer molecules, though experimentally convenient, adds another layer of complexity to the modeling approach and has only been systematically characterized in a few cases [12].

Recent measurements [11, 24, 28–32] have also often focussed on the variability or “noise” associated with transcriptional regulation. Although there has been great recent interest in this gene expression variability, we argue that a crucial quantitative prerequisite to fully dissecting the properties of genetic networks is a viable description of their mean response, and any conceptual frameworks used to describe the noise must first be consistent with these mean responses.

In this work we test these thermodynamic models of transcriptional regulation by generating *parameter-free predictions* for the level of gene expression as a function of the regulatory tuning variables of the simple repression architecture. We show significant agreement between the theoretical description and the measurements over multiple orders of magnitudes of the inputs and outputs of the system. We conclude that through thermodynamic models we can accurately predict the level of regulation due to simple repression, opening the door to the design of synthetic genetic circuits where the level of gene expression can be tuned theoretically and to the better interpretation of the transcriptional response of naturally occurring circuits.

7.2 Theory and Experimental Design

Though our analysis should be relevant generically for simple repression, the reasoning behind our experiments is based upon a series of earlier measurements and calculations on the level of repression in the specific case of the *lac* operon [33, 34]. In particular, we consider the case where there is only a single specific binding site for the Lac repressor (see figure 7.1). The wild-type *lac* operon was rewired such that only the main operator was present and then, in turn, different strains were constructed in which the strength of that main operator was systematically weakened according to the progression Oid to O1 to O2 to O3, as shown in figure 7.12.

Thermodynamic models assume that the processes leading to transcription initiation by RNAP are in quasiequilibrium. This means that we can use the tools of statistical mechanics to describe the binding

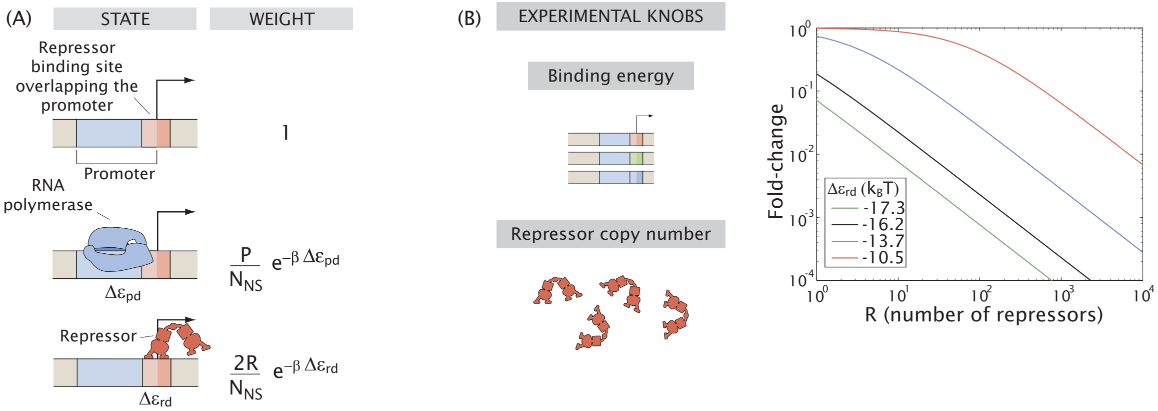


Figure 7.1: The simple repression motif. (A) States and weights of the thermodynamic model describing this regulatory motif. We assume that Lac repressor sterically excludes RNA polymerase from the promoter, though that assumption is not critical to our analysis. P and R are the number of RNA polymerase and Lac repressor molecules inside the cell, respectively. N_{NS} is the number of non-specific sites, which we assume to be the size of the genome. $\Delta\varepsilon_{pd}$ and $\Delta\varepsilon_{rd}$ are the difference in energy between being specifically and non-specifically bound for RNA polymerase and Lac repressor, respectively. The difference in color in the repressor binding site denotes an overlap of the binding site with the promoter. (B) The tuning variables that can be varied in the model and controlled experimentally are the binding strength (by changing the Lac repressor operator sequence) and the concentration of Lac repressor (by changing its mRNA ribosomal binding site). The effect of tuning these parameters on the fold-change in gene expression is shown in the graphs. Note that stronger repressor binding corresponds to a larger fold-change. For a detailed derivation of the expression and discussion of the assumptions used see the Supplementary Information.

of RNA polymerase and TFs to DNA. Further, the level of gene expression is assumed to be proportional to the probability that RNA polymerase is bound to the promoter of interest [14, 35]. This probability is determined, in turn, by the interactions between polymerase and the promoter and competition for those binding sites by repressors. In figure 7.1(A) we show the thermodynamic states and weights corresponding to a minimal model of the simple repression regulatory motif. In this simplified model the promoter can be found in only one of three states: i) empty, ii) occupied by RNA polymerase, and iii) occupied by Lac repressor. The partition function for this system is obtained by summing over the statistical weights of each of these states and is given by

$$Z = \underbrace{1}_{\text{promoter empty}} + \underbrace{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_{pd}}}_{\text{RNA polymerase bound}} + \underbrace{\frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_{rd}}}_{\text{LacI bound}}, \tag{7.2}$$

where P is the number of RNA polymerase molecules, R is the number of Lac repressor tetramers, and $N_{NS} \approx 5 \times 10^6$ is the number of non-specific DNA sites (the length of the genome), corresponding to the reservoir for both molecules. β is $(\mathrm{k_B}T)^{-1}$ with $\mathrm{k_B}$ being the Boltzmann constant and T the absolute temperature. The energies $\Delta\varepsilon_{pd}$ (RNA polymerase-DNA) and $\Delta\varepsilon_{rd}$ (repressor-DNA) correspond to the difference between specific and non-specific binding for RNA polymerase and Lac repressor, respectively, where we make the simplifying assumption of a homogeneous nonspecific background. The factor of 2 in front of the concentration of Lac repressor stems from the fact that this molecule is a tetramer, a dimer of dimers, with two binding heads. Therefore, $2R$ corresponds to the number of binding heads inside the cell. For a complete derivation of these terms, please refer to [15, 36] and the Supplementary Information.

The probability of finding RNA polymerase bound to the promoter is then given by

$$p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}}{Z}, \tag{7.3}$$

where Z is the partition function defined in equation 7.2. A much more convenient quantity to measure is the fold-change or relative change in gene expression due to the presence of the transcription factor, namely,

$$\text{fold-change} = \frac{p_{bound}(R \neq 0)}{p_{bound}(R = 0)} = \frac{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}} + \frac{2R}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}}. \tag{7.4}$$

The great advantage of this quantity is that it is easily accessible both theoretically and experimentally. It is unitless and can be measured by comparing the levels of gene expression (in any arbitrary or absolute units) when Lac repressor is present and absent. We define this fold-change in gene expression with respect to the absence of transcription factor and not with respect to a state where the transcription factor is fully induced, such as in the presence of saturating concentrations of IPTG. Using inducers would require us to consider the induction process explicitly [12]. In the case of a weak promoter such as *lacUV5* used in this work ([16] and Supplementary Information) the term $\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}$ is much smaller than one. This results in the fold-change collapsing to the simpler form

$$\text{fold-change} = \frac{1}{1 + \frac{2R}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}}. \tag{7.5}$$

This last expression serves as the basis of our experimental design where we identify two tuning variables that can be controlled experimentally in a systematic fashion: the binding energy and the concentration of Lac repressor. In figure 7.1(B) we show the predicted fold-change as a function of these two experimentally accessible parameters. Alternatively, the binding of Lac repressor can be described by a dissociation constant, the concentration of Lac repressor for which the fold-change in gene expression is 1/2. This approach is explained in the Supplementary Information. Throughout the text we report both binding energies and dissociation constants.

Earlier hints as to how simple repression plays out quantitatively were offered by Oehler et al. [33, 34] who measured the fold-change in gene expression for constructs bearing each one of the four operators and for two different concentrations of Lac repressor. Using equation 8.2 or equivalent expressions [16, 37] the binding energy of Lac repressor to each one of the operators can be estimated. These binding energies are shown in figure 7.2(A). It must be noted though that these original measurements were not performed with the intention of the kind of quantitative dissection advocated here and that therefore the uncertainties in the parameters are substantial.

If the binding energies obtained from the Oehler et al. data are to be more than an exercise in data fitting they need to be used to generate predictions that will help explore the range of applicability of thermodynamic models of transcriptional regulation. Our ambition was to follow these intriguing earlier measurements to their logical conclusions by generating parameter-free predictions for the fold-change in gene expression as a function of the repressor concentrations and to contrast them with a new set of experiments aimed at testing those same predictions over a wide range of repressor copy numbers.

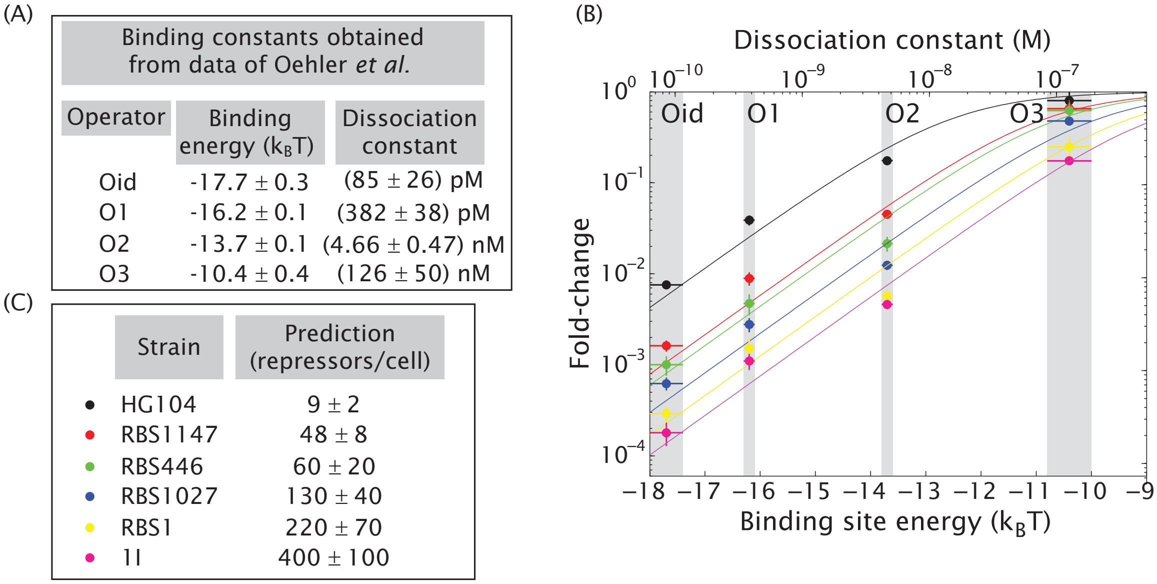
For the measurements reported here, we created roughly 30 strains of bacteria where we systematically tuned the concentration of repressor using a recently developed scheme for controlling ribosomal binding strength [38]. Though this scheme provides a rough expectation for the number of repressors in each one of those strains we had no precise or accurate *a priori* knowledge of the actual intracellular numbers of Lac repressors. These strains bear reporter constructs regulated by simple repression such as the ones shown in figure 7.12, for which we measure the fold-change in gene expression. If we are to believe the input-output function from equation 8.2, once we know the binding energy of the operator in question there is a direct and unequivocal relation between the fold-change in gene expression and the number of repressor molecules. Testing these predictions requires an accurate and precise quantification of the absolute levels of repressor inside the cell. In fact, we view this approach as a way to count molecules by inference by looking at levels of gene expression and passing these levels of expression through the theoretical filter of equation 8.2.

In the following sections we test these parameter-free predictions over a wide range of both expression and repressor concentration and show that they largely jibe with our experimental observations. The logic advocated here is that if equation 8.2 is shown to be predictive it will open the door to creating synthetic gene regulatory circuits whose level of gene expression can be precisely tuned *a priori* and to being able to predict the regulation of a particular promoter by just looking at its regulatory sequence. In addition, a predictive understanding of the input-output relation of these architectures will serve as a jumping off point for the design and understanding of more complex circuits such as those involving DNA looping, cooperative repression, etc. [16].

7.3 Results

Equation 8.2 represents a provocatively simple expression purporting to describe the response of a bacterial cell to a wide variety of perturbations such as altering the DNA target sites (with the K_d's changing by three orders of magnitude or, equivalently, $\Delta\varepsilon_{rd}$ changing by 6 $k_{\mathrm{B}}T$ [16, 37, 39, 40]) and repressor copy numbers (with the copy numbers changing by several orders of magnitude). If we take this equation seriously, it implies that once we have determined the parameter $\Delta\varepsilon_{rd}$ (or equivalently the *in vivo* K_d's), there is a quantitative relation between the fold-change in gene expression and the corresponding concentration of Lac repressor. Namely, once we know one quantity we can *predict* the other.

In order to exploit equation 8.2 we designed *lacUV5* promoters with a single binding site for Lac repressor



(A)

Binding constants obtained from data of Oehler *et al.*

Operator	Binding energy (k_BT)	Dissociation constant
Oid	-17.7 ± 0.3	(85 ± 26) pM
O1	-16.2 ± 0.1	(382 ± 38) pM
O2	-13.7 ± 0.1	(4.66 ± 0.47) nM
O3	-10.4 ± 0.4	(126 ± 50) nM

(B)

(C)

	Strain	Prediction (repressors/cell)
●	HG104	9 ± 2
●	RBS1147	48 ± 8
●	RBS446	60 ± 20
●	RBS1027	130 ± 40
●	RBS1	220 ± 70
●	1I	400 ± 100

Figure 7.2: Single-site binding energies and prediction of the number of repressors for different strains. (A) The operator binding energies and dissociation constants are deduced from the data by Oehler et al. [34] using equation 8.2. The error bars are calculated assuming an error in the fold-change measurement of 30% and assuming no error in the number of repressor molecules. (B) The fold-change in gene expression is measured for all four operators in six different strain backgrounds. Using the binding energies from (A) we fit the data to equation 8.2 in order to make a parameter-free prediction of the number of repressors present in each strain shown in (C). Errors in the predictions represent the standard error of the corresponding fit.

at the wild-type position of O1. These promoters bore either Oid, O1, O2 or O3 and controlled the expression of the enzymatic reporter gene LacZ (see Materials and Methods and figure 7.12). We integrated each one of these simple repression constructs such as the one shown in in figure 7.1(A) in the chromosome of a strain bearing no Lac repressor and in six different strains that we systematically designed to express different constitutive levels of Lac repressor. As mentioned above, though we had a qualitative expectation about the concentration of Lac repressor present in each strain we had no previous quantitative information about that magnitude.

7.3.1 Taking the repressor census through thermodynamic models

We measured the fold-change in gene expression of our simple repression constructs bearing the operators Oid, O1, O2 or O3 in the six different strain backgrounds we created. For a given strain if we fit equation 8.2 to data of fold-change in gene expression as a function of the operator binding energy we can obtain a prediction for the number of Lac repressors it harbors. The fits and resulting predictions are shown in figure 7.2(B) and (C). The corresponding absolute values measured for each strain are shown in figure 7.13.

Since the majority of our strains were created for this particular work, the resulting predicted cellular concentrations cannot be compared to any external standard. However, strain HG104 expresses wild-type levels of repressor from the native *lacI* gene. Indeed, for this strain we predict 9 ± 2 repressor tetramers per cell, comparable to the previous and, to our knowledge, only available absolute measurement [41].

In order to bring the predictions of the model for simple repression to fruition we need to directly measure the number of repressors in each one of our six strains. We measured the *in vivo* concentration of Lac repressor in these six strains by performing quantitative immunoblots from cell lysates such as those shown in figure 7.3(A). In order to get an absolute count of the amount of Lac repressor in each strain a series of dilutions of a purified Lac repressor standard of a known concentration was used (figure 7.3(B)). Quantification of the immunoblots luminescence was performed using a cooled CCD camera. Care was taken to account for spatial non-uniformities in the light collection as depicted by figure 7.3(C). We can reliably detect a wide range of purified Lac repressor standard using our immunoblots (as low as 50 pg, corresponding to approximately 5 repressors per cell). This increases our confidence in the method as a way of precisely quantifying protein concentrations in bulk even at very low levels (see Materials and Methods and figure 7.3(D)).

Our predictions for the number of Lac repressors in each strain can now be compared to the direct measurements of this quantity which are shown in figure 7.4(A). In figure 7.4(B) we compare the predictions and direct measurements explicitly. The direct measurements are comparable to the predictions within experimental error, giving us confidence that the proposed input-output function from equation 8.2 appropriately describes the input-output properties of the simple repression regulatory motif. This suggests in turn that once we know the binding energy for an operator we have predictive power. Though this analysis yielded results that are largely consistent between theory and experiment, it appears that we systematically underestimate the number of repressors in the two strains with the highest concentrations. The reader is referred to the Supplementary Information for a further discussion of these two strains. More generally, since the original measurements we used to obtain the binding energies [34] were not intended to determine the binding energies as input into a predictive model, given their uncertainties, it is interesting to see to what extent they jibe with measurements specifically designed to obtain them.

7.3.2 Direct determination of the *in vivo* Lac repressor binding energies

The scheme for exploring the limits and validity of the thermodynamic model advocated in the previous section is based on the previous knowledge of the binding energy of Lac repressor to its operator DNA. As noted above, these binding energies were obtained from previous experimental results, which correspond to data that was not taken with the objective of exploring the experimental results in terms of equation 8.2. In this sense, it is somewhat surprising that the resulting predictions match so well with our experimental data.

Having shown in the previous section that the input-output function corresponding to equation 8.2 can successfully account for all our experimental observations, one might still wonder if the binding energies used in the model could be more precisely constrained. One scheme to achieve a better quantification of the binding energies is to fit the fold-change in gene expression as a function of the number of repressors for a

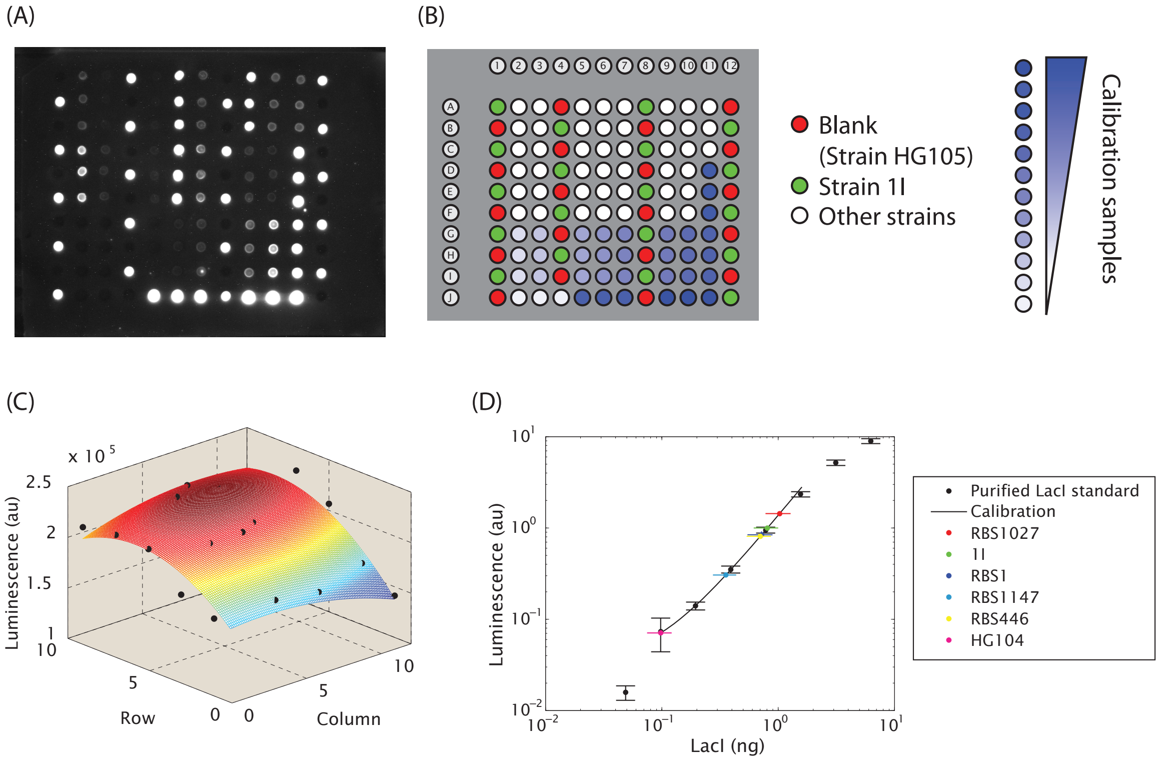


Figure 7.3: Immunoblots for the measurement of the *in vivo* concentration of Lac repressor. (A) Typical luminescence image obtained from an immunoblot. (B) Map of the samples loaded on the membrane shown in (A). The blank (HG105) and 1I samples are used to create a normalization map by subtracting the blank luminescence from all samples and dividing by 1I. White spots correspond to the cell lysates measured and the blue spots correspond to the different concentrations of purified Lac repressor standard. (C) Normalization map generated by fitting a 2D polynomial to 1I samples scattered around the membrane (black dots) after removing the blank. This map was used to account for non-uniformities in the collection of luminescence from the membrane. (D) Luminescence vs. quantity of LacI loaded. The calibration samples are used to construct a power law fit. The luminescence of the measured samples are shown as well. The unknown amounts of repressor loaded are determined by using the calibration curve. Samples 1I and RBS1 have been diluted 1:8 to match them to the dynamic range of the assay and therefore appear in the figure as having less signal within a spot (see Supplementary Information).

(A)

Strain	Direct measurement (repressors/cell)
HG104	11 ± 2
RBS1147	30 ± 10
RBS446	62 ± 15
RBS1027	130 ± 20
RBS1	610 ± 80
1I	870 ± 170

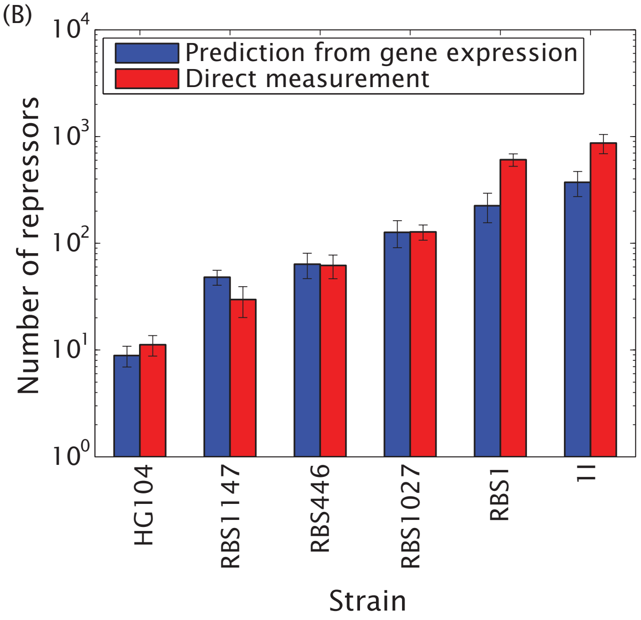


Figure 7.4: Experimental and theoretical characterization of repressor copy number. (A) Immunoblots were use to measure the cellular concentration of Lac repressor in six strains with different constitutive levels of Lac repressor. Each value corresponds to an average of cultures grown on at least three different days. The error bars are the standard deviation of these measurements. (B) The fold-change measurements in figure 7.2 were combined with the binding energies obtained from figure 7.2(A) (derived from previous experimental results [34]) in order to predict the number of Lac repressors per cell in each one of the six strains used in this work. These predictions were examined experimentally by counting the number of Lac repressors using quantitative immunoblots.

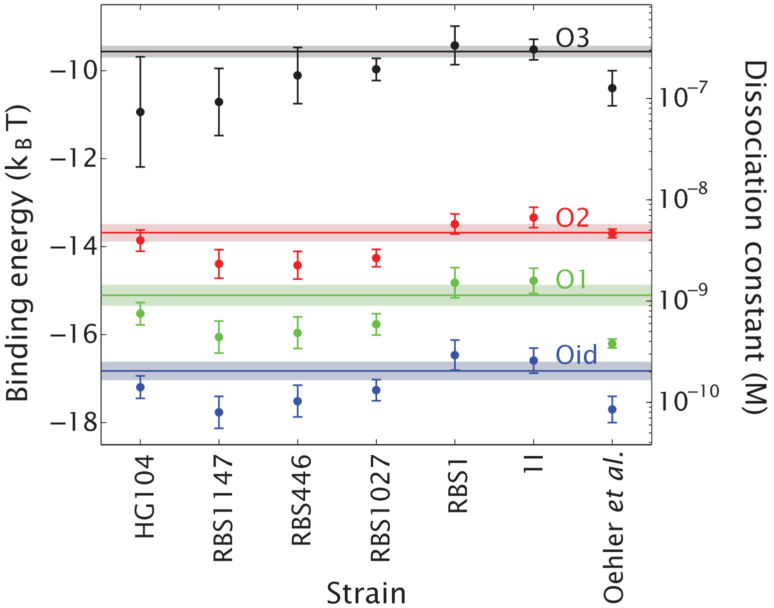


Figure 7.5: Determination of the *in vivo* binding energies. For each strain we combine the measurements of the fold-change in gene expression with its corresponding repressor concentration and solve equation 8.2 to obtain an estimate of the binding energies (dots). The energies obtained from the Oehler et al. data [34] are also shown. The lines correspond to using all measurements of the fold-change in gene expression with their corresponding repressor concentration to fit equation 8.2 in order to obtain the best possible estimate for the binding energies. This fit is shown in figure 7.15. The results of this approach are shown as horizontal lines and the shaded region captures the uncertainty.

particular construct bearing one of the operators. Implementation of this concept is shown in figure 7.15, where we are now combining all of our measurements in order to determine the best values of the different *in vivo* binding energies. On the other hand, one might chose to use the information about fold-change and repressor copy number for one particular strain in order to derive the different binding energies. This can be done, in turn, for all strains created for this work. In figure 7.5 we compare such fits with the binding energies that can be obtained from analyzing a single strain. Additionally, we show the energies from figure 7.2(A) for comparison. These multiple approaches for obtaining the binding energies, all leading to essentially comparable results (see, for example, figure 7.14), increases our confidence in the simple model of equation 8.2 and in the minimalist modeling philosophy used to obtain it as a quantitative and predictive tool.

It is common in the theoretical treatment of experiments on transcriptional regulation to include a constant level of expression dubbed the “leakiness”. This is usually understood as a low level of activity that is independent of any regulation. The reader is referred to the Supplementary Information for a more detailed description of leakiness where we show that the values obtained for the binding energies do not change significantly for reasonable values of the leakiness.

7.4 Discussion

Theoretical models of gene expression, especially in bacteria, have reached a very high level of sophistication. Similarly, measurements of gene expression have come to the point where they are both reproducible and quantitative enough to serve as the basis for explicit attempts at confronting theory and experiment and to explore the merits of these theoretical perspectives as a conceptual framework for describing regulatory response. Indeed, such measurements have now reached the point where in our view it is no longer appropriate to use just words to describe them — they call for a theoretical response that is commensurate with the level of quantitative detail in the experiments themselves. To that end, we have undertaken a detailed study of one of the most important and fundamental regulatory building blocks found in living organisms from all three domains of life, namely, simple repression. Simple repression and its positive regulation counterpart, namely simple activation, serve as the paradigmatic building blocks of the much richer regulatory strategies that are used both in the growing list of natural and synthetic networks now being explored.

In recent years, the governing equations characterizing the transcriptional response of these elementary regulatory building blocks and much more complicated assemblies of them have been worked out in detail using the ideas of statistical mechanics. The work described here provides a template for the kind of rich interplay between theory and experiment that should be demanded of these other networks as well. In particular, the governing equations describing regulatory architectures feature certain key tuning variables which serve to elicit different biological responses. In the experiments described here, we have explored two of the elementary tuning parameters that govern the simple repression motif, namely, the strength of the transcription factor binding sites and the molecular counts of the repressors themselves. We have shown that an input-output for simple repression obtained from thermodynamic assumptions, which depends on those two tuning parameters, can indeed predict in a parameter-free manner the regulatory outcome over roughly four orders of magnitude in the transcriptional output.

Using the thermodynamic model approach coupled tightly with precise measurements we have been able to perform a systematic quantitative dissection of the input-output relation for simple repression and believe that similar analyses should be carried out for each of the other governing equations describing key regulatory motifs. As a byproduct of these measurements, we have been able to make a precise determination of the *in vivo* binding energies for DNA-repressor interactions. In addition, these results provide a census of the repressor content for Lac repressor in *E. coli* over a large dynamic range (roughly two orders of magnitude in repressor counts). The predictive power revealed by this model based on a few parameters is one of the first steps towards having a standardized description of a regulatory architecture based on its microscopic parameters [2, 3]. Harkening back to the electronic circuit analogy, the results presented here are analogous to illustrating that for a resistor there is a value for the resistance which is necessary and sufficient to predict the current given the voltage. In our case specification of the binding energy $\Delta\varepsilon_{rd}$ is necessary and sufficient to predict the fold-change in gene expression given the number of repressors.

Further characterization of this architecture should explore the role of promoter copy number and operator position since these architectural features too are known to alter the expression profile as well [42–44]. In addition, with these insights in hand for the case of simple repression in the *lac* operon, it is now important to examine a suite of similar architectures in *E. coli* and other bacteria with the idea being to explore the extent to which the successes found in this case can be expected to apply to other genes.

7.5 Materials and Methods

7.5.1 DNA constructs and strains

The construction of all plasmids and strains is described in detail in the Supplementary Information.

In short, plasmids pZS25O1+11, pZS25O2+11, pZS25O3+11, and pZS25Oid+11 have a *lacUV5* promoter controlling the expression of a LacZ reporter. Care was taken to delete the O2 binding site present in the wild-type lacZ coding region [45]. These plasmids are shown schematically in figure 7.12.

Plasmids pZS3*1-lacI expresses Lac repressor off of a $\mathrm{p_{LtetO\text{-}1}}$ promoter [46]. The ribosomal binding site of this construct was weakened following [38] using Site Directed Mutagenesis (Quikchange II, Stratagene). In table 7.5 we show the predicted strength from the model and the corresponding concentration of Lac repressor once the constructs were chromosomally integrated. We can see that even though the predicted and measured values do not correlate too well the constructs chosen span a wide range of expression levels. This does not necessarily contradict the results reported in [38] as they claim they can predict the RBS strength within a factor of 2.3.

The *E. coli* strains used in this experiment are shown in table 7.3. Chromosomal deletions were generated using the protocol developed by Datsenko and Wanner [47]. HG105 is wild-type *E. coli* (MG1655) with a complete deletion of the *lacIZYA* genes. HG104 is also wild-type *E. coli* with a deletion of the *lacZYA* genes. We therefore expect strain HG104 to express wild-type levels of Lac repressor.

Reporter constructs and Lac repressor constructs were integrated into the *galK* and *ybcN* regions using recombineering [48]. The corresponding primers and a detail of the targeted chromosomal positions are shown in table 7.4. The reporter constructs were then combined with the different strains expressing varying amounts of Lac repressor using P1 transduction (openwetware.org/wiki/Sauer:P1vir_phage_transduction). All integrations and transductions were confirmed by PCR amplification of the replaced chromosomal region and by sequencing.

7.5.2 Growth conditions and gene expression measurements

Strains to be assayed were grown overnight in 5 ml LB plus 30 μg/ml of kanamycin and chloramphenicol (when needed) at 37 C and 300 RPM shaking. The cells were then diluted 1:4000 to 1:1000 into 4 ml of M9 minimal medium + 0.5% glucose in triplicate culture tubes. Antibiotics were not added at this step. These cells were grown for 6 to 9 hours until an OD600 of approximately 0.3 was reached after which they were once again diluted 1:10 and grown for 3 more hours to 0.3 OD600 for a total of more than 10 cell divisions.

At this point cells were harvested and their level of gene expression measured. Our protocol for measuring LacZ activity is basically a slightly modified version of the one described in [49, 50]. Details are given in the Supplementary Information.

7.5.3 Measuring *in vivo* Lac repressor concentration

Cell lysates of our different strains bearing Lac repressor were obtained as described in the Supplementary Information. Calibration samples using a known concentration of purified Lac repressor (courtesy of Stephanie Johnson) diluted in lysate of HG105 strain (strain without Lac repressor) were used.

A nitrocellulose membrane was prepared for sample loading and afterwards blocked and treated with Anti-LacI primary monoclonal antibody and HRP-linked secondary antibody as discussed in the Supplementary Information. 2 μl of each sample were spotted on the membrane in a pattern similar to that of a 96-well plate. The resulting drops had a typical size of 3 mm. All samples were loaded in triplicate with the exception of samples 1I and HG105. Both of them were loaded on the order of 20 times on different positions of the membrane in order to obtain a spatial standard that would allow for corrections of non-uniformities in the light collection (see below).

The membrane was dried and developed with Thermo Scientific SuperSignal West Femto Substrate (Thermo Scientific) and imaged in a BioRad VersaDoc 3000 system with an exposure of five minutes. A typical raw image of one of the membranes is shown in figure 7.3(A) and the corresponding loading map can be seen in figure 7.3(B). Custom Matlab code was written to detect the spots and calculate their total luminescence. The luminescence coming from the HG105 blank samples was fitted to a 2nd degree polynomial, which was in turn subtracted from all other luminescence values. After this another 2nd degree polynomial was fitted to the 1I samples, resulting in a typical surface such as the one shown in figure 7.3(C). Notice that differences of up to 25% could be observed between different positions on the membrane. This last polynomial was used to normalize the intensity of all other samples.

The luminescence corresponding to the calibration samples was overlaid with the luminescence from the strains. The calibration samples were fitted to a power law using only the calibration data points in the range of the samples that were to be measured. An example of this calibration is shown in figure 7.3(C). For additional details please refer to the Supplementary Information.

Finally, the amount of Lac repressor found in a spot was related to the number of Lac repressors molecules per cell by calibration of the OD readings of the original cultures to cell density. The calibration between mass detected on the membrane and the corresponding intracellular number of Lac repressors depends on the concentration of cells in the cultures assayed and the volume recovered from the various concentration and lysis steps. As such, there is no calibration factor. However, on average a detected mass of 12 pg within a spot would correspond to 1 Lac repressor tetramer per cell. Please refer to equation 7.35 for more details of this calibration. This whole procedure was repeated for four sets of strains grown on different days.

7.6 Supplementary Information

7.6.1 Theoretical background

In the following sections we explore the theoretical background leading to the different predictions explored in the main text. We start by introducing thermodynamic models in general and arrive at an expression for the fold-change in gene expression due to repression by Lac repressor.

7.6.1.1 "Thermodynamic models" of transcriptional regulation

Thermodynamic models of transcriptional regulation are based on computing the probability of finding RNA polymerase (RNAP) bound to the promoter and how the presence of transcription factors (TFs) modulates this probability. These models and their application to bacteria are reviewed in [15, 16].

These models make two key assumptions. First, the models assume that the processes leading to transcription initiation by RNAP are in quasiequilibrium. This means that we can use the tools of statistical mechanics to describe the binding of RNA polymerase and TFs to DNA. Second, they assume that the level of gene expression of a gene is proportional to the probability of finding RNAP bound to the corresponding promoter.

We start by analyzing the probability that RNAP will be bound at the promoter of interest in the absence of any transcription factors. We assume that the key molecular players (RNAP and TFs) are bound to the DNA either specifically or non-specifically. In particular, this question has been addressed experimentally in the context of RNAP [51] and the Lac repressor [52, 53] our two main molecules of interest in this paper. The reservoir for RNAP is therefore the background of non-specific sites. In order to determine the contribution of this reservoir we sum over the Boltzmann weights of all the possible configurations. For P RNAP molecules inside the cell with N_{NS} non-specific DNA sites we get

$$Z^{NS}(P; N_{NS}) = \frac{N_{NS}!}{P!(N_{NS}-P)!} e^{-\beta\varepsilon_{pd}^{NS}} \simeq \frac{(N_{NS})^P}{P!} e^{-\beta\varepsilon_{pd}^{NS}}, \tag{7.6}$$

where $\beta = 1/k_BT$. The first factor in the first expression accounts for all the possible configurations of RNAP on the reservoir. This is shown diagrammatically in figure 7.6. The second factor assigns the energy of binding between RNAP and non-specific DNA, ε_{pd}^{NS} (the subscript *pd* stands for RNA polymerase-DNA interaction), and as a theoretical convenience that may have to be revised in quantitatively dissecting real promoters, is taken to be the same for all non-specific sites. A more sophisticated treatment of this model to account for the differences in the non-specific binding energy has been addressed by [54]. Finally, the last expression corresponds to assuming that $N_{NS} \gg P$, a reasonable assumption given that the *E. coli* genome is ~ 5 Mbp long and that the number of σ^{70} RNAP molecules, the type of RNAP we are interested in for the purposes of this paper, is on the order of a thousand [55].

We calculate the probability of finding one RNAP bound to a promoter of interest in the presence of this non-specific reservoir. Two states are considered: either the promoter is empty and P RNAPs are in the reservoir or the promoter is occupied leaving $P-1$ RNAP molecules in the reservoir. The corresponding total partition function is

$$Z(P;N_{NS}) = \underbrace{Z^{NS}(P;N_{NS})}_{\text{Promoter unoccupied}} + \underbrace{e^{-\beta\varepsilon^{S}_{pd}} Z^{NS}(P-1;N_{NS})}_{\text{Promoter occupied}}, \tag{7.7}$$

where, in analogy to the non-specific binding energy, we have defined ε^{S}_{pd} as the binding energy between RNAP and the promoter. Our strategy in these calculations is to write the total partition function as a sum over two sets of states, each of which has its own partial partition function. The probability of finding the promoter occupied, p_{bound} is then

$$p_{bound}(P) = \frac{e^{-\beta\varepsilon^{S}_{pd}} Z^{NS}(P-1;N_{NS})}{Z^{NS}(P;N_{NS}) + e^{-\beta\varepsilon^{S}_{pd}} Z^{NS}(P-1;N_{NS})} = \frac{1}{1+\frac{N_{NS}}{P}e^{\beta\Delta\varepsilon_{pd}}}, \tag{7.8}$$

with $\Delta\varepsilon_{pd} = \varepsilon^{S}_{pd} - \varepsilon^{NS}_{pd}$, the difference in energy between being bound specifically and non-specifically. With these results in hand we can now turn to regulation by Lac repressor.

7.6.1.2 Simple repression by Lac repressor

In its simplest form, repression is carried out by a transcription factor that binds to a site overlapping the promoter. This causes the steric exclusion of RNAP from that region, decreasing the level of gene expression. Additionally, these transcription factors might be multimeric resulting in the presence of two DNA binding heads on the protein and leading to DNA looping if extra binding sites are present. In the case of Lac repressor, for example, the protein is already in its multimeric form before binding to DNA [56].

We begin by analyzing the case of repressors that require binding only to a single site to repress expression for the case of a repressor with only one binding head. This case study will allow us to develop key concepts like the role of non-specific binding which will be useful when addressing the case of repression by Lac repressor tetramers.

Repression by Lac repressor dimers

We will use the simpler case of a repressor with just one binding head to build some key concepts. In analogy to section 7.6.1.1 for the case of RNAP we consider Lac repressor to be always bound to DNA, either specifically or non-specifically. This assumption is consistent with the available experimental data [53]. Our aim is to examine all of the different configurations available to P RNA polymerase molecules, R LacI dimers and N_{NS} non-specific sites. If the binding energy of RNAP and the LacI head to non-specific DNA are ε^{NS}_{pd} and ε^{NS}_{rd}, respectively, the non-specific partition function becomes

$$Z^{NS}(P,R_2) = \underbrace{\frac{N_{NS}^{P}}{P!} e^{-P\beta\varepsilon^{NS}_{pd}}}_{Z^{NS}(P)} \underbrace{\frac{N_{NS}^{R_2}}{R_2!} e^{-R_2\beta\varepsilon^{NS}_{rd}}}_{Z^{NS}(R_2)}, \tag{7.9}$$

where we have assumed that both LacI dimers and RNAP are so diluted in the reservoir that they do not interact with each other and we use the notation R_2 with the subscript 2 as a reminder that we are considering the case of dimers.

Our model states that we can find three different situations when looking at the promoter: 1) both sites can be empty, 2) one RNAP can be taken from the reservoir and placed on its site and 3) a LacI dimer can be taken from the reservoir and placed on the main operator. These states and their corresponding normalized weights, which we derive below, are shown in figure 7.7(A). This model assumes that LacI sterically excludes RNA polymerase from the promoter, which is supported by the results from [57]. However, it can be easily modified to accommodate a state where both LacI and RNAP are bound simultaneously, for example.

Notice, however, that we are not considering a state where both RNAP and a LacI dimer are bound to the promoter region at the same time [57]. The total partition function is

$$Z_{total}(P,R_2) = \underbrace{Z^{NS}(P,R_2)}_{\text{promoter free}} + \underbrace{Z^{NS}(P-1,R_2)e^{-\beta\varepsilon^S_{pd}}}_{\text{RNAP on promoter}} + \underbrace{Z^{NS}(P,R_2-1)e^{-\beta\varepsilon^S_{rd}}}_{\text{LacI dimer on Om}}, \tag{7.10}$$

where ε^S_{pd} and ε^S_{rd} are the binding energies of RNAP and a Lac repressor head to their specific sites, respectively. We factor out the term corresponding to having all molecules in the reservoir and define $\Delta\varepsilon_{pd} = \varepsilon^S_{pd} - \varepsilon^{NS}_{pd}$ and $\Delta\varepsilon_{rd} = \varepsilon^S_{rd} - \varepsilon^{NS}_{rd}$ as the energy gain of RNAP and dimeric LacI when switching from a non-specific site to their respective specific sites, respectively. The probability of finding RNAP bound to the promoter is given by

$$p_{bound} = \frac{\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_{pd}}}{1 + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_{pd}} + \frac{R_2}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}}}. \tag{7.11}$$

This expression can be rewritten as

$$p_{bound} = \frac{1}{1 + \frac{N_{NS}}{P\cdot F_{reg}(R_2)}e^{\beta\Delta\varepsilon_{pd}}}, \tag{7.12}$$

where we have defined the regulation factor

$$F_{reg}(R_2) = \frac{1}{1 + \frac{R_2}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}}}. \tag{7.13}$$

Notice that in the absence of repressor ($R_2 = 0$), p_{bound} turns into equation 7.8. The regulation factor can be seen as an effective rescaling of the number of RNAP molecules inside the cell [15] and, in the case of repression, it is just the probability of finding an empty operator.

One of the key assumptions in the thermodynamic class of models is that the level of gene expression is linearly related to p_{bound}. This allows us to equate the fold-change in gene expression to the fold-change in promoter occupancy

$$\text{fold-change}(R_2) = \frac{p_{bound}(R_2 \neq 0)}{p_{bound}(R_2 = 0)}. \tag{7.14}$$

If we substitute p as shorthand for $\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_{pd}}$ in the expression for p_{bound}, we find

$$\text{fold-change}(R_2) = \frac{p+1}{p + \frac{1}{F_{reg}(R_2)}}. \tag{7.15}$$

The fold-change becomes independent of the details of the promoter in the case of a weak promoter, where $p \ll 1, \frac{1}{F_{reg}(R_2)}$, which permits us to write the approximate expression

$$\text{fold-change}(R_2) \simeq F_{Reg}(R_2) = \left(1 + \frac{R_2}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}}\right)^{-1}. \tag{7.16}$$

In the case of the *lac* promoter if one considers *in vitro* binding energies of RNAP to the promoter, p has the approximate value $\sim 10^{-3}$ [15]. The case of the *lacUV5* promoter used in this work is explored in section 7.6.1.4, where we show that though it is a stronger promoter than the wild-type *lac* promoter, p is still a small value. Repression always bears a regulation factor smaller than one, suggesting that we can use the weak promoter approximation for the *lacUV5* promoter.

In much the same way done in this work, Oehler et al. [34] created different constructs by varying the identity of the Lac repressor binding site. For each one of these constructs they measured the fold-change in gene expression as a function of the concentration of LacI dimers inside the cell.

In figure 7.7(B) we present a fit of their measured fold-change as a function of the number of Lac repressor molecules inside the cell. This fit is made by determining the parameters in equation 7.16. Notice that for each construct there is only one unknown: the *in vivo* binding energies, $\Delta\varepsilon_{rd}$. The results are summarized in table 7.1.

The non-specific reservoir for Lac repressor tetramers

We now consider the differences in the case where experiments are performed using tetramers rather than dimers (as in the present study). When dealing with Lac repressor tetramers only one head has to be bound to the DNA. In principle, it's not clear what the state of the other head will be. For example, that extra head could be "hanging" from the DNA without establishing contact with DNA. Another option is that the extra head will also be exploring different non-specific sites. For the purposes of this section we will assume that the second head can also bind to DNA.

Even though only one head bound to the operator is necessary for repression we will see that it is important to account for the presence of the second head. In analogy to the dimer case, we will assume that both Lac repressor binding heads are bound to DNA at all times, either specifically or non-specifically. This choice is arbitrary and the final results will not depend on the particular model for the state of the second head. We work with this particular formulation of the problem since it is both concrete and analytically tractable and makes the counting of the accessible states more transparent.

The model for the non-specific reservoir is depicted in figure 7.8. For LacI dimers we assumed that the molecules were exploring all possible non-specific sites. For the case of tetramers, in contrast, LacI will be exploring all possible DNA loops between two different non-specific sites. We start by considering only one

LacI molecule. We count the possible ways in which we can arrange the two heads on different non-specific sites on the DNA. We label the site where one of the heads binds i, the other site j. For every choice of sites an energy ε_{rd}^{NS} is gained for each head that is non-specifically bound. A cost in the form of a looping free energy $F_{loop}(i,j)$ is also paid for bringing sites i and j together. The sum over all nonspecific states can be written as

$$Z^{NS}(R_4=1)=\frac{1}{2}\underbrace{\sum_{i=1}^{N_{NS}}e^{-\beta\varepsilon_{rd}^{NS}}}_{\text{head 1, site } i}\underbrace{\sum_{j=1}^{N_{NS}}e^{-\beta\varepsilon_{rd}^{NS}}}_{\text{head 2,site } j}\underbrace{e^{-\beta F_{loop}(i,j)}}_{\text{Looping between sites } i \text{ and } j}. \tag{7.17}$$

Note that a factor of $\frac{1}{2}$ has been introduced in order not to over-count loops. This is equivalent to assuming that the two binding heads on a repressor are indistinguishable. Our model assumes that the binding of a tetramer head is independent of the state of the other head. Therefore, the interaction between a head and DNA are the same in the tetramer and dimer case.

Since the bacterial genome is circular we can chose a particular binding site for the first head, i_0, and sum over all possible positions for the second head. This can now be done for the different N_{NS} positions that can be chosen for i_0, resulting in

$$Z^{NS}(R_4=1)\simeq\frac{1}{2}\underbrace{N_{NS}}_{\text{choices for } i}e^{-\beta 2\varepsilon_{rd}^{NS}}\sum_j e^{-\beta F_{loop}(i_0,j)}. \tag{7.18}$$

Finally, we bury the term $\sum_j e^{-\beta F_{loop}(i_0,j)}$ into an effective non-specific looping free energy $e^{-\beta F_{loop}^{NS}}$. We will discuss different models for F_{loop}^{NS} and their distinctive predictions elsewhere [36]

In order to obtain the partition function for R_4 tetramers (where now the subscript 4 is a reminder that the repressor is a tetramer) we assume that all repressors are independent and indistinguishable. We therefore extend the partition function to the case of R_4 non-interacting tetramers in the reservoir by computing

$$Z^{NS}(R_4)=\frac{\left[Z^{NS}(R_4=1)\right]^{R_4}}{R_4!}=\frac{1}{2^{R_4}}\frac{(N_{NS})^{R_4}}{R_4!}e^{-\beta R_4\,2\varepsilon_{rd}^{NS}}e^{-\beta R_4\,F_{loop}^{NS}}, \tag{7.19}$$

where the binding energy is still defined as in section the previous section.

From this point on we will only consider Lac repressor tetramers. As a result, for notational compactness we replace R_4 with R. We obtain the complete non-specific partition function by multiplying the factor corresponding to repressors with a factor corresponding to RNAP being bound non-specifically shown in equation 7.9 resulting in

$$Z^{NS}(P,R)=\frac{(N_{NS})^P}{P!}e^{-\beta P\varepsilon_{pd}^{NS}}\,\frac{1}{2^R}\frac{(N_{NS})^R}{R!}e^{-\beta R\,2\varepsilon_{rd}^{NS}}e^{-\beta R\,F_{loop}^{NS}}, \tag{7.20}$$

which now allows us in the next section to address the case of repression by tetramers.

Repression by Lac repressor tetramers

We begin by taking one head of one Lac repressor tetramer out of the non-specific reservoir shown in equation 7.19 and binding it specifically to the operator. This can be easily done by going back to equation 7.17. We label the position on the genome corresponding to the specific site i_0. We will choose only those terms in the summation corresponding to the binding site of interest. Since either one of the heads can reach the position labeled by i_0 we obtain the following partition function for a single tetramer bound to a specific site

$$Z_R^{O,NS} = \frac{1}{2} e^{-\beta \varepsilon_{rd}^S} e^{-\beta \varepsilon_{rd}^{NS}} \left(\sum_{i=1}^{N_{NS}} e^{-F_{loop}(i,i_0)} + \sum_{j=1}^{N_{NS}} e^{-F_{loop}(i_0,j)} \right). \tag{7.21}$$

Because both sums are identical we can reduce this to

$$Z_R^{O,NS} = e^{-\beta \varepsilon_{rd}^S} e^{-\beta \varepsilon_{rd}^{NS}} \sum_{j=1}^{N_{NS}} e^{-F_{loop}(i_0,j)} = e^{-\beta \varepsilon_{rd}^S} e^{-\beta \varepsilon_{rd}^{NS}} e^{-\beta F_{loop}^{NS}}. \tag{7.22}$$

We are now ready to calculate the total partition function. We will consider the three states from figure 7.1. The weights corresponding to the first two states will be the same as in the LacI dimer case. The third state corresponds to the partition function term we just calculated. The total partition function is then

$$Z_{total}(P,R) = Z^{NS}(P,R) + Z^{NS}(P-1,R) e^{-\beta \varepsilon_{pd}^S} + Z^{NS}(P,R-1) \times Z_R^{O,NS}. \tag{7.23}$$

After rewriting these equations using equation 7.22, and using the weak promoter approximation we get a fold-change

$$\text{fold-change}(R) \simeq \left(1 + 2 \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}} \right)^{-1}. \tag{7.24}$$

The last term corresponds to having $R-1$ repressors in the reservoir and having 1 repressor with one head bound specifically. Even though the contribution from the non-specific loops just vanished, we see that there is a factor-of-two difference in front of the number of LacI tetramers. This is different from the fold-change in gene expression for dimers shown in equation 7.14. It can be easily understood if we think about the actual number of binding heads that are now present. In the case of dimers we have R_2 binding heads whereas for tetramers there are $2R_4$ binding heads inside the cell. As a result, no information about the non-specific looping background can be obtained by doing the experiment described in the main text. We see that as long as the number of binding heads is the same the fold-change will not vary. Interestingly, this is one of the conclusions from the data by Oehler et al. [34]. They compared repression for two different numbers of monomers of each kind of LacI, such that $2R_4 = R_2$. The fold-change in gene expression obtained for each monomer concentration is comparable for dimers and tetramers as long as this condition is met. An alternative way to look at this is by comparing the binding energies obtained for dimers and tetramers. These two set of energies, obtained from equations 7.16 and 7.24, are shown in table 7.1.

7.6.1.3 Connecting $\Delta\varepsilon_{rd}$ to K_d

We can also describe the fold-change in perhaps the more familiar language of dissociation constants [16]. We think of the two reactions shown in figure 7.9 where the DNA can either be bound by RNA polymerase or by Lac repressor. In steady state we can relate the concentrations of the different molecular players to the respective dissociation constants through

$$\frac{[P][D]}{[P-D]} = K_P, \tag{7.25}$$

and

$$\frac{[R][D]}{[R-D]} = K_d. \tag{7.26}$$

In these equations we have defined $[P]$ and $[R]$ as the concentrations of RNA polymerase and Lac repressor that are not bound to the promoter, respectively. The concentrations of their respective protein DNA complexes are $[P-D]$ and $[R-D]$. $[D]$ is the concentration of free DNA. Finally, K_P and K_d are the dissociation constants for RNA polymerase and repressor, respectively.

We want to determine p_{bound}, the probability of finding the promoter occupied by RNA polymerase. This can also be expressed as the fraction of DNA molecules occupied by RNA polymerase and given by

$$p_{bound} = \frac{[P-D]}{[D]+[R-D]+[P-D]}. \tag{7.27}$$

If we divide by $[D]$ and use equations 7.25 and 7.26 we arrive at

$$p_{bound} = \frac{[P]/K_P}{1+[R]/K_d+[P]/K_P}. \tag{7.28}$$

By comparing this expression to, for example, equation 7.3 we can relate the repressor binding energy $\Delta\varepsilon_{rd}$ to the tetramer dissociation constant through

$$\frac{[R]}{K_d} = \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_{rd}}. \tag{7.29}$$

Throughout the text we express the binding energies also in the language of dissociation constants. To do this we assume an *E. coli* volume of 1 fl such that a repressor per cell corresponds to a concentration of 1.7 nM.

7.6.1.4 Weak promoter approximation for the lacUV5 promoter

A key assumption leading to the simple expression for the fold-change in gene expression from equation 8.2 is that the weight corresponding to RNA polymerase being bound to the promoter is much smaller than one, meaning that the promoter will be unoccupied. Mathematically, we express this as $\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_{pd}} \ll 1$.

Following [15] we can write the binding energy as

$$\Delta\varepsilon_{pd} = \varepsilon_{pd}^{S} - \varepsilon_{pd}^{NS} = \frac{K_d^S}{K_d^{NS}}, \tag{7.30}$$

where K_d^S and K_d^{NS} are the dissociation constants of RNA polymerase to specific and non-specific DNA, respectively. *In vitro* values for the non-specific dissociation constant are about $K_d^{NS} = 10,000$ nM [58], whereas the specific dissociation constant for the *lacUV5* promoter has been measured to be $K_d^S = 6$ nM [59] and 80 nM [60]. This corresponds to a binding energy range between -4.8 and -7.4 $k_{\mathrm{B}}T$. In exponentially-growing *E. coli* there are around 500 σ^{70} RNA polymerase molecules available [55]. this results in a range for the factor $\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_{pd}}$ of 0.01–0.16. Therefore we conclude that not neglecting the term corresponding to RNA polymerase binding to the promoter from our expression for the fold-change would only result in a small correction at the most.

7.6.2 Sensitivity of the predictions

In this work we used the data by Oehler et al. to obtain the binding energies which in turn were used to generate predictions. This was done because we intended to test the predictions generated by the thermodynamic model. On the other hand, we could combine all our available data for the fold-change in gene expression with the corresponding data on the number of Lac repressor in each strain in order to obtain the best possible estimate for the Lac repressor binding energies. The corresponding fit and resulting energies are shown in figure 7.15.

In order to get a better sense for how well this fit was constraining the values of the binding energies we wished to analyzed the “sensitivity” of the fit. In order to do this we plotted the data corresponding to the binding site O1 and overlaid it with curves for the fold-change in gene expression where we have chosen different values for the binding energy. In figure 7.10 we show the data for the O1 binding site together with its best fit and several other curves with different choices of the binding energy. It is clear from this figure that the fit is constraining the value of the binding energy relatively well (within less than 1 k_BT) and that the error in the parameter resulting from the fit captures this.

7.6.3 Repression for strains RBS1 and 1I

In the main text we hint multiple times at a slight discrepancy between our theoretical predictions and the results measured for the fold-change in strains RBS1 and 1I. We do not believe that this discrepancy is due to a problem with the determination of the concentration of Lac repressor because we were able to reliably detect higher and lower concentrations of the purified standard than those corresponding to these two strains. Another alternative is that we didn’t quantify the level of gene expression correctly. Indeed, the measurements for Oid correspond to the lowest levels of gene expression quantified in this work. For example, could there be some constant transcription level or “leakiness” that cannot be repressed by Lac repressor? However, the shift is also present in the other operators where the levels of gene expression are such that a constant "leakiness" would have a negligible effect. Additionally, the measurements of these two strains for all other operators are well between the range of the rest of the data which shows no such systematic shift. We are then forced to conclude that the discrepancy, if real and not just an unfortunate experimental systematic error unaccounted for, is due to the fact that these strains have a much higher level of Lac repressor. This line of logic would lead us to conclude that affinity of Lac repressors to DNA can somehow be affected if its intracellular number is too high. However, further experimentation will be necessary in order to confirm this assertion.

7.6.4 Accounting for leakiness

One interesting property of equation 8.2 is that it predicts that the fold-change in gene expression will go down indefinitely as the number of repressors is increased. However, at some point one would expect to have some constant level that is, in principle, independent of any regulation. This is called "leakiness" and is usually attributed to transcription that is independent of the promoter of interest. Such non-desired transcription could stem, for example, from RNA polymerase escaping from a nearby promoter and generating a transcript.

We wish to determine if our results are being contaminated by such leakiness and if so, what its effect on the estimation of the binding energies would be. The smallest absolute value of LacZ activity detected in our strains corresponds to binding site Oid in strain 1I. This combination has an activity of about 1 MU. This activity level sets a bound on the maximum value of the leakiness: since we can measure activities down to 1 MU the leakiness cannot be any higher than that and, in the worst possible case, it would be equal to 1 MU.

The fold-change in gene expression was calculated throughout this work using the following formula

$$\text{fold-change} = \frac{\text{expression}(R \neq 0)}{\text{expression}(R = 0)}. \tag{7.31}$$

However, if there was leakiness in our measurements this would mean that we are overestimating the expression measurements. If *leak* corresponds to the value of this leakiness then the corrected fold-change in gene expression is

$$\text{fold-change} = \frac{\text{expression}(R \neq 0) - leak}{\text{expression}(R = 0) - leak}. \tag{7.32}$$

Here we have made the implicit assumption that the leakage does not depend on the presence of Lac repressor. Correcting our measurements for leakiness would then result in lower values of the fold-change. In order to determine how much of a difference this correction could make to our calculation of the binding energies we performed an analysis analogous to the one shown in figure 7.15 for different proposed values of leakiness ranging between 0 and 1 MU. The results of these different fits are shown in figure 7.11(A). It is clear from this figure that there would not be a significant change in the binding energies for any of the considered values of leakiness. In figure 7.11 we show the relative change in binding energy between the worst case scenario (leakiness of 1 MU) and the case where we do not correct for leakiness. It is clear that even in this extreme case the corrections to the binding energies are negligible. We conclude that leakiness, if present, would not be affecting our results in any measurable way.

7.6.5 Supplementary materials and methods

7.6.5.1 Plasmids

Plasmid pZS22-YFP was kindly provided by Michael Elowitz. The EYFP gene comes from plasmid pDH5 (University of Washington Yeast Resource Center [11]). The main features of the pZ plasmids are located between unique restriction sites [46]. The sequence corresponding to the lacUV5 promoter [61] between positions -36 and +21 was synthesized from DNA oligos and placed between the EcoRI and XhoI sites of pZS22-YFP in order to create pZS25O1+11-YFP. Note that we follow the notation of Lutz and Bujard [46] and assign the promoter number 5 to the lacUV5 promoter. The O1 binding site in pZS25O1+11-YFP was changed to O2, O3 and to Oid using Site Directed Mutagenesis (Quikchange II, Stratagene), resulting in pZS25O2+11-YFP, pZS25O3+11-YFP and pZS25Oid+11-YFP. These plasmids are shown diagrammatically together with the promoter sequence in figure 7.12.

The *lacZ* gene was cloned from *E. coli* between the KpnI and HindIII sites of all the single-site constructs mentioned in the previous paragraph. The O2 binding site inside the *lacZ* coding region was deleted without changing the LacZ protein [33] using Site Directed Mutagenesis. Successful mutagenesis was confirmed by sequencing the new constructs around the mutagenized area.

After we had generated these constructs and integrated them on the *E. coli* chromosome we determined that the different LacZ constructs had acquired some mutations. On average there were three different point mutations in each construct, though pZS25O3+11-lacZ lost both the KpnI and HindIII sites. All these constructs still expressed functional LacZ. This problem did not present itself in the case of the YFP constructs. We attribute this higher number of mutations in part to possible problems in the PCR amplification of the *lacZ* coding region.

Every time the fold-change in gene expression is calculated the expression of a strain is normalized by the expression of another strain bearing the exact same mRNA sequence. Therefore, we do not believe that the different mRNA sequences and potential different absolute LacZ activities have a considerable effect on the fold-change. This is in part also supported by the fact that our experimental data and theoretical predictions match reasonably well. If there is an effect on the fold-change due to the differences in the coding region it seems to be of the same magnitude as the experimental error.

A construct bearing the same antibiotic resistance, but no reporter, was created by deleting YFP from one of our previous constructs. This construct serves for determining the spontaneous hydrolysis or background of our enzymatic measurements.

Plasmid pZS21-lacI was kindly provided by Michael Elowitz. This plasmid has kanamycin resistance. The chloramphenicol resistance gene flanked by FLIP recombinase sites was obtained by PCR from plasmid pKD35. The insert was placed between the SacI and AatII sites of pZS21-lacI to generate pZS3*1-lacI. For this work we wished to have additional concentrations than those provided by pZS3*1-lacI, for which we mutated the ribosomal binding regions. These new ribosomal binding regions were designed using a recently developed thermodynamic model of translation initiation [38]. First, the original RBS ("WT") was deleted using Site Directed Mutagenesis (Quikchange II, Stratagene) using primer 15.29 and its reverse complementary. This primer deleted the sequence between the EcoRI site and the transcription start. From here we proceeded to add new ribosomal binding sequences by mutagenesis using primers 15.2, 15.31, 15.37 and 15.39. All the primer sequences are shown in table 7.2. These primers gave rise to new ribosomal binding regions named RBS1, RBS446, RBS1027 and RBS1147.

7.6.5.2 Strains

Chromosomal integrations were performed using recombineering [48]. Primers used for these integrations are shown in table 7.4. The reporter constructs were integrated into the galK region [28] of strain HG105 using primers HG6.1 and HG6.3. Note that our reporter gene was integrated in the opposite direction of the neighboring in order to avoid spurious read through of the LacZ coding region by RNA polymerase molecules transcribing from nearby promoters. Constructs expressing Lac repressor with the different RBS were integrated into the phage-associated protein *ybcN* [62] using primers HG11.1 and HG11.3.

This resulted in strains HG105::ybcn<>3*1-lacI, HG105::ybcn<>3*1RBS1-lacI, HG105::ybcn<>3*1RBS446-lacI, HG105::ybcn<>3*1RBS1027-lacI and HG105::ybcn<>3*1RBS1147-lacI. For simplicity we call these strains 1I, RBS1, RBS446, RBS1027 and RBS1147, respectively. The reporter constructs were then combined with the different strains expressing varying amounts of Lac repressor using P1 transduction (openwetware.org/wiki/Sauer:P1vir_phage_transduction). All integrations and transductions were confirmed by PCR amplification of the replaced chromosomal region and by sequencing.

7.6.5.3 β-galactosidase assay

Our protocol for measuring LacZ activity is basically the one described in [49, 50] with some slight modifications as follows. A volume of the cells between 2.5 μl and 200 μl was added to Z-buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol, pH 7.0) for a total volume of 1 ml. The volume of cells was chosen such that the yellow color would develop in no less than 15 minutes. For the case of the no-reporter constructs 200 μl of cell culture was used. Additionally, we included a blank sample with 1 ml of Z-buffer. The whole assay was performed in 1.5 ml Eppendorf tubes.

In order to lyse the cells, 25 μl of 0.1% SDS and 50 μl of chloroform were added and the mixture was vortexed for 10 s. Finally, 200 μl of 4 mg/ml 2-Nitrophenyl β-D-galactopyranoside (ONPG) in Z-buffer were added to the solution and its color, related to the concentration of the product ONP, monitored visually. Once enough yellow developed in a tube the reaction was stopped by adding 200 μl of 2. 5M Na_2CO_3 instead of adding 500 μl of a 1 M solution as done in other protocols. At this point the tubes were spun down at

$> 13,000$ g for three minutes in order to reduce the contribution of cell debris to the measurement.

200 μl of solution was read for OD420 and OD550 on a Tecan Safire2 and blanked using the Z-buffer sample. The OD600 of 200 μl of each culture was read with the same instrument. The absolute activity of LacZ was measured in Miller units using the formula

$$\mathrm{MU} = 1000 \frac{\mathrm{OD}_{420} - 1.75 \times \mathrm{OD}_{550}}{t \times v \times \mathrm{OD}_{600}} 0.826, \tag{7.33}$$

where t is the reaction time in minutes and v is the volume of cells used in ml. The factor of 0.826 is not present in the usual formula used to calculate Miller units. It is related to using 200 μl Na_2CO_3 as opposed to 500 μl. When using 500 μl, the final volume of the reaction is 1.725 ml (1ml Z-buffer, 25 μl 0.01% SDS, 200 l ONPG, 500 μl Na_2CO_3). However, when using only 200 μl of Na_2CO_3 the total volume is 1.425 ml. The factor of 0.826 adjusts for the difference in concentration of ONP.

All reactions were performed at room temperature. No significant difference in activity was observed with respect to performing the assay at 25C in an incubator.

7.6.5.4 Measuring *in vivo* Lac repressor concentration

The Lac repressor purification protocol used in this work is an adaptation of the one published in [63]. The strains to be assayed were first grown to saturation in LB + 20 μg/ml of chloramphenicol. They were then diluted 1:40,000 into 50 ml of M9 minimal medium + 0.5% glucose and grown to an OD600 of approximately 0.6. Cells were spun down (6,000g for 10 minutes) and resuspended in 36 μl of breaking buffer (BB, 0.2 M Tris-HCl, 0.2 M KCl, 0.01 M magnesium acetate, 5% glucose, 0.3 mM DTT, 50 μg/L PMSF, 50 mg/100 ml lysozyme, pH 7.6) per ml of culture and per OD. Typically, around 45 ml of culture would be spun down and resuspended in 900 μl of BB. Cells were slowly frozen by placing them at -20C, after which they were slowly thawed on ice. At this point 4 μl of a 2000 Kunitz/ml DNase solution (Sigma) and 40 μl of a 1 M $MgCl_2$ solution were added and the samples were incubated at 4C with mixing for 4 hours. Samples were frozen, thawed and incubated with mixing at 4C two more times after which they were spun down at 15,000 g for 45 minutes. At this point the supernatant was obtained and its volume measured. The pellet was subsequently resuspended with 900 μl of BB and spun down again. This will serve as a control that most Lac repressor was in the original supernatant. The luminescence of these sample resuspensions were compared with respect to the luminescence of the samples corresponding to the first spin. On average, the resuspension signal would be about 12% of the first spin signal. However, some samples showed signals as high as 35%. We chose to discard any data coming from samples showing a resuspension signal higher than 20%.

Additionally to the cell lysates calibration samples were prepared before performing a measurement. Purified Lac repressor (courtesy of Stephanie Johnson) was diluted into lysate of strain HG105 to different concentrations. The concentration of purified repressor in our stock solution was determined by spectroscopy using the available extinction coefficient [64]. In order to have all samples within the dynamic range of our methods (see below) cell lysates corresponding to strains 1I and RBS1 were diluted 1:8 in HG105 lysate.

A nitrocellulose membrane was prewetted in TBS (20 mM Tris-HCl, 500 mM NaCl, pH7.5) for 10 minutes and then left to air dry. After loading the samples the immunoblots were blocked using blocking solution which consists of 5% dry milk and 2% BSA in TBST (20 mM Tris Base, 140 mM NaCl, 0.1% Tween 20, pH 7.6) with mixing at room temperature for one hour. After that the membrane was incubated in 1:1000 dilution of Anti-LacI monoclonal antibody (from mouse, Millipore) in blocking solution at 4C overnight. The membrane was subsequently incubated in a 1:2000 dilution of HRP-linked anti-mouse secondary antibody (GE Healthcare) for one hour at room temperature. Finally, the membrane was washed by incubating in TBST for 5 minutes twice and by a final incubation of 30 minutes.

As described in the text, we obtain the total luminescence corresponding to each spot using Matlab image analysis custom code. This information is stored in a matrix $Lum(x,y)$, where the coordinates on the membrane are given by x and y. The values corresponding to the HG105 blank sample are them fitted to a 2nd-degree 2D polynomial. This polynomial can be represented as $Background(x,y)$. Finally, we can also fit such a polynomial to the luminescence of the samples corresponding to strain 1I. This results in the polynomial $1I(x,y)$. In figure 7.3(C) we plot the polynomial $1I(x,y) - Background(x,y)$. The normalized luminescence matrix is then calculated in the following way

$$Lum_{norm}(x,y) = \frac{Lum(x,y) - Background(x,y)}{1I(x,y) - Background(x,y)}. \tag{7.34}$$

All further analysis is then done on the normalized matrix $Lum_{norm}(x,y)$.

The calibration standards are fitted to a power law $\text{LacI}_{\text{lum}} = A \times \text{LacI}_{\text{mass}}^{B} + C$, where LacI_{lum} is the luminescence collected from the spots on the membrane and $\text{LacI}_{\text{mass}}$ is their corresponding masses. We are interested in obtaining an interpolation between the calibration samples in order to get an estimate of the amount of Lac repressor loaded in each spot on the membrane. Therefore, we perform the fit on only the calibration data that is directly in the range of our unknown samples, as shown by the calibration line in figure 7.3(D).

Once the amount of Lac repressor in each spot was obtained the corresponding number of Lac repressors per cell were calculated. As an example, we will consider the case where there is one repressor tetramer per cell and estimate the expected amount of repressor on the membrane. We typically start with a 45 ml culture at an OD600 of 0.6. This, in turn, is concentrated down to 900 μl after the purification process. 2 μl of these concentrated cells are loaded on the membrane. In this case, we can now calculate the amount loaded on the membrane resulting in

$$\begin{aligned} N_{\text{cells loaded}} &= \underbrace{0.8\times 10^{9}\text{cells/ml}}_{\text{OD600 to cell density calibration}} \times \underbrace{0.6}_{\text{OD600}} \times \\ & \underbrace{45\text{ ml}}_{\text{culture volume}} \times \underbrace{\frac{2\ \mu\text{l}}{900\ \mu\text{l}}}_{\text{final purified volume and amount loaded}} \\ &= 48\times 10^{6}\text{ cells.} \end{aligned} \tag{7.35}$$

The calibration of OD600 to cell density was performed by plating serial dilutions of a culture at a known OD600 and counting colonies. The molecular weight of a tetramer is 154kDa. This results in a mass of around 12 pg in a spot. Of course, there is an uncertainty associated with this calculate of the number of cell loaded which will propagate into the measurement of the number of repressors per cell. However, this uncertainty stems from errors in measuring volumes and in calibrating the OD600 readings and are no larger than 5 to 10 %. On the other hand, the day-to-day variation of the reading were on the order of 20 to 30 %. As a result we chose to report only the day-to-day variation as our error in the measurement of the intracellular concentration of Lac repressor.

7.6.6 Supplementary figures and tables

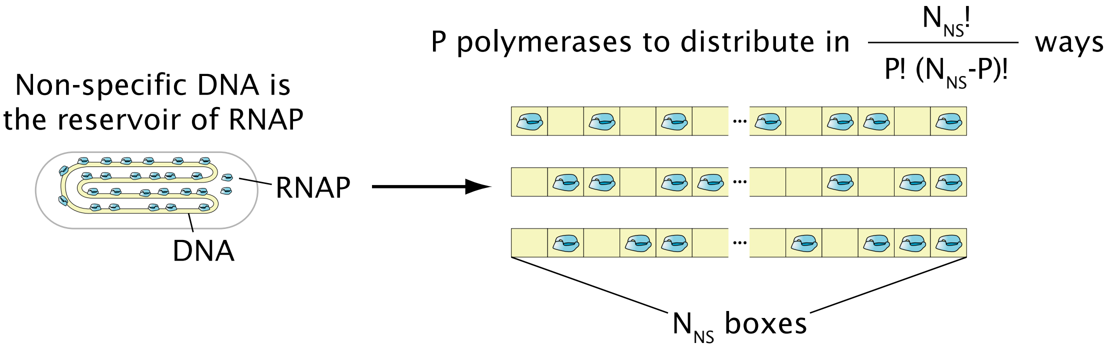

P polymerases to distribute in $\frac{N_{NS}!}{P!\,(N_{NS}-P)!}$ ways
Non-specific DNA is the reservoir of RNAP
RNAP
DNA
N_{NS} boxes

Figure 7.6: Model for the RNA polymerase reservoir. The non-specific sites on the genome are assumed to be the reservoir for RNAP. Different arrangements of RNAP on this reservoir are shown.

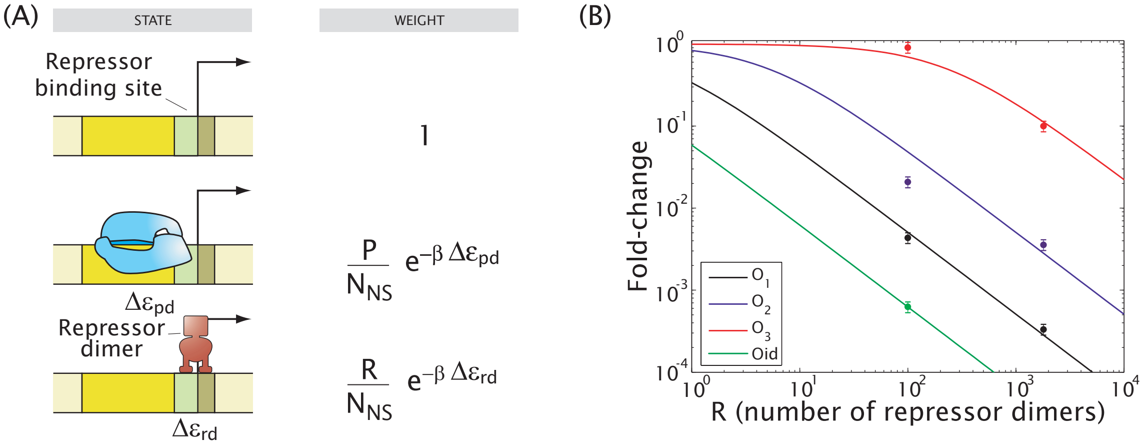


Figure 7.7: Single-site repression by LacI dimers. (A) Schematic listing of the different states and their respective weights when RNAP and the dimeric repressor have overlapping sites. (B) Repression for four different strengths of the main repressor binding site (Om) as a function of the number of dimers inside the cell. The binding energy of dimeric Lac repressor to each operator is calculated by fitting each data set to the repression expression from equation 7.16 and presented in table 7.1.

Table 7.1: Single-site binding energies for repressor dimers and tetramers. The energies are obtained using the data by Oehler et al. [34] and equations 7.16 and 8.2 for the dimers and tetramers, respectively. The error bars are calculated assuming an error in the fold-change measurement of 30%.

Operator	Dimers (k_BT)	Tetramers (k_BT)
Oid	-18.2 ± 0.3	-17.7 ± 0.3
O_1	-16.1 ± 0.2	-16.2 ± 0.1
O_2	-13.7 ± 0.5	-13.7 ± 0.1
O_3	-10.0 ± 0.4	-10.4 ± 0.4

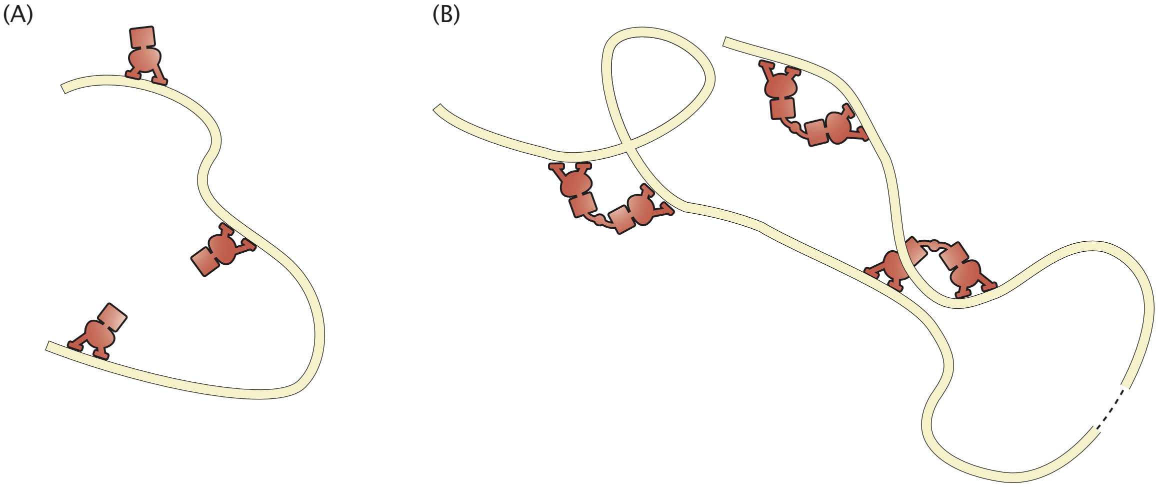


Figure 7.8: Model for the non-specific looping background. Possible states of non-specific DNA bound by Lac repressor. (A) Dimers will explore all available non-specific sites. (B) Tetramers explore all possible loops between non-specific sites.

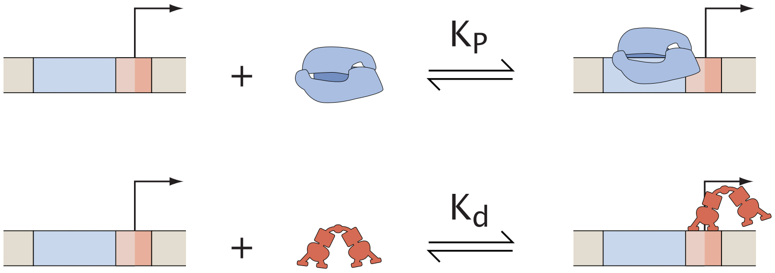


Figure 7.9: Repression as a set of chemical reactions. The two reactions involved in regulation by simple repression are shown. K_P and K_d are dissociation constants. These reactions are also described by equations 7.25 and 7.26.

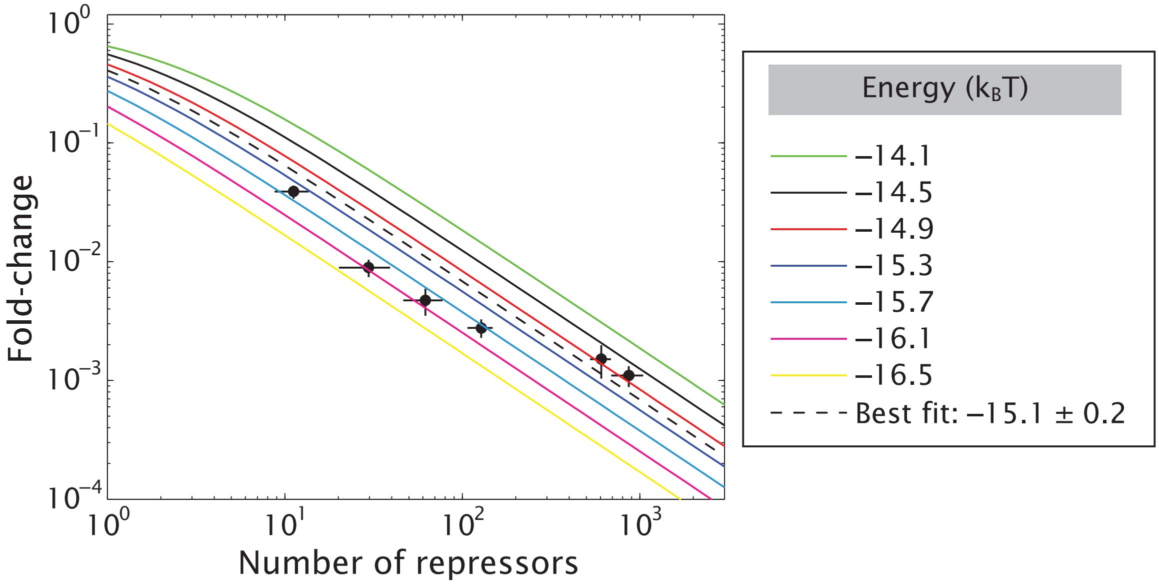
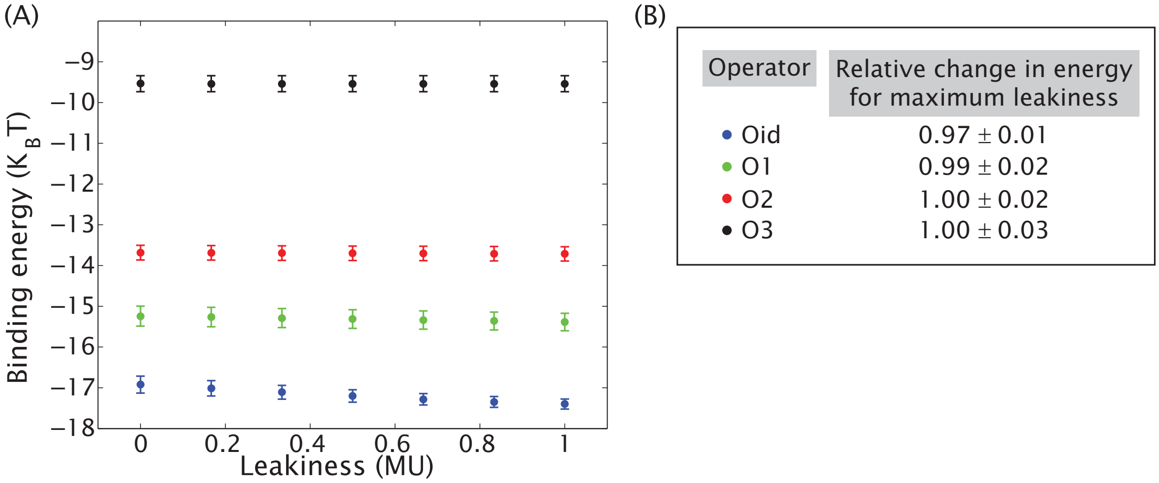


Figure 7.10: Sensitivity in the determination of the binding energies. The data for binding site O1 is shown with its best fit along with several other choices of the binding energy parameter which reveal how the positions of the curves depend upon this choice. Visual inspection of the curves constrains the value of the binding energy to within less than 1 k_BT of the fit value.



Operator	Relative change in energy for maximum leakiness
● Oid	0.97 ± 0.01
● O1	0.99 ± 0.02
● O2	1.00 ± 0.02
● O3	1.00 ± 0.03

Figure 7.11: Potential effects of leakiness on the calculation of binding energies. (A) A variable leakiness in the level of gene expression was assumed and the fold-change in gene expression was reanalyzed using equation 7.32. The resulting binding energies are shown as a function of the assumed leakiness. (B) Relative change in binding energies for each operator corresponding to the case without any assumed leakiness and to the worst possible leakiness of 1 MU.

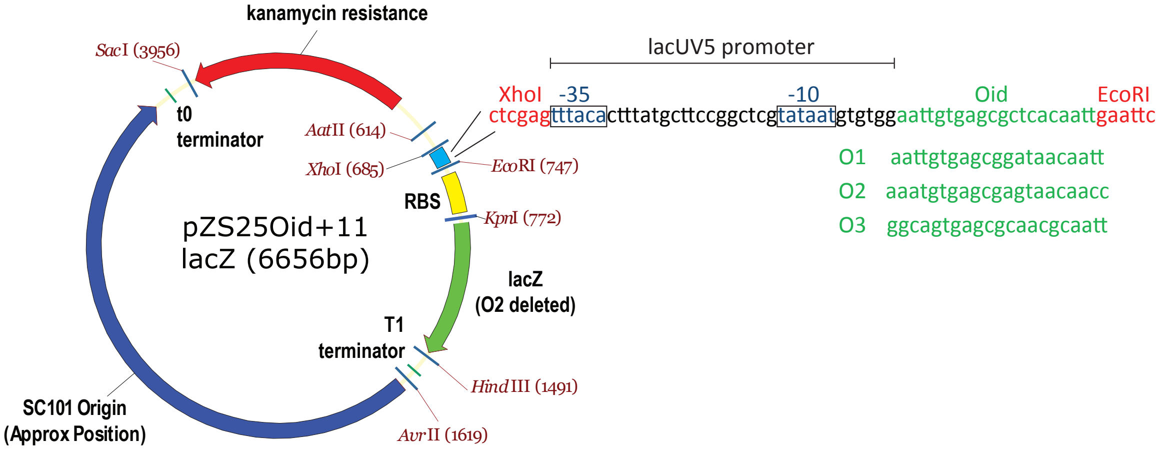


Figure 7.12: Plasmid diagram and promoter sequence. The main features of the plasmid pZS25Oid+11-lacZ are shown flanked by unique restriction sites (the features are not to scale). The particular promoter sequence based on the lacUV5 promoter is shown together with the sequences of the different Lac repressor binding sites used.

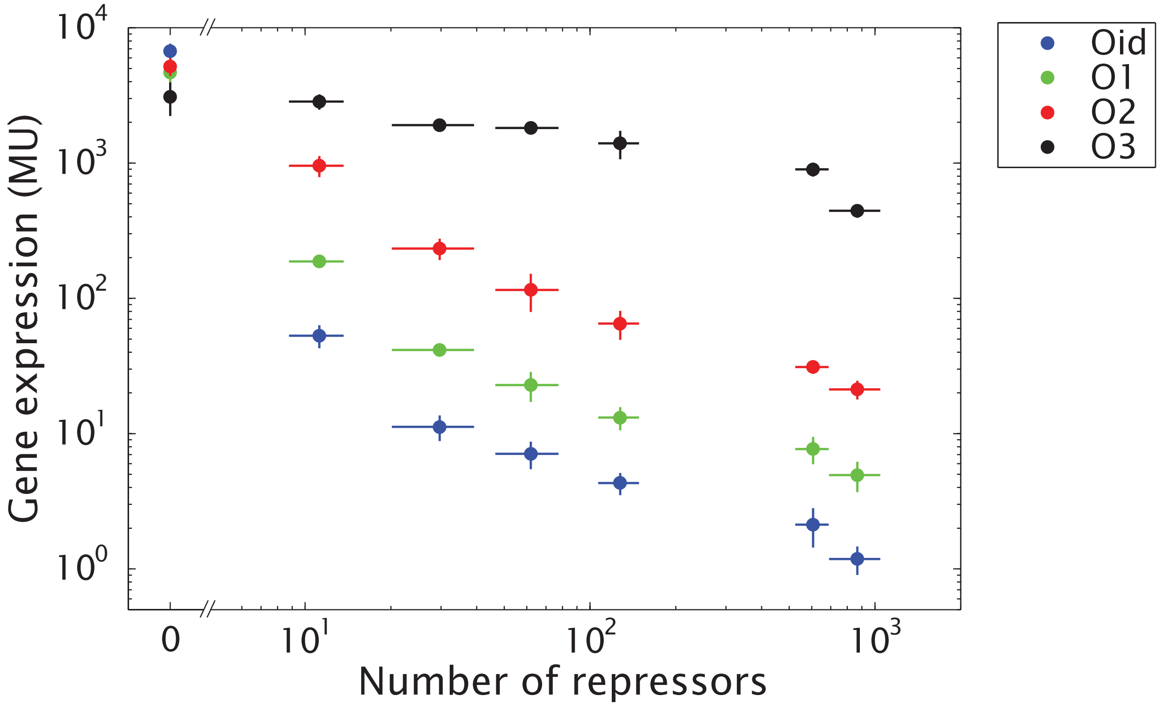


Figure 7.13: Average absolute levels of expression. The absolute levels of expression corresponding to our different constructs in the different strain backgrounds are shown in Miller units. By the ratio of the activity of a given construct in a given strain with respect to the activity of the same construct in strain HG105 we calculate the fold-change in gene expression. Note that throughout the repression values correspond to the average of the repression measured on different days. In this case we plot the average of the absolute expression of each strain and construct over different days. The error bars correspond to the standard deviation of the repeats.

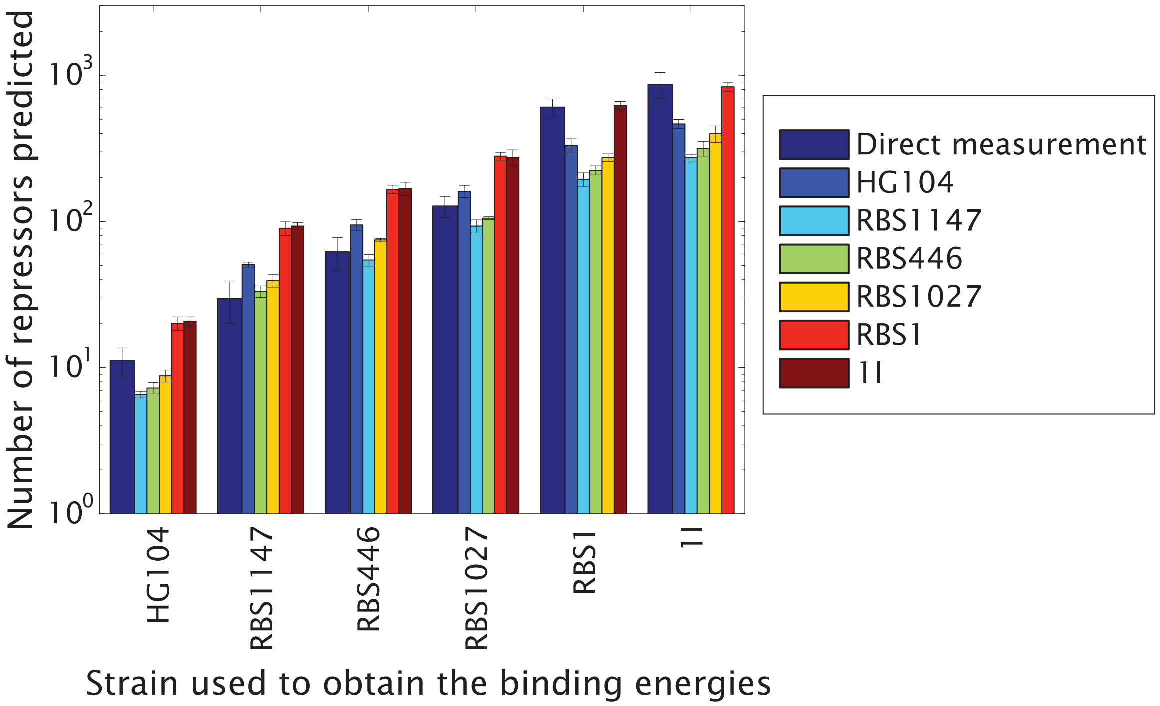


Figure 7.14: Different ways of calculating the binding energies give comparable predictions. For each strain noted by a group of bars the binding energies were obtained by taking the number of repressors obtained through immunoblots as a given and combining this with the fold-change measurements for the same strain. With these binding energies we predict the number of repressors for all the remaining strains. For comparison, the actual direct measurement done using immunoblots is also included

Table 7.2: Primers sequences. These primers and their respective reverse complement were used to modify the RBS of the different constructs. The inserted RBS regions are denoted by capitalized bases.

Primer number and name	Sequence	Description
15.29-RBSDelete	gacgcactgaccgaattcatggtgaatgtgaaaccag	Delete the RBS from pZS3*1-lacI
15.2-tetR-RBS1	cgcactgaccgaattcattaaagaTTT gaaaggtaccatatggtg	
15.31-RBS446	cgcactgaccgaattc TCTAGACAGTATAGAGTAGAGAGACTAA atggtgaatgtgaaac	
15.37-RBS1027	cgcactgaccgaattc TCTAGATATTTAAGAGGACAATACTGG atggtgaatgtgaaac	
15.39-RBS1147	cgcactgaccgaattc TCCCCACATTAAACAGGGAAGACTGG atggtgaatgtgaaac	

Table 7.3: List of *E. coli* strains used throughout this experiment. Chromosomal positions correspond to the sequence in GenBank accession no. U00096.

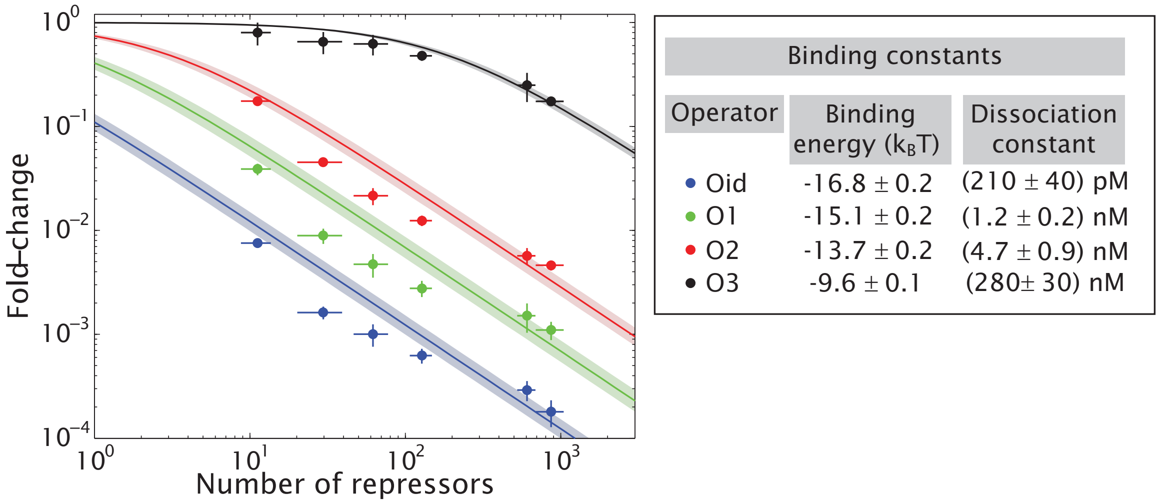
Strain	**Genotype**	**Derived from**	**Comment**
HG104	$\Delta lacZYA$	MG1655	Deletion from 360,483 to 365,579
HG105	$\Delta lacZYA$, $\Delta lacI$	MG1655	Deletion from 360,483 to 366,637

Table 7.4: Primers used throughout this work. For integration primers, lowercase indicates the portion of the primer that is homologous to the *E. coli* gene where the integration is made and uppercase indicates primer homology to the plasmid where PCR was carried out. Chromosomal positions correspond to the sequence in GenBank accession no. U00096.

Primer	Sequence	Comment
HG6.1	gtttgcgcgcagtcagcgatatccattttcgcgaatccggagtg taagaaACTAGCAACACCAGAACAGCC	Integration of the *lacZ* reporter constructs into the *galK* gene between positions 1,504,078 and 1,505,112.
HG6.3	ttcatattgttcagcgacagcttgctgtacggcaggcaccagct cttccgGGCTAATGCACCCAGTAAGG	
HG11.1	acctctgcggaggggaagcgtgaacctctcacaagacggcatca aattacACTAGCAACACCAGAACAGCC	Integration of lacI constructs into the *ybcN* gene between positions 1,287,628 and 1,288,047.
HG11.3	ctgtagatgtgtccgttcatgacacgaataagcggtgtagccat tacgccGGCTAATGCACCCAGTAAGG	

Table 7.5: Predicted and measured strength of the different ribosomal binding sequences used to generate constitutive levels of Lac repressor. The ribosomal binding sequence denoted as "WT" corresponds to the original found in pZS3*1-lacI [46]. The measured strength corresponds to the resulting level of repressor once these constructs are integrated on the chromosome. The predicted strengths are calculated form [38]. Both the predicted and measured strengths are normalized by this RBS.

RBS	Normalized predicted strength (au)	Normalized measured strength (repressors/cell)
"WT"	1	1 ± 0.3
RBS1	0.88	0.7 ± 0.2
R1027	0.58	0.6 ± 0.1
R446	0.25	0.25 ± 0.07
R1147	0.64	0.64 ± 0.03



Binding constants

Operator	Binding energy (k_BT)	Dissociation constant
● Oid	-16.8 ± 0.2	(210 ± 40) pM
● O1	-15.1 ± 0.2	(1.2 ± 0.2) nM
● O2	-13.7 ± 0.2	(4.7 ± 0.9) nM
● O3	-9.6 ± 0.1	(280± 30) nM

Figure 7.15: Obtaining the binding energies. Using all measurements of the fold-change in gene expression with their corresponding repressor concentration we fit equation 8.2 to obtain the best possible estimate for the binding energies. The results of the fits are expressed in units of $k_{\mathrm{B}}T$.

Bibliography

[1] H. G. Garcia and R. Phillips. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A*, 2011. (*Under review*).

[2] D. Endy. Foundations for engineering biology. *Nature*, 438(7067):449–53, 2005.

[3] C. A. Voigt. Genetic parts to program bacteria. *Curr Opin Biotechnol*, 17(5):548–57, 2006.

[4] S. Ben-Tabou De-Leon and E. H. Davidson. Gene regulation: Gene control network in development. *Annu Rev Biophys Biomol Struct*, 36:191, 2007.

[5] T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek. Probing the limits to positional information. *Cell*, 130(1):153–64, 2007.

[6] R. S. Cox Iii, M. G. Surette, and M. B. Elowitz. Programming gene expression with combinatorial promoters. *Mol Syst Biol*, 3:145, 2007.

[7] I. S. Peter and E. H. Davidson. Modularity and design principles in the sea urchin embryo gene regulatory network. *FEBS Lett*, 583(24):3948–58, 2009.

[8] M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–8, 2000.

[9] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–8, 2002.

[10] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci U S A*, 100(13):7702–7, 2003.

[11] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–5, 2005.

[12] T. Kuhlman, Z. Zhang, M. H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *escherichia coli*. *Proc Natl Acad Sci U S A*, 104(14):6043–8, 2007.

[13] S. Kaplan, A. Bren, A. Zaslaver, E. Dekel, and U. Alon. Diverse two-dimensional input functions control bacterial sugar genes. *Mol Cell*, 29(6):786–92, 2008.

[14] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003.

[15] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[16] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[17] S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Trevino-Quintanilla, and J. Collado-Vides. Regulondb (version 6.0): Gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic Acids Res*, 36:D120–4, 2008.

[18] U. Alon. *An introduction to systems biology: Design principles of biological circuits.* Chapman & hall/crc mathematical and computational biology series. Chapman & Hall/CRC, Boca Raton, FL, 2007.

[19] B. Alberts. *Molecular biology of the cell.* Garland Science, New York, 5th edition, 2008.

[20] J. Q. Wu and T. D. Pollard. Counting cytokinesis proteins globally and locally in fission yeast. *Science*, 310(5746):310–4, 2005.

[21] S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–41, 2003.

[22] S. Takamori, M. Holt, K. Stenius, E. A. Lemke, M. Gronborg, D. Riedel, H. Urlaub, S. Schenck, B. Brugger, P. Ringler, S. A. Muller, B. Rammner, F. Grater, J. S. Hub, B. L. De Groot, G. Mieskes, Y. Moriyama, J. Klingauf, H. Grubmuller, J. Heuser, F. Wieland, and R. Jahn. Molecular anatomy of a trafficking organelle. *Cell*, 127(4):831–46, 2006.

[23] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1):117–24, 2007.

[24] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying *e. Coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329:533, 2010.

[25] J. Malmstrom, M. Beck, A. Schmidt, V. Lange, E. W. Deutsch, and R. Aebersold. Proteome-wide cellular protein concentrations of the human pathogen leptospira interrogans. *Nature*, 460(7256):762–5, 2009.

[26] A. E. Mayo, Y. Setty, S. Shavit, A. Zaslaver, and U. Alon. Plasticity of the cis-regulatory input function of a gene. *PLoS Biol*, 4(4):e45, 2006.

[27] N. J. Guido, X. Wang, D. Adalsteinsson, D. Mcmillen, J. Hasty, C. R. Cantor, T. C. Elston, and J. J. Collins. A bottom-up approach to gene regulation. *Nature*, 439(7078):856–60, 2006.

[28] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002.

[29] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.

[30] D. Zenklusen, D. R. Larson, and R. H. Singer. Single-rna counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*, 15(12):1263–71, 2008.

[31] L. Cai, N. Friedman, and X. S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62, 2006.

[32] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–6, 2008.

[33] S. Oehler, E. R. Eismann, H. Kramer, and B. MÜLler-Hill. The three operators of the lac operon cooperate in repression. *EMBO J*, 9(4):973–9, 1990.

[34] S. Oehler, M. Amouyal, P. Kolkhof, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. Quality and position of the three *lac* operators of *e. Coli* define efficiency of repression. *EMBO J*, 13(14):3348–55, 1994.

[35] G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A*, 79(4):1129–33, 1982.

[36] H. G. Garcia, J. Boedicker, L. Bintu, P. Wiggins, J. Kondev, and R. Phillips. DNA looping and gene regulation: The physics of biological action at a distance. 2011. (*In preparation*).

[37] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[38] H. M. Salis, E. A. Mirsky, and C. A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol*, 27(10):946–50, 2009.

[39] R. B. Winter and P. H. Von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. the *escherichia coli* repressor–operator interaction: Equilibrium measurements. *Biochemistry*, 20(24):6948–60, 1981.

[40] W. T. Hsieh, P. A. Whitson, K. S. Matthews, and R. D. Wells. Influence of sequence and distance between two operators on interaction with the *lac* repressor. *J Biol Chem*, 262(30):14583–91, 1987.

[41] W. Gilbert and B. Muller-Hill. Isolation of the lac repressor. *Proc Natl Acad Sci U S A*, 56(6):1891–1898, 1966.

[42] M. Lanzer and H. Bujard. Promoters largely determine the efficiency of repressor action. *Proc Natl Acad Sci U S A*, 85(23):8973–7, 1988.

[43] S. J. Elledge and R. W. Davis. Position and density effects on repression by stationary and mobile DNA-binding proteins. *Genes Dev*, 3(2):185–97, 1989.

[44] S. Ryu, N. Fujita, A. Ishihama, and S. Adhya. Galr-mediated repression and activation of hybrid lacuv5 promoter: Differential contacts with rna polymerase. *Gene*, 223(1–2):235–45, 1998.

[45] J. MÜLler, S. Oehler, and B. MÜLler-Hill. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol*, 257(1):21–9, 1996.

[46] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in *escherichia coli* via the lacr/o, the tetr/o and arac/i1-i2 regulatory elements. *Nucleic Acids Res*, 25(6):1203–10, 1997.

[47] K. A. Datsenko and B. L. Wanner. One-step inactivation of chromosomal genes in *escherichia coli* k-12 using pcr products. *Proc Natl Acad Sci U S A*, 97(12):6640–5, 2000.

[48] S. K. Sharan, L. C. Thomason, S. G. Kuznetsov, and D. L. Court. Recombineering: A homologous recombination-based method of genetic engineering. *Nat Protoc*, 4(2):206–23, 2009.

[49] J. H. Miller. *Experiments in molecular genetics.* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1972.

[50] N. A. Becker, J. D. Kahn, and L. J. Maher Iii. Bacterial repression loops require enhanced DNA flexibility. *J Mol Biol*, 349(4):716–30, 2005.

[51] W. Runzi and H. Matzura. *in vivo* distribution of ribonucleic acid polymerase between cytoplasm and nucleoid in escherichia coli. *J Bacteriol*, 125(3):1237–9, 1976.

[52] P. H. Von Hippel, A. Revzin, C. A. Gross, and A. C. Wang. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: i. the *lac* operon: Equilibrium aspects. *Proc Natl Acad Sci U S A*, 71(12):4808–12, 1974.

[53] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'conner, D. W. Noble, and P. H. Von Hippel. Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound *escherichia coli lac* repressor *in vivo.* *Proc Natl Acad Sci U S A*, 74(10):4228–32, 1977.

[54] U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc Natl Acad Sci U S A*, 99(19):12015–20, 2002.

[55] M. Jishage and A. Ishihama. Regulation of rna polymerase sigma subunit synthesis in *escherichia coli*: Intracellular levels of sigma 70 and sigma 38. *J Bacteriol*, 177(23):6832–5, 1995.

[56] J. K. Barry and K. S. Matthews. Thermodynamic analysis of unfolding and dissociation in lactose repressor protein. *Biochemistry*, 38(20):6520–8, 1999.

[57] P. J. Schlax, M. W. Capp, and M. T. J. Record. Inhibition of transcription initiation by lac repressor. *J Mol Biol*, 245(4):331–50, 1995.

[58] M. T. J. Record, W. Reznikoff, M. Craig, K. Mcquade, and P. Schlax. Escherichia coli rna polymerase (sigma70) promoters and the kinetics of the steps of transcription initiation. In F. C. Neidhardt, R. CURTISS III, J. L. INGRAHAM, E. C. C. LIN, K. BROOKS LOW, B. Magasanik, W. S. REZNIKOPP, M. Riley, M. SCHAECHTER, and H. E. UMBARGER, editors, *In escherichia coli and salmonella cellular and molecular biology*, pages 792–821. ASM Press, Washington, DC, 1996.

[59] H. Buc and W. R. Mcclure. Kinetics of open complex formation between escherichia coli rna polymerase and the lac uv5 promoter. Evidence for a sequential mechanism involving three steps. *Biochemistry*, 24(11):2712–23, 1985.

[60] D. L. Matlock and T. Heyduk. A real-time fluorescence method to monitor the melting of duplex DNA during transcription initiation by rna polymerase. *Anal Biochem*, 270(1):140–7, 1999.

[61] B. MÜLler-Hill. *The lac operon: A short history of a genetic paradigm.* Walter de Gruyter, Berlin, New York, 1996.

[62] G. Posfai, G. Plunkett 3rd, T. Feher, D. Frisch, G. M. Keil, K. Umenhoffer, V. Kolisnychenko, B. Stahl, S. S. Sharma, M. De Arruda, V. Burland, S. W. Harcum, and F. R. Blattner. Emergent properties of reduced-genome *escherichia coli*. *Science*, 312(5776):1044–6, 2006.

[63] J. Xu and K. S. Matthews. Flexibility in the inducer binding region is crucial for allostery in the escherichia coli lactose repressor. *Biochemistry*, 48(22):4988–98, 2009.

[64] A. P. Butler, A. Revzin, and P. H. Von Hippel. Molecular parameters characterizing the interaction of escherichia coli lac repressor with non-operator DNA and inducer. *Biochemistry*, 16(22):4757–68, 1977.

Chapter 8

Using Fluctuations to Characterize a Regulatory Network at the Single-Cell Level

This chapter is the reproduction of a current paper draft in its very early stages. As a result it should be viewed as a snapshot of our current efforts.

Recent developments in the analysis of gene expression have made it possible to query distributions in the level of gene expression rather than just population averages on their means. In this paper, we use fluctuations in partitioning during cell division as a way to explore both the means and variability in gene expression for a ubiquitous regulatory motif found in bacteria, namely simple repression. One of the outcomes of this method is the ability to identify the usually unknown calibration factor linking fluorescence units to repressor number. This allows us, in turn, to directly count transcription factors. With the repressor count in hand, it is then possible to make a systematic characterization of the governing equation resulting from thermodynamic and stochastic models of transcription for several different regulatory architectures. Using thermodynamic models we show that the *in vivo* binding energies of the repressor to DNA are consistent with previous measurements performed in bulk. Additionally, we qualitatively confirm predictions based on simple stochastic model that state that the variance in simple repression scales with the strength of the binding of the repressor to its operator DNA.

8.1 Introduction

Regulatory biology remains one of the most fertile areas for the quantitative dissection of biological systems, with two broad classes of examples coming from the study of cell signaling and gene regulation [1–5]. With increasing regularity, these systems are examined in tandem using both theoretical ideas with precise "governing equations" that are thought to constitute a predictive first approximation to their behavior, and using precision measurements whose ambition is to explicitly test the validity of these models. In the context of bacterial chemotaxis, there has been a long tradition of this direct interplay between theory and experiment

[6–9]. Similarly, the study of gene expression in bacteria has enjoyed a close interplay between the so-called thermodynamic models which predict the mean level of expression as a function of architectural parameters characterizing the regulatory motif of interest, and quantitative measurements which can now be performed at the single-cell level [5, 10–14].

One of the key tunable parameters in the governing equations describing different regulatory motifs is the number of copies of the transcription factors in question. For example, in the context of the simple repression regulatory motif considered here, there is a simple linear relationship between the repression and the number of repressors. As such, knowledge of the number of transcription factors is a key prerequisite for testing models of the regulatory response. However, at best, it is an extremely laborious process to construct the different strains that make it possible to tune the repressor concentration and to perform the measurements yielding the actual counts of transcription factors [14–18]. In an earlier paper, we have adopted this strategy permitting an examination of the level of expression over more than four orders of magnitude in fold-change and nearly two orders of magnitude in repressor number [14]. It is extremely appealing to have alternative methods that do not require new strain construction for every measurement and even more importantly, if the quantitative dissection of biological systems is to be put on a solid footing, different methods must yield the same results in a reproducible fashion.

In our earlier dissection of the simple repression regulatory motif, we resorted to an approach using quantitative immunoblots to measure the number of repressors [14]. However, like with all methods, there are always associated uncertainties and experimental limitations. One such limitation is the fact that suitable antibodies with high enough affinity are needed to carry out the detection of the proteins of interest. In immunoblots the amount of repressor in a cell lysate is quantified. However, it is not necessarily true that all repressors inside the lysed cells are in the lysate itself. Some of the protein may be left behind in the fractions that lead to the the cell extract. These issues would result in an underestimate of the amount of protein that is actually inside the cell resulting in the conclusion that immunoblots probably provide a lower bound on the total intracellular number. As a result, we are interested in a completely independent means of characterizing the same regulatory motif.

An extremely intriguing alternative strategy was recently developed that is based upon the clever use of fluctuations in the partitioning of transcription factors during cell division [17, 19]. It is this method that serves as the centerpiece of the present work. There is a great tradition of the use of fluctuations as the basis of key measurements. At the end of his book “Atoms”, Jean Perrin provides a table of more than 15 independent ways to determine Avogadro’s number and many of them are based on exploiting fluctuations [20]. Further, one of the points made by Perrin in his analysis of these independent means of determining Avogadro’s number is that together, they provided great confidence in the underlying hypothesis of atomism. Of course, biology has a similar tradition of the use of fluctuations for the purposes of exploring processes such as bacterial evolution as exemplified in the Luria-Delbruck experiment [21].

The idea of the fluctuation-based counting methods is to explore the asymmetries in partitioning of the molecules or molecular complexes of interest during the cell division process. In the simplest scenario, it is *assumed* that the partitioning between daughter cells is random, corresponding effectively to each molecule making a coin flip. As a result, the distribution of transcription factors should be binomial. This simple fact alone is enough to determine the unknown calibration factor, α, relating the total fluorescence intensity of a cell, I_{tot}, to the number of fluorescently labelled transcription factors it harbors, N_{tot}, through the expression $I_{\text{tot}} = \alpha N_{\text{tot}}$. In particular, by observing the fluorescence of the two daughters, captured in the quantities I_1 and I_2, it can be shown that

$$\langle (I_1 - I_2)^2 \rangle = \alpha I_{\text{tot}}. \tag{8.1}$$

This relation follows from the properties of the binomial distribution as shown in the Supplementary Information ("Derivation of calibration factor") and features the unknown parameter α which characterizes the number of photons emitted per fluorophore.

Of course, the use of a method like this carries with it a number of intrinsic assumptions about both the cellular processes in question and the nature of our measurements. In particular, the method relies on the assumption that the fluctuations due to partitioning are the dominant source of the partitioning error as revealed by differences in fluorescence intensity of the two daughter cells. However, the method will fail when the variations coming from other sources are larger than the fluctuations due to the partitioning [17–19]. One such source is the experimental uncertainty in quantifying a total level of fluorescence within a cell. In addition, proteins which live in the cytoplasm are influenced by variations in relative daughter cell size [18], cells with a larger volume are proportionally more likely to inherit proteins from the parent cell. However, for transcription factors which spend most of their time bound to the chromosome, the effective partitioning should be identical in every daughter because of the equal segregation of DNA between them. Lac repressor, has been shown in various occasions to reside mostly on the DNA [22–25]. Nevertheless, demonstrating the independence of fluorescence fluctuations from volume fluctuations is an essential control when assuming fair binomial partitioning. We present this control in figure 8.9.

Additionally, equation 8.1 only applies if there is no new production of the repressor in question over the cell cycle. If this assumption is not met, the calibration can still be obtained, but the formula is different, as discussed in the Results section. Finally, the method assumes that all repressor molecules are labeled with a bright fluorescent protein. This is certainly not the case as fluorescent proteins can be misfolded giving rise to a dark species. We expect this to occur in approximately 20% of the molecules resulting in an underestimation of the number of repressors [26].

Interestingly, exploring lineages even when the partitioning is not random is useful. In such cases, as shown for example in a recent paper describing the partitioning of carboxysomes during division of cyanobacteria, the distribution reveals an active mechanism of partitioning of the macromolecular assemblies of interest [27].

Thermodynamic models of transcriptional regulation have proven a valuable tool when dissecting gene regulatory motifs [10, 12–14, 28]. They predict the relative change in mean gene expression levels as a function of the concentrations of the relevant transcription factors and their binding energies to DNA and their interaction energies with each other. However, these analyses of the mean expression level only give information about binding energies or dissociation constants. They do not provide any information about the microscopic rates involved in the transcription process.

Recently developed experimental techniques now make it possible to query the levels of gene expression at the single-cell level. Such measurements can be performed either by reading out the level of mRNA or proteins and sometimes are even carried out with single molecule resolution [17, 29–38]. Measurement of the higher moments of the distribution of expression levels opens the door to a dissection of regulatory networks in terms of their rates [32, 34, 37, 39, 40].

When dissecting gene regulatory networks in terms of the higher moments of the protein distribution it is important to know what the sources of variability in protein production are. For example, when quantifying the variability of a gene regulatory input-output relation as a function of the concentration of the relevant transcription factor it is important to decouple the inherent variability in transcriptional regulation from the variability in concentration of the transcription factor of interest over a cell population. As such, the method presented here has the advantage that the rate of production is measured for a known concentration of repressor at the single-cell level. This results in the elimination of a significant component of the "extrinsic" contribution to the noise. One of our aims in this paper is to go beyond the earlier use of the fluctuation method to actually shed light on the correspondence between recent theoretical predictions about the noise and measurements where the promoter architecture is varied systematically by changing the affinity of the repressor operator DNA sequence.

8.2 Results

8.2.1 Introduction to the circuit and the dilution method

Many measurements of gene regulation in bacteria are based on growing macroscopic cultures under conditions where it is certain that there is exponential growth [14–16, 28, 41]. Once the cells have reached this stage of growth, they are queried for their level of expression. On the other hand, the measurements advocated here consist in taking a movie of an *E. coli* microcolony as it grows under a microscope [42]. As a result, it is not clear that such proper environmental conditions exist when taking movies of growing bacteria, and more importantly, cells themselves are subjected to a wide range of conditions and densities and it is important to develop a picture of the level of expression under all such conditions. In this paper, we begin to explore these questions by comparing the behavior of networks of interest following both protocols for growing cells and to explore the relation between their repression values.

One of the motifs that we examine is shown in figure 8.1(A). In particular, the motif corresponds to the so-called "simple repression" regulatory motif, with the addition that the repressor which carries out

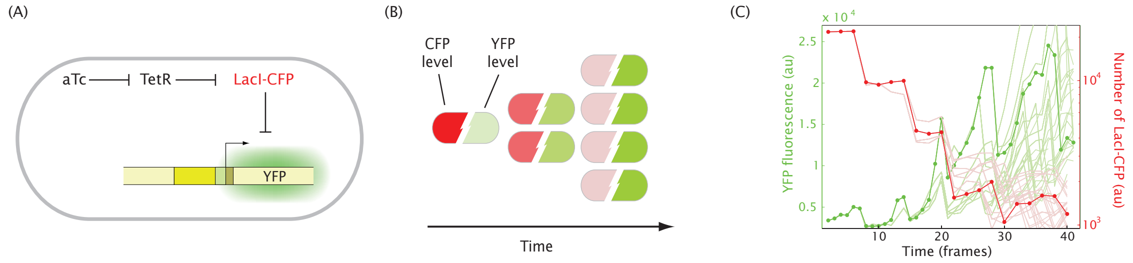


Figure 8.1: Dilution circuit. (A) In the presence of aTc, Tet repressor cannot repress the production of LacI-CFP resulting in high levels of repressor and low levels of the downstream gene YFP in the cell. (B) Once aTc is washed away, production of new repressor is turned off. As the cells divide, the LacI-CFP proteins get diluted resulting in an ever increasing production rate of YFP. (C) Fluorescence trace over time for a lineage of bacteria. As expected from the circuit in which the Lac repressors are diluted out over cell cycles, fluorescence of LacI-CFP (red) decreases at each cell division. At the same time, expression of the regulated YFP (green) increases indicated by the increasing slope. At each cell division two new lineages are created corresponding to each of the daughter cell. These new parallel lineages are shown using dimmer colors in the background.

this repression (Lac repressor) can itself be controlled by the presence or absence of an inducer (aTc) which controls a second repressor (Tet repressor). The steady-state response of this circuit characterized using traditional methods during exponential growth is reported in the Supplementary Information ("Steady-state characterization of the regulatory circuit"). These controls are useful in establishing the correct function of the regulatory circuit.

With these controls in hand, we turn to the analysis of the movies themselves. In this case, individual bacteria are followed through subsequent divisions and the level of fluorescence in two different fluorescent channels are monitored simultaneously as shown in figure 8.1(B,C). As expected, as the level of repressor is diluted through sequential cell divisions, its ability to repress the production of downstream target genes is reduced. As a result, there is an ever increasing rate of production of the product gene.

8.2.2 Calibrating the total number of repressors per cell

By far the most common way to elicit a change in the regulatory response of a genetic circuit is to resort to small inducer molecules that are added to the media and interact with the transcription factors of interest [28, 43–45]. However, such approaches add an extra layer of complexity to the analysis of regulatory motifs because we now need to account both for the transport of the inducer into the cell and for the interaction of the inducer with the transcription factor. Finally, we also require a quantification of how the activity of the transcription factor depends upon its binding to these inducers [28, 45, 46].

An alternative scenario is to tune the *number* of transcription factors rather than their activity. This approach gives direct access to parameters such as the binding energies of transcription factors to their operators and their interaction energies with each other [10, 12, 13]. We recently employed such an approach to dissect the same regulatory network [14]. The strategy employed in this case to control the number of repressor molecules was to generate a library of strains with different constant constitutive levels of Lac repressor. In this case the number of Lac repressors was quantified using quantitative immunoblotting.

Of course, for this approach to work one needs to be able to both control the number of transcription factors and to measure how many of them are present. Measuring the absolute concentration of transcription factors through immunoblotting depends upon the availability of suitable antibodies. Alternatively, the transcription factor of interest can be fused to a protein sequence or tag for which there is an antibody available [47–50]. Still, this technique is inherently a bulk technique and does not allow for protein quantification at the single-cell level.

One way around this is through fusions of the transcription factor to a fluorescent protein. In principle, this can permit monitoring of the intracellular concentration at the single-cell level. In this case, the fluorescence needs to be calibrated in order to convert arbitrary units on a camera into an absolute count of molecules. In this context one common approach for calibrating absolute number of fluorescent molecules has been to measure the mean intensity of a single copy of the fluorescent molecule [37, 38, 51] or of a bulk solution of purified fluorophore [6, 38, 52–54]. In this work we use instead the recently developed calibration method based on fluctuations in protein partitioning during cell division [17, 19]. The potential advantage of this method over those described above is that the calibration is performed in the same experiment. As a result there is no cross-calibration to be done with respect to an independent measurement of a sample that serves as a standard of concentration.

One potential caveat of this approach is that the fusion to a fluorescent protein might affect the function of the transcription factor in question. Unfortunately, there is no straightforward way to determine if this is the case, short of performing some sort of control using the same transcription factor but in the absence of the fluorescent fusion. We have recently published a dissection of the same regulatory network using wild-type transcription factor without any modifications such as fusions [14]. Repeating this protocol with our fluorescent fusion would allow us to directly compare the absolute number of the two different species of repressor.

With these caveats in mind, once we have obtained a series of image sequences and performed the relevant lineage identification, we can plot the fluorescence intensity of the mother cell during each cell division and the difference in the intensities of the two daughters. Such data is shown in figure 8.2. As noted already in equation 8.1, through the accumulation of a collection of triplets $(I_1, I_2, I_{\mathrm{tot}})$, we can fit the unknown calibration factor α. In figure 8.2 we show the results for our calibration factor using this method. As can be seen in the figure, the individual cell divisions (corresponding to the data points) are subject to very large partitioning errors, precisely as would be found in direct simulations of this division process. On the other hand, experimental errors not related to the binomial partitioning such as, for example, the uncertainty in quantifying fluorescence can contaminate the calculation of the calibration.

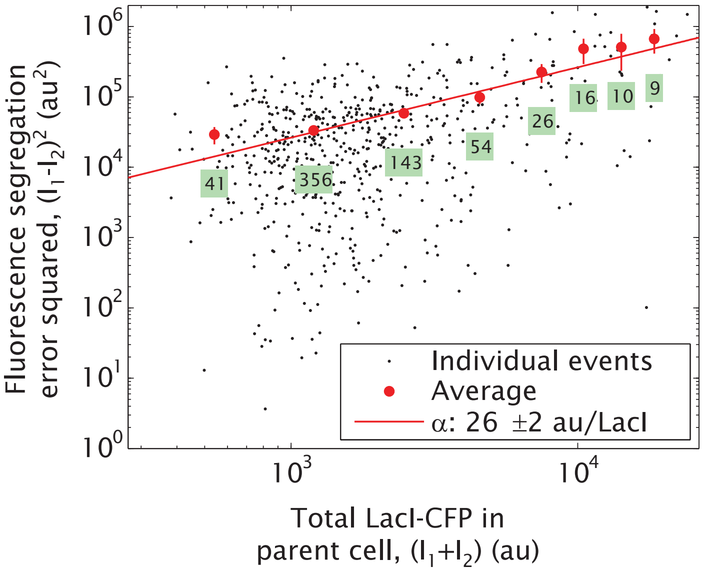


Figure 8.2: Calibration of fluorescence of LacI-CFP molecules. The square of the error in fluorescence partitioning between two daughter cells is plotted as a function of the fluorescence of the mother cell for a representative data set. Each black point represents a specific division event. We bin these points resulting in the red, averaged data points which are fitted to equation 8.1 in order to obtain the calibration factor α relating CFP fluorescence to the absolute number of LacI-CFP molecules inside the cell. The numbers below each red point correspond to the number of individual events that were used to calculate each average.

8.2.3 Measuring the single-cell input-output function

One of the main thrusts of our efforts has been to set up measurements such that it is possible to test the governing equations associated with a variety of different regulatory architectures. For the case of simple repression that we consider here, the fold-change in gene expression is given by

$$\text{fold-change} = \frac{1}{1+\frac{R}{N_{NS}}e^{-\Delta\varepsilon_{rd}/k_BT}}, \tag{8.2}$$

where R is the number of repressor dimers present in the cell, N_{NS} is the size of the non-specific reservoir (which we take here to be the whole *E. coli* chromosome such that $N_{NS} = 5 \times 10^6$) and $\Delta\varepsilon_{rd}$ is the binding energy of repressor to its operator. Experimentally it is defined as the ratio of the level of expression in the presence of repressor to the level of expression of the same regulatory architecture in the absence of the transcription factor such that

$$\text{fold-change} = \frac{\text{gene expression}(R \neq 0)}{\text{gene expression}(R = 0)}. \tag{8.3}$$

In earlier work, Oehler et al. [16] measured the repression for two different operator concentrations (though there were large uncertainties in these measurements). We complemented and improved those measurements by characterizing the regulatory output for several more discrete repressor concentrations [14]. Here, we use the dilution method as a way to more or less continuously titrate the number of repressors and thereby to explore the simple repression function over a reasonable fraction of its full dynamic range.

The fold-change in gene expression shown in equation 8.3 is defined in the context of steady-state measurements, where the mean fluorescence per cell is quantified for strains with and without repressor. However, measurements such as those shown in figure 8.1(C) give a rate of increase in the regulated gene (YFP) as a function of the amount of LacI-CFP in the cell. Over a cell cycle the steady-state level, $\langle \mathrm{YFP} \rangle$, and the average rate of production of YFP, r_{YFP} are related by

$$\langle \mathrm{YFP} \rangle = r_{\mathrm{YFP}}/\beta. \tag{8.4}$$

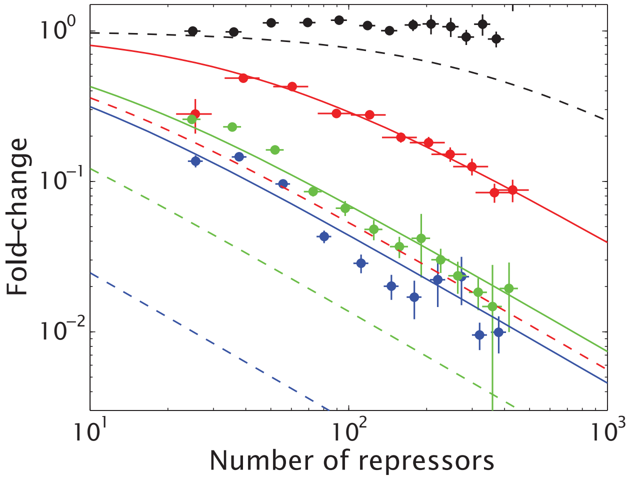
Here β is the decay rate of YFP which we assume to be constant and, due to the long lifetime of YFP, we assume it to be determined by the time for cell division. We see that the fold-change in mean level of YFP expression is equivalent to the fold-change in the rate of expression of YFP

$$\text{fold-change} = \frac{\langle \mathrm{YFP} \rangle (R \neq 0)}{\langle \mathrm{YFP} \rangle (R = 0)} = \frac{r_{\mathrm{YFP}}(R \neq 0)}{r_{\mathrm{YFP}}(R = 0)}. \tag{8.5}$$

This equivalence assumes that there is well-defined single rate of expression throughout the cell cycle. This is a helpful assumption when comparing single-cell measurements to bulk measurements as we will do in the remainder of this section. However, it is clear that at some point in the cell cycle the number of promoters inside the cell doubles due to the duplication of the genome [55]. As such a realistic analysis of the regulatory network will have to take this subtlety into consideration.

In figure 8.3 we show the mean fold-change in gene expression (calculated from the ratio of the rates of expression) as a function of the number of repressors for different realizations of the simple repression motif where the strength of the operator is changed systematically. The data corresponding to each construct can be fit to equation 8.2 in order to obtain an *in vivo* binding energy. Interestingly, we can compare the values obtained for the binding energies to our previous measurements obtained in bulk using immunoblotting to quantify the number of repressors per cell. The comparison between the results from both techniques reveals a non-negligible systematic difference between them of about 2 to 3 k_BT. Additionally, the binding site O3 appears to be too weak to exert any considerable repression. As a result we are unable to determine its binding energy.

One potential explanation for the systematic discrepancy between the two techniques could be related to an error in the calibration of the absolute fluorescence of a single LacI-CFP molecule. From equation 8.2 we see that a difference in energy between 2 and 3 k_BT would correspond to an underestimation of the number of repressor molecules by a factor of 7 to 20. Though there is an uncertainty associated with the calibration factor α from equation 8.1, we view it as unlikely that there would be such a systemic shift in the calibration. Another plausible explanation is that the binding activity of Lac repressor is quantitatively affected by the fact that it is fused to a fluorescent protein. This conclusion is also supported by the steady-state characterization of the regulatory network shown in the Supplementary Information



Binding energies (k_BT)		
Operator	From dilution ———	From immunoblots – – – –
● Oid	-13.90 ± 0.09	-16.8 ± 0.2
● O1	-13.42 ± 0.06	-15.1 ± 0.2
● O2	-11.72 ± 0.05	-13.7 ± 0.2
● O3	N/A	-9.6 ± 0.1

Figure 8.3: Repression as a function of concentration of LacI. By fitting the data corresponding to each construct to equation 8.2 we can obtain the corresponding binding energy. Those fits are denoted by solid lines and their results are shown in the table. In that table we also show the binding energies obtained recently by counting repressors using immunoblots in bulk [14] rather than fluctuations in single cells. Notice a systematic difference of about 2 to 3 k_BT between the two methods. The binding energy for O3 cannot be determined because there is not a significant repression over the range of repressor concentrations assayed.

("Steady-state characterization of the regulatory circuit"). Finally, it is also feasible that the binding energies obtained through immunoblots have been overestimated. This could occur if the measurement of the Lac repressor concentration done using immunoblots were only detecting a percentage of the total real amount of intracellular Lac repressor. As a result we would conclude that the few repressors that were detected repress very strongly, leading to the overestimation of the binding energies. All these different effects could certainly conspire to give the observed difference between the binding energies obtained through immunoblots and the fluctuation method.

8.2.4 Variability in expression

In the previous sections we saw how the fluctuation method is useful in providing an absolute count of the number of transcription factors. Such an absolute calibration is key for experimental efforts aimed at validating the input-output functions for the mean level of gene expression predicted by thermodynamic models [10, 12–14]. However, the circuit utilized in this work also presents a significant advantage when measuring higher moments of the protein distribution.

One strategy to quantify the variability in the protein expression is to measure the steady-state levels of gene expression of our network as a function of the concentration of aTc in the media. Each aTc concentration results in a different steady-state mean level of LacI-CFP in the cells. As a result of such an experiment we can quantify the variability in gene expression as a function of the concentration of repressor. However, such an approach has a potential inherent drawback related to the fact that there is variability in the expression of the repressor itself. Not only is each cell continuously producing LacI-CFP, but variation in this production rate is to be expected from cell to cell. As a result, the variation in the levels of the downstream YFP level is not only related to the inherent "intrinsic" stochasticity of the transcription process itself, but also to "extrinsic" factors related, among others, to fluctuations in the repressor concentration. An approach to decouple the intrinsic fluctuations from their extrinsic counterparts based on measuring correlations between identical constructs has been recently developed and used on many occasions [17, 30, 31, 40, 56].

An alternative scheme for decoupling the fluctuations in repressor concentration from fluctuations in the downstream production of YFP is to shut down the production of repressor such that its level stays constant over a cell cycle. This is precisely the scenario we have over the course of one of our movies. For each one of these two scenarios, the steady-state and the video microscopy measurement, we can calculate the fold-change in the variance with respect to a strain lacking repressor. For the first scenario, the fold-change will be calculated using the steady-state level of expression whereas for the second scenario the fold-change is calculated based on the rates of YFP expression as discussed previously in relation to equation 8.5. In figure 8.4 we show a comparison of the fold-change in the variance with respect to the fold-change in mean levels of gene expression for both approaches. We observe what seems to be a slight systematic shift of the data obtained through the steady-state approach or no more than 25 %. As a result we estimate that the contribution of fluctuations in repressor copy number in the steady-state approach to the "extrinsic" noise will be of order 25 %.

Though variability in gene expression has been widely regarded as a common strategy used in decision making in bacteria [57–61], to our knowledge there is only a very limited set of examples where the transcriptional noise has been systematically characterized as a function of the promoter regulatory architecture [37]. We perform such a characterization by measuring the fold-change in the variance as a function of the fold-change in mean gene expression for our different simple repression architectures. In figure 8.5(A) we show the results of such measurements, where a slight systematic effect can be appreciated. For the same fold-change in mean level the resulting fold-change in the variance is related to the strength of the binding site in question. Qualitatively we can account for this trend through the use of stochastic models of gene expression [39, 40], as shown in figure 8.5(B). The model used to generate the theoretical curve in figure 8.5(B) is extremely simple in its nature. It assumes that the rate of transcription and translation is constant when the repressor is not occupying its operator and that when it is bound transcription gets completely shut down. As such, it does not incorporate any "extrinsic" sources of noise that could lead to a variability in the rates of transcription and translation themselves. Another approximation in this model is that we assume that there is one copy of the promoter of interest inside the cell. This will certainly be true for the initial stages of the cell cycle, but once the chromosome is replicated the cell will have two promoters. Our failure to account for these issues implies that the comparison between theory and experiment must for now remain largely qualitative. Nevertheless, it is intriguing that the same trend seems to be observed in both the experimental data and the theoretical expectation.

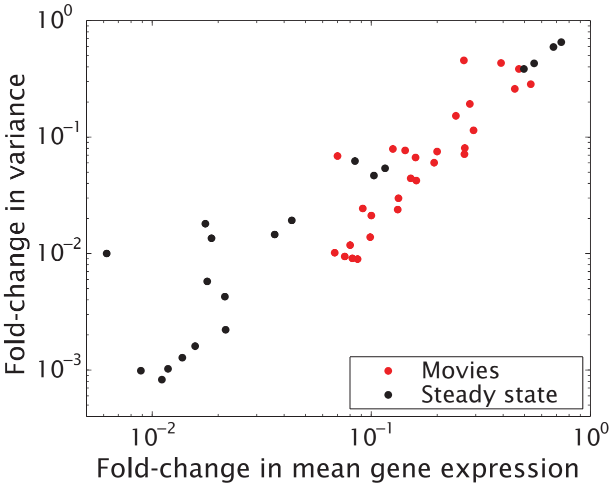


Figure 8.4: Contribution of fluctuations in repressor concentration to the steady-state variation of the network. The fold-change in the variance in the rate of expression measured with respect to a strain lacking LacI is plotted as a function of the fold-change in the mean of the rate of expression for data obtained through video microscopy. This is compared to the fold-change in variance in the steady-state levels of expression measured as a function of the fold-change in the mean steady-state level of gene expression. This comparison reveals that fluctuations in repressor can at most increase the variability in the downstream levels of YFP by 25 %.

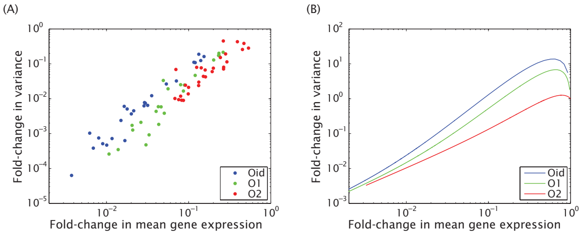


Figure 8.5: Noise in the simple repression motif. (A) The fold-change in variance in the rate of expression measured with respect to a strain lacking LacI is plotted as a function of the fold-change in the mean of the rate of expression for different operators. A small systematic difference between the operators is observed, where the strongest operator, Oid, has a larger fold-change in variance than the weaker O1 and O2 for the same fold-change in mean gene expression. (B) The qualitative trends observed in (A) can be reproduced using a stochastic model of transcriptional regulation. The parameters used for generating this plot are presented and discussed in [40].

8.2.5 Limits and generality of the method

Though the method described here is elegant, it also suffers from several key limitations. In particular, since the measurements depend upon tracking lineages of cell division there are bounds on both the maximal initial concentrations of transcription factors and on the limits of detectability as the number of repressors continue to decrease as a result of successive cell divisions. On the other hand, for high enough number of repressors, the fold-change in gene expression will be so low that the absolute rate of expression of the downstream YFP gene will be below the detection limit. These two limitations are illustrated in figure 8.6. In fact, as discussed in more detail in the Supplementary Information, these issues impact the range over which one can actually hope to use this method as a *general* tool for the quantitative analysis of gene expression.

In addition, it is rarely feasible to completely "shut off" production of the transcription factor during dilution. If production events are common between the division event and the subsequent fluorescence measurement, the measured fluorescence in the daughters is not entirely the result of binomial partitioning. We confirmed the existence of such "leakage" in the repression of the production of LacI-CFP by comparing the rate of YFP production in a strain without Lac repressor to the rate corresponding to our dilution circuit in the absence of the inducer aTc. We observed a small fold-change in gene expression which can be attributed to a production of 3 to 5 LacI-CFP molecules per cell cycle.

As a result of this production, the fluorescence in the daughters will be the result both of the partitioning and the production of new labeled proteins. If the variance associated with the production is relatively small (compared with the mean), the calibration factor takes a new form,

$$\left\langle (I_1 - I_2)^2 \right\rangle = I_{\mathrm{tot}} + 2(\sigma_{\mathrm{p}}^2 - < N_{\mathrm{p}} >), \tag{8.6}$$

where $< N_{\mathrm{p}} >$ and σ_{p}^2 are the first and second moments of the protein production distribution and full details can be found in the Supplementary Information ("Accounting for repressor production in the calculation of the calibration factor"). The linear relationship in equation 8.1 is preserved since I_{tot} still equals αN_{tot}, but now the y-intercept of the linear fit is related to the specific details of the production mechanism. If we fit the data in figure 8.2 with the new model posed by equation 8.6 allowing for a non-zero y-intercept we obtain $\alpha = (23 \pm 4)$ au/LacI and $2(\sigma_{\mathrm{p}}^2 - < N_{\mathrm{p}} >) = (7200 \pm 7900)$ au. Clearly, if there is a production of LacI-CFP this method is not able to detect it in a reliable manner.

8.3 Discussion

The advent of governing equations that are purported to describe both the mean levels of expression as well as the noise for a host of different regulatory architectures has placed new demands on the theory-experiment interplay in regulatory biology. An interesting case study to test these approaches is afforded by the simple repression motif in bacteria. By changing architectural parameters such as the binding affinity of repressors

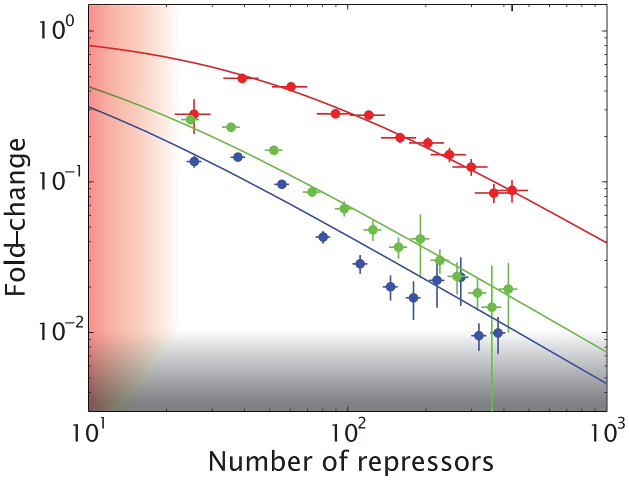


Figure 8.6: Limitations of the fluctuation method. Our method is limited in terms of the detection of the repressor fusion by the autofluorescence. This establishes a minimum number of repressors we can measure reliably denoted schematically by the red shaded region. When the number of repressors reaches a certain threshold the rate of expression of the downstream gene will be below what can be detected reliably, resulting in a maximum fold-change that can be measured. This limitation is denoted schematically by the gray shaded region. The data and fits correspond to those shown in figure 8.3.

for their DNA target sites, we have a family of predictions (and associated strains) which can be tested directly. One of the other key tuning variables in eliciting different regulatory responses is the number of repressors that mediate repression. To test these ideas, we need to explicitly count the number of repressors. In earlier work, we showed how quantitative immunoblots can be used to effect such a count. However, true confidence in these results requires independent measurements on the same regulatory motifs.

To that end, in this paper we have made systematic use of a recently introduced fluctuation method for taking the repressor census and thereby checking the governing equation for the simple repression regulatory motif. The outcome of this analysis is once again consistent with the governing equation for simple repression that emerges from the thermodynamic models. In particular, the scaling of the fold-change with repressor number appears to follow the trends dictated by that equation. On the other hand, there are also interesting numerical discrepancies between our two independent schemes for quantifying the simple repression motif. These discrepancies are manifested in the fact that the *in vivo* binding energies obtained using the fluctuation scheme are systematically shifted with respect to those obtained using immunoblots. One possible explanation for the discrepancy in binding energies could be that the binding of LacI-CFP is just different from the binding of wild-type Lac repressor. This by no means reduces the applicability of the methods presented here for dissecting gene regulatory networks at the single-cell level. However, is it important to keep in mind that we are potentially characterizing a transcription factor fusion that could have different properties from the its wild-type version.

We also go beyond previous applications of the fluctuation method to explore variability in gene expression. In particular, we characterize the noise in transcriptional regulation as a function of a systematic modification in the regulatory architecture through changes in the binding site strength. Though our simple model based on stochastic models of transcriptional regulation [39, 40] cannot yet account for the quantitative differences in the noise of each architecture, it predicts trends that are consistent with our measurements. To our knowledge, this is one of the first characterizations of transcriptional noise in bacteria to explore the effect of systematic variation in promoter architecture on the resulting noise profile. In order to achieve a quantitative understanding of the observed behavior the model will have to be improved to include the contribution of extrinsic noise and the variation in promoter copy number throughout the cell cycle. Alternatively, the contribution from extrinsic factors to the noise could be factored out by correlating fluctuations in two identical promoters each expressing a different color of fluorescent protein [30].

8.4 Materials and Methods

8.4.1 Plasmids and strains

For an independent control of the repression system, the wild-type *lac* operon, including the *lacI* gene controlling Lac repressor were deleted from wild-type *E. coli* (MG1655) to create HG105 as described in [14]. Plasmids pZS3*1-lacI expresses Lac repressor off of a $\mathrm{p_{LtetO\text{-}1}}$ promoter [62]. We follow the notation form Lutz and Bujard for the antibiotic resistance (3 corresponding to chloramphenicol resistance). However, in the 3* version we have replaced the original chloramphenicol resistance gene with the same resistance gene coming from plasmid pKD3 [63]. The advantage of the latter gene is that it is flanked by FRT sites which can be recognized by FLP recombinase in order to excise the resistance if needed. A lacI-CFP fusion was obtained from plasmid pLAU53 ([64], kindly provided by Paul Wiggins) and ligated into a pZS3* vector between its KpnI and HindIII restriction sites. This repressor cannot tetramerize due to the deletion of the last 11 amino acids of its sequence. Additionally, it has a C-terminal fusion to CFP.

Our simple repression constructs have been described elsewhere [14, 51]. In short, they consist of a *lacUV5* promoter with only one binding site. This binding site can either be O1, O2, O3 or Oid as shown in figure 8.10.

Explain the chromosomal integration. The YFP gene was chromosomally integrated by transduction with the P_{lacUV5} promoter and a single repressor binding site (operator O1, O2, O3, or Oid).

The TetR repressors that were responsible for the regulation of the LacI-CFP repressors were expressed from the plasmid pZS3*pN25-tetR. This consists in a $\mathrm{P_{N25}}$ promoter controlling the TetR gene and was obtained by PCR from the chromosome of DH5αZ1 [62].

8.4.2 Gene expression measurements

Steady-state measurements are performed as described in [51]. For the movies, cultures were grown overnight in 5 ml of LB in the presence of 20 μg/mL of chloramphenicol at 37°C and diluted 1 : 4000 in M9 + 0.5% glucose minimal media with 100 ng/mL anhydrotetracycline (aTc) to induce the production of LacI-CFP.

The diluted cultures were grown at 37°C until they reached an OD600 $\approx 0.6-0.8$ and then they were washed three times with in media to remove the inducer. They were then diluted to give about 1 cell per field of view when placed on a 1.5% low melting point M9+0.5% Glucose agar pad. Growth of cells was observed by fluorescence microscopy at 37°C. The cell doubling time was 67 min with a standard deviation of about 21 min. An automated Olympus fluorescent microscope was controlled by the software Micro-Manager, and multiple fields of view were recorded simultaneously. An exposure of 800 ms for both the CFP and YFP channels was used. A total of 40 frames of subsequent exposures were programmed to be taken at an interval of 12 minutes. Fluorescence images of the CFP channel were acquired only on alternate frames to reduce photobleaching.

8.4.3 Data analysis

Data analysis was performed using the Matlab code "schnitzcells" kindly provided my Michael Elowitz [17]. This code segments cells in a movie and tracks their lineages.

8.5 Supplementary Information

8.5.1 Derivation of calibration factor

It is of interest to have a simple derivation of the relation between fluorescence intensity and repressor number. In particular, to exploit the connection of partitioning to the binomial distribution, we need some key averages. For example, we have

$$< N_1 >= Np, \tag{8.7}$$

where p is the probability of a protein partitioning into cell 1, and which in the present context we assume has the value $p = 1/2$. By similar reasoning, we can arrive at

$$\mathrm{var}(N_1) = Np(1-p) \tag{8.8}$$

which can be used to rewrite this as

$$\mathrm{var}(N_1) =< N_1^2 > - < N_1 >^2 \Rightarrow < N_1^2 >= Np(1-p) + N^2p^2. \tag{8.9}$$

As a result of these expressions for the averages, we have

$$
\begin{aligned}
< (N_1 - N_2)^2 > &= 4 < N_1^2 > -4N < N_1 > +N^2 && (8.10)\\
&= 4(Np(1-p) + N^2p^2) - 4N(Np) + N^2 && (8.11)\\
&= 4\left(N\frac{1}{4} + N^2\frac{1}{4}\right) - 4N\left(N\frac{1}{2}\right) + N^2 && (8.12)\\
&= N. && (8.13)
\end{aligned}
$$

How can we use this to relate the measured fluorescence intensity of a cell to the number of fluorescent proteins in that cell? Let's assume that the intensity in a cell I can be written as $I = \alpha N$, where α is some calibration factor that converts from number of proteins to intensity. Now we can exploit the assertion that $I = \alpha N$ to write

$$
\begin{aligned}
< (N_1 - N_2)^2 >= N \Rightarrow < \left(\frac{I_1}{\alpha} - \frac{I_2}{\alpha}\right)^2 > &= \frac{I_{tot}}{\alpha} && (8.14)\\
\frac{1}{\alpha^2} < (I_1 - I_2)^2 > &= \frac{I_{tot}}{\alpha} && (8.15)\\
\Rightarrow \sqrt{< (I_1 - I_2)^2 >} &= \sqrt{\alpha I_{tot}}. && (8.16)
\end{aligned}
$$

This gives us the relationship between the fluctuations in the difference between the intensities of two daughter cells and the total intensity in the original mother cell, $I_{tot} = I_1 + I_2$. We can determine the unknown calibration factor α by taking time-lapse movies of dividing bacteria, tracing lineages to determine which pairs of daughter cells came from which mother cells, and for each mother+daughters set plotting $\sqrt{< (I_1 - I_2)^2 >}$ vs. I_{tot}, the intensity of the mother cell.

8.5.2 Steady-state characterization of the regulatory circuit

Though the method described here is predicated upon the measurement of the level of gene expression as a function of time by tracking cell lineages, we first wanted to ensure that the relation between inducer concentrations and the fluorescence in the two channels corresponded to our expectations. Here we characterize the steady-state behavior of the regulatory circuit. The simple repression construct characterized here corresponds to having the binding site O2.

Steady-state measurements are performed as described in the Materials and Methods section by growing the cells at different concentrations of aTc over multiple generations. The resulting profile of the level of LacI-CFP as a function of the concentration of aTc is shown in figure 8.7(A). The level of LacI-CFP determines the level of YFP, which is shown in figure 8.7(B) as a function of the aTc concentration.

One of the most convenient ways to characterize the quantitative response of a circuit like that used here is by appealing to the fold-change or the repression. Effectively, these quantities provide a measure of the extent to which the transcription factor of interest, in this case Lac repressor, alter the expression profile of the circuit. In particular, the fold-change in gene expression is computed as the ratio of the level of expression in the presence of the repressor of interest to its level in the absence of the repressor. Since our readout of the state of the circuit is in terms of fluorescence, this ratio is constructed as

$$\text{Fold-change} = \frac{\text{Fluorescence of YFP([aTc])}}{\text{Fluorescence of YFP}(R=0)}. \tag{8.17}$$

In this equation we are dividing the level of YFP expression for a given concentration of aTc by the fluorescence of a strain bearing no LacI-CFP. By using the data shown in figure 8.7 and measuring an additional strain without LacI-CFP we can then plot the fold-change in gene expression as a function of the LacI-CFP concentration measured in arbitrary fluorescence units. The results are shown in figure 8.8.

From the governing equation for the simple repression regulatory circuit [12, 13], we expect that the fold-change in gene expression will be described by the equation

$$\text{Fold-change} = \left[1 + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_{rd}}\right]^{-1}, \tag{8.18}$$

where R is the number of LacI-CFP molecules per cell and N_{NS} is the number of non-specific binding sites available to the repressor. We take this to be the size of the *E. coli* genome of 5×10^6 bp. $\Delta\varepsilon_{rd}$ is the *in vivo* binding energy of Lac repressor to its operator DNA. In a recent study we determined this *in vivo* binding energy of wild-type Lac repressor to O2 to be $(-13.7 \pm 0.2)\ k_B T$.

The functional form predicted by equation 8.18 cannot be compared directly to the data shown in figure 8.8 due to the fact that we do not know the absolute number of LacI-CFP molecules per cell. We only know the resulting fluorescence determined in the CFP channel. However, we can rewrite the equation in the following form

$$\text{Fold-change} = \left[1 + \frac{\alpha^{-1} I}{N_{NS}} e^{-\beta \Delta\varepsilon_{rd}}\right]^{-1}. \tag{8.19}$$

Here, we have replaced the absolute number of repressors by $\alpha^{-1} I$. I is the measured intensity in the CFP channel, while α is the calibration factor relating absolute number of molecules and fluorescence. Notice that since we know the binding energy we can now fit the data in figure 8.8 in order to obtain the value of α. We obtain a calibration factor of $(1.0 \pm 0.2) \times 10^3$ arbitrary fluorescent units per LacI-CFP molecule. We can now use this calibration factor to go back to the LacI-CFP data expressed in arbitrary units in figure 8.8 and convert it to an absolute number of repressors. This is shown as an alternative top x-axis in figure 8.8.

Once we calibrate the data in figure 8.8 in terms of the absolute number of LacI-CFP molecules using the value we determined for α we find an interesting outcome. The calibration predicts that we can reliably detect two LacI-CFP molecules per cell. Since LacI-CFP is a dimer this would correspond to the detection of four individual CFP molecules. It is highly unlikely for this to be the case, as recent measurements using similar microscopy conditions have shown a lower detection limit of about ten fluorescent proteins per cell

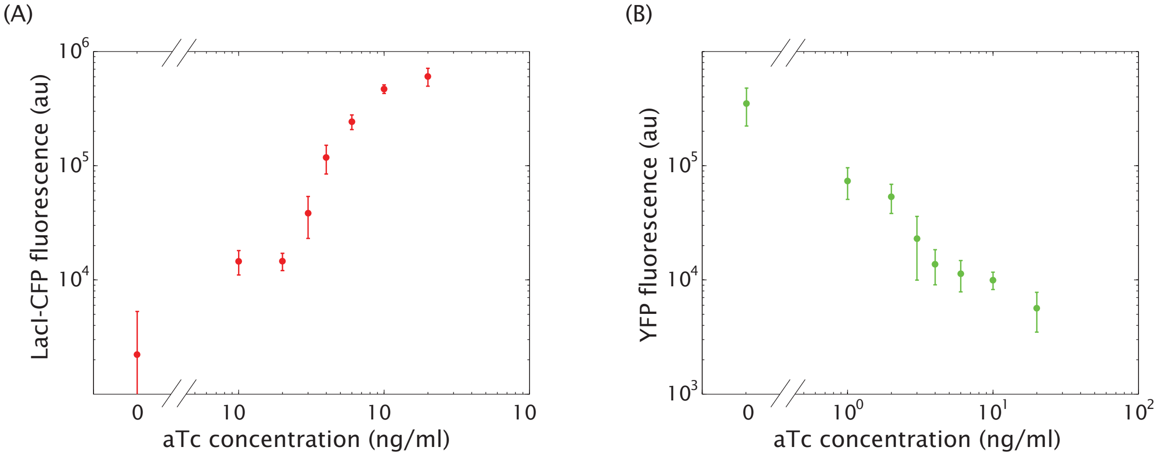


Figure 8.7: Response of the regulatory circuit as function of aTc concentration. Cells are grown in the presence of various concentrations of aTc and their levels of LacI-CFP and YFP quantified using microscopy. aTc determines the intracellular concentration of LacI-CFP which, in turn, determines the intracellular concentration of YFP.

[51]. These measurements in question were performed using YFP as a reporter. The autofluorescence of the cell at the CFP wavelengths is expected to be even larger than at the YFP wavelength [65]. As a result we would expect the lower bound for YFP of ten molecules per cell to be even higher for CFP.

This potential discrepancy could be easily solved if we consider a different binding energy. For example, if the binding energy was lower by 2 k_BT such that $\Delta\varepsilon_{rd} = -11.7\ k_BT$ the calibration α would increase such that now the lowest absolute level of LacI-CFP detected would be on the order of 15, which would correspond to the detection of 30 CFP molecules. We conclude that if we adopt a view where there is a difference in the binding energies obtained for wild-type Lac repressor and for the LacI-CFP fusion we can account for the behavior and concentrations obtained from our regulatory circuit in steady-state. Interestingly, we will draw similar conclusions when inferring the *in vivo* binding energies from the dilution experiment.

8.5.3 Accounting for repressor production in the calculation of the calibration factor

The above formulation provides the basis for measuring the calibration factor between fluorescence counts and number of proteins for an experiment where the production of protein is completely suppressed through tight regulation. Now we derive a general method where the production of protein during the dilution process is accounted for and a measure of the absolute production can be obtained along with the desired calibration factor. For now, we will treat the production process generally, assuming we know both the average number of proteins produced and the variance in this production. Using this general formalism, we would like a similar expression to equation 8.1, however the total number of proteins in each daughter is now the result of a combination of the initial partitioning upon division *and* the production process. We begin with the

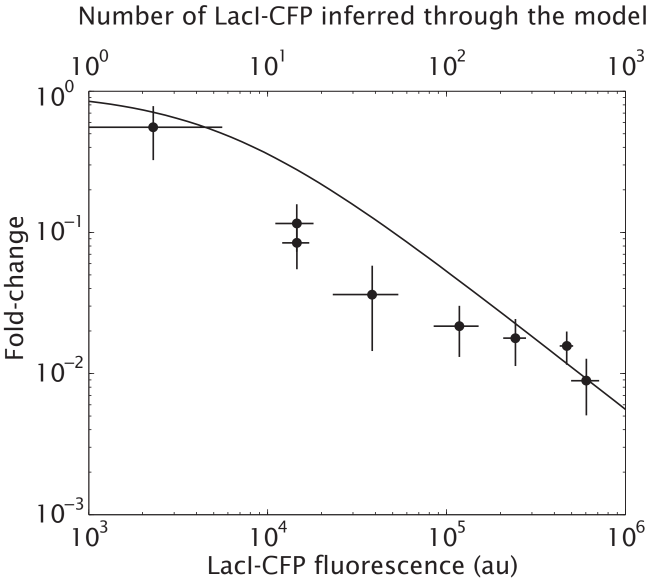


Figure 8.8: Fold-change as a function of the LacI-CFP fluorescence. The fold-change in gene expression is shown as a function of the corresponding LacI-CFP fluorescence (lower x-axis). Using previous knowledge of the *in vivo* binding energy [14] we can fit the data to equation 8.19 in order to obtain an absolute calibration between the arbitrary fluorescent units and the absolute number of LacI-CFP molecules. The result of such a calibration is shown as an alternative x-axis on the top. It is interesting to note that the calibration factor obtained from these measurements implies that we can see on the order of two LacI-CFP molecules per cell.

same statement of the variance in the number of proteins found in daughter 1,

$$\sigma^2 \quad = \quad < (N_1 - < N_1 >)^2 > . \tag{8.20}$$

The average is now $< N_1 >= N_T^b/2+ < N_1^p >$ where N_1^p is the number of proteins measured which were a result of production. Making this substitution, the above equation can be rewritten,

$$\sigma^2 \quad = \quad < \left(N_1 - \frac{N_T^b}{2} - < N_1^p > \right)^2 > . \tag{8.21}$$

With a bit of manipulation,

$$\sigma^2 \quad = \quad < \left(\frac{N_1 - N_2}{2} + \frac{N_1^p}{2} + \frac{N_2^p}{2} - < N_1^p > \right)^2 > . \tag{8.22}$$

The average production in both daughters is the same so we write $< N_1^p > = < N_1^p > /2 + < N_2^p > /2$ and then distribute the square,

$$\sigma^2 \quad = \quad < \left(\frac{N_1 - N_2}{2} \right)^2 + \left(\frac{N_1^p - < N_1^p >}{2} \right)^2 + \left(\frac{N_2^p - < N_2^p >}{2} \right)^2 >, \tag{8.23}$$

since the cross terms cancel out; the mixed N and N_p terms are equal and opposite while the N_1^p and N_2^p terms are statistically independent. Each N^p term contributes a factor of the variance due to the production process, σ_p^2, divided by 4, and thus we find,

$$\sigma^2 = \left\langle \left(\frac{N_1 - N_2}{2} \right)^2 \right\rangle + \frac{\sigma_p^2}{2}. \tag{8.24}$$

Now, we can also write the variance for the entire processes as the sum of the variances from each statistically independent process,

$$\begin{aligned} \sigma^2 &= \sigma_{\mathrm{b}}^2 + \sigma_{\mathrm{p}}^2, & (8.25) \\ \sigma^2 &= \frac{N_T^b}{4} + \sigma_{\mathrm{p}}^2. & (8.26) \end{aligned}$$

From equations 8.24 and 8.26 we arrive at,

$$\left\langle \left(\frac{N_1 - N_2}{2} \right)^2 \right\rangle = \frac{N_T^b}{4} + \frac{\sigma_{\mathrm{p}}^2}{2}. \tag{8.27}$$

Now, this formulation is not yet useful to experiments because it contains N_T^b which is the total number of proteins which were distributed to the daughters through binomial statistics and does not equal the measured quantity $N_1 + N_2$. Instead, the measured $N_1 + N_2 = N_T^b + N_T^p$. We do not know N_T^p and without taking more data, it is inaccessible to us, so we approximate it with the average production. Thus the final expression is,

$$\left\langle \left(\frac{N_1 - N_2}{2} \right)^2 \right\rangle = \frac{N_T}{4} + \frac{\sigma_{\mathrm{p}}^2 - < N_1^p >}{2}. \tag{8.28}$$

We now need a realistic model for the first two moments of the protein production distribution. The gamma distribution is used to represent a process where events occur randomly, uncorrelated in time (i.e., they are Poissonian), but the number produced during each event (often called a “burst”) follows an exponential distribution. Such a distribution is characterized by two parameters: a is the Poissonian parameter which characterizes the number of production events per time and b is the average number of proteins produced in each burst. The average number of proteins produced in this model is $< N_1^P >= ab$ and the variance is $\sigma_p^2 = ab^2$. Putting this information into equation 8.27 we arrive at,

$$\left\langle \left(\frac{N_1 - N_2}{2} \right)^2 \right\rangle = \frac{N_T}{4} + \frac{ab(b-1)}{2}. \tag{8.29}$$

We now have a means to measure both the calibration factor AND the separate production parameters a and b; the slope of the plot when fitting data such as shown in figure 8.2 would correspond to the calibration factor, the y-intercept would be equal to $ab(b-1)/2$ and the steady-state protein level would be $(3/2)ab$ (on

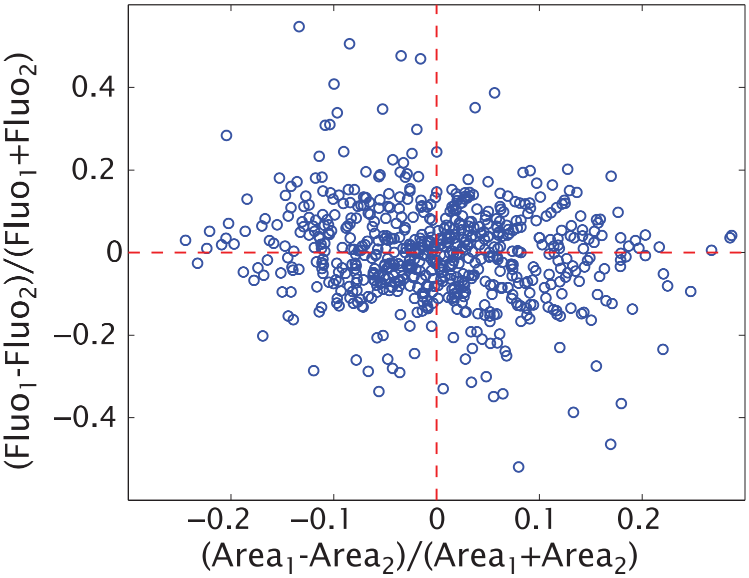


Figure 8.9: Contribution of differences in cell size to the fluorescence partitioning. In cells bearing LacI-CFP, but no production of the transcription factor, the relative difference in the fluorescence of the two daughter cells upon cell division is measured. These measurements are shown as a function of the relative difference in the cell area of the daughter cells. The lack of a strong correlation suggests that the partitioning of fluorescent is indeed determined by a binomial distribution and not to the particular fluctuations in cell size upon cell division.

average ab right after division and $2ab$ immediately preceding the subsequent division).

8.5.4 Supplementary figures

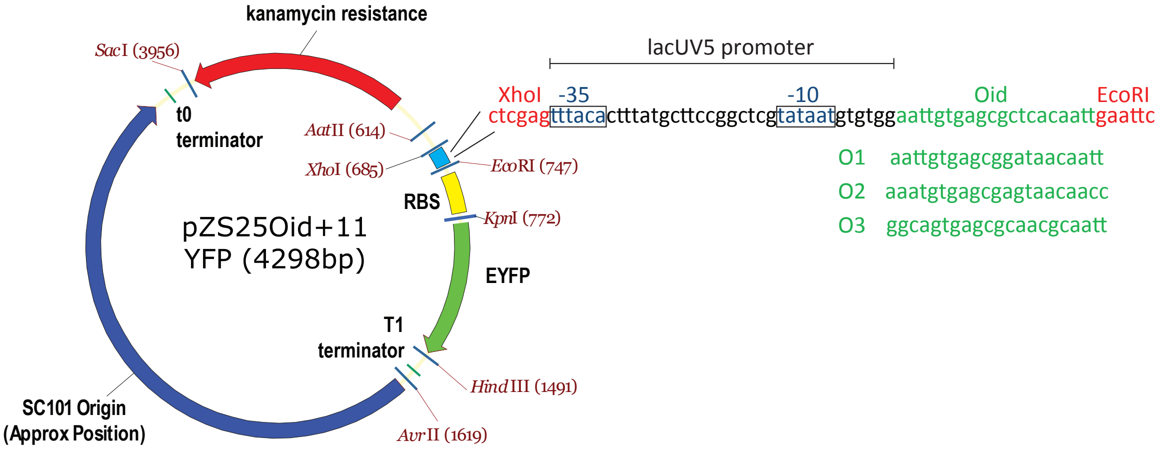


Figure 8.10: Plasmid diagram and promoter sequence. The main features of the plasmids pZS25O1+11-YFP are shown flanked by unique restriction sites. The particular promoter sequence based on the *lacUV5* promoter is shown together with the sequences of the different Lac repressor binding sites used.

Bibliography

[1] W. A. Lim. The modular logic of signaling proteins: Building allosteric switches from simple binding domains. *Curr Opin Struct Biol*, 12(1):61–8, 2002.

[2] M. Ptashne and A. Gann. *Genes and signals.* Cold Spring Harbor Laboratory Press, New York, 2002.

[3] R. P. Bhattacharyya, A. Remenyi, B. J. Yeh, and W. A. Lim. Domains, motifs, and scaffolds: The role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu Rev Biochem*, 75:655–80, 2006.

[4] D. Kentner and V. Sourjik. Use of fluorescence microscopy to study intracellular signaling in bacteria. *Annu Rev Microbiol*, 64:373–90, 2010.

[5] H. G. Garcia, A. Sanchez, T. Kuhlman, J. Kondev, and R. Phillips. Transcription by the numbers redux: Experiments and calculations that surprise. *Trends Cell Biol*, 2010.

[6] V. Sourjik and H. C. Berg. Binding of the *escherichia coli* response regulator chey to its target measured *in vivo* by fluorescence resonance energy transfer. *Proc Natl Acad Sci U S A*, 99(20):12669–74, 2002.

[7] J. E. Keymer, R. G. Endres, M. Skoge, Y. Meir, and N. S. Wingreen. Chemosensing in escherichia coli: Two regimes of two-state receptors. *Proc Natl Acad Sci U S A*, 103(6):1786–91, 2006.

[8] Y. Tu, T. S. Shimizu, and H. C. Berg. Modeling the chemotactic response of escherichia coli to time-varying stimuli. *Proc Natl Acad Sci U S A*, 105(39):14855–60, 2008.

[9] D. Greenfield, A. L. Mcevoy, H. Shroff, G. E. Crooks, N. S. Wingreen, E. Betzig, and J. Liphardt. Self-organization of the escherichia coli chemotaxis network imaged with super-resolution light microscopy. *PLoS Biol*, 7(6):e1000137, 2009.

[10] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003.

[11] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[12] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[13] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[14] H. G. Garcia and R. Phillips. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A*, 2011. (*Under review*).

[15] S. Oehler, E. R. Eismann, H. Kramer, and B. MÜLler-Hill. The three operators of the lac operon cooperate in repression. *EMBO J*, 9(4):973–9, 1990.

[16] S. Oehler, M. Amouyal, P. Kolkhof, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. Quality and position of the three *lac* operators of *e. Coli* define efficiency of repression. *EMBO J*, 13(14):3348–55, 1994.

[17] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–5, 2005.

[18] S. W. Teng, Y. Wang, K. C. Tu, T. Long, P. Mehta, N. S. Wingreen, B. L. Bassler, and N. P. Ong. Measurement of the copy number of the master quorum-sensing regulator of a bacterial cell. *Biophys J*, 98(9):2024–31, 2010.

[19] N. Rosenfeld, T. J. Perkins, U. Alon, M. B. Elowitz, and P. S. Swain. A fluctuation method to quantify in vivo fluorescence data. *Biophys J*, 91(2):759–66, 2006.

[20] J. Perrin. *Atoms*. Ox Bow Press, Woodbridge, CT, 1990.

[21] S. E. Luria and M. Delbruck. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943.

[22] P. H. Von Hippel, A. Revzin, C. A. Gross, and A. C. Wang. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: i. the *lac* operon: Equilibrium aspects. *Proc Natl Acad Sci U S A*, 71(12):4808–12, 1974.

[23] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'conner, D. W. Noble, and P. H. Von Hippel. Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound *escherichia coli lac* repressor *in vivo*. *Proc Natl Acad Sci U S A*, 74(10):4228–32, 1977.

[24] S. M. Law, G. R. Bellomy, P. J. Schlax, and M. T. J. Record. *in vivo* thermodynamic analysis of repression with and without looping in *lac* constructs. Estimates of free and local *lac* repressor concentrations and of physical properties of a region of supercoiled plasmid DNA *in vivo*. *J Mol Biol*, 230(1):161–73, 1993.

[25] J. Elf, G. W. Li, and X. S. Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–4, 2007.

[26] M. H. Ulbrich and E. Y. Isacoff. Subunit counting in membrane-bound proteins. *Nat Methods*, 4(4):319–21, 2007.

[27] D. F. Savage, B. Afonso, A. H. Chen, and P. A. Silver. Spatially ordered dynamics of the bacterial carbon fixation machinery. *Science*, 327(5970):1258–61, 2010.

[28] T. Kuhlman, Z. Zhang, M. H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *escherichia coli*. *Proc Natl Acad Sci U S A*, 104(14):6043–8, 2007.

[29] E. Bertrand, P. Chartrand, M. Schaefer, S. M. Shenoy, R. H. Singer, and R. M. Long. Localization of ash1 mrna particles in living yeast. *Mol Cell*, 2(4):437–45, 1998.

[30] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002.

[31] J. M. Raser and E. K. O'shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–4, 2004.

[32] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.

[33] W. J. Blake, G. Balazsi, M. A. Kohanski, F. J. Isaacs, K. F. Murphy, Y. Kuang, C. R. Cantor, D. R. Walt, and J. J. Collins. Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell*, 24(6):853–65, 2006.

[34] D. Zenklusen, D. R. Larson, and R. H. Singer. Single-rna counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*, 15(12):1263–71, 2008.

[35] A. Raj, P. Van Den Bogaard, S. A. Rifkin, A. Van Oudenaarden, and S. Tyagi. Imaging individual mrna molecules using multiple singly labeled probes. *Nat Methods*, 5(10):877–9, 2008.

[36] L. Cai, N. Friedman, and X. S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62, 2006.

[37] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–6, 2008.

[38] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying *e. Coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329:533, 2010.

[39] A. Sanchez and J. Kondev. Transcriptional control of noise in gene expression. *Proc Natl Acad Sci U S A*, 105(13):5081–6, 2008.

[40] A. Sanchez, H. G. Garcia, D. Jones, R. Phillips, and J. Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput Biol*, 2011. (*Under review*).

[41] S. Klumpp, Z. Zhang, and T. Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1366–75, 2009.

[42] J. C. Locke and M. B. Elowitz. Using movies to analyse gene circuit dynamics in single cells. *Nat Rev Microbiol*, 7(5):383–92, 2009.

[43] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci U S A*, 100(13):7702–7, 2003.

[44] N. J. Guido, X. Wang, D. Adalsteinsson, D. Mcmillen, J. Hasty, C. R. Cantor, T. C. Elston, and J. J. Collins. A bottom-up approach to gene regulation. *Nature*, 439(7078):856–60, 2006.

[45] S. Kaplan, A. Bren, A. Zaslaver, E. Dekel, and U. Alon. Diverse two-dimensional input functions control bacterial sugar genes. *Mol Cell*, 29(6):786–92, 2008.

[46] J. A. Megerle, G. Fritz, U. Gerland, K. Jung, and J. O. Radler. Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophys J*, 95(4):2103–15, 2008.

[47] T. Borggrefe, R. Davis, A. Bareket-Samish, and R. D. Kornberg. Quantitation of the rna polymerase ii transcription machinery in yeast. *J Biol Chem*, 276(50):47150–3, 2001.

[48] S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–41, 2003.

[49] J. Q. Wu and T. D. Pollard. Counting cytokinesis proteins globally and locally in fission yeast. *Science*, 310(5746):310–4, 2005.

[50] J. Q. Wu, C. D. Mccormick, and T. D. Pollard. Counting proteins in living cells by quantitative fluorescence microscopy with internal standards. *Methods Cell Biol*, 89:253–73, 2008.

[51] H. G. Garcia, H. J. Lee, J. Q. Boedicker, and R. Phillips. The limits and validity of methods of measuring gene expression for the testing of quantitative models. *Biophys J*, 2011. (*Under review*).

[52] K. Hirschberg, C. M. Miller, J. Ellenberg, J. F. Presley, E. D. Siggia, R. D. Phair, and J. Lippincott-Schwartz. Kinetic analysis of secretory protein traffic and characterization of golgi to plasma membrane transport intermediates in living cells. *J Cell Biol*, 143(6):1485–503, 1998.

[53] D. W. Piston, G. H. Patterson, and S. M. Knobel. Quantitative imaging of the green fluorescent protein (gfp). *Methods Cell Biol*, 58:31–48, 1999.

[54] T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek. Probing the limits to positional information. *Cell*, 130(1):153–64, 2007.

[55] H. Bremer and P. P. Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. In N. F. e. al., editor, *In escherichia coli and salmonella cellular and molecular biology*, pages 1553–1569. ASM Press, Washington DC, 1996.

[56] T. L. To and N. Maheshri. Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science*, 327(5969):1142–5, 2010.

[57] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: From theories to phenotypes. *Nat Rev Genet*, 2005.

[58] N. Maheshri and E. K. O'shea. Living with noisy genes: How cells function reliably with inherent variability in gene expression. *Annu Rev Biophys Biomol Struct*, 36:413–34, 2007.

[59] A. Raj and A. Van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–26, 2008.

[60] R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65–8, 2008.

[61] A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–73, 2010.

[62] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in *escherichia coli* via the lacr/o, the tetr/o and arac/i1-i2 regulatory elements. *Nucleic Acids Res*, 25(6):1203–10, 1997.

[63] K. A. Datsenko and B. L. Wanner. One-step inactivation of chromosomal genes in *escherichia coli* k-12 using pcr products. *Proc Natl Acad Sci U S A*, 97(12):6640–5, 2000.

[64] I. F. Lau, S. R. Filipe, B. Soballe, O. A. Okstad, F. X. Barre, and D. J. Sherratt. Spatial and temporal organization of replicating *escherichia coli* chromosomes. *Mol Microbiol*, 49(3):731–43, 2003.

[65] G. S. Harms, L. Cognet, P. H. Lommerse, G. A. Blab, and T. Schmidt. Autofluorescent proteins in single-molecule research: Applications to live cell imaging microscopy. *Biophys J*, 80(5):2396–408, 2001.

Chapter 9

Concentration and Length Dependence of DNA Looping in Transcriptional Regulation

This chapter is a reproduction of reference [1].

In many cases, transcriptional regulation involves the binding of transcription factors at sites on the DNA that are not immediately adjacent to the promoter of interest. This action at a distance is often mediated by the formation of DNA loops: Binding at two or more sites on the DNA results in the formation of a loop, which can bring the transcription factor into the immediate neighborhood of the relevant promoter. These processes are important in settings ranging from the historic bacterial examples (bacterial metabolism and the lytic-lysogeny decision in bacteriophage), to the modern concept of gene regulation to regulatory processes central to pattern formation during development of multicellular organisms. Though there have been a variety of insights into the combinatorial aspects of transcriptional control, the mechanism of DNA looping as an agent of combinatorial control in both prokaryotes and eukaryotes remains unclear. We use single-molecule techniques to dissect DNA looping in the *lac* operon. In particular, we measure the propensity for DNA looping by the Lac repressor as a function of the concentration of repressor protein and as a function of the distance between repressor binding sites. As with earlier single-molecule studies, we find (at least) two distinct looped states and demonstrate that the presence of these two states depends both upon the concentration of repressor protein and the distance between the two repressor binding sites. We find that loops form even at interoperator spacings considerably shorter than the DNA persistence length, without the intervention of any other proteins to prebend the DNA. The concentration measurements also permit us to use a simple statistical mechanical model of DNA loop formation to determine the free energy of DNA looping, or equivalently, the J-factor for looping.

9.1 Introduction

The biological significance of DNA is primarily attributed to the information implicit in its sequence. Still, there are a wide range of processes for which DNA's physical basis as a stiff polymer also matters [2]. For example, the packaging of DNA into nucleosomes appears to select for sequence motifs that are particularly flexible [3, 4]. In the setting of transcriptional regulation, there are a host of regulatory architectures both in prokaryotes and eukaryotes which require the interaction of sequences on the DNA that are not adjacent [5–8]. These interactions are mediated by DNA-binding proteins, which have to deform the DNA. In eukaryotes, action of transcription factors over long distances seems the rule rather than the exception. One of the most transparent examples of DNA looping is in bacteria where some repressors and activators can bind at two sites simultaneously, resulting in a DNA loop. This effect was first elucidated in the context of the arabinose operon [9]. It is an amusing twist of history that the two regulatory motifs considered by Jacob and Monod, namely, the switch that makes the decision between the lytic and lysogenic pathways after phage infection [10] and the decision making apparatus associated with lactose digestion in bacteria [6, 11], both involve DNA looping as well.

To understand the physical mechanism of the biological action at a distance revealed by DNA looping, it is necessary to bring both *in vitro* and *in vivo* experiments as well as theoretical analyses to bear on this important problem. Over the last few decades there have been a series of impressive and beautiful experiments from many quarters that inspired our own work. In the *in vivo* context, it is especially the work of Müller-Hill and coworkers that demonstrates the intriguing quantitative implications of DNA looping for regulation [12]. In their experiments, they tuned the length of the DNA loop in one base pair increments and measured the resulting repression. More recently, these experiments have been performed with mutant bacterial strains that were deficient in architectural proteins such as HU, IHF and H-NS [13, 14]. On the *in vitro* side, single molecule experiments using the tethered-particle method [15–23] have also contributed significantly [24–29]. The idea of these experiments is to tether a piece of DNA to a microscope coverslip with a bead attached to the end. The DNA construct has the relevant binding sites (operators) for the protein of interest along the DNA and when one of these proteins binds, it shortens the length of the tether. As a result of the shorter tether, the Brownian motion of the bead is reduced. Hence, the size of the random excursions of the bead serves as a reporter for the status of the DNA molecule (i.e., looped or unlooped, DNA-binding protein present or not).

In addition to single-molecule studies, *in vitro* biochemical assays have also shed important light on the interactions between transcription factors and their DNA targets. Both filter binding assays and electrophoretic mobility shift assays have been widely used to study how variables dictating DNA mechanics such as length and degree of supercoiling, alter the looping process [30–34].

One of the missing links in the experimental elucidation of these problems is systematic, single-molecule experiments which probe the length, repressor concentration and sequence dependence of DNA looping. Such experiments will complement earlier *in vivo* work, which has already demonstrated how DNA length and repressor concentration alter repression [12]. Our view is that such systematic experiments will help clarify the way in which both length and sequence contribute to the probability of DNA looping, and begin to elucidate the mechanisms whereby transcription factors act over long genomic distances. Further, such experiments can begin to shed light on broader questions of regulatory architecture and the significance of operator placement to transcriptional control. To that end, we have carried out experiments that probe the DNA looping process over a range of concentrations of repressor protein and for a series of different loop lengths. In addition, intrigued by the sequence preferences observed in nucleosomal DNA, we have made looping constructs in which these highly bendable nucleosomal sequences are taken out of their natural eukaryotic context and are inserted between the operators that serve as binding sites for the Lac repressor (the results of those experiments will be reported elsewhere). The point of this exercise is to see how the looping probability depends upon these tunable parameters, namely, length, repressor concentration and sequence.

Our key results are: (1) The concentration dependence of looping as a function of repressor concentration (a "titration" curve) can be described by a simple equilibrium statistical-mechanics model of transcription factor-DNA interactions. The model predicts a saturation effect, which agrees with our experimental observations. (2) By measuring this effect, we were able to isolate the free energy change of looping (that is, separate it from the binding free energy change), obtaining an experimental measurement of its value for a range of different lengths in an uncluttered, *in vitro*, setting. (3) Systematic measurement of looping free energy as a function of interoperator spacing hints at the same modulations seen in analogous *in vitro* work on cyclization [4, 35], and *in vivo* work on repression [12, 13]. (4) Clear experimental signature of multiple looped states, consistent with theory expectations [36–39] and other recent experiments [27, 29]. In the remainder of the paper, we describe a series of experiments that examine both the length and concentration dependence of DNA looping induced by the Lac repressor. A companion paper gives extensive details about our theoretical calculations [40].

9.2 Results

As argued above, one of our central concerns in performing these experiments was to have sufficient, systematic data to make it possible to carry out a thorough analysis of the interplay between theories of transcriptional regulation (and DNA looping) [41–44], and experiment. To that end, we have carried out a series of DNA looping experiments using the tethered-particle method [24] for loop lengths ranging from 300 to 310 bp in one base pair increments as well as several representative examples for lengths below 100 bp. The experiments described here use DNA constructs harboring two different operators, symmetric operator *Oid* and primary natural operator $O1$ as Lac repressor binding sites. In addition, we have explored how the looping trajectories depend upon the concentration of Lac repressor. The particular experimental details are described in the "Materials and Methods" section.

A typical experimental trace resulting from these measurements is shown in figure 9.1. (Representative examples of experimental traces from all of the lengths and concentrations considered throughout the paper as well as examples of rejected traces are shown in the Supporting Information.) As seen in the figure, as with other recent work [27, 29], there are clearly two distinct looped states as seen both in the trajectory and the histogram. Control experiments with one of the two binding sites removed show only the highest peak, which further supports the idea that the two lower peaks indeed indicate looped configurations. One hypothesis is that these two looped states correspond to two different configurations of the Lac repressor molecule and its attendant DNA, which we will refer to as the "open" and "closed" configurations. Direct interconversion between the two looped species suggested the two distinct looped states are indeed due to different conformations of Lac repressor protein [27]. An alternate hypothesis is that the two peaks reflect different DNA topologies [45–47]. Although this hypothesis does not obviously accommodate the apparent observation of direct interconversion, nevertheless we will present data from Monte Carlo simulations of DNA chain conformations that show that it *can* quantitatively explain the observed multi-peak structure observed in the data.

9.2.1 Concentration dependence

In order to extract quantities such as the free energy of looping associated with repressor binding (or equivalently, a J-factor for looping, essentially the concentration at which in a solution of DNA with sticky ends, the probability of forming circles and dimers is equal) and to examine how the propensity for looping depends upon the number of repressors, we needed looping data at a number of different concentrations. At very low concentration, we expect that there will be negligible looping because neither of the operators will be bound by Lac repressor. At intermediate concentrations, the equilibrium will be dominated by states in which a single repressor tetramer is bound to the DNA at the strong operator, punctuated by transient looping events. In the very high concentration limit, each operator will be occupied by a tetramer (see figure 9.2 below), making the formation of a loop nearly impossible.

This progression of qualitative behavior is indeed seen in figure 9.3, which shows data from eight distinct concentrations of Lac repressor, as well as a single-operator control in which the DNA lacks a secondary operator. Throughout this work we define *sequence length* or *loop length* as the end-to-end distance between the operators as shown in figure 9.15. These curves correspond to a sequence length of 306 bp and are generated by summing the normalized histograms from *all* of the individual trajectories for each concentration that pass our bead selection criteria (bead selection criteria are discussed in detail in the Supporting Information). A key feature of these data is the way in which the two looped states are turned off as the concentration of Lac repressor is increased to very high levels. This phenomenon is expected since the Lac repressor exists always as tetramers under the conditions used here [48, 49], and competition for binding at the second operator between loose Lac repressor and Lac repressor bound to the other operator is stronger as the concentration of Lac repressor increases. However, the two different looped species have slightly different responses at high repressor concentrations. For example, at 1 nM concentrations, the intermediate looped state has become very infrequent, whereas the shortest looped state remains competitive. Similar concentration dependence of Lac repressor-mediated DNA looping was studied previously [25] at 4 pM, 20 pM and 100 pM. Those experiments revealed that looping is suppressed as the concentration goes up.

One way to characterize the looping probability as a function of concentration is shown in figure 9.4. There are various ways to obtain data of the sort displayed in this plot. First, by examining the trajectories, we can simply compute the fraction of time that the DNA spends in each of the different states, with the looping probability given by the ratio of the time spent in either of the looped states to the total elapsed time. Of course, to compute the time spent in each state, we have to make a thresholding decision about when each transition has occurred. This can be ambiguous, because trajectories sometimes undergo rapid jumps back and forth between different states; it is not unequivocally clear when an apparent transition is real, and when it is a random fluctuation without change of looping state. A second way of obtaining the looping probability is to use figure 9.3 and to compute the areas under the different peaks and to use the ratios of areas as a measure of looping probability. This method, however, does not properly account for possible variation between different beads, because they are all added up into one histogram. A third alternative is to obtain the looping probability for each *individual* bead, by plotting its histogram and calculating the area under that subset of the histogram corresponding to the looped states. We used this last method to calculate the mean looping probability and the standard error for each construct, which is shown in figure 9.4.

These results can also be explored from a theoretical perspective using the tools of statistical mechanics [43, 44, 50]. The goal of a statistical mechanical description of this system is to compute the probability of the various microstates available to the repressor-DNA system as shown in figure 9.2. The simplest model posits 5 distinct states [24, 25, 27]: Both operators empty, *Oid* occupied by repressor without looping, $O1$ occupied by repressor without looping, *Oid* and $O1$ separately occupied by single repressors and the looped state (the subtleties associated with the statistical weight of the looped state are described in the Supporting Information). The model does not take into account the effect of non-specific binding of Lac repressor to non-operator DNA, because a simple estimate reveals that the vast majority of repressors are free in solution rather than bound nonspecifically to the tethered DNA. We argue that this effect is negligible because the equilibrium association constant of Lac repressor to non-operator DNA at conditions similar to ours is around $10^6 \sim 10^7$ M^{-1} [51–57], which is roughly six orders of magnitude less than the corresponding quantity for specific binding [31, 58–63]. Given such association constants, the ratio between non-specifically-bound Lac repressor and the free Lac repressor in solution is given as

$$\begin{aligned} \frac{[RD]}{[R]} &= K_{NS} \times [D] \\ &\approx 2 \times 10^{-5}, \end{aligned}$$

where $[RD]$ is the concentration of non-specifically-bound Lac repressor, $[R]$ is the concentration of Lac repressor in solution, and $[D]$ is the DNA concentration, which is around 2 pM in our experiment. For $[R] = 200$ nM, we have $[RD] \approx 4$ pM, which is far smaller than the concentration of Lac repressor in solution.

It is convenient to describe the probability of the various states using both the language of microscopic binding energies (and looping free energies) and the language of equilibrium constants (and J-factors). From a microscopic perspective, the key parameters that show up in the model are the standard free energy changes for repressor binding to the two operators, $\Delta\epsilon_{id}$ and $\Delta\epsilon_1$, the looping free energy ΔF_{loop} and the concentration of repressor $[R]$. The binding energy here contains two components. One is the standard positional free energy required for bringing one Lac repressor molecule to its DNA binding site at 1 M concentration of Lac repressor. The other is the rotational entropy loss times $-T$, plus the interaction free energy due to the physical contact upon protein binding [43, 44, 64]. The associated free energy with each configuration gives the statistical weights of the equilibrium probability (listed in the middle column of figure 9.2). For example, to obtain the probability of the looped state, we construct the ratio of state (v) in the figure to the sum over all five states, as given by

$$\begin{aligned} p_{\text{loop}} \quad = \quad & \left[8\frac{[R]}{1\ \text{M}}e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id}+\Delta F_{\text{loop}})}\right] \\ & \left[1+4\frac{[R]}{1\ \text{M}}\left(e^{-\beta\Delta\varepsilon_1}+e^{-\beta\Delta\varepsilon_{id}}\right)+16\left(\frac{[R]}{1\ \text{M}}\right)^2 e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id})}+\right. \\ & \left. 8\frac{[R]}{1\ \text{M}}e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id}+\Delta F_{\text{loop}})}\right]^{-1}, \end{aligned} \tag{9.1}$$

where $\beta = 1/k_BT$ and the temperature is in degrees Kelvin. As detailed in the Supporting Information and can be read off from the right column in figure 9.2, this microscopic description is conveniently rewritten in terms of the equilibrium constants and J-factor for looping as

$$p_{\text{loop}} = \frac{\frac{1}{2}\frac{[R]J_{\text{loop}}}{K_1K_{id}}}{1+\frac{[R]}{K_1}+\frac{[R]}{K_{id}}+\frac{[R]^2}{K_1K_{id}}+\frac{1}{2}\frac{[R]J_{\text{loop}}}{K_1K_{id}}}. \tag{9.2}$$

Here J_{loop} is the average of the individual J factors corresponding to different loop topologies. These topologies can be classified according to the orientation of each one of the operators with respect to the two Lac repressor binding heads as shown in figure 9.5. We define the state variables α and β that describe the orientation of $O1$ and Oid, respectively, and that can adopt a value of either 1 or 2. The average J_{loop} is then

$$J_{\text{loop}} = \frac{1}{4}\sum_{\alpha,\beta} J_{\text{loop},\alpha,\beta}. \tag{9.3}$$

An alternative to this scheme is to construct the ratio $p_{\text{unloop}}/p_{\text{loop}}$. In the limit where the strongest operator, *Oid*, is always occupied, this ratio takes the simple, linear form

$$p_{\text{ratio}} = \frac{2K_1}{J_{\text{loop}}} + \frac{2[R]}{J_{\text{loop}}}. \tag{9.4}$$

This permits the determination of the J-factor as the slope of a linear fit of the form without necessarily a need to obtain K_1. Below we discuss the validity of this particular model. For the remaining data points at loop lengths L other than 306 bp, where no titration was done, we can use the relation

$$J_{\text{loop}}(L) = \frac{p_{\text{ratio}}(306\text{ bp})}{p_{\text{ratio}}(L)} J_{\text{loop}}(306\text{ bp}). \tag{9.5}$$

Just like in the titration case, this relation allows to obtain J_{loop} without knowing K_1, as long as we know at least one value of J_{loop} and its corresponding p_{ratio}.

The data shown in figure 9.4 can be fit in several different ways as suggested by the three different formulae characterizing the looping probability given above. The fit shown in figure 9.4 is a full non-linear fit in which the parameters K_1, K_2 and J_{loop} are treated as fitting parameters. Alternatively, using this same data of figure 9.4, we can actually obtain the looping free energy, as well as the binding energies by fitting the data to equation 9.1. Note that these two descriptions are equivalent and each depends upon three unknown parameters. Once one set of parameters is known, in principle, the complementary parameters are also known. We find it convenient to work in terms of both languages because in some discussions it is useful to talk in terms of looping free energies, and in other contexts, in terms of the looping J-factor. Finally, we can fit the data corresponding to LacI concentrations of 10 pM and higher using the linear model from equation 9.4. The results of these different fits are shown in table 9.1. These results are usefully contrasted with results of other experiments on the *lac* operon, which are also summarized in table 9.1. We see from the table that the non-linear model fails to constrain the value of K_{id} reliably. In the case of the $O1$ binding constants we see a difference of almost two orders of magnitude with published dissociation constants, which translates into a difference of roughly 4 k_BT in the binding energy.

One of the challenges of single-molecule experiments like those described here is that the concentration of protein introduced into the system may not correspond to the actual concentration "seen" by the DNA that is tethered to the surface. For example, some of the protein might be lost as a result of nonspecific binding to the microscope coverslip. From the linear model shown in equation 9.4 it follows that any error in the concentration will translate linearly into an error in J_{loop} and K_1. Therefore, in order for the above discrepancy to be explained solely by surface effects on the LacI concentration we would have to have a difference of between one and two orders of magnitude between the concentration of the stock that flowed into the chamber and the actual free concentration within the chamber.

Once the parameters that characterize the model are in hand, we can plot the probability of all five possible states as a function of the Lac repressor concentration as shown in figure 9.6. This figure reveals that at the concentrations we normally use ($[R] = 100$ pM), the system is dominated by the looped state and the state with single occupancy of *Oid*. A detailed discussion of the significance of the looping free energies (or the J-factors) will follow later in the paper once we have explored the question of the length dependence of DNA looping in the *lac* operon.

9.2.2 Length dependence

9.2.2.1 1bp resolution for a whole helical turn: $L_{\text{loop}} =$ 300 bp to 310 bp

The beautiful *in vivo* repression experiments of [12] demonstrate that the length of the DNA loop formed by Lac repressor strongly affects the probability of loop formation (especially for loop lengths less than 150 bp). In particular, those authors (and others) [13, 14, 65, 66] have observed “phasing”: The relative orientations of the two operators changes the ease with which repressor can loop. Similar phasing effects have been observed in *in vitro* cyclization assays [4, 35, 67, 68]. What has not been clear is how to concretely and quantitatively relate these results on DNA mechanics from the *in vivo* and *in vitro* settings. Our idea was to systematically examine the same progression of DNA lengths that have been observed *in vivo*, but now using TPM experiments. To that end, we have measured TPM trajectories for a series of interoperator spacings measured in 1 bp increments. The results of this systematic series of measurements for DNAs harboring operators spaced over the range $L_{\text{loop}} = 300 \sim 310$ bp are shown in figure 9.7 (as are the results for several shorter lengths to be discussed in the next section). Each plot shows the probability of the three states for a particular interoperator spacing.

The data can be converted into a plot of the dependence of the looping probability on interoperator spacing as shown in figure 9.8. This figure shows p_{loop} as a function of the DNA length between the two operators. The looping probability shows a weak dependence on the interoperator spacing but reveals no conclusive signature of phasing; to really detect such phasing with confidence, however, would require more measurements in single basepair increments. The maximum looping is achieved when the two binding sites are 306 bp apart, suggesting that at this distance, the two sites are in an optimal phasing orientation for binding of the two heads of Lac repressor. The ability to form stable out-of-phase (two binding sites are on the opposite side of the DNA) loops with only a small reduction in stability is consistent with previous studies [27]. The relatively stable looping over the entire helical repeat is also consistent with the relatively constant repression level *in vivo* for similar interoperator spacing [12].

As already indicated in table 9.1, the looping probability can be converted into a corresponding looping free energy based on the statistical mechanics model described above and culminating in equation 9.1. The results of such calculation are shown in figure 9.9. The measurements on length dependence permit us to go beyond the concentration dependence measurements by systematically exploring how the phasing of the two operators impacts the free energy of DNA looping. One might expect that when the two operators are on opposite sides of the DNA, additional twist deformation energy is required to bring the operators into good registry for Lac repressor binding. Our results show that the phasing effect imposes an energy penalty ΔF_{loop} that differs by only about 1.5 $k_{\text{B}}T$ between the in-phase and out of phase cases. An alternative interpretation of these same results on looping probability is offered by the J-factor for looping as shown in figure 9.10.

To get a feel for the energy scale associated with twist deformations, we perform a simple estimate. Twisting DNA for a torsional angle θ requires energy

$$\Delta F_t = k_B T \xi_{tp} \theta^2 / 2L \tag{9.6}$$

where ξ_{tp} is the torsional persistence length for double stranded DNA, which is around 250 bp [69–71]. L is the DNA length. For half a helical turn twist, $\theta = \pi$ and $L = 300$ bp. The energy introduced for half a helical turn is around 4.11 $k_B T$. Our experimentally determined looping energy difference between in-phase and out-of-phase DNA, about 1.5 $k_B T$, is indeed comparable in magnitude to this estimate. Our simple estimate is high, in part because it neglects the fact that in addition to twisting, a loop can writhe to accommodate a non-ideal operator phasing. Additionally, the observed small magnitude of our observed phasing modulation may reflect partially canceling out-of-phase contributions of different topologies [40], not a low free energy cost for twisting. Finally, the Lac repressor itself is flexible, and so can partially compensate for non-ideal phasing.

9.2.2.2 Sub-persistence length loops

One of the intriguing facts about the architecture of regulatory motifs that involve DNA looping is that often the loops formed in these systems have DNA lengths that are considerably shorter than the persistence length of DNA (i.e., 150 bp). For example, in the *lac* operon, one of the three wild-type loops has a length of 92 bp. However, this trend goes well beyond the *lac* operon as is seen for a variety of different architectures found in *E. coli*, for example [2]. As a result, it is of great interest to understand the interplay between transcriptional regulation and corresponding mechanical manipulations of DNA this implies.

So far, we have considered loops that are roughly twofold larger than the persistence length through our investigation of one full helical repeat between 300 and 310 bp. To begin to develop intuition for the mechanism of loop formation in the extremely short loops exhibited in many regulatory architectures, we have examined three different lengths: 89, 94 and 100 bp. One of the reasons that the examination of these loops is especially important is that it has been speculated that the *in vivo* formation of these loops either requires special supercoiling of the DNA or the assistance of helper proteins that prebend the DNA [2]. However, as indicated by the TPM results shown in figure 9.7(B), even in our controlled *in vitro* setting, where neither of these mechanisms can act, Lac repressor is nevertheless able to form DNA loops. The essence of these experiments is identical to those described earlier in the paper except that now the overall tether lengths are shorter so as to ensure that the loops are detectable. (Representative TPM trajectories for these lengths are shown in the Supporting Information.) It is clear from the histograms that of the three lengths we have investigated, loop formation is most favorable at 94 bp. Interestingly, it also appears that different loops are being formed for the in-phase and out-of-phase cases as evidenced by the changes of relative strengths among the looping peaks for the different constructs. The looping free energy and J-factor for looping for these short constructs are shown in figures 9.9(A) and 9.10(A).

9.2.3 Analysis of the TPM experiment

Both the observed length and sequence dependence of the formation of a repression complex are intriguing from the perspective of DNA mechanics. In particular, DNA is not a passive mechanical bystander in the process of transcriptional regulation. To better understand the experiments carried out here and how they might shed light on the interplay of transcription factors and their target DNA, we have appealed to two classes of models: i) statistical mechanics models of the probability of DNA-repressor complex formation which depends upon the looping free energy (these models were invoked earlier in the paper to determine the looping free energy) and ii) Monte Carlo simulations of the TPM experiment itself which include the energetics of the bent DNA and excluded volume interactions of the bead with the coverslip. Our Monte Carlo calculations allow us to compute how easily loops form, based on a mathematical model of DNA elasticity. For illustration, we have chosen a linear-elasticity model, that is, a model in the class containing the wormlike chain, but any other elastic theory of interest can be used with the same calculation strategy. Details of these calculations appear in [40].

One of the puzzles that has so far been unresolved concerning DNA mechanics at short scales is whether *in vivo* and *in vitro* experiments tell a different story. In particular, *in vivo* experiments, in which repression of a given gene is measured as a function of the interoperator spacing [12, 13], have the provocative feature that the maximum in repression (or equivalently the minimum in looping free energy) correspond to interoperator spacings that are shorter than the persistence length. Some speculate that this *in vivo* behavior results from the binding of helper proteins such as the architectural proteins HU, H-NS or IHF [2, 13, 14] or the control of DNA topology through the accumulation of twist. In the TPM measurements reported here, there are neither architectural proteins nor proteins that control the twist of the DNA. As a result, these experimental results serve as a jumping off point for a quantitative investigation of whether DNA at length scales shorter than the persistence length behaves more flexibly than expected on the basis of the wormlike chain model. To address this question, we performed a series of simulations of the probability of DNA looping for short, tethered DNAs like those described here using, a variant of the wormlike chain model to investigate the looping probability. Our theoretical model used *no fitting parameters;* the few parameters defining the model were obtained from other, non-TPM, experiments.

The fraction of time spent in the looped configuration is controlled by several competing effects. For example, suppose that a repressor tetramer is bound to the stronger operator, *Oid*. Shortening the interoperator spacing reduces the volume over which the other operator ($O1$) wanders relative to the second binding site on the repressor, increases the apparent local “concentration” of free operator in the neighborhood of that binding site, and hence enhances looping. But decreasing the interoperator spacing also has the opposite effect of discouraging looping, due to the larger elastic energy cost of forming a shorter loop. Moreover, a shorter overall DNA construct increases the entropic force exerted by bead–wall avoidance, again discouraging looping [72]. To see what our measurement of this looping equilibrium tells us, we therefore needed to calculate in some detail the expected local concentration of operator (the "looping J factor") based on a particular mathematical model of DNA elasticity. Traditionally, DNA has been modeled mathematically as a thin, elastic solid body with a classical Hooke-law elastic energy function. Because classical elasticity theory assumes that energy is a quadratic ("harmonic") function of strain, such models are collectively called "harmonic-elasticity" models; one example is the wormlike chain model. Accordingly, we used a harmonic-elasticity model, to see if it could adequately explain our results, or if, on the contrary, some non-harmonic model (for example the one proposed in [73, 74]) might be indicated.

To perform the required calculation, we modified the Gaussian sampling method previously used in [72, 75–77] (see section S6 and [40]). Our code generated many simulated DNA chains, applied steric constraints [72], and reported what fraction of accepted chain/bead configurations had the two operator sites at the correct relative position and orientation for binding to the tetramer, which was assumed to be rigidly fixed in the form seen in PDB structure 1LBG [78]. Once this fraction has been computed, it is straightforward to relate it to the looping J factor [40]. Note that the beauty of the looping J factor is that it is independent of the particular binding strengths of the different operators. To generate the simulated chains, we assumed a linear (harmonic, or wormlike-chain type) elastic energy function at the junctions in a chain of finite elements. Our energy function accounted for the bend anisotropy and bend–roll coupling of DNA, and yielded a value for the persistence length $\xi = 44\,\mathrm{nm}$ appropriate for our experiment's buffer conditions [79, 80]. Our model did not account for sequence dependence. We assume that this simplification is appropriate for comparison to our experimental results for the case of the 300 bp constructs and the 90 bp constructs with the sequence E8, but not with the sequence TA. The simulation treated the bead and the microscope slide as hard walls and accounted for bead–wall, bead–chain and wall–chain avoidance; we did not consider any interactions involving the repressor tetramer other than binding.

The symmetry of each LacI dimer implies four energetically equivalent ways for the two operators to bind when forming a loop, and hence four topologically distinct loop configuration classes [36–40]. We first asked whether this multiplicity of looped states could explain the general structure of the excursion distributions seen in figure 9.7. Accordingly, we made histograms of the distance between wall attachment point and bead center for our simulated chains. Figure 9.11 shows a subset of the same experimental data seen in figure 9.7, together with the simulation results. Although the correspondence is not perfect, it is clear that the simple physical model of looping outlined above can account for many features of the data, for example the locations of the looped peaks and their relative strengths, including the variation as loop length is changed. We acknowledge that we have no definitive reply to the argument that the apparent direct transitions between the B and M peaks of our distributions seem to require an open-to-closed conformational switch in the tetramer [27]. We merely point out that the existence of three peaks in the distribution, with the the observed locations, is not by itself conclusive evidence of such a switch. (Indeed, Villa et al. have argued that the opening transition does not occur [81].)

We were also interested to see if the high incidence of looping observed in our experiments on short (sub-persistence-length) loops was compatible with the hypotheses of harmonic DNA elasticity and fixed repressor geometry, or if on the contrary it demanded some modification to those hypotheses. Accordingly, we asked the simulation to compute the average J factor for loop lengths near 305 bp, and also for loop lengths near 95 bp. As discussed in reference [40], the result of the simulation was that the ratio of these quantities is $\bar{J}_{\mathrm{loop}}(95\,\mathrm{bp})/\bar{J}_{\mathrm{loop}}(305\,\mathrm{bp}) \approx 0.02$. In contrast, figure 9.10 shows that the experimental ratio is $\approx 0.35 \pm 0.1$, roughly 20-fold larger than the theoretical value. Our experimental results and those interpolated from our MC calculations for $\bar{J}_{\mathrm{loop}}$ as a function of loop length are shown in figure 9.12.

We conclude that the hypotheses of linear elasticity, a rigid protein coupler and a lack of non-specific DNA–repressor interactions cannot explain the high looping incidence seen in our experiments. (Special DNA sequences loop even more easily than the random sequences reported here.) One possible explanation, for which other support has been growing, is the hypothesis of DNA elastic breakdown at high curvature [73, 74, 82]. An alternative hypothesis is that for our shorter loops, *both* the lower and the intermediate peaks in our distributions of bead excursion correspond to the some alternative, “open” conformation of the repressor tetramer [36, 37, 46, 83–86]. To be successful, however, this hypothesis would have to pass the same quantitative hurdles to which we subjected our hypotheses.

9.3 Discussion

The regulatory regions on DNA can often be as large as (or even larger than) the genes they control. The relation between the biological mechanisms of transcriptional control and the physical constraints put on these mechanisms as a result of the mechanical properties of the DNA remains unclear. One avenue for clarifying action at a distance by transcription factors is systematic single-molecule experiments, which probe the dynamics of loop formation for different DNA architectures (i.e., different sequences, different transcription factor binding strengths, different distances between transcription factor binding sites) to complement systematic *in vivo* experiments that explore these same parameters. In this paper, we have described an example of such a systematic series of measurements, which begins to examine how the formation of transcription factor-DNA complexes depend upon parameters such as transcription factor concentration and the length of the DNA implicated in the complex.

In the case of the *lac* operon, our *in vitro* measurements demonstrate that the formation of the looped repressor-DNA complex does not require any helper proteins, nor does it call for supercoiling of the DNA (as appears to be required in other bacterial regulatory architectures [6, 7]). Further, we find that even in the absence of these mechanisms, which can only enhance the probability of loop formation, the formation of DNA loops by Lac repressor occurs more easily than would be expected on the basis of traditional views of DNA elasticity. A summary of the various measurements of short-length DNA cyclization and looping is shown in figure 9.12. The idea of this figure is to present the diversity of data that weighs in on the subject of short-length DNA elasticity. In particular, several sets of controversial measurements on DNA cyclization present different conclusions on the ease of this process at lengths of roughly 100 bp. Note that in addition, we have included both the theoretical cyclization J-factor and looping J-factor. The looping J-factor reveals that because of the less restrictive looping geometry (end points are not at same point in space and the tangents are not constrained to be equal), looping costs less free energy than does cyclization. TPM experiments like those presented here offer another avenue to resolve this issue, one that does not involve the complex ligase enzyme, the need to ensure a specific kinetic regime, nor other subtleties of the ligation reaction inherent in cyclization measurements. However, as seen in the figure, even here there are unexplained discrepancies between different TPM experiments which call for continued investigation. One observation from our own work that could have an important bearing on the differences in TPM results between different groups is that there is a substantial temperature dependence to the looping probabilities and different groups may be working at different temperatures.

Several intriguing mysteries remain which demand both further experimentation as well as theoretical analysis, e.g.: i) Why are the probabilities of DNA loop formation systematically higher than would be expected on the basis of traditional arguments about DNA elasticity, and ii) what is the significance of three repressor binding sites in the wild-type *lac* operon? To explore these questions, TPM experiments with different DNA sequences between the two operators, as well as with Lac repressor mutants that are less flexible, would go a long way towards clarifying the mechanisms at work and would provide a basis for examining the even richer action at a distance revealed in the eukaryotic setting.

9.4 Materials and Methods

9.4.1 Plasmid DNAs

Plasmid DNAs, bearing two Lac repressor binding sites spaced at a designed distance, are created using a point mutation method (QuikChange site-directed mutagenesis, Stratagene) on plasmid pUC19. Plasmid pUC19 was chosen as a starting template because it is not only a high copy plasmid but also contains two Lac repressor binding sites: $O1$ and $O3$. The procedure for creating two binding sites separated by the desired distance from template pUC19 is illustrated in figure 9.13(A). We first mutate six basepairs in the $O3$ site converting it to $O3^*$ in a way that eliminates the binding affinity for this site [87]. The resulting plasmid is called pUC19O1 indicating it only has a single $O1$ site. To construct another binding site on the pUC19O1 plasmid, we replace 20bp with the Lac repressor binding sequence Oid at a series of locations differing by 1bp increments in their distance from $O1$ using the mutagenesis method again. For some of the secondary site construction, we have to use either deletion or addition from already made plasmids with two designed binding sites. The details on primers and templates used in this process are listed in table 9.2. The final product contains two binding sites $O1$ and Oid spaced at the desired distance.

The short-loop DNA (89, 94 and 100 bp) was constructed in the following way. Plasmid pZS22-YFP was kindly provided by Michael Elowitz. The main features of the pZ plasmids are located between unique restriction sites [88]. The YFP gene comes from plasmid pDH5 (University of Washington Yeast Resource Center [89]).

A variant of the lacUV5 promoter [11] was synthesized and placed between the EcoRI and XhoI sites of pZS22-YFP in order to create pZS25'-YFP. This promoter included the -35 and -10 regions of the lacUV5 promoter, an AseI site between the two signals and a $O1$ operator at position -45 from the transcription start as shown in figure 9.14(A).

The random sequence E8-89 [35] was obtained by PCR from a plasmid kindly provided by Jonathan Widom. The primers used had a flanking AatII site and Oid operator upstream and a flanking $O1$ operator, -35 region and AseI site downstream. This PCR product was combined with the appropriate digest of pZS25'-YFP to give raise to pZS25'Oid-E89-$O1_{-45}$-YFP. This is shown schematically in figure 9.14(B). Finally, the different lengths used by Cloutier and Widom [4, 35] were generated from this template using site directed mutagenesis.

9.4.2 Construction of labeled DNAs

In TPM experiments, DNA is linked between the substrate and a bead. Two pairs of linkers: biotin-streptavidin and digoxigenin-anti-digoxigenin, are chosen to permit specific linkage of the DNA to a polystrene microsphere and glass coverslip, respectively. As illustrated in figure 9.13(B), PCR was used to amplify such labeled DNA with two modified primers. Each primer is designed to be about 20 bp in length and linked with either biotin or digoxigenin at the 5' end (Eurofins MWG Operon). In the case of the long sequence constructs, in order to optimize the PCR reaction linearized plasmids with an AatII cut are used as the template. Detailed information concerning the design of our PCR reactions is listed in table 9.3 and the constructs are shown schematically in figure 9.15. The PCR products were then purified by gel extraction (QIAquick Gel Extraction Kit, QIAGEN) and the concentration of the DNA was measured using quantitative DNA electrophoresis.

9.4.3 TPM sample preparation

TPM sample preparation involves assembly of the relevant DNA tethers and their associated reporter beads. Streptavidin-coated microspheres (Bangs lab) of diameter 490 nm served as our tethered particle. Prior to each usage, a buffer exchange on the beads was performed by three cycles of centrifugation and resuspension in TPB buffer (20 mM Tris-acetate, pH=8.0, 130 mM KCl, 4 mM $MgCl_2$, 0.1 mM DTT, 0.1 mM EDTA, 20 μg/ml acetylated BSA (Sigma-Aldrich), 80 μg/ml heparin(Sigma-Aldrich) and 0.3% biotin-free casein. Biotin-free casein colloidal buffer (5% casein colloid with 0.001% Merthiolate, RDI, Flanders, NJ) was used as a cassein source. This combination of reagents was chosen in an attempt to maximize sample yield and longevity, while minimizing non-specific adsorption of DNA and microspheres onto the coverslip.

Tethered particle samples were created inside a 20–40 μl flow cell made out of a glass slide with one hole near each end, glass coverslip, double-sided tape and tygon tubing. The coverslip and glass slide were cleaned with plasma cleaning for 4 minutes and then the flow cell was constructed as shown in figure 9.16(A). Two tygon tubes serving as an input and output were inserted into the holes on the glass slide and sealed with epoxy. A reaction chamber was created by cutting a channel on the double-sided tape, which glues the coverslip and glass slide together. Making the end of the channel round and as close to the holes of the glass slide as possible is important to avoid generating bubbles. The flow cell was then heated for about 20 seconds to seal securely.

For DNA tether assembly, the flow chamber was first incubated with 20 μg/mg polyclonal anti-digoxigenin (Roche) in PBS buffer for about 25 minutes, and then rinsed with 400 μl wash buffer (TPB buffer with no casein) followed by 400 μl of TPB buffer. 250 μl of labeled DNA in TPB buffer with about 2 pM concentration was flushed into the chamber and incubated for around 1 hour. After washing with 750 μL TPB buffer to remove any unbound DNA, a 10 pM solution of beads were introduced into the chamber and incubated for 20 minutes. Finally, unbound microspheres were removed by flushing the chamber with 1 mL TPB buffer. For looping experiments, 0.5 mL$\sim$ 1 mL LRB buffer (10 mM Tris-Hcl, pH 7.4, 200 mM KCl, 0.1 mM EDTA, 0.2 mM DTT, 5% DMSO and 0.1% biotin-free casein) containing the desired concentration of Lac repressor (a kind gift from Kathleen Matthews' lab) was then flushed into the chamber and incubated about 15 minutes before observation. Although we were able to measure the overall concentration of Lac repressor used in the experiments, the more important quantity is the concentration of active repressor which we were unable to successfully measure other than through the looping assay itself. Each flow cell preparation would typically allow to acquire data on ten tethers.

9.4.4 Data acquisition and processing

The motion of the bead is recorded through a Differential Interference Contrast (DIC) microscope at 30 frames per second. The position of the bead is tracked in the x-y plane using a cross-correlation method [90] and recorded as raw data for further analysis. Such raw positional data are subject to a slow drift due to vibrations of the experimental apparatus. A drift correction is then applied using a high pass first-order Butterworth filter at cutoff frequency 0.1Hz [25]. From the filtered data, $R^2(t)$ is then calculated as $x(t)^2 + y(t)^2$ and a running average $\sqrt{< R^2(t) >}$ is obtained using a Gaussian filter at cutoff frequency 0.033 Hz [25, 91], which corresponds to the standard deviations of the filter's impulse response time of 4 s. The traces shown in this paper are all obtained in this way.

Figures

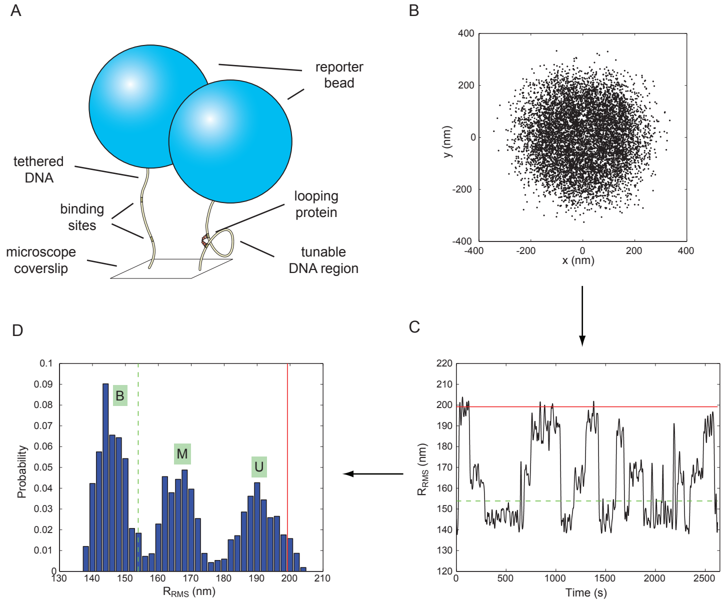


Figure 9.1: Different representations of TPM data. (A) Schematic of the TPM experiment. (B) Scatter plot of drift-corrected positional data. Each dot corresponds to the instantaneous projected position of the bead at a particular instant in time. (C) Running average of Gaussian filtered RMS motion over an effective window of 4 seconds. R is the distance from the bead center (dots in panel (B)) to the tether attachment point (centroid of all dots in panel (B)). Red (solid) and green (dashed) lines represent naively expected motion, based on calibration measurements [92] in the absence of any DNA binding protein, for 901 bp DNA and an imagined DNA for which 305+20.5 bp (the center to center distance between operators) are subtracted off of the full length 901 bp tether. (Figure 9.11 gives a more precise prediction of the expected excursions in looped states.) (D) Histogram of the RMS motion. Different peaks correspond to looped (labeled B, bottom, and M, middle) or unlooped (labeled U) states. The lines shown here are the same as those shown in (C). The presence of Lac repressor results in a shift of the excursion of the unlooped state with respect to the excursion expected from the protein-free calibration curve. This is reflected in the fact that the U peak does not coincide with the red line. The DNA used here is pUC305L1 (see Materials and Methods section) with 100 pM Lac repressor. A detailed discussion of how to go from microscopy images of beads to traces and histograms like those shown here is given the Supporting Information.

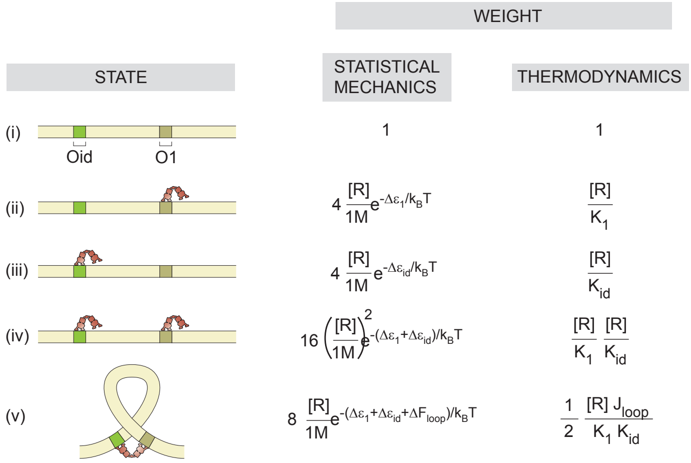

WEIGHT
STATE
STATISTICAL MECHANICS
THERMODYNAMICS
(i)
Oid
O1
1
1
(ii)
$4\frac{[R]}{1M}e^{-\Delta\varepsilon_1/k_BT}$
$\frac{[R]}{K_1}$
(iii)
$4\frac{[R]}{1M}e^{-\Delta\varepsilon_{id}/k_BT}$
$\frac{[R]}{K_{id}}$
(iv)
$16\left(\frac{[R]}{1M}\right)^2e^{-(\Delta\varepsilon_1+\Delta\varepsilon_{id})/k_BT}$
$\frac{[R]}{K_1}\frac{[R]}{K_{id}}$
(v)
$8\frac{[R]}{1M}e^{-(\Delta\varepsilon_1+\Delta\varepsilon_{id}+\Delta F_{loop})/k_BT}$
$\frac{1}{2}\frac{[R]\,J_{loop}}{K_1\,K_{id}}$

Figure 9.2: States and weights for the Lac repressor-DNA system [43]. Each of the five state classes shown in the left column has a corresponding statistical weight given by the product of the Boltzmann factor and the microscopic degeneracy of the state. All of the weights have been normalized by the weight of the state in which the DNA is unoccupied. State (v) is treated as a single looped state, even though there are multiple distinct looped configurations. The third column shows how to write these statistical weights in the language of equilibrium constants and *J*-factors. The derivation of these weights and the relation between the statistical mechanical and thermodynamic perspectives can be found in the Supporting Information.

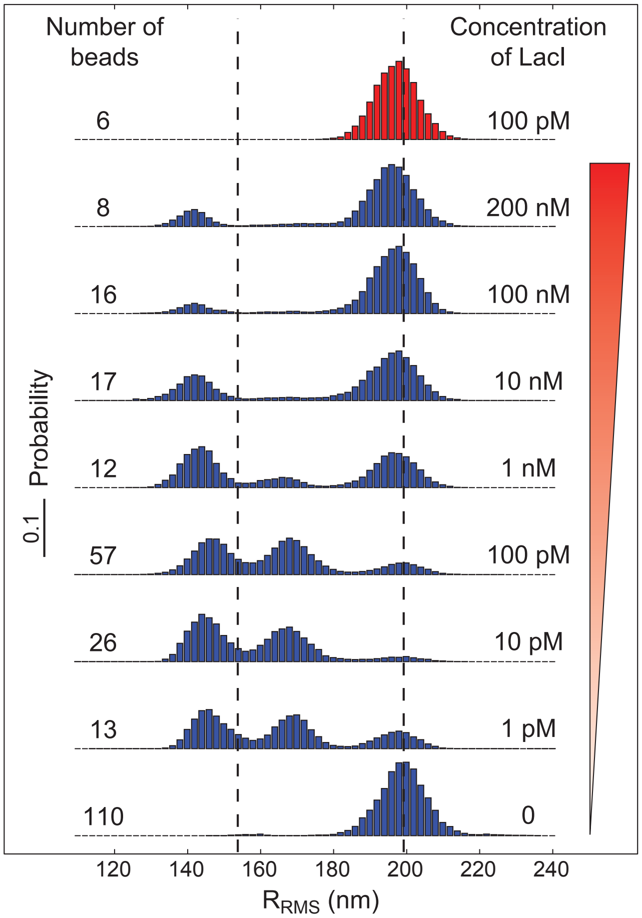


Figure 9.3: Concentration dependence of the distribution of bead excursions. The histograms show the distribution of RMS motions averaged over 4 seconds at different concentrations of Lac repressor. The blue histograms correspond to measurements for a length between operators of $L_{\mathrm{loop}} = 306$ bp (see figure 9.15), whereas the red histogram is a control where $O1$ has been deleted. The two dashed lines represent the naively expected motion, based on our calibration measurements [92]. (See figure 9.11 for a more precise prediction of the peak locations.)

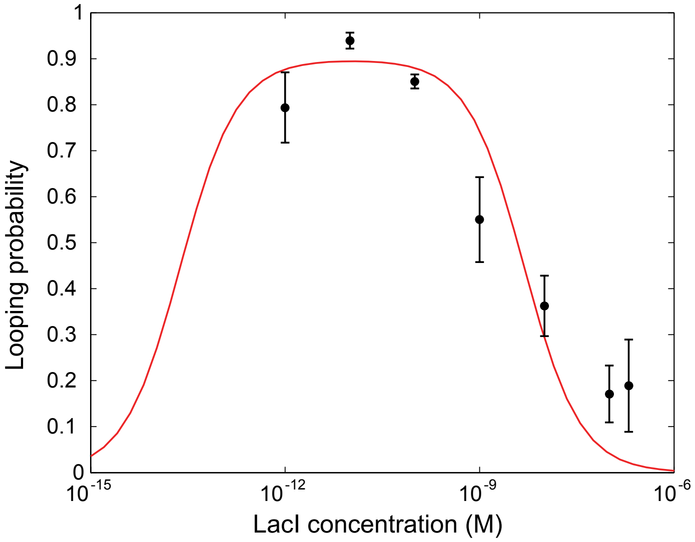


Figure 9.4: Looping probability p_{loop}, at different concentrations of Lac repressor. The DNA used in these experiments is 901 bp long and the loop length is $L_{\text{loop}} = 306$ bp. The vertical axis gives looping probability (fraction of time spent in either of the two looped states). The fraction of time spent in the looped states was calculated for each bead individually and the mean and standard error calculated for each construct. The curve is a fit to the experimental data using the statistical mechanics model described in the text. The obtained parameters are shown in table 9.1 under "Non-linear fit".

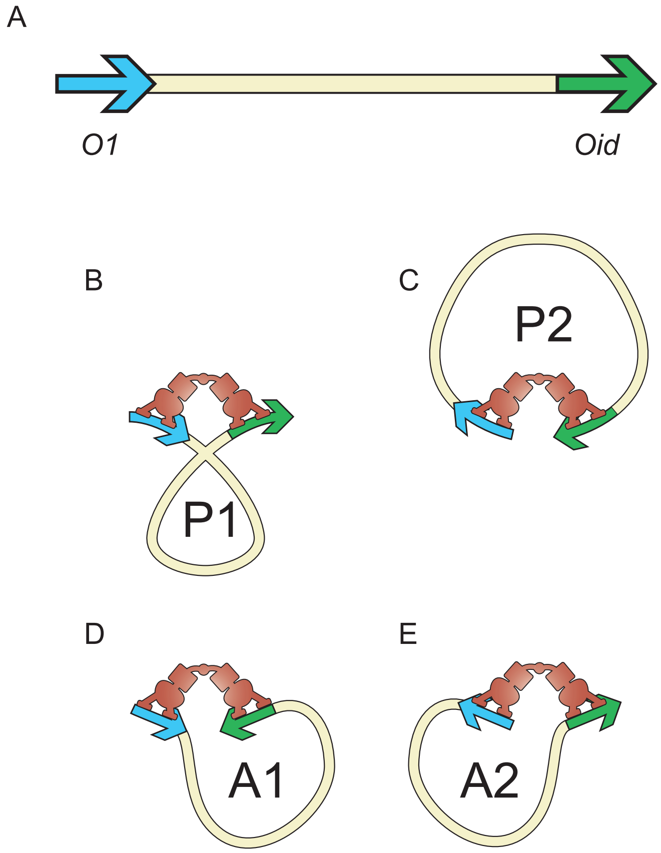


Figure 9.5: Schematic showing the different looping topologies associated with binding of Lac repressor. (A) Orientation of the two operators on the DNA. Choice of labeling orientation is arbitrary. (B)–(E) two parallel (P1 and P2) and antiparallel (A1 and A2) orientations of the DNA when subjected to Lac repressor mediated looping. We adopted the naming conventions given in references [37, 38].

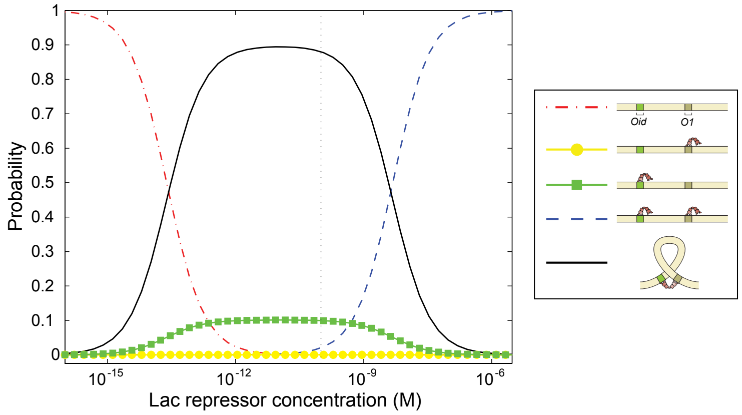


Figure 9.6: Probabilities for different states of Lac repressor and operator DNA. The curves show the probabilities of the five classes of microscopic states used in the statistical mechanics model based upon parameters shown in table 9.1. The vertical line corresponds to the concentration at which the loop length experiments in the remainder of the paper are performed.

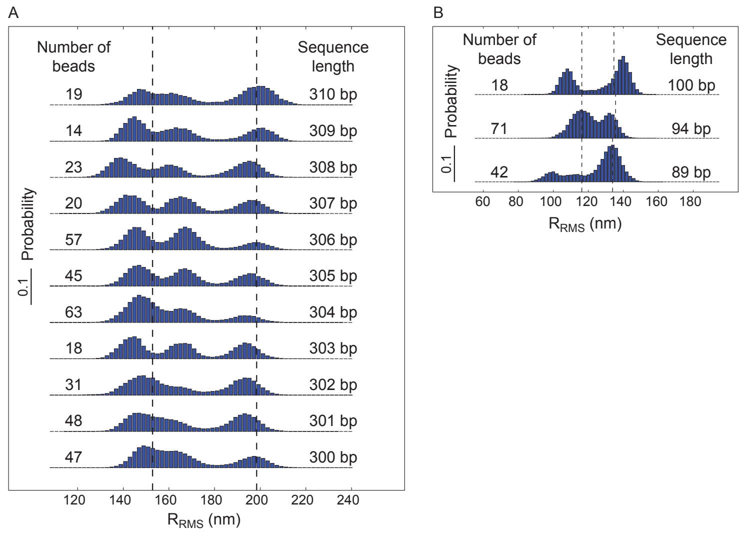


Figure 9.7: Length dependence of DNA looping. (A) Histogram of the tethered Brownian motion for DNAs with two Lac repressor binding sites spaced from $L_{\text{loop}} = 300$ bp (bottom) to 310 bp (top). (B) Histogram of the Brownian motion for DNAs with two Lac repressor binding sites spaced at $L_{\text{loop}} = 89$, 94 and 100 bp. The two dashed lines represent the naively expected motion based on our calibration measurements for the full-length tether and the same DNA when the center to center distance between operators is subtracted from the tether length. (Again see also figure 9.11.) Representative traces for each of the lengths shown here can be found in the Supporting Information.

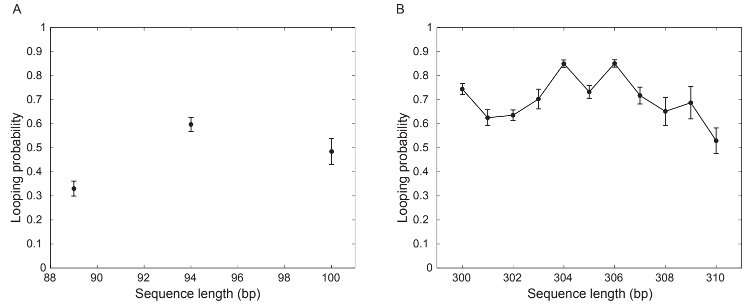


Figure 9.8: Looping probability p_{loop}, as a function of interoperator spacing. (A) Looping probability for short constructs. (B) Looping probability for one full helical repeat. These probabilities are obtained by averaging over the p_{loop} of each bead. The error bars correspond to the standard error associated with this magnitude. For more information see Supporting Information.

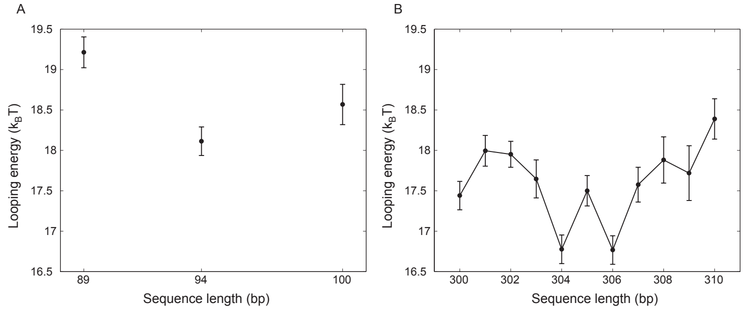


Figure 9.9: Length dependence of free energy of looping, defined via equation 9.1 with choice of reference concentration 1 M. (A) Looping free energy for short constructs. (B) Looping free energy for a full helical repeat.

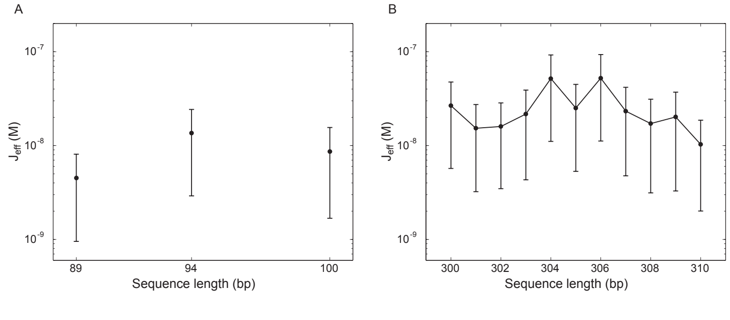


Figure 9.10: Looping J-factor resulting from TPM measurements. A) Effective J-factor for looping resulting from TPM data on short constructs. (B) Effective J-factor for looping resulting from TPM data on a full helical repeat.

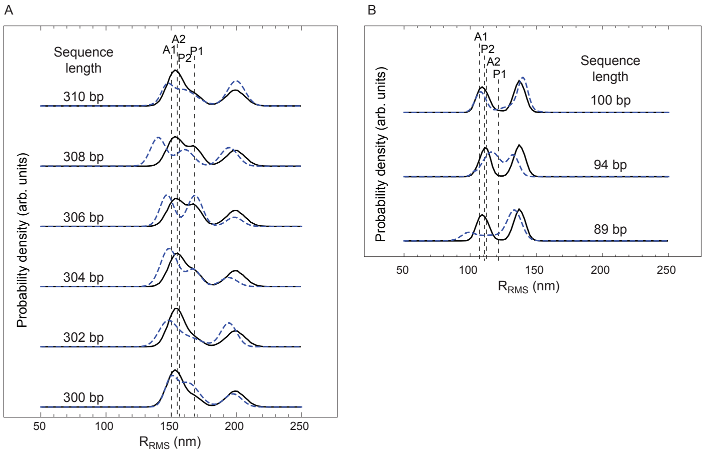


Figure 9.11: Theory and experiment for the probability density functions of RMS bead excursion for (A) our six "long chain" constructs and (B) our three "short chain" constructs. *Blue dashed curves* show the data in figure 9.7, represented as sums of three Gaussians. *Black curves* show our theoretically predicted distributions. Because our simulation results were not fits to the data, they did not reproduce perfectly the ratio of looped to unlooped occupancies. For visualization, therefore, we have adjusted this overall ratio by a factor common to all six curves. This rescaling does not affect the locations of the peaks, the relative weights of the two looped-state peaks, nor the dependences of weights on loop length L_{loop}, all of which are zero-fit-parameter predictions of our model. The model yields these histograms as the sum of five contributions, corresponding to the four looped topologies and the unlooped state. The topologies correspond to the different geometries shown in figure 9.5. The separate RMS displacements for each individual loop topology for the 89 bp case in (A) and for the 300 bp case in (B) are also shown, labeled according to the scheme in [37].

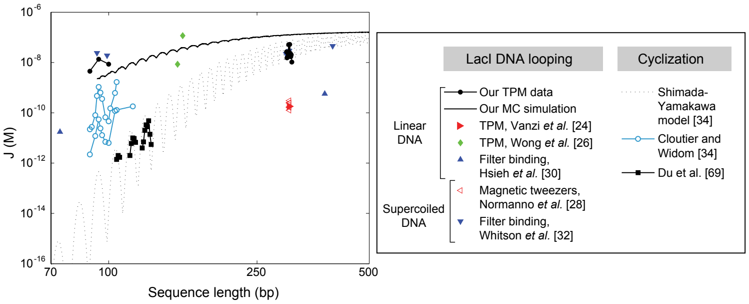


Figure 9.12: Effective J-factor from different experiments. Although the J factor obtained from cyclization experiments is not directly comparable to the looping J factor studied in this paper (due to the differences in geometry), we present the two quantities together as functions of loop length to summarize the work from many groups. Error bars have been omitted for clarity. The filter binding data is an order of magnitude estimate.

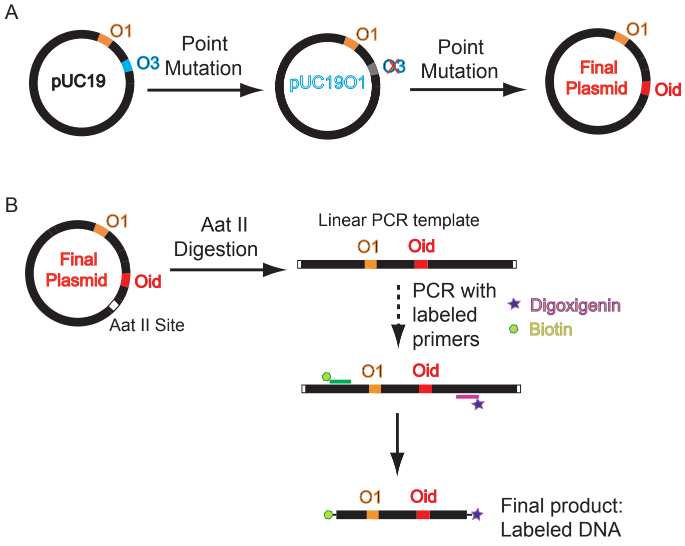


Figure 9.13: Synthesis of DNA construct. A) Schematic of the procedure for construction of the plasmid with two Lac repressor binding sites. (B) Schematic of the protocol for producing labeled DNA using a PCR reaction with labeled primers.

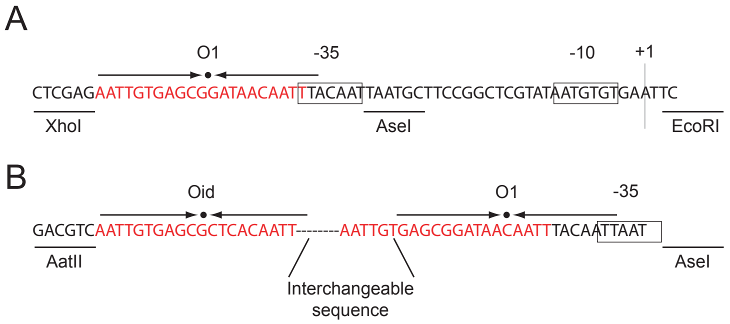

A
O1
-35
-10
+1
CTCGAGAATTGTGAGCGGATAACAATTTACAATTAATGCTTCCGGCTCGTATAATGTGTGAATTC
XhoI
AseI
EcoRI
B
Oid
O1
-35
GACGTCAATTGTGAGCGCTCACAATT--------AATTGTGAGCGGATAACAATTTACAATTAAT
AatII
AseI
Interchangeable sequence

Figure 9.14: Promoter regions of the different short loop constructs. (A) Promoter region of pZS25-YFP which has a variant of the lacUV5 promoter and an *O*1 operator upstream overlapping the -35 region. (B) Final construct that allows to insert arbitrary DNA sequences between a *Oid* and *O*1 operators.

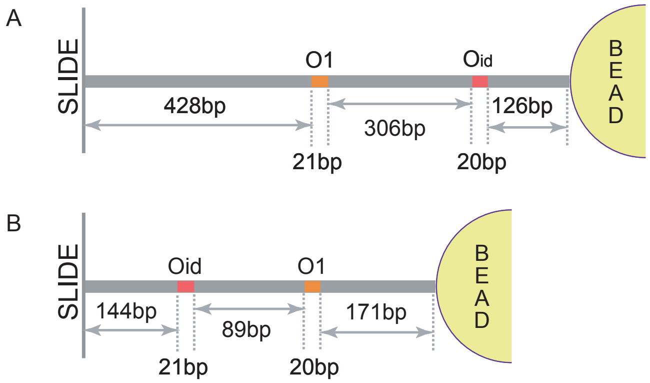


Figure 9.15: Examples of the tether constructs used. (A) In the long distance constructs *Oid* was displaced keeping the total construct length constant. (B) In the short distance constructs the sequence between the operators was altered, which results in each construct having a slightly different total length. (Drawings not to scale).

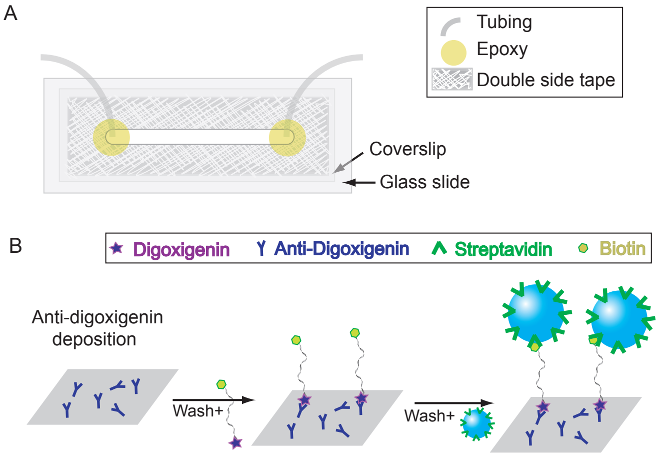


Figure 9.16: Illustration of TPM sample preparation. (A) Sketch of the flow cell. (B) The scheme for making DNA tethers.

Tables

Table 9.1: Results from the LacI titration experiments.

Parameter	Non-linear fit	Linear fit	Literature value
J_{loop}	8.6 ± 6.3 nM	52 ± 40 nM	See figure 9.12
ΔF_{loop}	$18.6 \pm 0.7\ k_BT$	$16.8 \pm 0.8\ k_BT$	N/A
K_1	0.49 ± 0.45 nM	3.0 ± 2.5 nM	$10 \sim 22$ pM [31, 58–63]
$\Delta\varepsilon_1$	$-20.0 \pm 0.9\ k_BT$	$-18.2 \pm 0.8\ k_BT$	$-23.2 \sim -24.0\ k_BT$
K_{id}	0.2 ± 2.3 pM	N/A	$2.4 \sim 8.3$ pM [93]
$\Delta\varepsilon_{id}$	$-28 \pm 9\ k_BT$	N/A	$-24.1 \sim -25.4\ k_BT$

The probability of looping as a function of Lac repressor concentration shown in figure 9.4 was fitted to the two non-linear models from equations 9.1 and 9.2. Both models were fit independently as a ways to check the robustness of the least-squares methods with respect to data reparametrization. A subset of the data corresponding to concentrations of LacI 10 pM and higher is fitted to the linear model shown in equation 9.4 and its statistical mechanics counterpart. See section S4 in the Supporting Information for a discussion of the different data fitting approaches. The literature values of the dissociation constants for $O1$ and Oid correspond to bulk binding assays performed in concentration ranges close to our TPM buffer conditions. The corresponding values for the binding energies of these operators are obtained from the dissociation constants using equations 9.11 and 9.17.

Table 9.2: Materials used in the mutagenesis process for creating plasmids with two Lac repressor binding sites.

Molecule	Primer	Template	Action	Resulting Molecule
pUC19O1	Mut0	pUC19	Replace	O1
pUC300	Mut1	pUC301	Delete 1bp	O1-300bp-Oid
pUC301	Mut2	pUC19O1	Replace	O1-301bp-Oid
pUC302	Mut3	pUC19O1	Replace	O1-302bp-Oid
pUC303	Mut4	pUC19O1	Replace	O1-303bp-Oid
pUC304	Mut5	pUC19O1	Replace	O1-304bp-Oid
pUC305	Mut6	pUC19O1	Replace	O1-305bp-Oid
pUC306	Mut7	pUC19O1	Replace	O1-306bp-Oid
pUC307	Mut8	pUC19O1	Replace	O1-307bp-Oid
pUC308	Mut9	pUC19O1	Replace	O1-308bp-Oid
pUC309	Mut10	pUC308	Add 1bp	O1-309bp-Oid
pUC310	Mut11	pUC308	Add 2bp	O1-310bp-Oid

Primer sequences(5' -> 3'):

Mut0: ctaactcacattaattgcgttgAgctcGAGgTTcgctttccagtc
Mut1: catacgagccggaa (G) cataaagtgtaaagc
Mut2: ctcggaaagaaca AATTGTGAGCGCTCACAATT aaggccaggaacc
Mut3: ctcggaaagaacat AATTGTGAGCGCTCACAATT aggccaggaaccg
Mut4: cggaaagaacatg AATTGTGAGCGCTCACAATT ggccaggaaccgt
Mut5: ggaaagaacatgt AATTGTGAGCGCTCACAATT gccaggaaccgta
Mut6: gaaagaacatgtg AATTGTGAGCGCTCACAATT ccaggaaccgtaa
Mut7: cggaaagaacatgtga AATTGTGAGCGCTCACAATT caggaaccgtaaaaag
Mut8: ggaaagaacatgtgag AATTGTGAGCGCTCACAATT aggaaccgtaaaaagg
Mut9: gaaagaacatgtgagc AATTGTGAGCGCTCACAATT ggaaccgtaaaaggc
Mut10: catacgagccggaag [C] cataaagtgtaaagc
Mut11: catacgagccggaag [CG] cataaagtgtaaagc

The capital letters in the primer sequences indicate the mutations. “()” indicates bp deletion and “[ ]” indicates bp addition. The inter-operator distance indicated here is the distance between two inner edges of the operators instead of the center-to-center distance that is commonly used in *in vivo* experiments [12–14, 87, 94].

Table 9.3: Materials used in amplifying labeled DNA using PCR.

Molecule	Template	Length(bp)	Resulting
pUC300L1	pUC300	900	Dig - 427bp-O1-300bp-Oid-132bp - Bio
pUC301L1	pUC301	901	Dig - 427bp-O1-301bp-Oid-132bp - Bio
pUC302L1	pUC302	901	Dig - 427bp-O1-302bp-Oid-131bp - Bio
pUC303L1	pUC303	901	Dig - 427bp-O1-303bp-Oid-130bp - Bio
pUC304L1	pUC304	901	Dig - 427bp-O1-304bp-Oid-129bp - Bio
pUC305L1	pUC305	901	Dig - 427bp-O1-305bp-Oid-128bp - Bio
pUC306L1	pUC306	901	Dig - 427bp-O1-306bp-Oid-127bp - Bio
pUC307L1	pUC307	901	Dig - 427bp-O1-307bp-Oid-126bp - Bio
pUC308L1	pUC308	901	Dig - 427bp-O1-308bp-Oid-125bp - Bio
pUC309L1	pUC309	902	Dig - 427bp-O1-309bp-Oid-125bp - Bio
pUC310L1	pUC310	903	Dig - 427bp-O1-310bp-Oid-125bp - Bio
E8-89	pZS25'Oid-E89-$O1_{-45}$-YFP	445	Dig - 144bp-Oid-89bp-O1-171bp - Bio
E8-94	pZS25'Oid-E94-$O1_{-45}$-YFP	450	Dig - 144bp-Oid-94bp-O1-171bp - Bio
E8-100	pZS25'Oid-E100-$O1_{-45}$-YFP	456	Dig - 144bp-Oid-100bp-O1-171bp - Bio

Primer sequences(5' -> 3'):

Plen901F: Dig - ACAGCTTGTCTGTAAGCGGATG

Plen901R: Bio - CGCCTGGTATCTTTATAGTCCTGTC

PF1: Dig - ATGCGAAACGATCCTCATCC

PR1: Bio - GCATCACCTTCACCCTCTCC

The inter-operator distances indicated here are the distance between two inner sides of the operators instead of center-to-center distance. Primers Plen901F and Plen901R were used for the long distance constructs. Primers PF1 and PR1 were used for the short distance constructs.

9.5 Supporting Information

9.5.1 Bead selection, data rejection and "representative data"

One of the most important challenges of these experiments (and perhaps any single-molecule experiment based upon watching the motions of beads tethered to single molecules) is devising systematic methods for deciding which beads are "qualified" and how to reject trajectories that are anomalous without biasing the results [18–20, 75]. To that end, we have attempted to institute a number of criteria for performing data selection that are indicated schematically in figures 9.17 and 9.18. The first attempt to "objectively" select qualified beads takes place by excising segments of the traces corresponding to the unlooped state and examining whether their motions are symmetric (i.e., jiggle in the x- and y- directions in the same way) as evidenced by the probability distribution for the x- and y- excursions. This screening permits us to select beads within a given field of view that are ostensibly properly tethered. Examples of these selection criteria are shown in figure 9.17 for the particular case where no protein is present. Typically, a fraction of roughly 20 $\sim$ 30% of the beads are rejected as a result of failure to exhibit proper symmetry or because they are stuck.

A more tricky question arises when we have to assess whether something went wrong during data acquisition that requires either all or part of a given TPM trajectory to be rejected. In some cases, the offending behavior is evident at the level of the bare images of the jiggling beads. For example, a given bead can become stuck to the surface or the DNA can break and the bead will disappear from the field of view. These events have a signature of spikes in the $\mathrm{R}_{\mathrm{RMS}}$ traces as shown in figure 9.18.

Figure 9.18 also shows an example of data that was kept with an offending region highlighted that was removed. Note that if the spike regions in trajectories were actually kept, it would have no bearing on histograms like those shown in figures 9.3 and 9.7 since the spikes will show up as features on the tails of the histograms. On the other hand, by excising certain pieces of trajectories, there can be some effect on the kinetic claims we would be able to make since these anomalies will cause errors in the dwell time measurements.

In none of the cases considered in this work were sticking events observed in any significant number. Assuming that sticking is mainly due to nonspecific interactions with the bead and the surface one would expect the shorter constructs to show the most sticking events. In order to control for this we performed TPM experiments using tethers of 351 bp in length in the absence of Lac repressor. This length is comparable to the length the short constructs (E889, E894 and E8100) would have if the sequence between the *lac* operators was removed. Out of 18 tethers characterized only 5 showed any sticking events. In those 5 traces, the sticking events corresponded to less than 4 % of the observation time for each bead (data not shown). In order to discard any contribution to the sticking events from the presence of the protein, Lac repressor was flowed in in the presence of 1 mM IPTG which serves to eliminate the binding of Lac repressor to the DNA (or at least drastically reduce it). The goal of this control is to see whether the presence of unbound protein

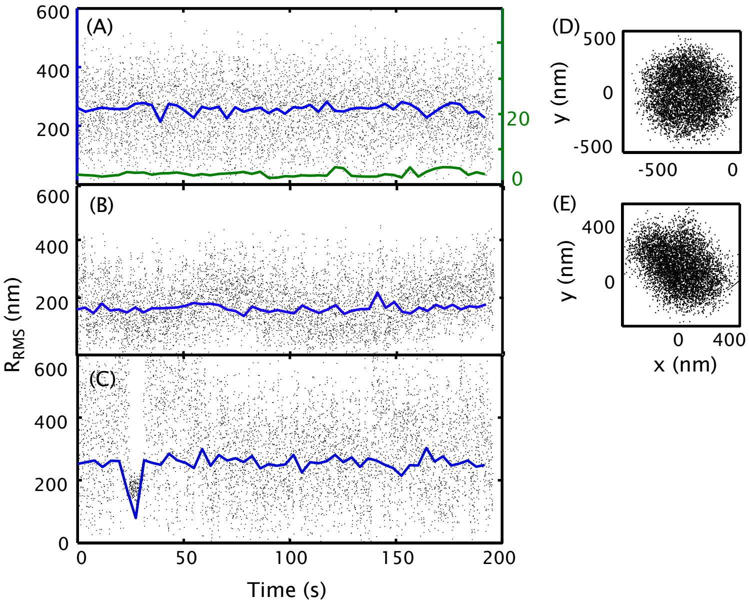


Figure 9.17: Conceptual description of data selection. All traces in this case are taken in the absence of Lac repressor and are used as the basis of choosing qualified beads for the looping study. (A) Experimental traces for a bead exercising symmetric motion (blue) and for a stuck bead (green). (B) Trajectory for a bead that exhibits non-symmetric motion. (C) Trajectory for a bead that exhibits a transient sticking event. (D) Positional data for a bead that exhibits symmetric motion. (E) Positional data corresponding to the trajectory shown in (B) and for which the motion is not symmetric.

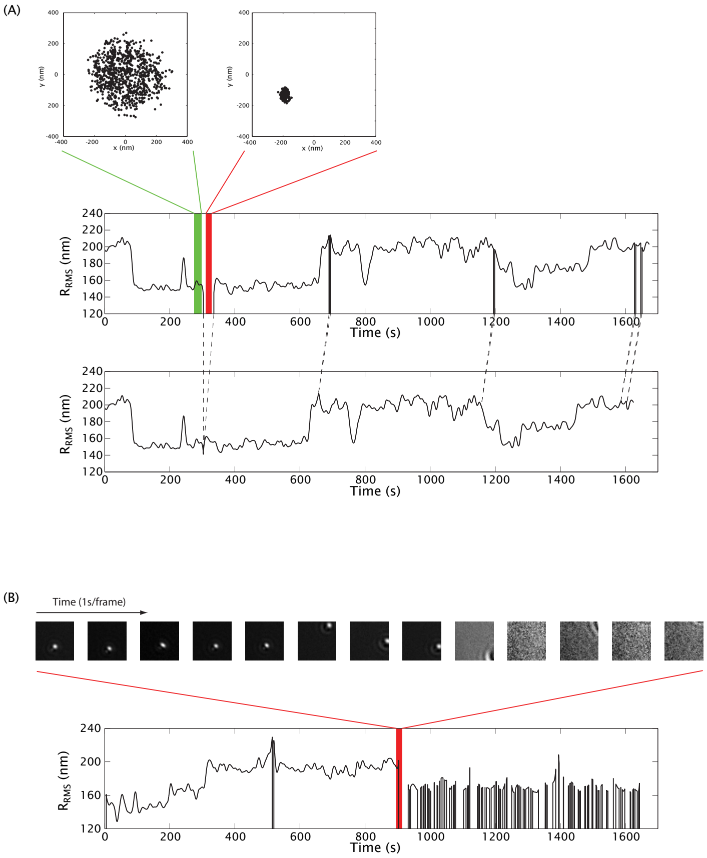


Figure 9.18: Transient sticking events and tether breaking. (A) A transient sticking event is revealed by a dramatic reduction in the movement of the bead and is associated with a spike in the $\mathrm{R_{RMS}}$ trace. These "offending" regions of the traces can be excised out and will not affect the resulting histogram, but might present an issue for any kinetic analysis as discussed in the text. (B) Signature of a tether breaking. Each movie frame is rescaled based on its maximum and minimum pixel value, which leads to overall differences in intensity between frames.

somehow induces unwanted sticking events. Out of the 7 tethers characterized all showed sticking events, but these corresponded to less than 1% of the time. Finally, there is still the chance that Lac repressor that is specifically-bound to the tether might contribute to sticking. In order to test this hypothesis we used a construct of this length with only one binding site. Here too (data not shown), there was no significant sticking lending further support for the idea that even for the short tethers, we are able to detect looping.

In order to produce histograms like those shown in figures 9.3 and 9.7 we have to sum over the histograms resulting from many individual trajectories. Figure 9.1 shows the connection between an individual TPM trace for a single bead and its corresponding motion histogram. However, since each trajectory has its own unique features, it is of interest to see how the smoothed histogram resulting from many individual trajectories emerges from the averaging process. Figure 9.19 shows the motion histogram obtained by averaging over the histograms from progressively larger numbers of individual trajectories.

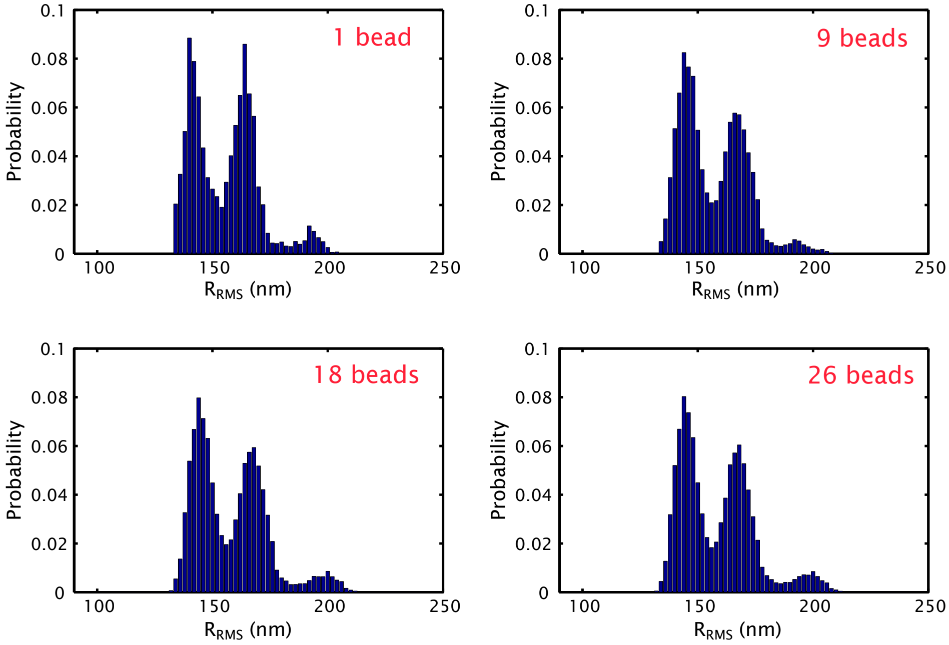


Figure 9.19: Effect of averaging on the data. These four histograms show the effects of including different numbers of beads in determining the overall average. Data obtained with pUC306L1 DNA in the presence of 10 pM Lac repressor.

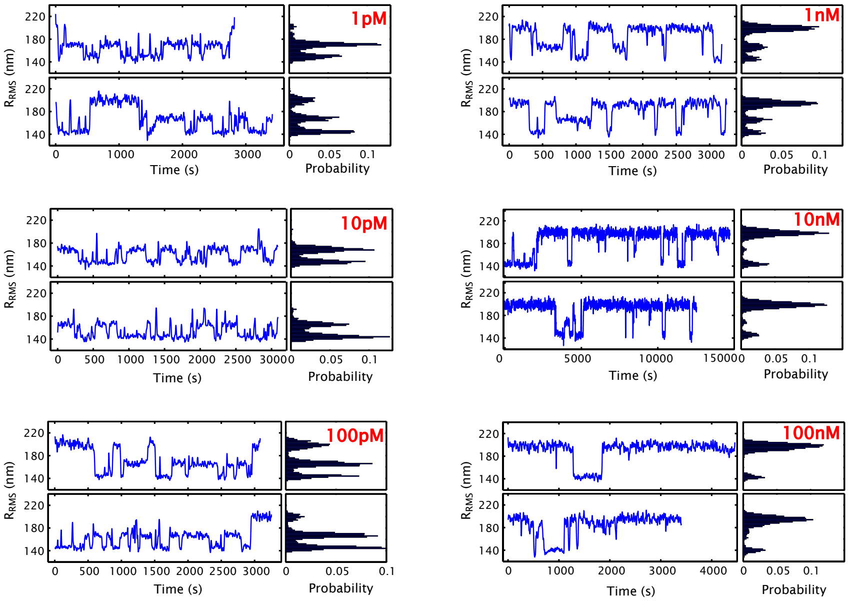


Figure 9.20: Concentration dependence of TPM trajectories. Representative examples of TPM trajectories. Typical TPM trajectories of the DNA tethered beads in the presence of different concentrations of Lac repressor varying from 1 pM to 100 nM. The total DNA length is 901 bp and the interoperator spacing is 306 bp.

Now that we have seen some of the pitfalls associated with TPM trajectories, we show "representative" examples of the individual trajectories culminating in figures 9.3 and 9.7. Figure 9.20 shows multiple examples of trajectories resulting from different concentrations of Lac repressor. Even at the level of visual inspection of these individual trajectories, it is evident that there are two distinct looping states and that the relative occupancies of the different looped and unlooped states depend upon the concentration of repressor. Similar results are shown in figures 9.21 and 9.22 which illustrate multiple individual trajectories for the case in which the interoperator spacing (rather than the Lac repressor) concentration is the experimental dial that we tune to vary the looping stability.

9.5.2 Data analysis and probabilities calculation

The data shown in figures 9.3 and 9.7 characterizes the results of many different TPM trajectories for each condition (Lac repressor concentration or interoperator spacing). We are interested in obtaining the probabilities associated with each state and to that end, we have tried a variety of different approaches to examine the sensitivity of the results to method of data analysis.

The first analysis we explored is based on directly looking at histograms such as those shown in figures 9.3 and 9.7. As mentioned in the main text, these histograms are the result of adding up the normalized contribution from each bead. One scheme for carrying this out is to fit the histogram to the sum of three Gaussians. The idea of such a fit is that there is a main peak associated with the unlooped state and then

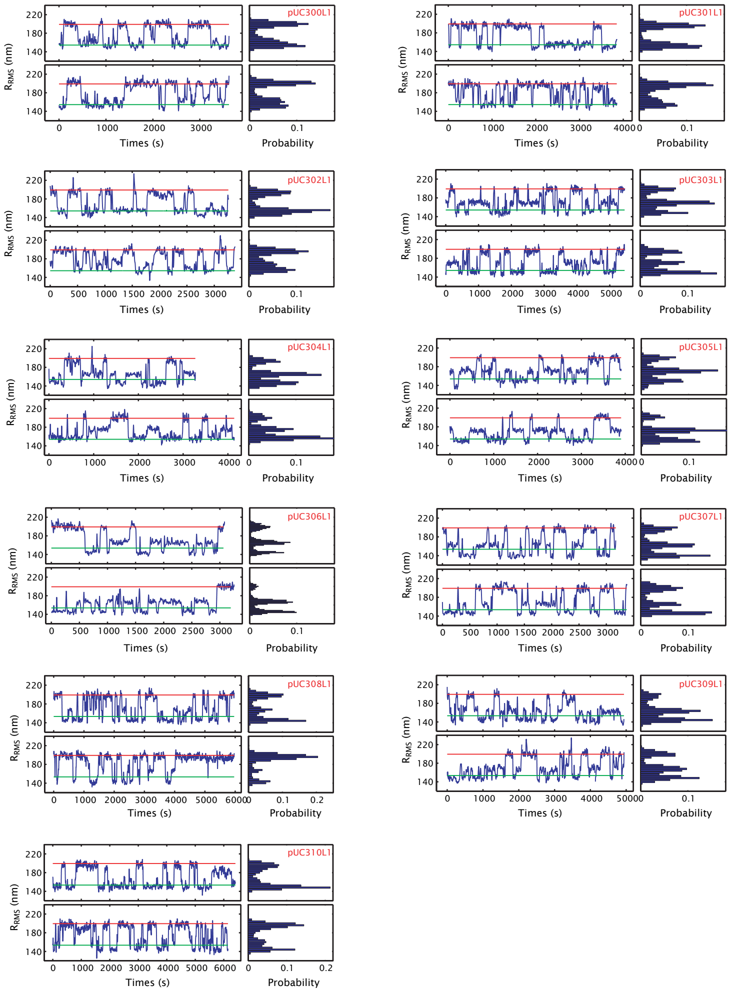


Figure 9.21: Length dependence of TPM trajectories. Typical TPM trajectories of the DNA tethered beads with interoperator spacing from 300 to 310 bp in 1 bp increments. The concentration of Lac repressor used in this set of experiments was 100 pM. The distance between the two operators is indicated in the naming of the construct.

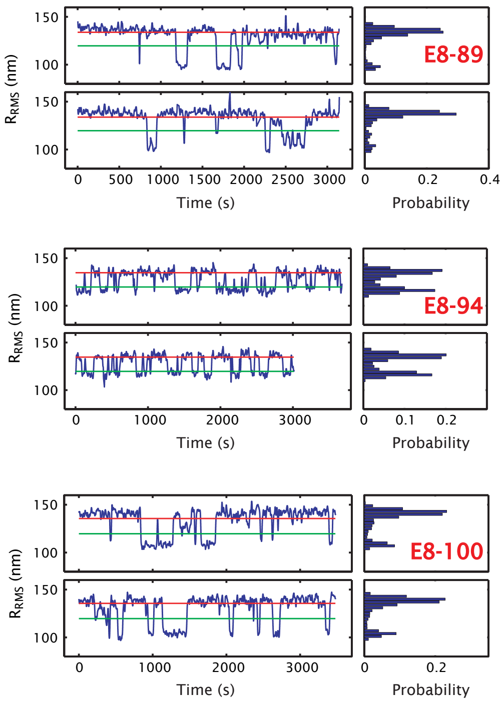


Figure 9.22: Typical TPM trajectories for DNA tethered beads with interoperator spacing of 89 bp, 94 bp and 100 bp. E8 refers to the particular sequence used in these experiments. The concentration of Lac repressor used to generate these trajectories is 100 pM. The red and green lines indicate the expected excursion for the unlooped and looped states, respectively, where the expected length of the looped state is based upon subtracting the interoperator spacing from the overall tether length.

two separate looping peaks, each of which is fit with its own Gaussian. With the fitting results in hand the area under each Gaussian can be computed, which leads to a probability assignment. We call this scheme "Gaussian Integral".

An alternative scheme is to define thresholds between the different states. The bins on either side of the thresholds are then added, giving the different probabilities. We explore two ways of calculating the thresholds: i) Finding the minimum between adjacent Gaussians from the fit described previously ("Gaussian Minimum"), and ii) finding the minimum in the histograms between peaks ("Histogram Minimum").

Finally, we have also explored the use of alternatives such as the Diffusive Hidden Markov Model ("dHMM"). The Diffusive Hidden Markov Model (DHMM) method is applied to do the kinetic analysis [95, 96] and for our present purposes permits us a different way to determine the looping probability by telling us the fraction of time spent in each of the distinct states. This method employs the concept of HMM and customizes it in a way suitable for TPM data, through which the rate constants are directly derived from the positional data obtained in the TPM experiments. To characterize the dynamical information of the beads in each state, control experiments are performed in the following ways: i) To obtain the information for the unlooped state, the bead's motion is observed in the absence of the DNA looping protein Lac repressor. ii) For the looped state, we monitor the bead's motion in the presence of a Lac repressor mutant V52C instead of Lac repressor itself. This mutant is designed to permit disulfide bond formation, which makes important contacts that are critical to DNA binding. As a result, V52C has increased affinity for DNA operators [97], leading to a measurement of primarily looped states. Such data containing only one type of looped state is selected to obtain the information that serves as input to the HMM model. One of the outcomes of the HMM analysis is an explicit statement about the amount of time spent in each of the states which can be used in turn to compute the looping probability.

One argument against the previously mentioned schemes is that they do not capture the variability inherent in single molecule experiments. Each tether will behave in a slightly different way, as is illustrated in figure 9.23 for construct pUC300L1. Notice that even though the two looped states were overlapping in figure 9.7 they are discernable in most individual traces. Figure 9.23(F) also shows a case where no call on the identity of the looped state could be made. For the long length constructs where this happened only a small fraction of the beads, between 2% and 6% would show this type of histogram. Identification of the individual loops becomes more problematic in the short length constructs. In this case around 10% of the beads would show this behavior.

The looping probabilities obtained using all these methods are shown in figure 9.24. We conclude that there is no significant variation in the results from any of the different approaches. In section 9.5.4 we show that the quantitative parameters extracted from these different looping probabilities do not differ significantly. Finally, figures 9.25 and 9.26 show the looping probability for each individual state in the cases where both states were discernible. Ultimately, it would be of great interest to use experiments like those described here to determine the looping free energies (or J_{loop}s for the different states. This is presented in

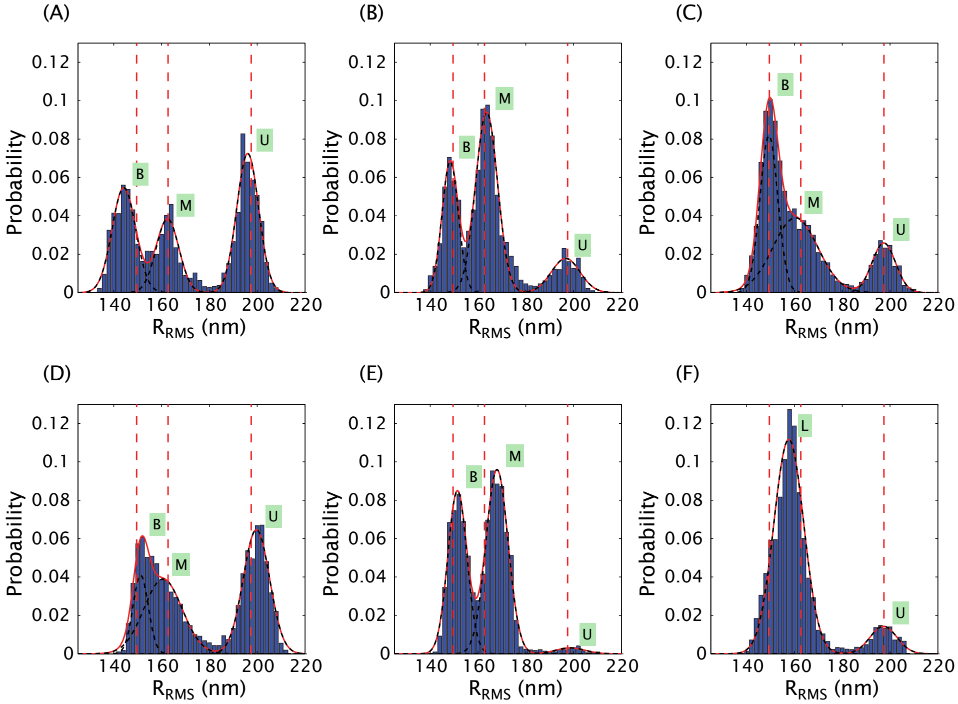


Figure 9.23: Rogues gallery of individual bead histograms. Three Gaussian fit to individual bead traces corresponding to the pUC300L1 construct. The vertical dashed lines correspond to the locations of the peaks as revealed by a three Gaussian fit to the corresponding histogram of figure 9.7. The black dashed line are the individual Gaussians, while the solid red line is their sum. (A–E) The peaks are labeled B (bottom loop), M (middle loop) and U (unlooped state). In the small fraction of cases that no decision about the identity of the looped state could be made the label L (looped state) is used.

section 9.5.5.

9.5.3 Theoretical analysis of looping

Statistical mechanics provides a powerful tool for dissecting the DNA-protein interactions that take place during transcriptional regulation. We find it convenient to derive the various expressions for binding probabilities using simple lattice models of DNA binding proteins and their DNA targets. These models can then be reinterpreted in the familiar language of equilibrium constants and effective J-factors. In this section, we sketch the derivations of the formulae used in the main body of the paper. An alternative derivation appears in [40].

9.5.3.1 Simple binding of Lac repressor

In a lattice model, we imagine the solution as discretized into a set of Ω boxes of volume v. The R repressors are free to occupy any of these distinct boxes which provide a simple and convenient basis for computing the entropic contribution to the overall free energy. A repressor in solution has an energy ε_{sol} which appears

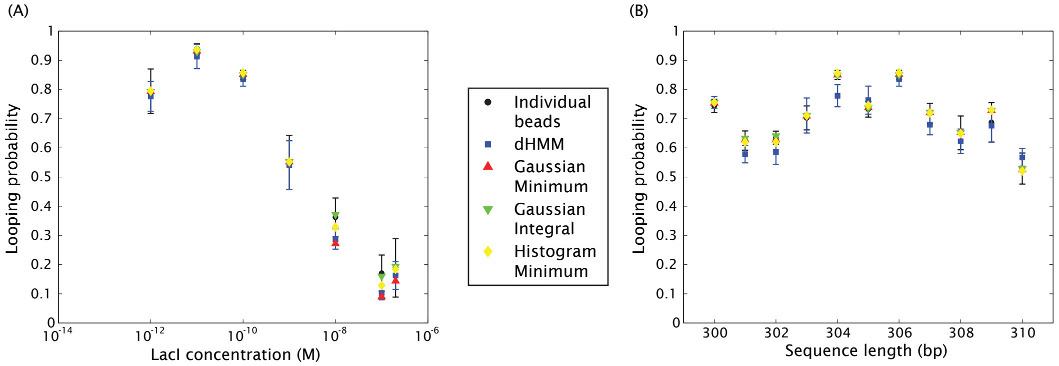


Figure 9.24: Different approaches for calculating the looping probability. The looping probability as a function of (A) concentration and (B) sequence length, calculated using the approaches described in the text.

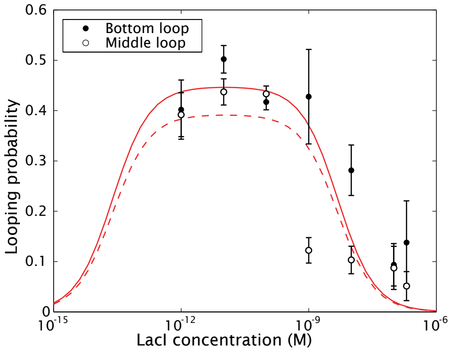


Figure 9.25: Individual loops vs. concentration. Probability of each looped state as a function of concentration. The lines are fits to the non-linear model from equation 9.20.

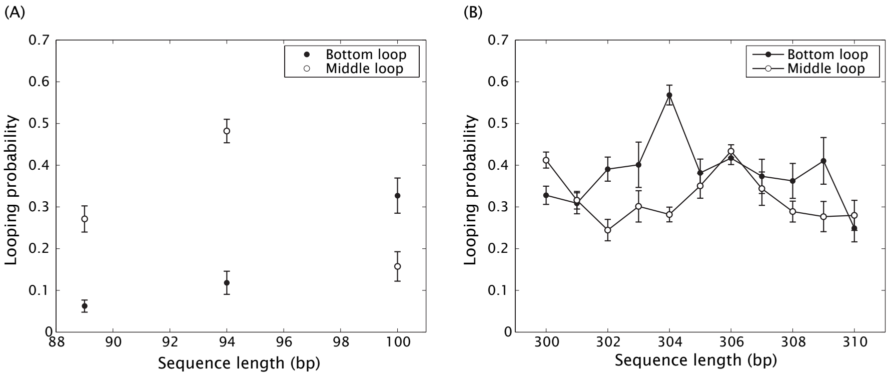


Figure 9.26: Individual loops vs. phasing. Probability of each looped state as a function of sequence length. (A) Short loops, (B) a full cycle at 300 bp.

in the Boltzmann factor. The configurational degrees of freedom (both translational and rotational) in this model are taken care of by assigning the molecules to one of the Ω boxes available in our lattice model of the solution and by noting that there is a factor of $\frac{8\pi^2}{\delta\omega}$ associated with its rotational degrees of freedom (4π for the directions in which the molecule can point on the unit sphere and 2π for the rotation around the protein's axis). The partition function of R repressors in the solution is

$$Z_{sol} = \binom{\Omega}{R} e^{-\beta R \varepsilon_{sol}} \left(\frac{8\pi^2}{\delta\omega}\right)^R . \tag{9.7}$$

Now we introduce a DNA molecule with one binding site. This case is appropriate when LacI is in excess of the DNA. When one Lac repressor from the solution binds to the operator it now has an energy ε_b associated with the binding itself and a "tether" energy ε_t associated with the extra binding head that is still in the solution. Next, we exploit the fact that we can choose either head to bind to the operator of interest and this head can bind in two distinct orientations, yielding a factor of 4 degeneracy in this state. The total partition function is

$$Z = Z_{sol}(R) + 4 Z_{sol}(R-1) e^{-\beta(\varepsilon_b + \varepsilon_t)}. \tag{9.8}$$

This translates into the following probability of binding

$$p_{bound} = \frac{4 \frac{\delta\omega}{8\pi^2} \frac{R}{\Omega} e^{-\beta\Delta\varepsilon}}{1 + 4 \frac{\delta\omega}{8\pi^2} \frac{R}{\Omega} e^{-\beta\Delta\varepsilon}}, \tag{9.9}$$

where we have defined $\Delta\varepsilon = \varepsilon_b + \varepsilon_t - \varepsilon_{sol}$.

We recover the usual formula when characterizing binding using dissociation constants

$$p_{bound} = \frac{[R]/K_d}{1 + [R]/K_d}, \tag{9.10}$$

if we make the identification

$$K_d = \frac{1}{4v} \frac{8\pi^2}{\delta\omega} e^{\beta\Delta\varepsilon}. \tag{9.11}$$

With this result in hand we are ready to address the more complex case of DNA looping.

9.5.3.2 DNA looping by Lac repressor

We now have two operators, each one with a binding energy ε_1 and ε_{id}, corresponding to the operators $O1$ and Oid, respectively. We consider the usual five classes of states that include: i) free operators, ii+iii) one of the operators occupied, iv) both operators occupied by different LacI molecules, and v) LacI looping both operators, which can happen in multiple configurations. The partition function is

$$\begin{aligned} Z &= Z_{sol}(R) + 4Z_{sol}(R-1)e^{-\beta\varepsilon_t}\left(e^{-\beta\varepsilon_1} + e^{-\beta\varepsilon_{id}}\right) \\ &\quad +16Z_{sol}(R-2)e^{-\beta(\varepsilon_1+\varepsilon_{id}+2\varepsilon_t)} + \\ &\quad +\sum_i Z_{sol}(R-1)e^{-\beta(\varepsilon_1+\varepsilon_{id}+F_{loop,i})}. \end{aligned} \tag{9.12}$$

The factors of 4 in the second and third term correspond to the degeneracy described above. The factor of 16 in the fourth term accounts for all of the different ways of binding two repressors independently. Here we defined $F_{loop,i}$ as the looping free energy associated with a particular configuration (orientation of operators with respect to the molecule). The sum in the last term includes all four possible loop topologies [37, 47] and the fact that we are thinking of the two binding heads of LacI as being distinguishable. Defining α and β as state variables that describe the orientation of $O1$ and Oid with respect to the binding heads (see figure 9.5, respectively, we can write the sum as

$$\sum_i = \sum_{\text{heads}} \sum_{\alpha,\beta}. \tag{9.13}$$

The sum over the heads results in a factor of two, since none of the terms inside the sum actually depend on that choice. We next define the overall looping energy ΔF_{loop} by

$$e^{-\beta\Delta F_{\text{loop}}} = \frac{1}{\sum_{\alpha,\beta} 1}\sum_{\alpha,\beta} e^{-\beta F_{\text{loop},\alpha,\beta}} = \frac{1}{4}\sum_{\alpha,\beta} e^{-\beta\Delta F_{\text{loop},\alpha,\beta}}. \tag{9.14}$$

Using the calculations and definitions from section 9.5.3.1 we arrive at the looping probability

$$\begin{aligned} p_{\text{loop}} &= \left[8\frac{R}{\Omega}\frac{\delta\omega}{8\pi^2}e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id}+\Delta F_{\text{loop}}+2\varepsilon_t-\varepsilon_{sol})}\right] \\ &\quad \left[1 + 4\frac{R}{\Omega}\frac{\delta\omega}{8\pi^2}\left(e^{-\beta\Delta\varepsilon_1} + e^{-\beta\Delta\varepsilon_{id}}\right) + 16\frac{R(R-1)}{\Omega^2}\left(\frac{\delta\omega}{8\pi^2}\right)^2 e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id})} + \right. \\ &\quad \left. 8\frac{R}{\Omega}\frac{\delta\omega}{8\pi^2}e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id}+\Delta F_{\text{loop}}+2\varepsilon_t-\varepsilon_{sol})}\right]^{-1}. \end{aligned} \tag{9.15}$$

Notice that the term that corresponds to looping has the energy $\Delta F_{\text{loop}} + 2\varepsilon_t - \varepsilon_{sol}$. In principle this is the parameter associated with looping, but it also includes information about the energetics of LacI when it is in solution and when it has only one head bound to the DNA. However, we can make the assumption that the energy associated with having half a LacI in solution, ε_t is half the energy of having a full LacI in solution, ε_{sol}. This is equivalent to saying that there is no change in the energetics of binding if the other head is already bound, that there is no allosteric cooperativity. If this is true then the parameter obtained from an experiment where p_{loop} is measured will actually be ΔF_{loop}.

Since we measure concentration of Lac repressor rather than absolute number of repressor molecules we want to rewrite this formula as a function of $[R]$ using the lattice definitions

$$\frac{R}{\Omega} = \frac{R}{\Omega v} v = [R]v. \tag{9.16}$$

The parameter v corresponds to the volume of a lattice site, which means that Ωv corresponds to the whole volume. We now make the choice of a standard concentration

$$\frac{1}{v}\frac{8\pi^2}{\delta\omega} = 1\ \text{M}, \tag{9.17}$$

which turns the looping probability from equation 9.15 into equation 9.1 which we repeat here for completeness

$$\begin{aligned} p_{\text{loop}} \;\; = \;\; & \left[8\frac{[R]}{1\ \text{M}} e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id}+\Delta F_{\text{loop}})}\right] \\ & \left[1 + 4\frac{[R]}{1\ \text{M}}\left(e^{-\beta\Delta\varepsilon_1} + e^{-\beta\Delta\varepsilon_2}\right) + 16\left(\frac{[R]}{1\ \text{M}}\right)^2 e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id})} + \right. \\ & \left. 8\frac{[R]}{1\ \text{M}} e^{-\beta(\Delta\varepsilon_1+\Delta\varepsilon_{id}+\Delta F_{\text{loop}})}\right]^{-1}. \end{aligned}$$

Finally, we make the connection to the thermodynamic formalism using equations 9.11 and by defining that

$$J_{\text{loop}} = \frac{1}{v}\frac{8\pi^2}{\delta\omega} e^{-\beta\Delta F_{\text{loop}}}. \tag{9.18}$$

The point here is to use simple binding to define the parameters K_1, K_{id} and cyclization to assign the parameter J_{loop} [98]. Here, we use a looping J_{loop} factor rather than the regular factor J factor to emphasize the fact that the boundary conditions are different from those present in cyclization, where J is clearly defined [99]. In this way, we appeal to these other experiments semantically and plug their definitions into the expression for the looping probability derived above. This results in equation 9.2, namely

$$p_{\text{loop}} = \frac{\frac{1}{2}\frac{[R]J_{\text{loop}}}{K_1 K_{id}}}{1 + \frac{[R]}{K_1} + \frac{[R]}{K_{id}} + \frac{[R]^2}{K_1 K_{id}} + \frac{1}{2}\frac{[R]J_{\text{loop}}}{K_1 K_{id}}},$$

where J_{loop} is the average of the individual J_{loop} factors over α and β as defined in equation 9.3.

In the case where we distinguish between bottom and middle looped states we can split J_{loop} into their corresponding looping J factors

$$J_{\text{loop}} = \frac{1}{2}\left(J_{\text{loop,B}} + J_{\text{loop,M}}\right). \tag{9.19}$$

In this case, for example, the probability of looping into the bottom state can be written as

$$p_{\text{loop}} = \frac{\frac{1}{4}\frac{[R]J_{\text{loop,B}}}{K_1 K_{id}}}{1 + \frac{[R]}{K_1} + \frac{[R]}{K_{id}} + \frac{[R]^2}{K_1 K_{id}} + \frac{1}{2}\frac{[R]J_{\text{loop}}}{K_1 K_{id}}}. \tag{9.20}$$

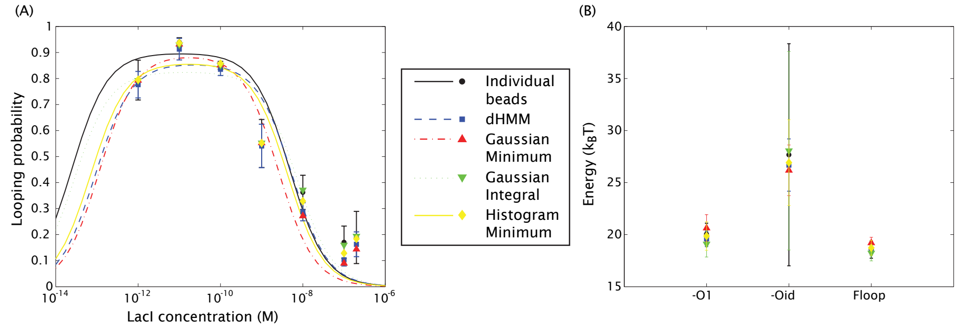


Figure 9.27: Alternative methods for fitting the looping probabilities. (A) Different schemes for determining the looping probability from the data result in slightly different fits for the concentration dependent data. (B) Results of the various fits performed in (A). Notice how the model cannot constrain the binding energy of *Oid* very accurately.

9.5.4 Comparison of theory and experiment

One of the important goals of this work is to demand a rich interplay between theories of transcriptional regulation and corresponding experiments. To that end, the entirety of the data presented in the paper is viewed through the prism of the statistical mechanics model described above.

One of the questions that we have examined is how the statistical mechanics fit depends upon the choice of how we analyze the data to determine the looping probability. Examples of different schemes for determining the looping probability and their allied fits are shown in figure 9.27. In the main body of the paper, we presented looping probabilities based upon Gaussian fits to the looping peaks. However, we have also explored the use of alternatives such as the Diffusive Hidden Markov Model.

Another point of curiosity concerns the extent to which our fits for the equilibrium constants and effective J-factor depends upon which points from figure 9.4 are actually used to make the fit. Figure 9.28 shows the fit to both K_1 and J_{loop} as a function of the particular model (non-linear or linear) and range of data points from figure 9.4 that are used in the fit. The key observation is that the final two data points (i.e., those at the largest concentrations of Lac repressor) lead to a systematic shift in the values for both K_1 and J_{loop} when fitting using the linear model from equation 9.4. Another interesting point revealed by figure 9.28(A) is that the full non-linear model fit results in a value for K_1 that is too large relative to the literature value by roughly a factor of 10, corresponding to a difference in binding energy of roughly 2 k_BT.

The dependence of our fits on the choice of data points included is also revealed in figure 9.29. In this case, we show the result of using equation 9.4 as the basis of the fit and including different subsets of the data from figure 9.4.

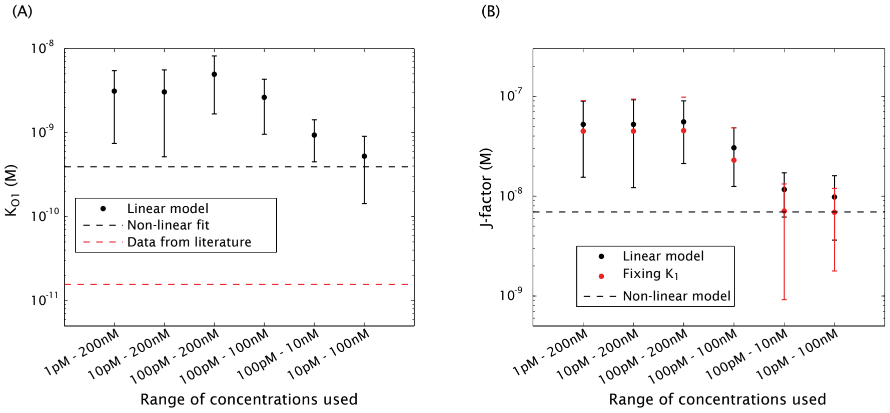


Figure 9.28: Sensitivity of fits to the method of data analysis. (A) Different fits to the value of K_1 using the linear model of equation 9.4 and different ranges of data points from figure 9.4. The results corresponding to the non-linear model of equation 9.2 are also shown. (B) Different fits to the value of J_{loop} using the linear and non-linear models as shown in (A). "Fixing K_1" corresponds to fixing the $O1$ dissociation constant to the literature value shown in table 9.1.

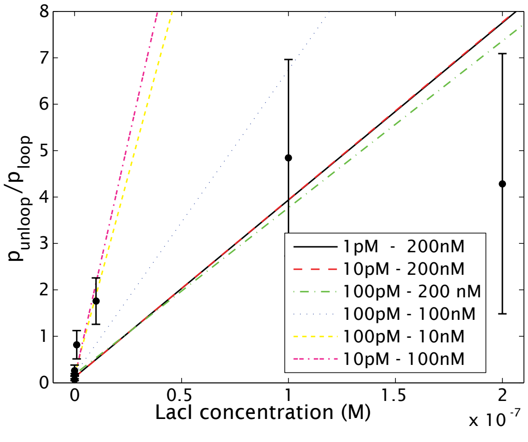


Figure 9.29: Sensitivity of linear fits to the range of data used. Different ranges of concentration from figure 9.4 are fit using the linear model of equation 9.4.

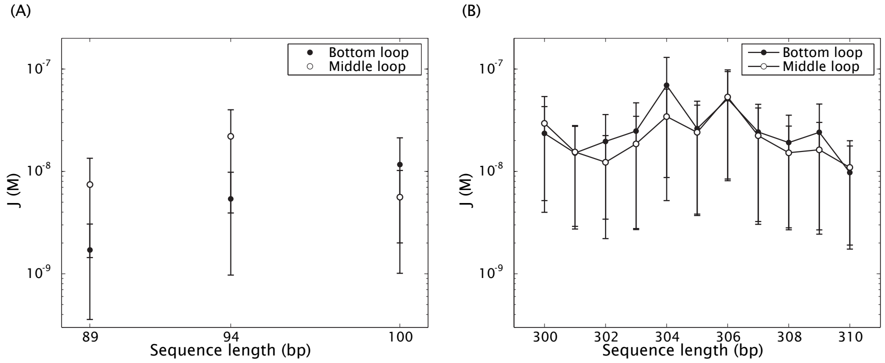


Figure 9.30: Individual loops J_{loop} as a function of sequence length. (A) Results for short constructs, (B) results for long constructs.

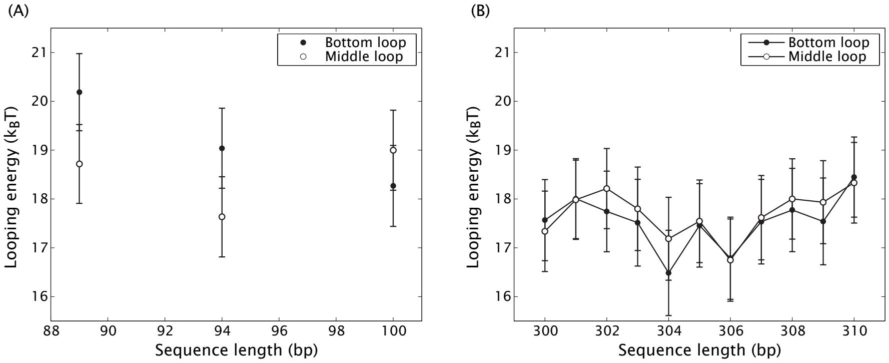


Figure 9.31: Individual loops energies as a function of sequence length. (A) Results for short constructs, (B) results for long constructs.

9.5.5 Individual looped states

In figures 9.25 and 9.26 we showed the looping probabilities corresponding to each individual loop: the bottom and middle loops. In order to analyze these results we can construct an individual loop ratio analogous to the one defined in equation 9.4. For the case of the bottom loop, for example, this is

$$p_{\text{ratio,B}} = \frac{p_{\text{loop,B}}}{p_{\text{unloop}}} = \frac{4K_1}{J_{\text{loop,B}}} + \frac{4[R]}{J_{\text{loop,B}}}. \tag{9.21}$$

Using an approach analogous to the one leading to equation 9.5 we obtain the looping J factors associated with each individual loop as shown in figure 9.30. In figure 9.31 we show their corresponding looping energies.

9.5.6 Monte Carlo simulation

Our mathematical model built on our previous work [72, 77, 100], which showed that a Gaussian-sampling simulation could accurately model the experimentally observed relation between DNA tether length and TPM bead motion by including an effective entropic stretching force from bead–wall repulsion. This technique is essentially a Monte Carlo evaluation of the equilibrium partition function of a chain. Instead of a Metropolis implementation, we simply generated many discretized chains using Gaussian distributions for each link's bending and twisting angles, then discarded any such chains that violated the global steric constraints. To compute looping J factors, we modified our previous code to monitor the separation and relative orientation of the operator centers in the generated chains, and found the fraction of all chains that met the conditions needed for looping. See [40] for more details.

To obtain the distributions of bead excursion shown in figure 9.11, we needed to make a correction before comparing to the experimental data. Our video camera gathers light for almost the entire 33 ms video frame time. This time scale is an appreciable fraction of the bead's diffusion time in the trap created by its tether, leading to a blurring of the bead image and an apparent reduction of bead RMS excursion. We measured this effect by looking at the apparent RMS excursion for a bead/tether system with many different shutter times, then corrected our numerically generated values for the position of the bead center to account for blurring [40].

In addition, we reduced our simulation data in a way that parallels what was done with the experimental data. The experiment takes data in the form of a time series for the projected location of the bead center (relative to its attachment), that is, $(x(t), y(t))$. We found the length-squared of these position vectors, R^2, then applied a Gaussian filter that essentially averaged over a 4-s window. To simulate equilibrium averages in this context, we harvested batches of N_{samp} independent simulated chains and found the standard deviation of excursion within each batch. From the resulting series of values for $R_{\mathrm{RMS}} = \sqrt{\langle R^2\rangle_{N_{\mathrm{samp}}}}$, we made a histogram representing the probability density function of R_{RMS}. To choose an appropriate value for N_{samp}, we found a characteristic time scale for bead diffusion from the time autocorrelation function of R_{RMS}, then divided the 4 s window into N_{samp} slots corresponding to the larger of the frame time, 33 ms, or the bead diffusion time [40].

Bibliography

[1] L. Han, H. G. Garcia, S. Blumberg, K. B. Towles, J. F. Beausang, P. C. Nelson, and R. Phillips. Concentration and length dependence of DNA looping in transcriptional regulation. *PLoS One*, 4(5):e5621, 2009.

[2] H. G. Garcia, P. Grayson, L. Han, M. Inamdar, J. Kondev, P. C. Nelson, R. Phillips, J. Widom, and P. A. Wiggins. Biological consequences of tightly bent DNA: The other life of a macromolecular celebrity. *Biopolymers*, 85(2):115–30, 2007.

[3] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–8, 2006.

[4] T. E. Cloutier and J. Widom. Spontaneous sharp bending of double-stranded DNA. *Mol Cell*, 14(3):355–62, 2004.

[5] S. Adhya. Multipartite genetic control elements: Communication by DNA loop. *Annu Rev Genet*, 23:227–50, 1989.

[6] R. Schleif. DNA looping. *Annu Rev Biochem*, 61:199–223, 1992.

[7] K. S. Matthews. DNA looping. *Microbiol Rev*, 56(1):123–36, 1992.

[8] R. W. Zeller, J. D. Griffith, J. G. Moore, C. V. Kirchhamer, R. J. Britten, and E. H. Davidson. A multimerizing transcription factor of sea urchin embryos capable of looping DNA. *Proc Natl Acad Sci U S A*, 92(7):2989–93, 1995.

[9] T. M. Dunn, S. Hahn, S. Ogden, and R. F. Schleif. An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proc. Natl. Acad. Sci. USA*, 81(16):5017–20, 1984.

[10] M. Ptashne. *A genetic switch: Phage lambda revisited.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 3rd edition, 2004.

[11] B. MÜLler-Hill. *The lac operon: A short history of a genetic paradigm.* Walter de Gruyter, Berlin, New York, 1996.

[12] J. MÜLler, S. Oehler, and B. MÜLler-Hill. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol*, 257(1):21–9, 1996.

[13] N. A. Becker, J. D. Kahn, and L. J. Maher Iii. Bacterial repression loops require enhanced DNA flexibility. *J Mol Biol*, 349(4):716–30, 2005.

[14] N. A. Becker, J. D. Kahn, and L. J. Maher 3rd. Effects of nucleoid proteins on DNA repression loop formation in *escherichia coli*. *Nucleic Acids Res*, 35(12):3988–4000, 2007.

[15] D. A. Schafer, J. Gelles, M. P. Sheetz, and R. Landick. Transcription by single molecules of rna polymerase observed by light microscopy. *Nature*, 352(6334):444–8, 1991.

[16] H. Yin, R. Landick, and J. Gelles. Tethered particle motion method for studying transcript elongation by a single rna polymerase molecule. *Biophys J*, 67(6):2468–78, 1994.

[17] F. Vanzi, S. Vladimirov, C. R. Knudsen, Y. E. Goldman, and B. S. Cooperman. Protein synthesis by single ribosomes. *RNA*, 9(10):1174–9, 2003.

[18] N. Pouget, C. Dennis, C. Turlan, M. Grigoriev, M. Chandler, and L. Salome. Single-particle tracking for DNA tether length monitoring. *Nucleic Acids Res*, 32(9):e73, 2004.

[19] S. Blumberg, A. V. Tkachenko, and J. C. Meiners. Disruption of protein-mediated DNA looping by tension in the substrate DNA. *Biophys J*, 88(3):1692–701, 2005.

[20] N. Pouget, C. Turlan, N. Destainville, L. Salome, and M. Chandler. Is911 transpososome assembly as analysed by tethered particle motion. *Nucleic Acids Res*, 34(16):4313–23, 2006.

[21] B. van den Broek, F. Vanzi, D. Normanno, F. S. Pavone, and G. J. Wuite. Real-time observation of DNA looping dynamics of Type IIE restriction enzymes NaeI and NarI. *Nucleic acids research*, 34(1):167–174, 2006.

[22] S. F. Tolic-Norrelykke, M. B. Rasmussen, F. S. Pavone, K. Berg-Sorensen, and L. B. Oddershede. Stepwise bending of DNA by a single tata-box binding protein. *Biophys J*, 90(10):3694–703, 2006.

[23] R. F. Guerra, L. Imperadori, R. Mantovani, D. D. Dunlap, and L. Finzi. DNA compaction by the nuclear factor-y. *Biophys J*, 93(1):176–82, 2007.

[24] L. Finzi and J. Gelles. Measurement of lactose repressor-mediated loop formation and breakdown in single DNA molecules. *Science*, 267(5196):378–80, 1995.

[25] F. Vanzi, C. Broggio, L. Sacconi, and F. S. Pavone. Lac repressor hinge flexibility and DNA looping: Single molecule kinetics by tethered particle motion. *Nucleic Acids Res*, 34(12):3409–20, 2006.

[26] C. Zurla, A. Franzini, G. Galli, D. D. Dunlap, D. E. A. Lewis, S. Adhya, and L. Finzi. Novel tethered particle motion analysis of cI protein-mediated DNA looping in the regulation of bacteriophage lambda. *J Phys–Cond Matt*, 18(14):S225–S234, 2006.

[27] O. K. Wong, M. Guthold, D. A. Erie, and J. Gelles. Interconvertible lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biol*, 6(9):e232, 2008.

[28] C. Zurla, T. Samuely, G. Bertoni, F. Valle, G. Dietler, L. Finzi, and D. D. Dunlap. Integration host factor alters laci-induced DNA looping. *Biophys Chem*, 128(2-3):245–52, 2007.

[29] D. Normanno, F. Vanzi, and F. S. Pavone. Single-molecule manipulation reveals supercoiling-dependent modulation of *lac* repressor-mediated DNA looping. *Nucleic Acids Res*, 36(8):2505–13, 2008.

[30] H. KräMer, M. NiemÖLler, M. Amouyal, B. Revet, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. *lac* repressor forms loops with linear DNA carrying two suitably spaced *lac* operators. *EMBO J*, 6(5):1481–91, 1987.

[31] W. T. Hsieh, P. A. Whitson, K. S. Matthews, and R. D. Wells. Influence of sequence and distance between two operators on interaction with the *lac* repressor. *J Biol Chem*, 262(30):14583–91, 1987.

[32] H. KräMer, M. Amouyal, A. Nordheim, and B. MÜLler-Hill. DNA supercoiling changes the spacing requirement of two *lac* operators for DNA loop formation with *lac* repressor. *EMBO J*, 7(2):547–56, 1988.

[33] P. A. Whitson, W. T. Hsieh, R. D. Wells, and K. S. Matthews. Influence of supercoiling and sequence context on operator DNA binding with *lac* repressor. *J Biol Chem*, 262(30):14592–9, 1987.

[34] J. A. Borowiec, L. Zhang, S. Sasse-Dwight, and J. D. Gralla. DNA supercoiling promotes formation of a bent repression loop in *lac* DNA. *J Mol Biol*, 196(1):101–11, 1987.

[35] T. E. Cloutier and J. Widom. DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc Natl Acad Sci U S A*, 102(10):3645–50, 2005.

[36] Y. Zhang, A. E. Mcewen, D. M. Crothers, and S. D. Levene. Statistical-mechanical theory of DNA looping. *Biophys J*, 90(6):1903–12, 2006.

[37] D. Swigon, B. D. Coleman, and W. K. Olson. Modeling the lac repressor-operator assembly: The influence of DNA looping on lac repressor conformation. *Proc Natl Acad Sci U S A*, 103(26):9879–84, 2006.

[38] M. Geanacopoulos, G. Vasmatzis, V. B. Zhurkin, and S. Adhya. Gal repressosome contains an antiparallel DNA loop. *Nat Struct Biol*, 8(5):432–6, 2001.

[39] A. Balaeff, L. Mahadevan, and K. Schulten. Modeling DNA loops using the theory of elasticity. *Phys Rev E Stat Nonlin Soft Matter Phys*, 73(3 Pt 1):031919, 2006.

[40] K. B. Towles, J. F. Beausang, H. G. Garcia, R. Phillips, and P. C. Nelson. First-principles calculation of DNA looping in tethered particle experiments. *Phys Biol*, 6(2):25001, 2009.

[41] G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A*, 79(4):1129–33, 1982.

[42] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003.

[43] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[44] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[45] A. M. Friedman, T. O. Fischmann, and T. A. Steitz. Crystal structure of *lac* repressor core tetramer and its implications for DNA looping. *Science*, 268(5218):1721–7, 1995.

[46] R. A. Mehta and J. D. Kahn. Designed hyperstable lac repressor. DNA loop topologies suggest alternative loop geometries. *J Mol Biol*, 294(1):67–77, 1999.

[47] S. Semsey, M. Y. Tolstorukov, K. Virnik, V. B. Zhurkin, and S. Adhya. DNA trajectory in the gal repressosome. *Genes Dev*, 18(15):1898–907, 2004.

[48] M. M. Levandoski, O. V. Tsodikov, D. E. Frank, S. E. Melcher, R. M. Saecker, and M. T. J. Record. Cooperative and anticooperative effects in binding of the first and second plasmid osym operators to a laci tetramer: Evidence for contributions of non-operator DNA binding by wrapping and looping. *J Mol Biol*, 260(5):697–717, 1996.

[49] J. K. Barry and K. S. Matthews. Thermodynamic analysis of unfolding and dissociation in lactose repressor protein. *Biochemistry*, 38(20):6520–8, 1999.

[50] L. Saiz, J. M. Rubi, and J. M. Vilar. Inferring the in vivo looping properties of DNA. *Proc Natl Acad Sci U S A*, 102(49):17642–5, 2005.

[51] P. L. Dehaseth, C. A. Gross, R. R. Burgess, and M. T. J. Record. Measurement of binding constants for protein-DNA interactions by DNA-cellulose chromatography. *Biochemistry*, 16(22):4777–83, 1977.

[52] R. B. O'gorman, M. Dunaway, and K. S. Matthews. DNA binding characteristics of lactose repressor and the trypsin-resistant core repressor. *J Biol Chem*, 255(21):10100–6, 1980.

[53] A. Revzin and P. H. Von Hippel. Direct measurement of association constants for the binding of *escherichia coli lac* repressor to non-operator DNA. *Biochemistry*, 16(22):4769–76, 1977.

[54] M. T. J. Record, P. L. Dehaseth, and T. M. Lohman. Interpretation of monovalent and divalent cation effects on the *lac* repressor-operator interaction. *Biochemistry*, 16(22):4791–6, 1977.

[55] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'conner, D. W. Noble, and P. H. Von Hippel. Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound *escherichia coli lac* repressor *in vivo. Proc Natl Acad Sci U S A*, 74(10):4228–32, 1977.

[56] A. C. Wang, A. Revzin, A. P. Butler, and P. H. Von Hippel. Binding of e. Coli lac repressor to non-operator DNA. *Nucleic Acids Res*, 4(5):1579–93, 1977.

[57] M. D. Barkley. Salt dependence of the kinetics of the *lac* repressor-operator interaction: Role of nonoperator deoxyribonucleic acid in the association reaction. *Biochemistry*, 20(13):3833–42, 1981.

[58] X. Zhang and P. A. Gottlieb. Thermodynamic and alkylation interference analysis of the *lac* repressor-operator substituted with the analogue 7-deazaguanine. *Biochemistry*, 32(42):11374–84, 1993.

[59] M. C. Mossing and M. T. J. Record. Thermodynamic origins of specificity in the *lac* repressor-operator interaction. Adaptability in the recognition of mutant operator sites. *J Mol Biol*, 186(2):295–305, 1985.

[60] N. Horton, M. Lewis, and P. Lu. *escherichia coli lac* repressor-*lac* operator interaction and the influence of allosteric effectors. *J Mol Biol*, 265(1):1–7, 1997.

[61] D. V. Goeddel, D. G. Yansura, and M. H. Caruthers. Binding of synthetic lactose operator dnas to lactose represessors. *Proc Natl Acad Sci U S A*, 74(8):3292–6, 1977.

[62] C. M. Falcon and K. S. Matthews. Glycine insertion in the hinge region of lactose repressor protein alters DNA binding. *J Biol Chem*, 274(43):30849–57, 1999.

[63] R. B. Winter and P. H. Von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. the *escherichia coli* repressor–operator interaction: Equilibrium measurements. *Biochemistry*, 20(24):6948–60, 1981.

[64] L. Saiz and J. M. Vilar. DNA looping: The consequences and its control. *Curr Opin Struct Biol*, 16(3):344–50, 2006.

[65] D. H. Lee and R. F. Schleif. In vivo DNA loops in aracbad: Size limits and helical repeat. *Proc Natl Acad Sci U S A*, 86(2):476–80, 1989.

[66] S. M. Law, G. R. Bellomy, P. J. Schlax, and M. T. J. Record. *in vivo* thermodynamic analysis of repression with and without looping in *lac* constructs. Estimates of free and local *lac* repressor concentrations and of physical properties of a region of supercoiled plasmid DNA *in vivo. J Mol Biol*, 230(1):161–73, 1993.

[67] D. Shore and R. L. Baldwin. Energetics of DNA twisting. i. Relation between twist and cyclization probability. *J Mol Biol*, 170(4):957–81, 1983.

[68] Q. Du, C. Smith, N. Shiffeldrim, M. Vologodskaia, and A. Vologodskii. Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proc Natl Acad Sci U S A*, 102(15):5397–402, 2005.

[69] M. T. Record, S.J. Mazur, P. Melancon, J. H. Roe, S. L. Shaner, and L. Unger. Double helical DNA: conformations, physical properties, and interactions with ligands. *Annual review of biochemistry*, 50:997–1024, 1981.

[70] T. Strick, J. Allemand, V. Croquette, and D. Bensimon. Twisting and stretching single DNA molecules. *Prog Biophys Mol Bio*, 74(1-2):115–40, 2000.

[71] J. D. Moroz and P. Nelson. Entropic elasticity of twist-storing polymers. *Macromolecules*, 31(18):6333–6347, 1998.

[72] D. E. Segall, P. C. Nelson, and R. Phillips. Volume-exclusion effects in tethered-particle experiments: Bead size matters. *Phys Rev Lett*, 96(8):088306–(1–4), 2006.

[73] J. Yan and J. F. Marko. Localized single-stranded bubble mechanism for cyclization of short double helix DNA. *Phys Rev Lett*, 93:108108–(1–4), 2004.

[74] P. A. Wiggins, P. C. Nelson, and R. Phillips. Exact theory of kinkable elastic polymers. *Phys. Rev. E*, 71:021909–(1–19), 2005.

[75] P. C. Nelson, C. Zurla, D. Brogioli, J. F. Beausang, L. Finzi, and D. Dunlap. Tethered particle motion as a diagnostic of DNA tether length. *J Phys Chem B*, 110(34):17260–7, 2006.

[76] L. Czapla, D. Swigon, and W. K. Olson. Sequence-dependent effects in the cyclization of short DNA. *J Chem Theory Comp*, 2(3):685–695, 2006.

[77] P. C. Nelson. Colloidal particle motion as a diagnostic of DNA conformational transitions. *Curr Op Colloid Intef Sci*, 12:307–313, 2007.

[78] M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, 271(5253):1247–54, 1996.

[79] T. R. Strick, V. Croquette, and D. Bensimon. Homologous pairing in stretched supercoiled DNA. *Proc Natl Acad Sci U S A*, 95:10579–10583, 1998.

[80] M. D. Wang, H. Yin, R. Landick, J. Gelles, and S. M. Block. Stretching DNA with optical tweezers. *Biophys J*, 72(3):1335–1346, 1997.

[81] E. Villa, A. Balaeff, and K. Schulten. Structural dynamics of the *lac* repressor-DNA complex revealed by a multiscale simulation. *Proc Natl Acad Sci U S A*, 102(19):6783–8, 2005.

[82] P. A. Wiggins, T. Van der Heijden, F. Moreno-Herrero, A. Spakowitz, R. Phillips, J. Widom, C. Dekker, and P. C. Nelson. High flexibility of DNA on short length scales probed by atomic force microscopy. *Nature Nanotech*, 1(2):137–141, 2006.

[83] G. C. Ruben and T. B. Roos. Conformation of *lac* repressor tetramer in solution, bound and unbound to operator DNA. *Microsc Res Tech*, 36(5):400–16, 1997.

[84] L. M. Edelman, R. Cheong, and J. D. Kahn. Fluorescence resonance energy transfer over approximately 130 basepairs in hyperstable lac repressor-DNA loops. *Biophys J*, 84(2 Pt 1):1131–45, 2003.

[85] M. A. Morgan, K. Okamoto, J. D. Kahn, and D. S. English. Single-molecule spectroscopic determination of lac repressor-DNA loop conformation. *Biophys J*, 89(4):2588–96, 2005.

[86] Y. Zhang, A. E. Mcewen, D. M. Crothers, and S. D. Levene. Analysis of in-vivo lacr-mediated gene repression based on the mechanics of DNA looping. *PLoS ONE*, 1:e136, 2006.

[87] S. Oehler, M. Amouyal, P. Kolkhof, B. Von Wilcken-Bergmann, and B. MÜLler-Hill. Quality and position of the three *lac* operators of *e. Coli* define efficiency of repression. *EMBO J*, 13(14):3348–55, 1994.

[88] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in *escherichia coli* via the lacr/o, the tetr/o and arac/i1-i2 regulatory elements. *Nucleic Acids Res*, 25(6):1203–10, 1997.

[89] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–5, 2005.

[90] J Gelles, B.J. Schnapp, and M.P. Sheetz. Tracking kinesin-driven movements with nanometre-scale precision. *Nature*, 331(4):450–453, 1988.

[91] D. Colquhoun and F. J. Sigworth. Fitting and statistical analysis of single-channel records. In B. Sakmann and E. Neher, editors, *Single-Channel Recording*, pages 483–587. Plenum Press, New York, 2nd edition, 1995.

[92] Motion of the bead is systematically characterized with various DNA lengths ranging from 200 bp to 3 kbp. Such DNA is then interpolated using a second-order polynomial function to served as a calibration curve. From this curve, for any given length DNA tether in that range, the amplitude of the motion of the DNA-tethered bead can be evaluated. Experimental data, and a theoretical model of the calibration, appear in [101].

[93] D. E. Frank, R. M. Saecker, J. P. Bond, M. W. Capp, O. V. Tsodikov, S. E. Melcher, M. M. Levandoski, and M. T. J. Record. Thermodynamics of the interactions of *lac* repressor with variants of the symmetric *lac* operator: Effects of converting a consensus site to a non-specific site. *J Mol Biol*, 267(5):1186–206, 1997.

[94] S. Oehler, E. R. Eismann, H. Kramer, and B. MÜLler-Hill. The three operators of the lac operon cooperate in repression. *EMBO J*, 9(4):973–9, 1990.

[95] J. F. Beausang, C. Zurla, C. Manzo, D. Dunlap, L. Finzi, and P. C. Nelson. DNA looping kinetics analyzed using diffusive hidden Markov model. *Biophys J*, 92(8):L64–6, April 2007.

[96] J. F. Beausang and P. C. Nelson. Diffusive hidden Markov model characterization of DNA looping dynamics in tethered particle experiments. *Physical Biology*, 4:205–219, 2007.

[97] C. M. Falcon, L. Swint-Kruse, and K. S. Matthews. Designed disulfide between N-terminal domains of lactose repressor disrupts allosteric linkage. *J Biol Chem*, 272(43):26818–21, October 1997.

[98] R. Phillips, J. Kondev, and J. Theriot. *Physical biology of the cell*. Garland Science, New York, 2009. (Illustrated by N. Orme; with problems, solutions, and editorial assistance of H. G. Garcia.).

[99] D. Shore, J. Langowski, and R. L. Baldwin. DNA flexibility studied by covalent closure of short fragments into circles. *Proc Natl Acad Sci U S A*, 78(8):4833–7, 1981.

[100] P. C. Nelson, C. Zurla, D. Brogioli, J. F. Beausang, L. Finzi, and D. Dunlap. Tethered particle motion as a diagnostic of DNA tether length. *J Phys Chem B*, 110(34):17260–17267, 2006.

[101] K. Towles, J. F. Beausang, H. G. Garcia, R. Phillips, and P. C. Nelson. First-principles calculation of DNA looping in tethered particle experiments. *Physical Biology, in press*, 2009.