

The Combinatorics of Transcriptional Regulation

Thesis by
Mattias Rydenfelt

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



California Institute of Technology
Pasadena, California

2014
(Defended 2014-04-28)

Acknowledgements

First and foremost I wish to thank my adviser Rob Phillips. When at the end of my second year in graduate school I realized that my current line of work in particle physics was not going anywhere and that it was time to start looking for a new lab, Rob warmly welcomed me into his group. This initially struck me with surprise, I had virtually no experience in molecular biology and would need to spend my first year in the lab bothering everyone with silly beginner questions. Three years later I am no longer the least surprised. Rob is the kind of scientist who is not going with the current, but is constantly looking for new and creative ways to attack interesting problems, and having a diverse research group best achieves this goal. I have never met a person in my life with a more genuine passion for science (or larger science book collection) than Rob.

Working in the Phillips lab has been a pleasure and I cannot imagine a more friendly and helpful environment in which to pursue research. I wish to thank all past and present Phillips lab members responsible for this: Stephanie Barnes, Nathan Belliveau, Maja Bialecka-Fornal, James Boedicker, Justin Bois, Rob Brewster, Yi-Ju Chen, Sidney Cox, Hernan Garcia, Christoph Haselwandter, Kelsey Homyk, Stephanie Johnson, Daniel Jones, Heun Jin Lee, Geoff Lovely, Gita Mahmoudabadi, Katie Miller, Manuel Razo and Franz Weinert. In particular I wish to thank my collaborators (and Ernie comrades) Rob Brewster and Franz Weinert, without whose beautiful experiments and hard work the theories presented in this thesis would remain untested.

I also wish to thank my thesis committee members Richard Murray, Gil Refael and David Tirrell. It has been an honor to be under the guidance of such exceptional scientists.

Finally I wish to thank my family and friends for all the special moments not only inside but also outside of the lab.

Abstract

The ability to regulate gene expression is of central importance for the adaptability of living organisms to changes in their internal and external environment. At the transcriptional level, binding of transcription factors (TFs) in the vicinity of promoters can modulate the rate at which transcripts are produced, and as such play an important role in gene regulation. TFs with regulatory action at multiple promoters is the rule rather than the exception, with examples ranging from TFs like the cAMP receptor protein (CRP) in *E. coli* that regulates hundreds of different genes, to situations involving multiple copies of the same gene, such as on plasmids, or viral DNA. When the number of TFs heavily exceeds the number of binding sites, TF binding to each promoter can be regarded as independent. However, when the number of TF molecules is comparable to the number of binding sites, TF titration will result in coupling (“entanglement”) between transcription of different genes. The last few decades have seen rapid advances in our ability to quantitatively measure such effects, which calls for biophysical models to explain these data. Here we develop a statistical mechanical model which takes the TF titration effect into account and use it to predict both the level of gene expression and the resulting correlation in transcription rates for a general set of promoters.

To test these predictions experimentally, we create genetic constructs with known TF copy number, binding site affinities, and gene copy number; hence avoiding the need to use free fit parameters. Our results clearly prove the TF titration effect and that the statistical mechanical model can accurately predict the fold change in gene expression for the studied cases. We also generalize these experimental efforts to cover systems with multiple different genes, using the method of mRNA fluorescence in situ hybridization (FISH). Interestingly, we can use the TF titration effect as a tool to measure the plasmid copy number at different points in the cell cycle, as well as the plasmid copy number variance.

Finally, we investigate the strategies of transcriptional regulation used in a real organism by analyzing the thousands of known regulatory interactions in *E. coli*. We introduce a “random promoter architecture model” to identify overrepresented regulatory strategies, such as TF pairs which coregulate the same genes more frequently than would be expected by chance, indicating a related biological function. Furthermore, we investigate whether promoter architecture has a systematic effect on gene expression by linking the regulatory data of *E. coli* to genome-wide expression censuses.

Publications

This thesis is based on the following publications:

- I. **Mattias Rydenfelt**, Robert Sidney Cox III, Hernan Garcia and Rob Phillips “Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration,” *Phys. Rev. E*. 89, 012702 (2014).
- II. Robert Brewster[†], Franz Weinert[†], Linda Song, Hernan Garcia, **Mattias Rydenfelt** and Rob Phillips “The transcription factor titration effect dictates level of gene expression,” *Cell* 156:6 (1312-1323) (2014).
- III. **Mattias Rydenfelt**, Hernan Garcia, Robert Sidney Cox III and Rob Phillips “The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*”, *about to be submitted for publication*.

These three articles are reproduced in Chapters 2,3, and 5 respectively. Chapter 4 presents preliminary results of a follow-up experiment to II, designed and conducted in collaboration with Robert Brewster.

Contents

Acknowledgements	iii
Abstract	iv
Publications	v
1 Introduction	5
1.1 Gene regulation: an overview	9
1.2 Quantitative modeling of transcriptional regulation	14
1.2.1 Statistical mechanical model of an unregulated promoter	15
1.2.2 Stochastic model of a regulated promoter	16
1.3 Experimental techniques to measure gene expression	17
1.4 Roadmap to thesis	19
References	22
2 Statistical mechanical model of coupled transcription from multiple promoters due to TF titration	28
2.1 Introduction	28
2.2 Underlying assumptions of thermodynamic model	30
2.3 Single promoter partition function	32
2.3.1 Simple repression	32
2.3.2 Repression with looping	34
2.3.3 Exclusive looping repression	37
2.4 Multiple promoter partition function	38
2.4.1 General set of promoters	38
2.4.2 Identical promoters	40
2.4.3 Simple repression	41
2.4.4 Repression with looping	42
2.4.5 Exclusive looping repression	42

2.5	Fold change	43
2.6	Transcriptional correlation	48
2.6.1	Toy model of transcriptional correlation	48
2.6.2	General theory	50
2.6.3	Two anticorrelated genes	51
2.7	Statistically distributed TF and promoter copy numbers	54
2.7.1	Fold change	54
2.7.2	Transcriptional correlation	55
2.8	Verifying the thermodynamic model of TF titration using Gillespie simulations	56
2.8.1	Simple repression	58
2.8.2	Repression with looping	59
2.8.3	Repression exclusively due to looping	62
2.8.4	Transcriptional correlation	62
2.9	Conclusion	63
2.A	Partition function for a set of promoters regulated by multiple low-copy TFs	65
2.B	Number of binding sites vs. TF copy number in <i>E. coli</i>	66
	References	68
3	The transcription factor titration effect dictates level of gene expression	75
3.1	Introduction	75
3.2	Results	77
3.2.1	Thermodynamic model	77
3.2.2	Fluorescent measurements of gene expression and absolute TF copy number	80
3.2.3	Gene copy number measured by qPCR	80
3.2.4	Determining sequence dependent TF binding energies	81
3.2.5	Simple thermodynamic model predicts expression level of single integrated gene copy	81
3.2.6	Predicting expression levels from plasmid constructs as a function of gene copy number	83
3.2.7	Simple thermodynamic model predicts expression levels from multiple integrated chromosomal gene copies	83
3.2.8	Predicting expression levels in complex TF binding landscapes	85
3.2.9	The influence of plasmid distribution on the repressor titration curves	87
3.2.10	Cell cycle dependence of the plasmid copy number and the resulting expression	89
3.3	Discussion	91
3.4	Experimental procedures	92

3.4.1	Gene expression measurements	92
3.4.2	Data analysis	93
3.5	Acknowledgments	93
3.A	Genetic elements and details of the dilution method	94
3.A.1	Dilution circuit	94
3.A.2	The one step dilution method	94
3.A.3	Physiological effect of repressor induction	96
3.A.4	Cell growth and detailed experimental procedure	96
3.B	Image segmentation and analysis	99
3.B.1	Flattening fluorescence images	99
3.B.2	Chromatic aberration correction	100
3.B.3	Autofluorescence correction	100
3.B.4	Correcting for crosstalk and cross bleaching	100
3.B.5	Correcting for photobleaching	101
3.C	Calibrating LacI-mCherry intensity to absolute copy number	102
3.C.1	Fairness of repressor partitioning	104
3.C.2	Derivation of calibration factor	104
3.C.3	Interpretation of $\langle (I_1 - I_2)^2 \rangle$	105
3.C.4	Photon counting noise	106
3.C.5	Limits in LacI-mCherry detection	106
3.C.6	Limits in YFP production detection	106
3.D	qPCR measurement of average plasmid copy number	107
3.E	The copy number of multiple chromosomal integrations strain	108
3.F	Additional theoretical details of the thermodynamic model	110
3.F.1	Equivalence of fold-change in steady-state measurements and video microscopy	110
3.F.2	Thermodynamic model in the limit $R \gg N$ or $R \gg N_c$	110
3.F.3	Accounting for chromosome replication in competitor theory	111
3.F.4	Thermodynamic model with plasmid distribution	113
3.F.5	Determining errors in theoretical predictions	114
3.G	Constructs and strains	115
3.H	Primers used in this study to create strains	119
	References	120
4	The transcription factor titration effect in a system of two coregulated genes	126
4.1	Introduction	126
4.2	Results	129

4.2.1	Thermodynamic model	129
4.2.2	Stochastic model	131
4.2.3	FISH measurements	133
4.3	Discussion	137
4.4	Methods	138
4.4.1	Strains	138
4.4.2	Growth	139
4.4.3	Single cell mRNA FISH	139
4.4.4	FISH data acquisition	139
4.4.5	Determining the absolute number of LacI-mCherry molecules	140
4.5	Acknowledgements	140
4.A	aTc induction curve	141
4.B	Plasmid structure	141
4.C	PCR primers	143
	References	145
5	The influence of promoter architectures and regulatory motifs on gene expression in <i>Escherichia coli</i>	148
5.1	Introduction	148
5.2	Models	150
5.2.1	Random promoter architecture model	150
5.2.2	Linear energy model of RNAP-DNA binding	154
5.3	Results	155
5.3.1	How many genes do TFs regulate?	155
5.3.2	How are activator and repressor binding sites configured?	160
5.3.3	Where are TF binding sites located?	163
5.3.4	How does promoter architecture relate to promoter strength?	168
5.4	Discussion	170
5.5	Acknowledgments	171
	References	173
6	Conclusion	179

Chapter 1

Introduction

One of the defining features of living organisms is their ability to adapt to changes in their environment. In humans, examples are obvious: depending on the weather we dress differently, we behave differently depending on our relationship to people around us, and we quickly pull away if one of our hands touches a hot plate. Though less obvious, adaptation plays an equally important role across all levels of the tree of life. Single celled organisms like bacteria can sense nutrient gradients through receptors in their cell membranes and move towards regions where they are more likely to grow, divide, and spread their genes [1]. Other bacteria hibernate and form spores [2] when the supply of nutrients is low, hoping that better times will come eventually. Fundamentally, however, adaptation in humans and bacteria are alike, in the sense that decisions are made at the levels of single cells. The human body is nothing but an ecosystem of different cell types and microorganisms working together. In fact, the majority of cells in the human body are not human but bacterial [3]. Any action of a multicellular organism therefore originates from the actions of individual cells, and to understand adaptation and decision making in detail, we need to understand how these mechanisms operate at the single cell level.

The desire to understand life – how it started, how it evolved, how organisms function and replicate – is probably as old as philosophical questions such as the origin of the universe and what matter is made of. Modern science has been able to answer many questions, both in cosmology and biology, that were previously thought to belong rather to religion or philosophy. When a person gets the flu, we now know that it is caused by the infection of a microscopic virus which replicates inside our bodies, rather than failing to sacrifice enough goats to the mountain gods. Perhaps upsetting to some supernaturalists, the most successful picture of life we have is that we are simply made up of a collection of particles which mindlessly obey the laws of physics. Air fills our lungs on a summer day because the pressure outside of our body is greater than inside, and our heart beats because of coordinated electrical stimulatory signals. For the heart, life is simple: it beats as long as it can, then it stops.

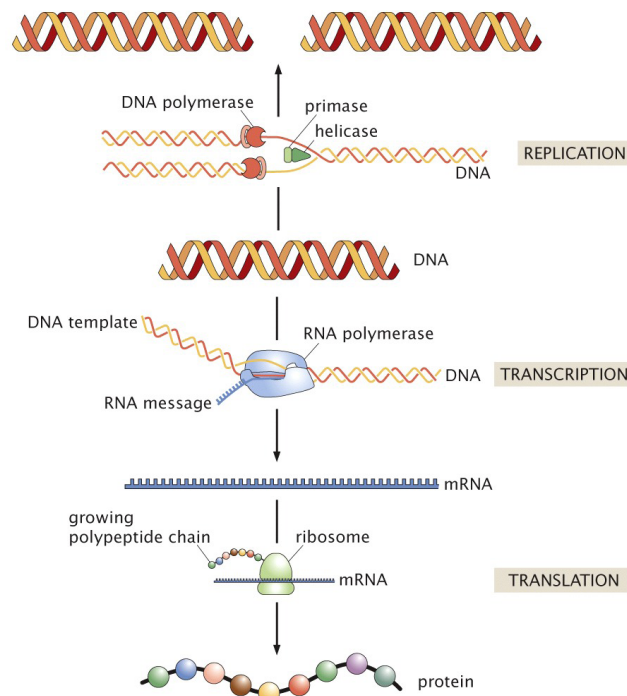


Figure 1.1: The central dogma of molecular biology establishes a connection between the DNA nucleotide sequence and protein amino acid sequence, through the transcription of DNA into mRNA by RNA polymerases (RNAP), and the translation of mRNA into proteins by ribosomes. (Adapted from “Physical Biology of the Cell”, *Garland Science*, 2013.)

The central dogma of molecular biology dictates how genetic information is converted into functional proteins

A naturalistic perspective opens up the door to gaining understanding of life, including adaptation and decision making, through systematic experimentation and hypothesis testing. The quest is, however, a challenging one. Cells are tiny (typically $1 - 10\mu\text{m}$ in size) membrane-bound bodies, and probing their inner workings requires ingenious experiments. Still, an enormous amount of molecular biology has been learned over the past century, and new important discoveries continue to be made. The isolation of DNA was already accomplished by the late 19th century by Friedrich Miescher, but at the time its crucial role to life was not yet understood. Over the next few decades the importance of DNA was gradually recognized, culminating in 1953 with the discovery of the double helix model of DNA by James Watson and Francis Crick [4], and the central dogma of molecular biology [Fig. 1.1] by Crick a few years later [5]. By then it was appreciated that essentially all information about an organism is stored in its DNA sequence, which gets passed along to its progeny. The DNA codes for genes which can be transcribed into mRNA by RNA polymerase (RNAP), and further translated into proteins by ribosomes. Proteins are the workhorses of the cell, providing all different kinds of functionality, like motility, regulation, and transport. Proteins either function by themselves or in

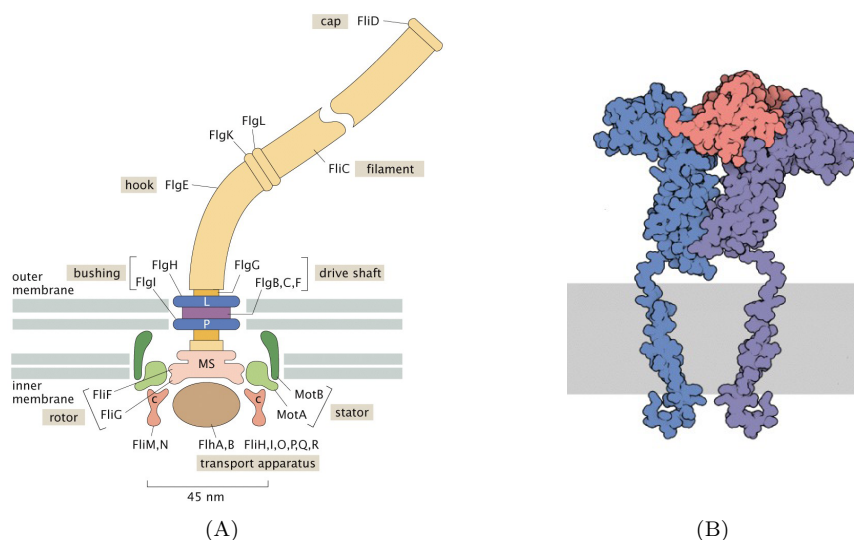


Figure 1.2: (A) Cartoon of a bacterial flagellum which is composed of around twenty five different proteins. (B) Protein structure of a human growth membrane receptor protein. (Adapted from “Physical Biology of the Cell”, *Garland Science*, 2013.)

complexes with other proteins, for example in the flagella [Fig. 1.2(A)] or in membrane receptors [Fig. 1.2(B)]. The biological role of a protein depends on its folding structure, which is determined by the minimum free energy configuration of its amino acid sequence. This structure can often switch between different conformations by interacting with a small molecule (e.g. a phosphate group), which plays an important role in cellular decision making.

DNA sequencing makes it possible to read the genetic information in organisms of all types, but making sense of it all is harder

Even with the discovery of the DNA structure it took several decades before it became possible to read (or sequence) the letters of life stored in the double helix. This changed with the advent of Sanger sequencing in 1977 [6, 7] and polymerase chain reaction (PCR), a DNA amplification technique, in 1983 [8]. These breakthroughs made it possible, in a practical manner, to sequence up to hundreds of base pairs at a time. When sequencing of the first human genome was completed in the early 2000s [9] it raised hopes that many serious diseases could soon be cured. However, some of these wishes have not been realized. A major difficulty with analyzing DNA sequences is the challenge of relating them to biological function. The millions of 'A', 'C', 'T' and 'G' letters that a biologist is presented with after a large scale sequencing experiment are all written in the same “font size”, and without important features highlighted. To use a computer science metaphor, understanding the role of a particular DNA sequence is like figuring out how to set the desktop background color on your computer by looking at the machine code of the operating system. While

some genes are critical to the survival of the organism, other genes can be considered as “helper genes”, useful only in particular contexts. The cell, however, has no “higher level” programming language, that we are aware of, to distinguish critical features from the secondary. In addition, large parts of the human genome consist of noncoding elements, including introns, retrotransposons [9], or retroviruses [10], which altogether make up some 98% of the genomic DNA [9]. Even sequences which have no function at all will remain in the genome, as long as they do not cause an evolutionary disadvantage for their host. In conclusion, the genome was not designed with comprehensibility in mind, only evolutionary fitness.

Although many important lessons in molecular biology have been learned over the last century, such as the central dogma, many questions still remain to be answered. For example, once a functional DNA sequence has been identified, how can it be altered, or how can its behavior be predicted from just looking at the sequence? By mutating the promoter region of a gene one can alter the regulation of a gene, or by mutating the coding sequence one can replace amino acids in the expressed protein to change its function. To connect, at the single base pair level, these modifications in DNA sequence to biological function is a challenging task, which can only be addressed with high resolution experiments together with quantitative biophysical models. With a detailed understanding of how to connect DNA sequence and function, one could create synthetic biological circuits with any desired properties, or modify the behavior of an existing circuit in an advantageous way. Such abilities could potentially have large impact on other fields, for example medicine, where broken genetic circuits are a common source of disease.

Model organisms are the harmonic oscillators of molecular biology

Using simple toy models as a starting point to attack new and more complex problems has been a favorite approach among physicists for centuries, including the ever so useful harmonic oscillator or the hydrogen atom. Similar ideas have in biology led many researchers to focus their attention on a mere handful of different model organisms including *Escherichia coli* (*E. coli*), *Bacillus subtilis*, *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans*, and humans; this despite the plethora of possible organisms to choose from. A model organism is chosen sometimes because of practical reasons, perhaps being easy to isolate and grow (e.g. *Escherichia coli*), or because it has unusual and interesting properties worth studying. The large amount of effort put into understanding these model organisms makes them even more appealing for other researchers to use, as many of the necessary tools and protocols (for example cloning techniques) have already been developed. Unlike the harmonic oscillator or the hydrogen atom, model organisms are not necessarily “simpler” than other organisms, which gives hope that the lessons learned in these species will also prove valuable when turning to other organisms. Perhaps the largest divide among model organisms falls between prokaryotes (e.g. bacteria) and eukaryotes. In terms of cell structure the prokaryotes are more

simplistic, lacking a nucleus or other membrane-bound organelles. Another major difference between the two groups is that transcription in procaryotes requires no preinitiation complex assembly, no mRNA export and no post-transcriptional modification [11]. For these reasons, prokaryotes like *Escherichia coli* are popular among quantitative biologists for studying gene regulation. In the remainder of this thesis we will only cover gene regulation in procaryotes, again with the unspoken hope that our efforts can later also become valuable in the study of eukaryotes.

1.1 Gene regulation: an overview

The amount produced of a protein of a certain kind depends on the rate at which the gene is being transcribed to mRNA and then translated into proteins. A large number of proteins leads to high activity of the corresponding biological function. For example, if a large number of flagella class proteins are produced, the motility of the cell will increase, independently of if this is the best strategy or not for the cell under the given conditions. Proteins simply obey the laws of physics, ignorant of what their actions will lead to.

The basal transcriptional and translational machinery permits a generous range of gene expression. Even for an unregulated gene, the expression can vary by at least three orders of magnitude [12], ranging from around 10 protein copies per cell up 10,000 in *E. coli* [13, 14]. While some proteins will always be in great demand, like ribosomal proteins, others are only needed in specific contexts. The *constitutive*, or unregulated, expression of genes cannot, however, be the full story. To respond to changes in their external or internal environment organisms must be able to activate or repress expression of genes at given times. Without this ability, they would produce the same configuration of proteins, at the same rates, at all times, independent of their surroundings. Nature would likely deem such an organism unfit. Already by the early 1900s it was known, from work by Duclaux [15], Dienert [16] and others, that certain metabolic enzymes in yeast are only expressed in conditions where they are actually needed – the first evidence of gene regulation before genes had even been discovered! A modern explanation for this observation was provided sixty years later with the seminal study of the *lac* operon by Jacob and Monod [17]. They were able to show that a DNA binding protein, the Lac repressor, in *E. coli* can turn off the production of enzymes responsible for the uptake and digestion of lactose when this sugar (or glucose, the favored energy source) is not present. The results of Jacob and Monod provided the key insight that there is not only a causal link from genes to proteins, but also from proteins to genes.

Genetic feedback regulation allows sophisticated dynamical behavior

The ability of proteins to control gene expression opens up the possibility for feedback regulation, which can lead to interesting dynamics in expression patterns, including bistability [18] or oscillatory

behavior [19]. As an example, consider a protein A with concentration c_A which activates its own transcription according to the Hill function

$$\frac{dc_A}{dt} = \beta_0 + \beta_A \frac{(c_A/K_d)^2}{1 + (c_A/K_d)^2} - \gamma c_A. \quad (1.1)$$

Here K_d specifies the dissociation constant of A to the promoter, and γ is the protein degradation rate. When two activators cooperatively bind to the promoter, as is indicated by the quadratic dependence of c_A in the Hill function, the transcription rate is enhanced from the basal level of β_0 to β_A . When the probability is small that two activators bind to the promoter, $(c_A/K_d)^2 \ll 1$, we can approximate Eq. (1.1) by

$$\frac{dc_A}{dt} \approx \beta_0 + \beta_A (c_A/K_d)^2 - \gamma c_A. \quad (1.2)$$

This self activating system is bistable, as there are two steady state solutions c_A^\pm to Eq. (1.1). To find these we set the left hand derivative equal to 0 and solve for c_A^\pm (valid when $\left(\frac{\gamma K_d}{2\beta_A}\right)^2 > \frac{\beta_0}{\beta_A}$)

$$\frac{c_A^\pm}{K_d} = \frac{\gamma K_d}{2\beta_A} \pm \sqrt{\left(\frac{\gamma K_d}{2\beta_A}\right)^2 - \left(\frac{\beta_0}{\beta_A}\right)}. \quad (1.3)$$

This example involves only a single gene, but with a higher number of genes there are an endless number of ways to connect them, and the dynamical response gets rapidly more complex. In general, with F number of DNA binding proteins, or TFs (TFs), and G independently transcribed genes there are in theory 2^{FG} number of ways to connect them into a genetic network. To realize this we first notice that each gene is regulated by a subset of the F TFs, and that there are in total 2^F possible subsets to choose. By assuming that we can make this choice independently for every gene, the multiplication rule of combinatorics tells us that there are $(2^F)^G = 2^{FG}$ possible networks. Using some representative input for prokaryotes, $F = 200$ and $G = 5000$, one would get around $10^{300,000}$ possible networks. In Fig. 1.3 we show the currently best known protein-protein interaction network in yeast [20]. Even for such a relatively simple organism, the genetic network is extraordinary complicated and nearly impossible to overview. It is an impressive fact that this machinery works at all. Gene expression is an inherently stochastic process [21, 22, 23], with long delays between transcription initiation to the finally matured protein. In *E. coli*, some decisions the bacteria make will only affect its daughters or granddaughters [24]. From the standard theory of control and dynamical systems, such systems would be very hard to control.

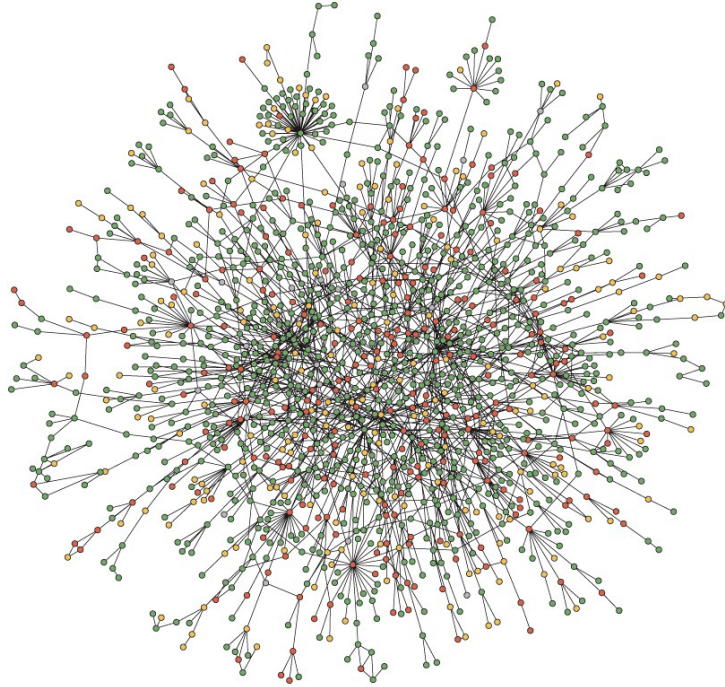


Figure 1.3: The yeast protein-protein interaction network. Each dot represents a single protein, and each line a genetic or biochemical interaction. (Adapted from H. Jeong et al., *Nature* 411:41, 2001.)

TFs affect how RNAP interacts with the promoter

TFs can enhance or repress the transcription rate of a gene by binding to a (more or less) conserved nucleotide sequence in the promoter region. To repress a gene, TFs commonly bind at positions which overlap the RNAP binding site [Fig. 1.4(D)], hence preventing it from accessing the promoter and initiate transcription of the gene. An example is the Lac repressor mentioned above, which represses transcription of the *lac* operon [27]. To activate a gene, TFs can instead bind upstream of the promoter and form a stabilizing complex with RNAP [Fig 1.4(C)], or alternatively mediate other steps in the transcription initiation process, such as open complex formation [28, 29] [Fig. 1.5] or promoter escape. Apart from interacting with RNAP, TFs can also interact with each other to *cooperatively* [30, 31, 32] regulate the expression of a gene [Fig. 1.4(F)]. Typically cooperativity leads to a sharper regulatory response as a function of TF concentration. As an example, the cI_2 activator binds two different sites [26] Fig. 1.4(F), such that one activator stabilizes not only RNAP binding to the promoter, but also the binding of yet another cI_2 activator in its vicinity. A different method of achieving a sharper regulatory response is for a TF to simultaneously bind two different binding sites by bending the intermediate DNA sequence into a loop [33, 34, 35] [Fig. 1.4(E)]. In this case each binding site essentially works as a “fishing hook” for the other, hence increasing the local concentration of the repressor around the promoter. Again, we can use the Lac repressor as

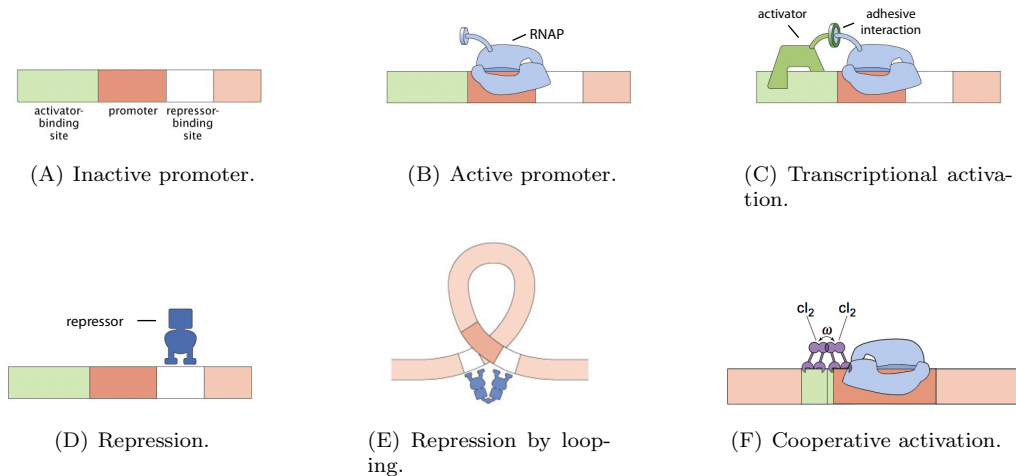


Figure 1.4: Different ways TFs can enhance or repress transcription of a gene [25, 26]. (A) Promoter with binding sites for an activator, a repressor, and an RNAP. (B) Promoter transcribed by RNAP. (C) An activator stabilizes the binding of RNAP to the promoter, increasing the transcription rate of the gene. (D) A repressor binds to a site which sterically excludes RNAP from accessing the promoter, hence preventing transcription of the gene. (E) A repressor simultaneously binds two sites by forming a DNA loop of the intermediate sequence. (F) Two activators cooperatively stabilizes the binding of RNAP to the promoter. Stronger cooperativity leads to a sharper regulatory response. ((A)-(E) adapted from “Physical Biology of the Cell”, *Garland Science*, 2013, (F) adapted from Bintu et al., *Curr. Opin. Genet. Dev.*, 2005.)

an example, which can form DNA loops between the O_1 , O_2 or O_3 binding sites.

DNA sequence determines RNAP and TF binding affinity

Like everything in a cell, the transcription and translation rates are determined by the DNA sequence. RNAP recognizes two upstream sequence elements located 35 bp and 10 bp upstream of the transcription start site and binds these to form a closed complex [Fig. 1.5]. The more closely the two recognition sequences match the (σ^{70}) consensus sequences TTGACA (-35 bp) and TATAAT (-10 bp), the more strongly RNAP binds to the promoter [36]. After closed complex formation, a series of steps lead to the melting of the DNA helix (open complex formation) and promoter escape, which indicates the starting point of mRNA synthesis [37, 38, 39]. Just like closed complex formation, the rate of these steps depends on the DNA sequence of the promoter. Finding a fully generalized model that describes all different steps of transcription initiation, and accurately predicts the transcription rate for any promoter sequence is a difficult but important unsolved problem in biophysics. However, under some restricting circumstances a fully generalized model is not necessary. An important example is when closed complex formation is the rate limiting step of transcription initiation, in which case the other rates effectively “drop out”, and gene expression scales linearly with the RNAP binding affinity to the promoter [40]. This assumption, called the occupancy hypothesis, will be frequently used throughout this work. Mapping DNA sequence to RNAP binding affinity is still,

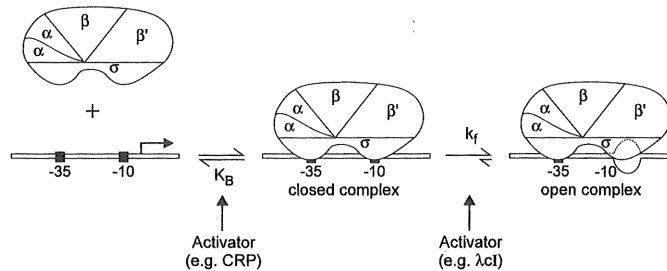


Figure 1.5: Transcription initiation described by a (simplified) two step process, initiated by the binding of RNAP to the promoter of a gene (with binding constant K_B), and the subsequent melting of the two DNA strands into an open complex formation. (Adapted from “The Bacterial Chromosome”, *ASM Press*, 2005.)

however, a challenging problem. A commonly used approximation to model the affinity of RNAP, or TFs, to DNA is to assume that each base pair in a binding sequence S contributes independently to the total binding energy $E(S)$

$$E(S) = \sum_{i=1}^L \sum_{j=A,C,T,G} M_{i,j} S_{j,i} = \text{Tr}(MS). \quad (1.4)$$

Here $S_{A/C/T/G,i} = 1$ if the identity of the base at nucleotide position i in the sequence is given by $A/C/T/G$ and otherwise $S_{A/C/T/G,i} = 0$, $M_{i,A/C/T/G}$ represents the energy contribution at position i for base $A/C/G/T$ respectively, and L is the length (in base pairs) of the binding sequence S . The energy matrix M can be determined from studying a large set of mutated promoters, and relating the identity of a base pair at a specific position to the resulting gene expression [41, 12, 36, 42, 43].

There are many different types of gene regulation

Apart from transcriptional regulation there are several other mechanisms to regulate a gene, targeting different stages in the transcription, translation or protein maturation chain, as well as the decay processes of mRNA or proteins. As each transcribed mRNA gets translated by ribosomes into proteins, the number of proteins produced per mRNA (the *burst size*) can vary significantly depending on how effectively ribosomes can access the ribosomal binding sequence (RBS) on the mRNA. Just like the binding of RNAP to a promoter can be regulated by TFs, also the binding of ribosomes can be regulated. One common method to achieve translational regulation in prokaryotes is through the interference of an antisense RNA, that can complementarily base pair with an mRNA and inhibit its translation [44]. In eukaryotes a similar strategy goes under the name of RNAi [45]. These methods might not only prevent ribosomes from binding the mRNA, but also initiate mRNA degradation.

Naively, RNA interference seems like an energetically costly regulatory strategy, as the cell

spends resources to produce mRNAs which never get translated. However, there can be other merits associated with RNAi, which makes it a key regulatory strategy in the immune response against viruses [46].

Understanding the precise role of all different regulatory strategies, and how they are orchestrated in a cell, is currently an active area of research. However, in this thesis we limit ourselves to the study of transcriptional regulation only.

1.2 Quantitative modeling of transcriptional regulation

Decision making in cells is implemented through the activation or repression of genes related to a particular biological action. The ability of RNAP to access and transcribe a promoter, perhaps with the influence of TFs, underlies the decision of whether a gene is turned on or off; hence any quantitative model of transcriptional regulation must address how RNAP interacts with the promoter, depending on TF concentration. Transcriptional regulation models typically fall into one or other of two categories: *thermodynamic* models [47, 48, 25] or *stochastic* models [49, 23], and we will illustrate the differences between these two models by considering two specific examples below.

Thermodynamic models are useful in situations where time dependence is not a key factor, such as predicting the average gene expression for a given concentration of TFs. These models generalize well, as we will show in Chapter 2, when we move from simple to more complicated systems that involve many different kinds of TFs that regulate many different targets. Thermodynamic models should, however, be used with care. The (quasi)equilibrium approximation rests on the assumption that TFs homogenize throughout the cell at a much faster rate than they are being produced. Moreover, the equilibrium approximation demands that the complicated process of transcription initiation can be effectively summarized by a single RNAP binding constant. A more thorough discussion about the underlying assumption of the thermodynamic model is presented in Chapter 2.

Stochastic models are more general than thermodynamic models and can, for example, be used to study the dynamical behavior of a genetic circuit over an extended period of time [50]. They also allow the study of variability (or noise) in mRNA or protein production [51, 52], which is believed to play an important role in both prokaryotic [21] and eukaryotic [53] decision making processes. A difficulty with stochastic models is that they require a large number of input rate parameters, such as binding and unbinding rates of TFs, most of which, in general, are unknown. Solving a large system of rate equations analytically can be very challenging, therefore stochastic simulations [54] are commonly used to analyze these systems numerically.

1.2.1 Statistical mechanical model of an unregulated promoter

Life is inherently an out of equilibrium process which, using the words of Erwin Schrödinger, “feeds on negative entropy” [55]. With every single action, an organism leaves the universe in a slightly more disordered state. Yet, the idea of thermodynamic equilibrium has played a profound role in the field of biological physics. Using statistical mechanics, scientists have successfully modeled the binding of ligands to biological receptors [56], the shape of cell membranes [57], among other examples. One reason that an equilibrium model can be successfully employed in certain biological systems, stems from the fact that a process of interest might occur at a much shorter timescale compared to the nonequilibrium “drift”. This approximation makes available to us a large set of theoretical tools developed in field of statistical mechanics over the last century and a half.

To introduce the thermodynamic model in the context of transcriptional regulation, we consider the simplest possible example, that of an unregulated promoter [25]. A number of RNAP (P) molecules can either bind nonspecifically over a (large) number of sites N_{NS} , or specifically to a promoter of interest. As a simplification we assume that all nonspecific binding sites can be treated to have the same (effective) RNAP binding energy ε_{NS} , which presumably is higher (i.e. corresponding to less strong binding) than the specific binding energy ε_S . Furthermore we assume that RNAP always bind to DNA, and do not roam freely in the cell [58, 59]. Our goal is to compute the probability p_{bound} for the promoter to be bound by RNAP, which by the occupancy hypothesis discussed above, should scale linearly with gene expression. According to the laws of thermodynamics, the probability $P(s)$ for a system in equilibrium to be in a microstate s is given by $P(s) = \frac{1}{Z} e^{-\beta E(s)}$, where $E(s)$ is the energy of s . The normalization constant Z , called the partition function, is defined such that the sum of over all microstates $\sum_s P(s) = 1$. We can break the partition function into two terms, corresponding to when the promoter is specifically bound by RNAP or not

$$Z = Z_{\text{bound}} + Z_{\text{unbound}}. \quad (1.5)$$

When the promoter is not bound by RNAP, the microstates correspond to all $\binom{N_{NS}}{P} = \frac{N_{NS}!}{P!(N_{NS}-P)!}$ ways of distributing P indistinguishable RNAP molecules on a reservoir of N_{NS} binding sites, which results in

$$Z_{\text{unbound}} = \binom{N_{NS}}{P} e^{-\beta P \varepsilon_{NS}}. \quad (1.6)$$

When the promoter is bound by RNAP, there are now $P - 1$ RNAP molecules left to bind the

nonspecific reservoir, and by the same argument as above

$$Z_{\text{bound}} = \binom{N_{NS}}{P-1} e^{-\beta((P-1)\varepsilon_{NS} + \varepsilon_S)}. \quad (1.7)$$

The probability for the promoter to be bound by RNAP is equal to the sum of the probabilities of all microstates where the promoter is bound by RNAP, which is determined by Eq. (1.6)-(1.7)

$$\begin{aligned} p_{\text{bound}} &= \frac{\binom{N_{NS}}{P-1} e^{-\beta((P-1)\varepsilon_{NS} + \varepsilon_S)}}{\binom{N_{NS}}{P-1} e^{-\beta((P-1)\varepsilon_{NS} + \varepsilon_S)} + \binom{N_{NS}}{P} e^{-\beta P \varepsilon_{NS}}} \\ &\approx \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}}. \end{aligned} \quad (1.8)$$

Here we assume that $N_{NS} \gg P$ (in *E. coli*: $N_{NS} \approx 5 \times 10^6$, $P \approx 1000$ [60]), and define the energy difference between specific and nonspecific binding as $\Delta \varepsilon_{pd} = \varepsilon_S - \varepsilon_{NS}$.

1.2.2 Stochastic model of a regulated promoter

As TFs diffuse around in the cell they randomly meet and associate to specific binding sites [61, 62]. The stronger a binding site is, the longer the TF remains bound. If we assume that the probability per unit time is constant for a free TF molecule to associate with a binding site, or for a bound TF to dissociate from it, we can assign to each of these processes a rate constant and calculate the probability for a binding site to be either bound or unbound.

To illustrate the stochastic model of transcriptional regulation we first consider binding of a TF F to a gene G



Here each TF molecule associates to an unbound gene with rate constant k_F^{on} , and once bound it dissociates from the gene with rate constant k_F^{off} . If we have only one gene copy, the steady-state probability for it to be bound $P(B)$ (or unbound $P(0) = 1 - P(B)$) can be calculated by considering the balance equation corresponding to Eq. (1.9)

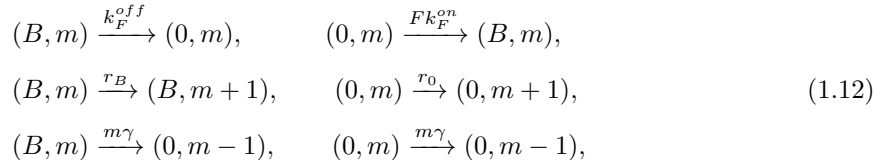
$$\frac{dP(B)}{dt} = F k_F^{on} P(0) - k_F^{off} P(B). \quad (1.10)$$

Setting the left hand side derivate to 0 gives us

$$P(B) = \frac{F k_F^{on}}{k_F^{off} + F k_F^{on}}. \quad (1.11)$$

We can make this toy example more interesting by adding the production mRNA. Let r_B be the

rate of mRNA production in the bound state, and r_0 the (basal) production rate in the unbound state. Further we assume that mRNA degrades with a rate constant γ . The possible state transitions of this system can be summarized by the following table



where (B, m) and $(0, m)$ denote states with m mRNAs, either in the bound or unbound state. To find the steady state probabilities $P(B, m)$ and $P(0, m)$ we first need to write down the balance equation of the two states [52]

$$\frac{d}{dt}P(B, m) = -(k_F^{off} + r_B + \gamma m)P(B, m) + k_F^{on}P(0, m) + r_BP(B, m - 1) + \gamma(m + 1)P(B, m + 1) \tag{1.13}$$

$$\frac{d}{dt}P(0, m) = -(k_F^{on} + r_0 + \gamma m)P(0, m) + k_F^{off}P(B, m) + r_0P(0, m - 1) + \gamma(m + 1)P(0, m + 1). \tag{1.14}$$

After a little bit of math one can show that the average number of mRNAs at steady state, where we set the left hand derivatives equal to 0, is given by [52]

$$\langle m \rangle = \frac{r_0}{\gamma} \frac{k_F^{on}}{k_F^{on} + k_F^{off}} + \frac{r_B}{\gamma} \frac{k_F^{off}}{k_F^{on} + k_F^{off}}. \tag{1.15}$$

We can also calculate the variance (or noise) in m at steady state in terms of the rate parameters.

Already computing the average number of mRNAs, $\langle m \rangle$, for the relatively simple system of Eq. (1.13) is not trivial. When moving to a more complicated system, involving several types of TFs that regulate many different genes, the calculations quickly get very demanding. In these cases Gillespie's algorithm [54] can be a useful tool to study the system numerically.

1.3 Experimental techniques to measure gene expression

To be able to quantitatively study gene regulation, the ability to measure mRNA and protein concentration, either in individual cells or in bulk, is essential. Only with a fruitful interplay between theory and experiment can our understanding of regulatory processes in biology continue to develop. Since the discovery of the double helix structure of DNA (1953), and transcriptional regulation through TFs (1961), the field of molecular biology has advanced by leaps and bounds. Summarizing all these advances would not only fall outside the expertise of the author, but require a thesis-length



Figure 1.6: Mice which have been genetically engineered to express green fluorescent protein (GFP). (Adapted from “Physical Biology of the Cell”, *Garland Science*, 2013.)

work on its own. Still, a few breakthrough discoveries relevant to our purposes deserve mention.

The isolation of green fluorescent protein (GFP) from the jellyfish *Aequorea victoria* [63], and subsequently the ability to fuse GFP (or some of its colorful analogs) to a protein of interest in different organisms [64, 65], has made it possible to quantify gene expression in single cells, and identify areas of a cell where different proteins are predominantly expressed [Fig. 1.6]. When GFP is subject to incident fluorescent light of a certain wavelength, electrons can get excited into a higher orbital, and as they return to a lower orbital light is emitted which can be directly observed in a microscope. The usefulness of fluorescent proteins in molecular biology can hardly be underestimated, and its discovery was awarded the 2008 Nobel Prize in Chemistry. For bulk measurements (as opposed to in single cells), protein expression can also be measured with other methods such as western blots [66] or mass spectrometry. The ability to measure the same signal using several different experimental methods provides a strong experimental control.

To measure RNA expression, several different methods exist, each with its own strengths and weaknesses. Fluorescence in situ hybridization (FISH) [67, 68] can be used to measure mRNA levels in single cells, by hybridizing multiple fluorescent probes with a complementary base pair sequence to the mRNA. We will use this method in Chapter 4 and refer to this chapter for a more detailed discussion. A disadvantage with mRNA FISH is that one cannot follow mRNA expression as a function of time, as the cells are “fixed” (killed) before hybridization. An alternative method, MS2 [69], addresses this problem (arguably by introducing some others), by modifying the gene sequence such that the transcript contains a binding region for fluorescent proteins. The mRNA level can then be quantified by the fluorescence intensity, and be continuously monitored in time. To measure mRNA/DNA levels for particular sequences in bulk, a popular technique is to use quantitative PCR

(qPCR) [70, 71]. During a PCR reaction, the two strands of DNA are separated by increasing the temperature, followed (at lower temperature) by the annealing of two specific primers that mark the beginning and the end of the region to be amplified. DNA polymerases, typically involved in DNA replication, assemble new second strands starting from the primers. By repeating this procedure over and over, the DNA concentration increases exponentially until the involved reagents run out. The initial concentration of DNA can be determined, at least relatively, by calibrating the measurement against a known DNA concentration. In order to use qPCR to measure mRNA levels, one must first convert the mRNA into complementary DNA (cDNA) using reverse transcriptase. A completely different method of measuring mRNA levels, which allows the measurement of hundreds of different mRNA types simultaneously, is based on the hybridization of cDNA to complementary sequences attached to a microarray [72]. By labeling the cDNA with fluorescent markers the amount of hybridization to the DNA probes can be measured. In medical research microarrays can be useful for determining expression profiles in different kinds of tissues, or linking expression profiles with diseases [73].

An increasingly popular alternative to microarrays is to use large-scale sequencing to determine the mRNA content of a sample, a method called RNA-seq [74]. With the development of next generation sequencing in the early 2000s the cost of sequencing (per base pair) dropped by an impressive four orders of magnitude in only a decade, making large scale sequencing a powerful and increasingly cost-effective tool. In RNA-seq, the resulting short sequencing reads are mapped back to a reference genome, and then used to estimate the abundance of different transcripts. Many experiments which at first sight seem to have little to do with sequencing have been formulated in a way that allows next generation sequencing methods to be used. For example, in ChIP-seq [75] one can identify sequences to which TFs predominately bind. First, DNA that is bound by TFs can be isolated through the use of antibodies, and after unlinking the TFs, the binding sites can be sequenced.

The above list presents some currently popular experimental techniques in molecular biology for the measurement of gene expression. This list is, however, far from complete and likely to become quickly outdated.

1.4 Roadmap to thesis

This thesis consists of two essentially independent parts, comprising Chapters 2-4 and Chapter 5 respectively. In the first part we predict and measure the regulatory effect of a TF that targets multiple promoters. In the second part we investigate how transcriptional regulation affects gene expression in *E. coli*, by combining data over the regulatory network in this organism with genome-wide expression measurements.

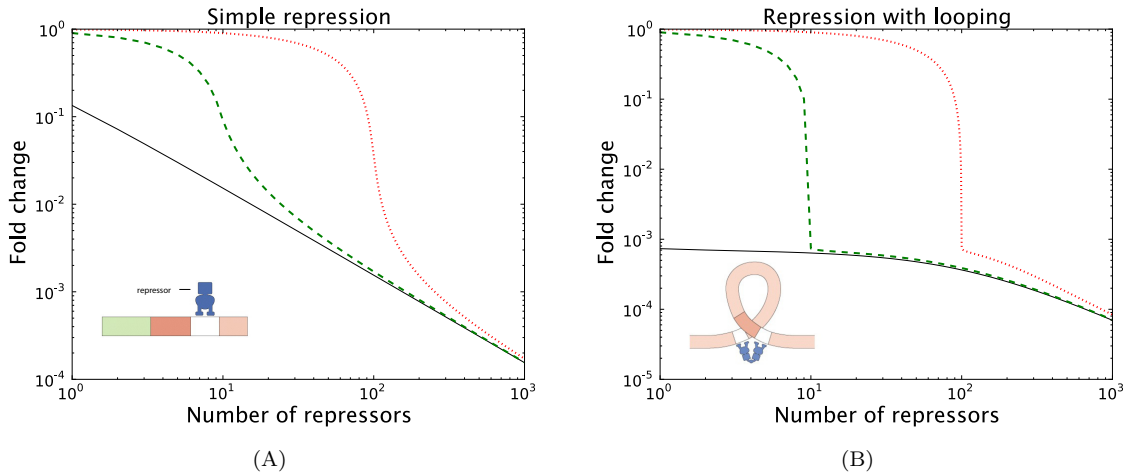


Figure 1.7: Predicted fold change in gene expression as a function of repressor copy number for gene copy numbers $N = 1$ (solid line), $N = 10$ (dashed line), and $N = 100$ (dotted line) for two different promoter architectures: (A) Simple repression promoter architecture. A repressor can bind to a single site overlapping the promoter, blocking RNAP from accessing the promoter. (B) Repression by looping promoter architecture. A repressor can simultaneously bind two sites, one of which overlaps the promoter, such that DNA gets looped.

Part I. In Chapter 2 we extend the thermodynamic model of transcriptional regulation presented in [25] by taking into account the fact that TFs can regulate not only one but multiple genes or gene copies. This includes the common scenario where a regulated gene is located on a plasmid. We predict the fold change, or the relative change in gene expression, as function of TF copy number (F) and gene copy number (N) for several different promoter architectures [Fig. 1.7]. We identify an interesting regime $N \approx F$, where the pool of TFs is large enough to simultaneously repress the transcription of all genes, resulting in a sharp drop in the regulatory function.

In Chapter 3 we verify the predictions of the previous chapter experimentally, by measuring the expression from multiple gene copies, located on a plasmid or on the chromosome, as a function of TF copy number. This experiment is carefully designed to eliminate the need to use free fit model parameters: the absolute number of TFs is measured by a binomial partitioning method [76], the plasmid copy number is measured using qPCR, and the affinities of the TF binding sites have been previously reported in the literature. The underlying motivation behind this experiment is to investigate the scope of the thermodynamic model. Previous experiments [77] have shown that the thermodynamic model can successfully describe transcriptional regulation at an individual promoter, but that should not be taken as a evidence that it will also be successful in a more complicated setting involving multiple gene targets. Any theoretical model has its regime of validity, including the beautiful theory of Newtonian mechanics, which breaks down either at the atomic scale, where quantum mechanics rules, or near the speed of light, where the theory of relativity comes into play. Little has been done to seek out the corresponding limitations of the much less mature theoretical

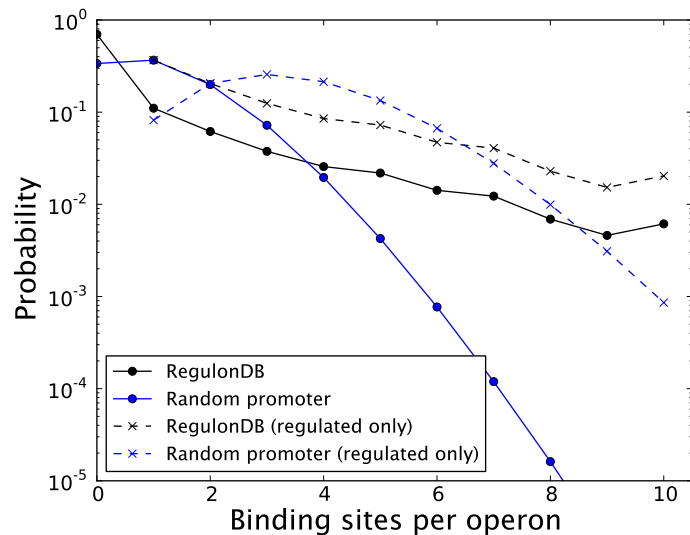


Figure 1.8: Distribution of number of TF binding sites per operon in RegulonDB 8.5 [78] and the random promoter architecture model. Shown separately (“regulated only”) are distributions after neglecting unregulated operons.

ideas in physical biology.

In Chapter 4 we extend these experimental efforts and study the coregulation of two (different) genes, using the method of mRNA FISH. This allows us to measure not only mRNA expression as a function of TF copy number, but also correlation in mRNA expression between the two genes.

Part II. In Chapter 5 we analyze the thousands of known regulatory interactions in *E. coli* to gain insights into the strategies of transcriptional regulation used in a real genetic network. Combining quantitative models of transcriptional regulation, such as presented in Part I, with genome-wide promoter architecture data, could become an important future method for predicting gene expression for a large set of genes in an organism. We take a step in this direction by connecting the known regulatory interactions in *E. coli* with genome-wide expression censuses. To identify overrepresented regulatory strategies [Fig. 1.8] we introduce a “random promoter architecture model”, where TF binding sites are randomly “sprinkled” over a given set of promoters. This model allows us to identify TF pairs that are frequently involved in the regulation of the same genes, signaling that these TFs have a biologically related function.

References

- [1] Adler, J., Hazelbauer, G. L. & Dahl, M. Chemotaxis toward sugars in *Escherichia coli*. *Journal of Bacteriology* **115**, 824–847 (1973).
- [2] Errington, J. Regulation of endospore formation in *Bacillus subtilis*. *Nature Reviews Microbiology* **1**, 117–126 (2003).
- [3] Savage, D. C. Microbial ecology of the gastrointestinal tract. *Annual Reviews in Microbiology* **31**, 107–133 (1977).
- [4] Watson, J. D., Crick, F. H. *et al.* Molecular structure of nucleic acids. *Nature* **171**, 737–738 (1953).
- [5] Crick, F. H. On protein synthesis. In *Symposia of the Society for Experimental Biology*, vol. 12, 138 (1958).
- [6] Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977).
- [7] Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441–448 (1975).
- [8] Arnheim, N. *et al.* Process for amplifying, detecting, and/or-cloning nucleic acid sequences (1987). US Patent 4,683,195.
- [9] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- [10] Nelson, P. N. *et al.* Human endogenous retroviruses: Transposable elements with potential? *Clinical & Experimental Immunology* **138**, 1–9 (2004).
- [11] Ptashne, M. & Gann, A. *Genes and Signals* (Cold Spring Harbor Laboratory Press, New York, 2002).
- [12] Brewster, R. C., Jones, D. L. & Phillips, R. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput. Biol.* **8**, e1002811 (2012).

- [13] Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
- [14] Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**, 117–24 (2007).
- [15] Duclaux, E. *Traide De Microbiologie* (Masson, 1899).
- [16] Diénert, M. F. Sur la fermentation du galactose et sur l'accoutumance des levures à ce sucre. *Ann. Inst. Pasteur* **14**, 139–189 (1900).
- [17] Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3**, 318–356 (1961).
- [18] Lai, K., Robertson, M. J. & Schaffer, D. V. The sonic hedgehog signaling system as a bistable genetic switch. *Biophysical Journal* **86**, 2748–2757 (2004).
- [19] Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–8 (2000).
- [20] Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
- [21] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).
- [22] Mettetal, J. T., Muzzey, D., Pedraza, J. M., Ozbudak, E. M. & van Oudenaarden, A. Predicting stochastic gene expression dynamics in single cells. *Proc Natl Acad Sci U S A* **103**, 7304–9 (2006).
- [23] Paulsson, J. Models of stochastic gene expression. *Physics of Life Reviews* **2**, 157–175 (2005).
- [24] Cairns, J. The chromosome of *Escherichia coli*. In *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 28, 43–46 (Cold Spring Harbor Laboratory Press, 1963).
- [25] Bintu, L. *et al.* Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev* **15**, 116–24 (2005).
- [26] Bintu, L. *et al.* Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev* **15**, 125–35 (2005).
- [27] Oehler, S., Eismann, E. R., Kramer, H. & Muller-Hill, B. The three operators of the *lac* operon cooperate in repression. *EMBO J* **9**, 973–9 (1990).

- [28] Niu, W., Kim, Y., Tau, G., Heyduk, T. & Ebright, R. H. Transcription activation at class II CAP-dependent promoters: two interactions between CAP and RNA polymerase. *Cell* **87**, 1123–1134 (1996).
- [29] Rhodius, V. A., West, D. M., Webster, C. L., Busby, S. J. & Savery, N. J. Transcription activation at class II CRP-dependent promoters: the role of different activating regions. *Nucleic acids research* **25**, 326–332 (1997).
- [30] Weiss, V., Claverie-Martin, F. & Magasanik, B. Phosphorylation of nitrogen regulator I of *Escherichia coli* induces strong cooperative binding to DNA essential for activation of transcription. *Proc Natl Acad Sci U S A* **89**, 5088–92 (1992).
- [31] Babic, A. C. & Little, J. W. Cooperative DNA binding by CI repressor is dispensable in a phage lambda variant. *Proc Natl Acad Sci U S A* **104**, 17741–6 (2007).
- [32] Dodd, I. B. *et al.* Cooperativity in long-range gene regulation by the lambda CI repressor. *Genes Dev* **18**, 344–54 (2004).
- [33] Krämer, H. *et al.* Lac repressor forms loops with linear DNA carrying two suitably spaced lac operators. *EMBO J* **6**, 1481–91 (1987).
- [34] Eismann, E. R. & Muller-Hill, B. lac repressor forms stable loops in vitro with supercoiled wild-type lac DNA containing all three natural lac operators. *J Mol Biol* **213**, 763–75 (1990).
- [35] Fried, M. G. & Hudson, J. M. DNA looping and lac repressor-CAP interaction. *Science* **274**, 1930–1; author reply 1931–2 (1996).
- [36] Mulligan, M. E., Hawley, D. K., Entriken, R. & McClure, W. R. *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic acids research* **12**, 789–800 (1984).
- [37] McClure, W. R. Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.* **54**, 171–204 (1985).
- [38] Reppas, N. B., Wade, J. T., Church, G. M. & Struhl, K. The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol. Cell* **24**, 747–757 (2006).
- [39] Hsu, L. M. Promoter clearance and escape in prokaryotes. *Biochim. Biophys. Acta* **1577**, 191–207 (2002).
- [40] Malan, T. P., Kolb, A., Buc, H. & McClure, W. R. Mechanism of CRP-cAMP activation of lac operon transcription initiation activation of the P1 promoter. *J Mol Biol* **180**, 881–909 (1984).

- [41] Kinney, J. B., Murugan, A., C. G. Callan, J. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**, 9158–63 (2010).
- [42] Brunner, M. & Bujard, H. Promoter recognition and promoter strength in the *Escherichia coli* system. *Embo J* **6**, 3139–44 (1987).
- [43] Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
- [44] Wagner, E. G. & Simons, R. W. Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.* **48**, 713–742 (1994).
- [45] Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
- [46] Stram, Y. & Kuzntzova, L. Inhibition of viruses by RNA interference. *Virus Genes* **32**, 299–306 (2006).
- [47] Weickert, M. J. & Adhya, S. The galactose regulon of *Escherichia coli*. *Mol Microbiol* **10**, 245–51 (1993).
- [48] Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A* **100**, 5136–41 (2003).
- [49] Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A* **98**, 8614–9 (2001).
- [50] Franco, E. *et al.* Timing molecular motion and production with a synthetic transcriptional clock. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E784–793 (2011).
- [51] Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–8 (2004).
- [52] Sanchez, A., Garcia, H. G., Jones, D., Phillips, R. & Kondev, J. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput. Biol.* **7**, e1001100 (2011).
- [53] Blake, W. J., KAern, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
- [54] Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
- [55] Schrodinger, E. *What is Life?: The Physical Aspects of Living Cell with Mind and Matter and Autobiographical Sketches* (Cambridge University Press, 1967).

- [56] Gilson, M. K., Given, J. A., Bush, B. L. & McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* **72**, 1047–1069 (1997).
- [57] Nelson, D. R., Piran, T. & Weinberg, S. *Statistical Mechanics of Membranes and Surfaces*, vol. 5 (World Scientific, 2004).
- [58] Kao-Huang, Y. *et al.* 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**, 4228–32 (1977).
- [59] Runzi, W. & Matzura, H. *In vivo* distribution of ribonucleic acid polymerase between cytoplasm and nucleoid in *Escherichia coli*. *J Bacteriol* **125**, 1237–9 (1976).
- [60] Jishage, M. & Ishihama, A. Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of sigma 70 and sigma 38. *J Bacteriol* **177**, 6832–5 (1995).
- [61] Elf, J., Li, G. W. & Xie, X. S. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316**, 1191–4 (2007).
- [62] Marklund, E. G. *et al.* Transcription-factor binding and sliding on DNA studied using micro- and macroscopic models. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19796–19801 (2013).
- [63] Shimomura, O., Johnson, F. H. & Saiga, Y. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, aequorea. *Journal of Cellular and Comparative Physiology* **59**, 223–239 (1962).
- [64] Prasher, D. C., Eckenrode, V. K., Ward, W. W., Prendergast, F. G. & Cormier, M. J. Primary structure of the aequorea victoria green-fluorescent protein. *Gene* **111**, 229–233 (1992).
- [65] Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).
- [66] Renart, J., Reiser, J. & Stark, G. R. Transfer of proteins from gels to diazobenzoyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure. *Proceedings of the National Academy of Sciences* **76**, 3116–3120 (1979).
- [67] Langer-Safer, P. R., Levine, M. & Ward, D. C. Immunological method for mapping genes on drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences* **79**, 4381–4385 (1982).
- [68] So, L. H. *et al.* General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.* **43**, 554–560 (2011).

- [69] Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–36 (2005).
- [70] Higuchi, R., Dollinger, G., Walsh, P. S. & Griffith, R. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology* **10**, 413–417 (1992).
- [71] VanGuilder, H. D., Vrana, K. E. & Freeman, W. M. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* **44**, 619 (2008).
- [72] Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
- [73] Augenlicht, L. H. & Koblin, D. Cloning and screening of sequences expressed in a mouse colon tumor. *Cancer Research* **42**, 1088–1093 (1982).
- [74] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).
- [75] Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–502 (2007).
- [76] Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962–5 (2005).
- [77] Garcia, H. G. & Phillips, R. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A* **108**, 12173–8 (2011).
- [78] Salgado, H. *et al.* RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* **41**, D203–213 (2013).

Chapter 2

Statistical mechanical model of coupled transcription from multiple promoters due to TF titration

2.1 Introduction

TFs with regulatory action at multiple promoter targets is the rule rather than the exception, with examples ranging from TFs like the cAMP receptor protein (CRP) in *E. coli* that regulates hundreds of different genes simultaneously, to situations involving multiple copies of the same gene, such as plasmids, retrotransposons, or highly replicated viral DNA. When the number of TFs heavily exceeds the number of binding sites, TF binding to each promoter can be regarded as independent. However, when the number of TF molecules is comparable to the number of binding sites, TF titration will result in correlation (“promoter entanglement”) between transcription of different genes.

In this chapter we develop a statistical mechanical model to show that when the number of TF molecules is comparable to the number of targets, depletion of the TF can result in nontrivial dependence of the regulatory effect on the relative abundance of targets and TF molecules. The existence of this effect has been previously explored in the context of ultrasensitive regulatory networks [1], as well as the impact of decoy binding sites on TF lifetimes and the response of particular genetic circuits [2, 3]. Here we present a generalized model of gene expression in the presence of TF competition. An advantage with this model is that any system of entangled promoters can be explicitly described in terms of its individual components. Moreover quantities of interest can be expressed analytically, which, for example, allows us to easily study the role of model parameters, explore certain limits of e.g. strong/weak TF binding, and efficiently compute TF titration curves without the need of running thousands of time-consuming Gillespie simulations.

A recent study asserts that half of the proteins in *E. coli* come in fewer than 10 copies [4] (30

for TFs), a number comparable to the gene copy number in many important biological situations, including plasmids [5], viral infections [6], gene duplications [7], (retro)transposons [8, 9, 10], rapid cell growth [11], and transfection of DNA into animal cells [12]. Even for some TFs the number of regular chromosomal binding sites could be large enough to titrate TFs (see Appendix 2.B). If this picture is correct, a quantitative understanding of TF titration due to multiple targets will be essential for making predictive models of transcription regulation. Such models could potentially also shed new light onto diseases where gene copy number abnormalities play a role, including cancers [13], neuropsychiatric diseases [14], and autoimmune disorders [12].

As case studies we use three specific promoter architectures, representing three different mechanisms of repressing a gene. All three of these examples have been studied extensively both experimentally and theoretically [15, 16, 17, 18, 19, 20]. The *simple repression* promoter architecture is arguably the most common nonconstitutive architecture in *E. coli* [21] and refers to a single TF binding site blocking RNAP from binding the promoter. For promoters with more than one binding site for a particular TF, 34% of these promoters have two binding sites separated by more than 100 bp [21], indicating a frequent scenario of facilitated *repression with DNA looping* [22, Table 1]. A famous example of this promoter architecture is the well-studied *lac* operon. In a variant of this promoter architecture, reminiscent of GalR repression at the P2 promoter [23], repression can *only* be achieved in the looped conformation. This *repression exclusively due to looping* promoter architecture has the interesting feature that the level of repression is not a monotonic function in number of TFs. Though we believe these three promoter architectures are both interesting and relevant, the particular choices are not central and the formalism presented here makes it possible to calculate the titration effect for any arbitrary regulatory architecture.

The organization of this chapter is as follows. In Sec. 2.2 we introduce the thermodynamic models and discuss their validity. In Sec. 2.3 we compute individual ($N = 1$) partition functions for the three important promoter architecture case studies. This will be an instructive exercise before turning to the more abstract treatment of Sec. 2.4, where we compute the partition function for a general set of promoters ($N \geq 1$). In Sec. 2.5 we benefit from the hard work of the previous two sections to make predictions of a quantity of great biological importance, namely the *fold change* in gene expression, a quantity directly accessible experimentally. In Sec. 2.6 we study correlation in transcription rates of different genes due to TF titration. In Sec. 2.7 we extend the work of previous sections to include the case when TF and promoter copy numbers are not fixed but rather fluctuating according to a statistical distribution. Finally, in Sec. 2.8 we use Gillespie simulations to verify the thermodynamic model and derive a relationship between the stochastic model rate constants and thermodynamic free energy parameters for the three specific promoter architectures considered.

2.2 Underlying assumptions of thermodynamic model

One of the most ubiquitous quantitative descriptions of transcription is founded upon the so-called thermodynamic models of regulation. In these models, the quantitative behavior of a given promoter is characterized in terms of the occupancy of that promoter by the transcription apparatus and a constellation of molecular partners such as TF and nucleosomes [24, 25, 26, 27]. One of the reasons for the success of these thermodynamic approaches is that in some cases the time scale associated with the production of mRNA is often much slower than the rate at which most proteins, such as TFs, move around within the cell [28] and bind or unbind DNA. For example, the effective (1D+3D [29]) diffusion constant of LacI has been measured as $D_{eff} = 0.4 \pm 0.02 \mu\text{m}^{-2}\text{s}^{-1}$ [28], which means that a LacI molecule can explore the full length of an *E. coli* cell in a few seconds. This should be compared to the significantly slower production rate of LacI which, averaged over the cell cycle, corresponds to around ~ 0.3 per min [30]. Thus, there is reason to believe that LacI, and probably other TFs, can significantly explore the DNA over the time scales at which LacI is produced, providing circumstantial support for a quasiequilibrium approximation. This separation of time scales permits the use of statistical mechanics at promoters that satisfy this condition in order to compute the probabilities of different configurations of TFs and RNAP on the promoter targets. The thermodynamic approach has been used far and wide for characterizing a host of different regulatory processes [24, 25, 26, 31, 32, 33, 34, 35, 36, 37]. Interestingly, this approach not only serves as a very powerful conceptual framework for predicting the behavior of different architectures, but even in those cases where it fails it is useful for suggesting new hypotheses [38, 39, 40, 41, 19, 42].

Of course, this thermodynamic approach is really only the simplest first idea that one can exploit, but at a deeper level it is just a caricature of the real complications of the transcription process and the next layer of sophistication involves using rate equations. However, even in those cases in which models of transcription are built using rate equations, they too essentially appeal to thermodynamic models through the functions describing the occupancy of TFs. Generically, in these cases one writes a rate of production for some protein as

$$\frac{dA}{dt} = -\gamma A + f_{occupancy}([\text{TF}]), \quad (2.1)$$

where $f_{occupancy}([\text{TF}])$ is an occupancy function that reflects the probability of occupancy of TF binding sites as a function of the concentration of these factors. To make the point concrete, consider the example of an activator that activates its own production. In this case, one typically writes a rate equation of the form

$$\frac{dA}{dt} = -\gamma A + r_0 + r_1 \frac{\left(\frac{A}{K_d}\right)^n}{1 + \left(\frac{A}{K_d}\right)^n}, \quad (2.2)$$

where the first term describes protein degradation and dilution from cell growth and the second term

describes basal production at a rate r_0 . The third term is a Hill function [43] relating production to the occupancy of the promoter by its activator. This is obtained using precisely the same statistical mechanics arguments that are common in thermodynamic models. The dissociation constant K_d is only meaningful in the context of equilibrium, and a rapid change in TF copy number cannot correspond to an instantaneous response in promoter occupancy. Therefore, one again needs to rest on the assumption of quasiequilibrium. The literature is replete with examples of both prokaryotic and eukaryotic transcription regulation based upon these kinds of occupancy-based rate equations [44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58], only further raising the stakes for exploring the limits and validity of this approach.

Using the thermodynamic formalism described above, we consider a (quasi) equilibrium system, where the number of RNAP (P), TFs (F) and target promoters (N) are fixed. The term *promoter* will be used either to refer to the RNAP binding site or the full promoter region, including TF binding sites, depending on context. The number of nonspecific binding sites N_{NS} is assumed to be much larger than the number of RNAPs, the number of TFs, and the number of promoters ($N_{NS} \gg P, F, N$). Representative values for these parameters in *E. coli* are given by $P \approx 10^3$ [59, 60, 61], $N_{NS} \approx 5 \times 10^6$ (the size of the *E. coli* genome), $F \approx 1 - 10^3$ [4, 62], and $N \approx 1 - 10^2$ [5, 6]. Unless stated otherwise we will use these given values of P and N_{NS} where concrete numbers are needed. Further we assume that TFs and RNAP are always bound to DNA and do not roam freely in the cell. This is justified in the cases of RNAP and the Lac repressor, for example, by studies using mini cells [63, 64], though this is not necessarily generically true. The results are easily adjusted to the case in which the TFs are free in the cytoplasm rather than nonspecifically bound. We furthermore assume that the promoters have no shared binding sites and that they do not interact except via the competition for TFs.

For each configuration of TFs and RNAP we associate a free energy and corresponding Boltzmann weight, which will determine the probability for the system to be in that particular state [24, 25, 26, 27]. The partition function (Z) is the sum of all these weights. Using the partition function the probability of finding RNAP bound to the promoter of interest can be calculated. This probability can in turn be related to the level of gene expression, a quantity accessible through the use of genetic reporters, or *fold change*, defined as the ratio of the level of gene expression in the presence vs the absence of a TF of interest, by assuming that the RNAP binding probability and gene expression are linearly related [26, 27]. Such a linear relationship has been observed *in vitro* between RNAP binding probability and open complex formation when RNAP binding is the rate limiting step in transcription initiation [65]. A fully generalized model of transcription initiation taking the rates of open complex formation, promoter escape as well as intermediate conformational changes into account [66, 67, 68] is beyond the scope of this thesis. Likewise, we assume that TFs act by modifying the RNAP binding affinity to the promoter. For repressors we can argue that this is indeed a

common mechanism of repression by noticing that almost half [69] of these operators overlap with the RNAP binding region spanning about 40bp upstream from the transcription start site, hence blocking RNAP from binding the promoter. In some cases also other mechanisms of transcriptional regulation, such as modulation of the promoter escape rate, can be rephrased in the thermodynamic language above, e.g., in the case of fast open complex formation. In general, however, the regulatory effect of a TF on transcription initiation depends in a complex way on the TF (un)binding rates and the rates of the various transcription initiation steps of the particular promoter, which again is beyond the scope of this thesis.

2.3 Single promoter partition function

2.3.1 Simple repression

Of the 795 transcription units reported in RegulonDB 7.1 [21] to have at least one TF interaction, 125 correspond to simple repressors [19], making it the most common promoter architecture in *E. coli*. The simple repressor has a single binding site overlapping the promoter such that RNAP cannot bind (or form an open complex which is mathematically equivalent in the context of our model) in the presence of repressor hence inhibiting transcription [see Fig. 2.1(A)]. A classic example of this regulatory motif are the well-studied *lac* operon mutants [15, 18].

The partition function for a simple repressor was derived in [26], but is for the sake of completeness recaptured here. We assume that when not bound to the promoter, RNAP can be found at any of N_{NS} nonspecific binding sites with a binding energy of ε_{pd}^{NS} . Treating the RNAP molecules as indistinguishable, there are $\binom{N_{NS}}{P}$ ways of arranging P RNAP molecules on this nonspecific reservoir. The partition function corresponding to this situation is

$$Z_P^{NS} = \binom{N_{NS}}{P} e^{-\beta P \varepsilon_{pd}^{NS}}. \quad (2.3)$$

As stated above, we assume that $N_{NS} \gg P$, which allows us to make the approximation $\binom{N_{NS}}{P} = \frac{N_{NS}!}{P!(N_{NS}-P)!} \simeq \frac{N_{NS}^P}{P!}$.

Assuming that the repressor has only one binding head, leaving the more complicated case of two binding heads to Sec. 2.3.2, the logic for finding the contribution of R repressor molecules to the total partition function imitates that for RNAP, namely,

$$Z_R^{NS} = \binom{N_{NS}}{R} e^{-\beta R \varepsilon_{rd}^{NS}}, \quad (2.4)$$

where ε_{rd}^{NS} is the nonspecific repressor binding energy. Again, assuming $N_{NS} \gg R$ allows us to approximate $\binom{N_{NS}}{R} \simeq \frac{N_{NS}^R}{R!}$.

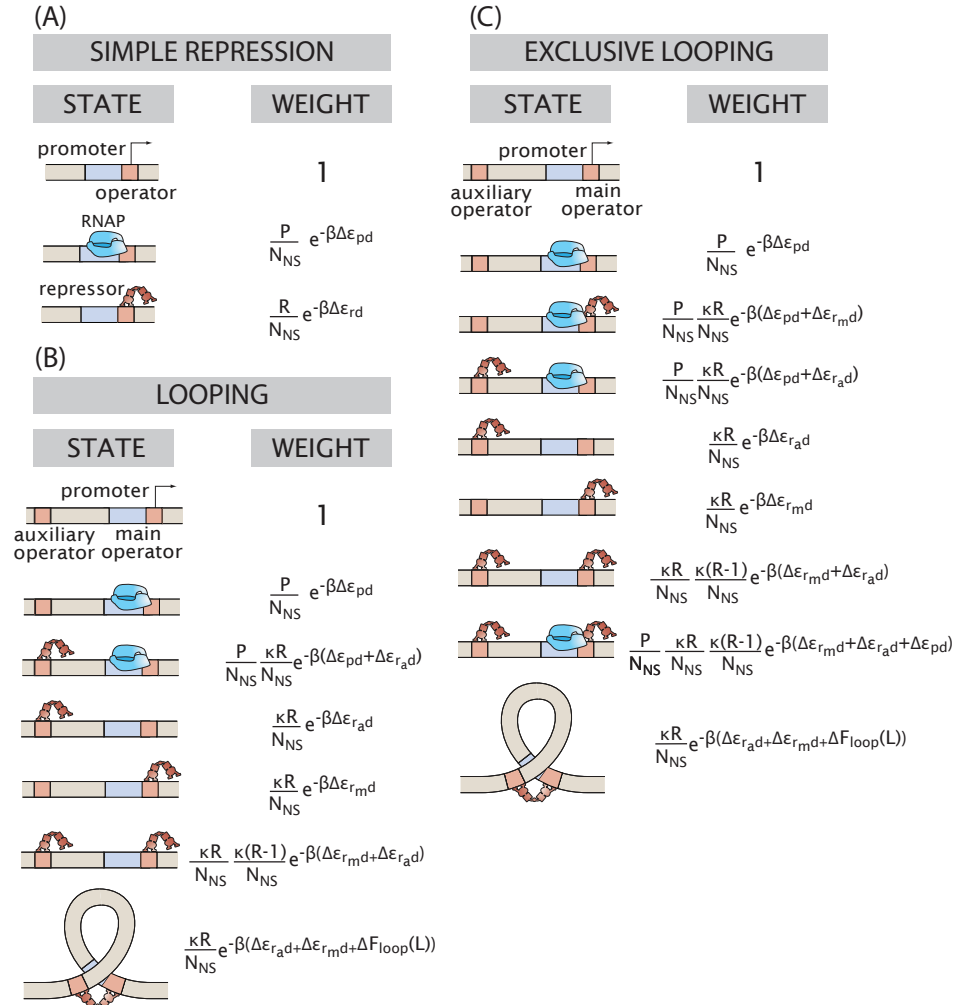


Figure 2.1: States and weights for the three studied promoter architectures (A) simple repression, (B) repression with looping, and (C) repression exclusively due to looping. The last two promoter architectures differ only by the addition of two states to the exclusive looping architecture (third and eighth from the top), corresponding to RNAP and the main operator being simultaneously bound.

Since the total number of nonspecific sites is in great excess with respect to both the number of repressors and RNAP, we can treat nonspecific binding of repressors and RNAP as independent, and hence the total nonspecific partition function is given by the product

$$Z^{NS} = Z_{\text{P}}^{NS} Z_{\text{R}}^{NS}. \quad (2.5)$$

We use this nonspecific partition function to find the overall partition function Z that accounts for binding to the promoter. The promoter can be found in three different states: empty, occupied by RNAP, or occupied by a repressor. As a consequence the overall partition function is given by

$$Z(P, R) = \underbrace{Z^{NS}(P, R)}_{\text{empty}} + \underbrace{Z^{NS}(P-1, R)e^{-\beta\varepsilon_{pd}^S}}_{\text{RNAP bound}} + \underbrace{Z^{NS}(P, R-1)e^{-\beta\varepsilon_{rd}^S}}_{\text{repressor bound}}. \quad (2.6)$$

The first term corresponds to an empty promoter, the second term corresponds to taking an RNAP molecule from the nonspecific reservoir and binding it to the promoter with a specific binding energy of ε_{pd}^S , and the third term similarly corresponds to taking a repressor from the nonspecific reservoir and binding it to the promoter with a specific binding energy ε_{rd}^S . If we normalize by $Z^{NS}(P, R)$ to assign the empty promoter weight 1, the partition function is given by

$$Z = 1 + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_{pd}} + \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{rd}}, \quad (2.7)$$

where we have defined the energy differences $\Delta\varepsilon_{rd} = \varepsilon_{rd}^S - \varepsilon_{rd}^{NS}$ and $\Delta\varepsilon_{pd} = \varepsilon_{pd}^S - \varepsilon_{pd}^{NS}$. The factors $\frac{R}{N_{NS}}, \frac{P}{N_{NS}}$ in the last two terms are of entropic origin and associated with the cost of forcing one molecule to stay on a particular site on the DNA, rather than letting it explore the full range of possible nonspecific sites.

2.3.2 Repression with looping

In repression with looping, RNAP is still excluded from the promoter by repressor binding to a main operator in the vicinity of the promoter. In this case, however, the repressors have two binding heads that can simultaneously bind the main operator and an *auxiliary* operator through the formation of a DNA loop, though the auxiliary operator does not block the promoter on its own (see Fig. 2.2). As a result, there is an increase of effective concentration of repressor in the vicinity of the main operator leading to an increase in repression [15, 16, 17, 70]. One of the most studied realizations of this promoter architecture is again based on modifications of the *lac* operon [15, 16].

To compute the nonspecific partition function for repressors with two binding heads, we begin with a single repressor molecule ($R = 1$). Then, invoking the assumption that the nonspecifically bound repressors are noninteracting, it is easy to generalize the result to any number of repressors

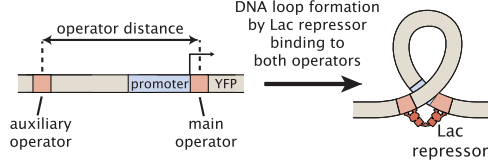


Figure 2.2: Repression through DNA looping. The repressor binds to the main and auxiliary operators simultaneously looping the intervening DNA.

($R > 1$). A single repressor molecule can be found in either a looped state, with both heads bound, or in a state with one head unbound. Each bound repressor head acquires a binding energy of ϵ_{rd}^{NS} , and for looped states there is an additional free energy cost (elastic plus entropic) $F_{loop}(i, j)$ of bringing two sites i and j together.

Taking every possible such configuration into account we find the nonspecific single repressor partition function

$$Z_R^{NS}(R=1) = \underbrace{\sum_{i=1}^{N_{NS}} e^{-\beta\epsilon_{rd}^{NS}}}_{\text{One head bound}} + \frac{1}{2} e^{-2\beta\epsilon_{rd}^{NS}} \underbrace{\sum_{i=1}^{N_{NS}} \sum_{j=1, j \neq i}^{N_{NS}} e^{-\beta F_{loop}(i,j)}}_{\text{Two heads bound}}. \quad (2.8)$$

The factor of $\frac{1}{2}$ in the second sum is necessary to avoid double counting of the looped states. To simplify this expression we assume translational invariance, such that the last sum over j is independent of i . This assumes that on average DNA “looks the same” everywhere, at least locally. Using this assumption we get

$$\begin{aligned} Z_R^{NS}(R=1) &= N_{NS} e^{-\beta\epsilon_{rd}^{NS}} \left(1 + \frac{1}{2} e^{-\beta\epsilon_{rd}^{NS}} \sum_{j=2}^{N_{NS}} e^{-\beta F_{loop}(1,j)} \right) \\ &\equiv N_{NS} e^{-\beta\epsilon_{rd}^{NS}} e^{-\beta F_{eff}^{NS}}. \end{aligned} \quad (2.9)$$

In the last step we defined the effective nonspecific free energy F_{eff}^{NS} . To extend $Z_R^{NS}(R=1)$ to an arbitrary number of repressors $R \geq 1$ we use a familiar result from statistical mechanics, namely,

$$Z_R^{NS}(R) = \frac{1}{R!} (Z_R^{NS}(R=1))^R \quad (2.10)$$

$$= \frac{N_{NS}^R}{R!} e^{-\beta R(\epsilon_{rd}^{NS} + F_{eff}^{NS})}, \quad (2.11)$$

which is applicable for indistinguishable and noninteracting repressors.

Finally, to find the total nonspecific partition function we combine $Z_R^{NS}(R)$ with the nonspecific

partition function for RNAP found in previous section [Eq. (2.3)] resulting in

$$\begin{aligned} Z^{NS}(R, P) &= Z_R^{NS}(R) Z_P^{NS}(P) \\ &= \frac{N_{NS}^R}{R!} \frac{N_{NS}^P}{P!} e^{-\beta R(\varepsilon_{rd}^{NS} + F_{eff}^{NS})} e^{-\beta P \varepsilon_{pd}^{NS}}. \end{aligned} \quad (2.12)$$

Our next task is to determine the weights for all states of the promoter shown in Fig. 2.1(B). As an example we show how to determine the weight for the state with only the main operator bound by a repressor with binding energy ε_{rmd}^S . For this state we need to consider all configurations for the second repressor head not bound to the main operator, as well as all configurations $Z_R^{NS}(R-1)$ for the remaining $R-1$ nonspecifically bound repressors. The weight associated with the specifically bound repressor is given by

$$\begin{aligned} Z_R^{NS}(R=1, \text{one repressor head bound to main operator}) \\ &= e^{-\beta \varepsilon_{rmd}^S} \left(1 + e^{-\beta \varepsilon_{rd}^{NS}} \sum_{j=2}^{N_{NS}} e^{-\beta F_{loop}(1,j)} \right) \\ &\equiv e^{-\beta \varepsilon_{rmd}^S} e^{-\beta \tilde{F}_{eff}^{NS}}, \end{aligned} \quad (2.13)$$

where we have introduced another useful effective free energy \tilde{F}_{eff}^{NS} (note the absent factor of $\frac{1}{2}$), which allows us to express the weight associated with the nonspecifically bound or free hanging repressor head simply as $e^{-\beta \tilde{F}_{eff}^{NS}}$.

Using the same normalization condition as above we find the Boltzmann weight for the state with only the main operator bound

$$\begin{aligned} \text{Weight} \left(\text{Diagram} \right) &= \frac{e^{-\beta \varepsilon_{rmd}^S} e^{-\beta \tilde{F}_{eff}^{NS}} Z_R^{NS}(R-1) Z_P^{NS}(P)}{Z_R^{NS}(R) Z_P^{NS}(P)} \\ &= \frac{R}{N_{NS}} e^{-\beta(\varepsilon_{rmd}^S - \varepsilon_{rd}^{NS})} e^{-\beta(\tilde{F}_{eff}^{NS} - F_{eff}^{NS})} \\ &= \frac{\kappa R}{N_{NS}} e^{-\beta \Delta \varepsilon_{rmd}}. \end{aligned} \quad (2.14)$$

For convenience we introduce the following notation

$$\left\{ \begin{array}{l} \Delta \varepsilon_{rmd} = \varepsilon_{rmd}^S - \varepsilon_{rd}^{NS}, \\ \Delta \varepsilon_{rad} = \varepsilon_{rad}^S - \varepsilon_{rd}^{NS}, \\ \Delta \varepsilon_{pd} = \varepsilon_{pd}^S - \varepsilon_{pd}^{NS}, \\ \Delta F_{loop} = F_{loop}^S - (\tilde{F}_{eff}^{NS} - \varepsilon_{rd}^{NS}), \\ \kappa = e^{-\beta(\tilde{F}_{eff}^{NS} - F_{eff}^{NS})}, \\ p = \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}. \end{array} \right. \quad (2.15)$$

Here $\Delta\varepsilon_{r_{md}}$ and $\Delta\varepsilon_{r_{ad}}$ correspond to the main and auxiliary operators, respectively. From the definitions of \tilde{F}_{eff}^{NS} and F_{eff}^{NS} it is easy to see that κ is always a number between 1 and 2. If $\tilde{F}_{eff}^{NS} \gg 1$ there is a large probability of nonspecific loop formation and $\kappa \simeq 2$. On the other hand, if $\tilde{F}_{eff}^{NS} \simeq 1$ then there is just a small probability of nonspecific loop formation and $\kappa \simeq 1$. Thus, κ can be viewed as a parameter related to how many repressor heads are effectively bound nonspecifically to DNA.

Using the same method we can compute the weights for all other states, and by adding these weights together we get the single promoter partition function

$$Z = 1 + p + p \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta \Delta\varepsilon_{r_{ad}}} + \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta \Delta\varepsilon_{r_{ad}}} + \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta(\Delta\varepsilon_{r_{ad}} + \Delta\varepsilon_{r_{md}} + \Delta F_{loop})} \\ + \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta \Delta\varepsilon_{r_{md}}} + \frac{\kappa R}{N_{NS}} \frac{\kappa(R-1)}{N_{NS}} e^{-\beta(\Delta\varepsilon_{r_{ad}} + \Delta\varepsilon_{r_{md}})}. \quad (2.16)$$

Here the states are listed in the same order as in Fig. 2.1(B).

2.3.3 Exclusive looping repression

For repression due exclusively to looping the situation is similar to the previous section but with the difference that RNAP is considered to be blocked from binding the promoter *only* in the looped state. Hence, it is not enough for just the main operator to be occupied to achieve repression. Such a model of repression is reminiscent of the mechanism of galactose metabolism repression by GalR at the P_2 promoter [23] and the arabinose metabolism AraC repression at the P_C promoter in the absence of arabinose [71].

For this promoter architecture terms need to be added to the partition function of Eq. (2.16) corresponding to states with the main or auxiliary operator bound by repressor and the promoter bound by RNAP. In Fig. 2.1(C) these states are given by the third and eighth states from the top. After taking these new states into account we find the single promoter partition function

$$Z = 1 + p + p \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta \Delta\varepsilon_{r_{md}}} + p \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta \Delta\varepsilon_{r_{ad}}} + \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta \Delta\varepsilon_{r_{ad}}} \\ + \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta \Delta\varepsilon_{r_{md}}} + (1+p) \frac{\kappa R}{N_{NS}} \frac{\kappa(R-1)}{N_{NS}} e^{-\beta(\Delta\varepsilon_{r_{ad}} + \Delta\varepsilon_{r_{md}})} \\ + \left(\frac{\kappa R}{N_{NS}} \right) e^{-\beta(\Delta\varepsilon_{r_{ad}} + \Delta\varepsilon_{r_{md}} + \Delta F_{loop})}, \quad (2.17)$$

where again the states have been listed in the same order as in Fig. 2.1(C).

2.4 Multiple promoter partition function

The simplest example of computing the total partition function for a set of promoters with individual partition functions $Z^{(1)}, Z^{(2)}, \dots, Z^{(N)}$ is when these are *independent*. In our model this happens when the promoters are unregulated or regulated by a TF whose copy number F greatly exceeds the number of promoters ($F \gg N$). Then if one TF binds to a promoter, the number of remaining available TFs is left essentially unchanged, and hence the other promoters are unaffected. By a familiar result from statistical mechanics the total partition function Z^{tot} for a system of independent promoters is given by

$$Z^{tot} = Z^{(1)} Z^{(2)} \dots Z^{(N)} \quad \text{for } F \gg N. \quad (2.18)$$

The complications associated with computing the partition function for a set of promoters regulated by the same TFs originate from the fact that at low TF copy numbers the promoters get “entangled.” For entangled promoters, binding of one TF molecule to a promoter directly influences the TF binding probability to another promoter, due to an effective decrease in the number of available TFs. In the following sections we extend Eq. (2.18) and derive the total partition function for a general set of promoters without making any assumptions about the number of TFs or promoters. While this generality leads to somewhat more abstract derivations, it has the benefit of allowing us to apply the results to a wide range of interesting problems.

2.4.1 General set of promoters

We start by deriving the total partition function for a general set of, potentially different, promoters under control of a single type of TF (F). In Appendix 2.A we generalize to regulation with an arbitrary number of TF types.

First we introduce the notation needed to make these calculations. Let f_n and p_n denote the number of TFs and RNAP bound to promoter $n \in \{1, \dots, N\}$, respectively. Here f_n is constrained by the number of binding sites and the total number of TFs in the cell, namely, $\sum_n f_n \leq F$ and p_n is always either 0 or 1. Let s_n denote the state of promoter n (e.g., empty promoter, operator 1 occupied, operator 2 unoccupied, etc.), and let $F(s_n)$ and $P(s_n)$ denote number of TFs and RNAP bound at promoter n for state s_n . To compute the total partition function we take every allowed state into account by summing over the variables f_n and p_n , as well as the variables s_n for all states compatible with the choice (f_n, p_n) . For each choice $\{f_n\}$ and $\{p_n\}$ there will be $F - \sum_i f_i$ TFs and $P - \sum_i p_i$ RNAPs left for nonspecific binding on the DNA “reservoir,” and the statistical weight associated with these are given by the nonspecific partition functions $Z_F^{NS}(F - \sum_i f_i)$ and $Z_P^{NS}(P - \sum_i p_i)$, which we assume to have the forms $Z_F^{NS}(F) = \frac{N_F^F}{F!} e^{-\beta F \epsilon_{fd}^{NS}}$ and $Z_P^{NS}(P) = \frac{N_P^P}{P!} e^{-\beta P \epsilon_{pd}^{NS}}$, in accordance with our results for the simple repressor (Sec. 2.3.1) and

repression by looping architecture (Sec. 2.3.2). The parameter ε_{fd}^{NS} is assumed to be independent of F . The specifically bound TFs and RNAP to promoter n will acquire a free energy $E(s_n)$ for state s_n . Since there might be many possible states s_n for a given choice (f_n, p_n) we need to sum over all states s_n compatible with this choice, to find the specific part $\sum_{s_n} e^{-\beta E(s_n)} \Big|_{F(s_n)=f_n, P(s_n)=p_n}$ of the statistical weight for promoter n . If there are no states s_n for a given (f_n, p_n) , the sum over s_n is set equal to 0. This is, for example, the case for the simple repressor which cannot have both TF and RNAP specifically bound at the same time ($f_n = p_n = 1$) due to steric exclusion. The specific part of the weight for different promoters “commute,” meaning that we can simply multiply these parts together. The promoter entanglement is fully contained inside the F dependent factorial terms, which motivates the order we have chosen to carry out the summations (f_n, p_n, s_n) . Using a normalization where the state with N empty promoters is assigned weight 1, the total partition function is given by

$$\begin{aligned} Z^{tot} &= \sum_{\substack{f_1, \dots, f_N \\ \sum_i f_i \leq F}} \sum_{p_1, \dots, p_N} \frac{Z_F^{NS}(F - \sum_i f_i) Z_P^{NS}(P - \sum_i p_i)}{Z_F^{NS}(F) Z_P^{NS}(P)} \prod_{n=1}^N \sum_{\substack{s_n \\ F(s_n)=f_n \\ P(s_n)=p_n}} e^{-\beta E(s_n)} \\ &= \sum_{\substack{f_1, \dots, f_N \\ \sum_i f_i \leq F}} \sum_{p_1, \dots, p_N} \frac{F!}{N_{NS}^{\sum_i f_i} (F - \sum_i f_i)!} \frac{P!}{N_{NS}^{\sum_i p_i} (P - \sum_i p_i)!} \prod_{n=1}^N \sum_{\substack{s_n \\ F(s_n)=f_n \\ P(s_n)=p_n}} e^{-\beta \Delta E(s_n)}, \end{aligned} \quad (2.19)$$

where on the second line we have defined $\Delta E(s_n) = E(s_n) - f_n \varepsilon_{fd}^{NS} - p_n \varepsilon_{pd}^{NS}$.

We now use the “high RNAP copy number” assumption $P \gg N$ to make further progress on Eq. (2.19) by approximating $\binom{P}{P-i} \simeq \frac{P^i}{i!}$ for i specifically bound RNAP, resulting in

$$\begin{aligned} Z^{tot} &\simeq \sum_{\substack{f_1, \dots, f_N \\ \sum_i f_i \leq F}} \sum_{p_1, \dots, p_N} \frac{F!}{N_{NS}^{\sum_i f_i} (F - \sum_i f_i)!} \prod_{n=1}^N \sum_{\substack{s_n \\ F(s_n)=f_n \\ P(s_n)=p_n}} \left(\frac{P}{N_{NS}} \right)^{p_n} e^{-\beta \Delta E(s_n)} \\ &= \sum_{\substack{f_1, \dots, f_N \\ \sum_i f_i \leq F}} \frac{F!}{N_{NS}^{\sum_i f_i} (F - \sum_i f_i)!} Z_{f_1}^{(1)} Z_{f_2}^{(2)} \dots Z_{f_N}^{(N)} \\ &= \sum_{f_1=0}^{\min(B_1, F)} \sum_{f_2=0}^{\min(B_2, F-f_1)} \dots \sum_{f_N=0}^{\min(B_N, F-\sum_{i=1}^{N-1} f_i)} \frac{F!}{N_{NS}^{\sum_i f_i} (F - \sum_i f_i)!} Z_{f_1}^{(1)} Z_{f_2}^{(2)} \dots Z_{f_N}^{(N)}. \end{aligned} \quad (2.20)$$

Here B_n is the number of TF binding sites on promoter n , and $Z_{f_n}^{(n)}$ has been defined as

$$Z_{f_n}^{(n)} \equiv \sum_{p_n} \sum_{\substack{s_n \\ F(s_n)=f_n \\ P(s_n)=p_n}} \left(\frac{P}{N_{NS}} \right)^{p_n} e^{-\beta \Delta E(s_n)}. \quad (2.21)$$

A key observation is that the single promoter partition functions $Z^{(n)}$ are precisely given in terms

of the $Z_i^{(n)}$ factors,

$$Z^{(n)} = \sum_{i=0}^{B_n} \frac{F!}{N_{NS}^i (F-i)!} Z_i^{(n)}, \quad (2.22)$$

which implies that once the single promoter partition functions are known, the total partition function for the set of promoters can be directly obtained from Eq. (2.20), independently of promoter architectures.

2.4.2 Identical promoters

Evaluating the total partition function for a general set of promoters can be computationally expensive. In Eq. (2.20) there are N summation indices $\{f_i\}$ and if these are not constrained by the number of TFs ($F \geq \sum_{i=1}^N B_i$) there are $\prod_{i=1}^N (1 + B_i)$ different terms in the summation. As the number of promoters N increases this number grows exponentially, and computing the partition function presents a great challenge. In the important special case of N identical promoter copies each with partition function

$$Z = \sum_{i=0}^B \frac{F!}{N_{NS}^i (F-i)!} Z_i, \quad (2.23)$$

however, the computational cost can be significantly reduced.

One way to keep track of the total number of bound TFs is to introduce numbers $\{k_i\}$, where k_i denotes the number of promoter copies occupied by i TFs, with the additional constraints $\sum_{i=0}^B k_i = N$ and $\sum_{i=0}^B i k_i \leq F$. To compute the partition function we first need to find the number of possible arrangements given numbers $\{k_i\}$, or the “degeneracy”. As an example for $k_0 = N$ there is only one choice (all promoters empty), but for $k_0 = N - 1, k_1 = 1$ there are N different choices, corresponding to N different ways of choosing a single promoter to be occupied by one TF (assuming $B, F \geq 1$). Here we treat the promoters as distinguishable physical objects which is a valid assumption since the promoters have the additional intrinsic degrees of freedom (e.g., position) that separate them. Starting with empty promoters, there are $\binom{N}{k_0}$ ways of choosing k_0 promoters without bound TF. From the remaining $N - k_0$ promoters we choose k_1 promoters with exactly one TF bound; this can be done in $\binom{N-k_0}{k_1}$ ways. Repeating this procedure B times gives us the degeneracy, namely,

$$\begin{aligned} \text{degeneracy}\{k_i\} &= \binom{N}{k_0} \binom{N-k_0}{k_1} \dots \binom{N-\sum_{i=0}^{B-1} k_i}{k_B} \\ &= \binom{N}{k_0, k_1, \dots, k_B}, \end{aligned} \quad (2.24)$$

where $\binom{N}{k_0, k_1, \dots, k_B} = \frac{N!}{k_0! k_1! \dots k_B!}$ is the multinomial coefficient. To find the total partition function

Z^{tot} we need to sum over all allowed values of $\{k_0, k_1, \dots, k_B\}$ and take the degeneracy into account. Using otherwise the same weights as in Eq. (2.20) we find the total partition function for identical promoter copies,

$$\begin{aligned}
Z^{tot} &= \sum_{\substack{k_0, k_1, \dots, k_B \\ \sum_i k_i = N \\ \sum_i i k_i \leq F}} \binom{N}{k_0, k_1, \dots, k_B} \frac{F!}{N_{NS}^{\sum_i i k_i} (F - \sum_i i k_i)!} \prod_{i=0}^B Z_i^{k_i} \\
&= \sum_{k_B=0}^{\min(N, \lfloor F/B \rfloor)} \cdots \sum_{k_j=0}^{\min(N - \sum_{i=j+1}^B k_i, \lfloor (F - \sum_{i=j+1}^B i k_i) / j \rfloor)} \cdots \sum_{k_1=0}^{\min(N - \sum_{i=2}^B k_i, F - \sum_{i=2}^B i k_i)} \\
&\quad \times \binom{N}{k_0, k_1, \dots, k_B} \frac{F!}{N_{NS}^{\sum_i i k_i} (F - \sum_i i k_i)!} \prod_{i=0}^B Z_i^{k_i}. \tag{2.25}
\end{aligned}$$

Here k_0 is assigned the implicit value $k_0 = N - \sum_{i=1}^B k_i$ and $\lfloor \cdot \rfloor$ denotes the floor function ¹.

When the indices $\{k_i\}$ are not constrained by the number of TFs ($F \geq NB$), corresponding to the most computationally expensive case, the number of terms in the summation of Eq. (2.25) equals the number of nonnegative integer solutions to the equation

$$k_0 + k_1 + \dots + k_B = N. \tag{2.26}$$

This is a classical problem from combinatorics with the number of solutions given by $\binom{N+B}{N} \approx N^B/B!$ which grows *polynomially* with number of promoters N . Intuitively, we can understand the polynomial dependence from the fact that there are B different indices (not counting $k_0 = N - \sum_{i=1}^B k_i$), each of which can take N different values. Hence, the partition function for identical promoter copies can be computed for much higher values of promoter copies N than permitted by the general formula [Eq. (2.20)].

2.4.3 Simple repression

We now use our general results [Eq. (2.25)] to compute the partition function for multiple copies of the specific promoter architectures considered in Sec. 2.3, starting with simple repression. From the single promoter partition function Z [Eq. (2.7)] one can easily termwise identify $Z_0 = 1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}$ and $Z_1 = e^{-\beta \Delta \varepsilon_{rd}}$, where $Z = Z_0 + \frac{R}{N_{NS}} Z_1$. These factors are needed to compute the total partition function for multiple promoter copies.

Plugging Z_0, Z_1 into the general formula of Eq. (2.25) gives us the total partition function for N

¹The floor function $\lfloor x \rfloor$ is the largest integer not greater than x , e.g. $\lfloor 1.8 \rfloor = \lfloor 1.2 \rfloor = 1$.

promoters,

$$Z^{tot} = \sum_{k_1=0}^{\min(N,R)} \binom{N}{k_1} \frac{R!}{N_{NS}^{k_1} (R-k_1)!} e^{-\beta k_1 \Delta \varepsilon_{rd}} (1+p)^{N-k_1}, \quad (2.27)$$

where $p = \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}$. The summation in Eq. (2.27) can be carried out explicitly to yield a closed form expression of the partition function in terms of the Tricomi confluent hypergeometric function [72, 73].

2.4.4 Repression with looping

From the single promoter partition function for repression with looping [Eq. (2.16)] we identify the following Z_0, Z_1, Z_2 factors

$$\begin{cases} Z_0 = 1 + p, \\ Z_1 = \kappa \left(e^{-\beta \Delta \varepsilon_{rmd}} + (1+p) e^{-\beta \Delta \varepsilon_{rad}} \right. \\ \quad \left. + e^{-\beta (\Delta \varepsilon_{rmd} + \Delta \varepsilon_{rad} + \Delta F_{loop})} \right) \\ Z_2 = \kappa^2 e^{-\beta (\Delta \varepsilon_{rmd} + \Delta \varepsilon_{rad})}, \end{cases} \quad (2.28)$$

where $Z = Z_0 + \frac{R}{N_{NS}} Z_1 + \frac{R(R-1)}{N_{NS}^2} Z_2$. With the help of these we get the total promoter partition function for N promoter copies from Eq. (2.25)

$$\begin{aligned} Z^{tot} &= \sum_{k_2=0}^{\min(N, \lfloor R/2 \rfloor)} \sum_{k_1=0}^{\min(N-k_2, R-2k_2)} \binom{N}{k_2, k_1, N-k_2-k_1} \frac{R!}{N_{NS}^{k_1+2k_2} (R-k_1-2k_2)!} \\ &\quad \times \kappa^{k_1+2k_2} \left(e^{-\beta \Delta \varepsilon_{rmd}} + (1+p) e^{-\beta \Delta \varepsilon_{rad}} + e^{-\beta (\Delta \varepsilon_{rmd} + \Delta \varepsilon_{rad} + \Delta F_{loop})} \right)^{k_1} \\ &\quad \times (1+p)^{N-k_1-k_2} e^{-\beta k_2 (\Delta \varepsilon_{rmd} + \Delta \varepsilon_{rad})}. \end{aligned} \quad (2.29)$$

2.4.5 Exclusive looping repression

Again, using the single promoter partition function [Eq. (2.17)] we identify the Z_0, Z_1, Z_2 factors for the exclusive looping repression architecture,

$$\begin{cases} Z_0 = 1 + p \\ Z_1 = \kappa \left((1+p) \left(e^{-\beta \Delta \varepsilon_{rmd}} + e^{-\beta \Delta \varepsilon_{rad}} \right) + e^{-\beta (\Delta \varepsilon_{rmd} + \Delta \varepsilon_{rad} + \Delta F_{loop})} \right) \\ Z_2 = \kappa^2 (1+p) e^{-\beta (\Delta \varepsilon_{rmd} + \Delta \varepsilon_{rad})}. \end{cases} \quad (2.30)$$

By plugging these factors into Eq. (2.25) we find the total partition function for N promoter copies

$$\begin{aligned}
Z^{tot} = & \sum_{k_2=0}^{\min(N, \lfloor R/2 \rfloor)} \sum_{k_1=0}^{\min(N-k_2, R-2k_2)} \binom{N}{k_2, k_1, N-k_2-k_1} \frac{R!}{N_{NS}^{k_1+2k_2} (R-k_1-2k_2)!} \\
& \times \kappa^{k_1+2k_2} \left((1+p) (e^{-\beta \Delta \varepsilon_{rmd}} + e^{-\beta \Delta \varepsilon_{rad}}) + e^{-\beta (\Delta \varepsilon_{rmd} + \Delta \varepsilon_{rad} + \Delta F_{loop})} \right)^{k_1} \\
& \times (1+p)^{N-k_1} e^{-\beta k_2 (\Delta \varepsilon_{rmd} + \Delta \varepsilon_{rad})}. \tag{2.31}
\end{aligned}$$

2.5 Fold change

In order to create a bridge between experimental measurements and the thermodynamic model a key assumption is made stating that the level of expression of a gene is proportional to the probability of RNAP being bound to the promoter of the gene, or in the case of multiple gene copies, the expression is proportional to the average number of promoters bound by RNAP. Using this assumption we can predict the *fold change*, defined as the ratio of level of gene expression in the presence vs. absence of a certain TF, which is a quantity commonly measured by experiments. We start by computing the fold change for a set of identical promoter copies and then move to the case with a general set of promoters. In Supplemental Material ² we show how to perform these computations using MATHEMATICA.

By assuming the number of RNAP molecules to be much bigger than the number of promoter copies, any state with i promoters bound by RNAP will have a weight of the form $\propto p^i$, with $p = \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}$. Here $\Delta \varepsilon_{pd}$ is the energy difference between specific and nonspecific RNAP binding to the promoter. Using this observation one can show that the expectation value for the number of promoters bound by RNAP is given by

$$\text{Occupancy} = p \frac{\partial}{\partial p} \ln Z^{tot}. \tag{2.32}$$

Equation (2.32) together with the partition function derived in the previous section allows us to compute the fold change f , defined as the ratio between occupancy in the presence and absence of a TF,

$$f = \frac{\text{Occupancy}(F)}{\text{Occupancy}(F=0)}. \tag{2.33}$$

In the particular case of simple repression, plugging the partition function [Eq. 2.27] into Eqs. (2.32)-

²See Supplemental Material at: <http://dx.doi.org/10.1103/PhysRevE.89.012702>

(2.33) leads, after a bit of algebra, to

$$f = \frac{1+p}{N} \frac{\sum_{k_1=0}^{\min(N,R)} \binom{N}{k_1} \frac{R!}{N_{NS}^{k_1} (R-k_1)!} e^{-\beta k_1 \Delta \varepsilon_{rd}} (N-k_1) (1+p)^{N-k_1-1}}{\sum_{k_1=0}^{\min(N,R)} \binom{N}{k_1} \frac{R!}{N_{NS}^{k_1} (R-k_1)!} e^{-\beta k_1 \Delta \varepsilon_{rd}} (1+p)^{N-k_1}}. \quad (2.34)$$

For weak promoters ($p \ll 1$) we can simplify this expression somewhat by dropping the last factor in the nominator and denominator. The summation can again be expressed in closed form using the Tricomi confluent hypergeometric function and a corresponding differentiation rule [74].

In Fig. 2.3, we show fold change as a function of number of repressors (R) for the three different promoter architectures considered in Sec. 2.3. This figure shows the importance of TF titration as there can exist order of magnitude differences in predicted fold change for $N = 1$ vs $N \geq 1$ promoter copies. For the simple repressor [Fig. 2.3(A)], with $R < N$ the fold change will never be less than $\frac{1}{N}$, corresponding to a situation where all promoters but one are “turned off.” However, as soon as $R \geq N$ all promoters can be repressed, which yields a steep decline in fold change around $R \approx N$, at least when the operators are strong enough to have high repressor binding probability (as is the case in Fig. 2.3). For weak operators the move across the “boundary” $R \approx N$ is uneventful and no such steep response occurs (see Fig. 2.4).

In the exclusive looping repression architecture [Fig. 2.3(C)], the fold change exhibits a sharp trough near $R \approx N$. This is explained by the fact that at high repressor copy number the operators will be bound by repressors separately (an unrepressed state), hence avoiding having to pay the energy cost of bending the DNA, and for low repressor copy number ($R < N$) the fold change is again never less than $\frac{1}{N}$. The observed trough corresponds to the middle range between these two extremes.

Finally, the repression with looping architecture [Fig. 2.3(B)] is a combination of the simple repression and exclusive looping repression architectures. Since both of these architectures show steep response around $R \approx N$ the repression by looping architecture will share this feature, as is apparent from Fig. 2.3(B). The free energy cost ΔF_{loop} of forming DNA loops is critical for this behavior. If ΔF_{loop} is increased such that it exceeds the binding energy of both operators, $\Delta F_{loop} > \max(|\Delta \varepsilon_{rmd}|, |\Delta \varepsilon_{rad}|)$, the auxiliary operator serves only to titrate repressors and the fold change will resemble the simple repression case. For all architectures, the fold change curves converge in the high TF copy number limit ($R \gg N$) independently of promoter copy number. In this limit the number of TFs available for binding is essentially constant and transcription from each promoter can be regarded as independent.

So far we assumed that all promoters are identical; however, for a general set of promoters there might be several different “output” proteins, each with its own associated fold change. By analogy

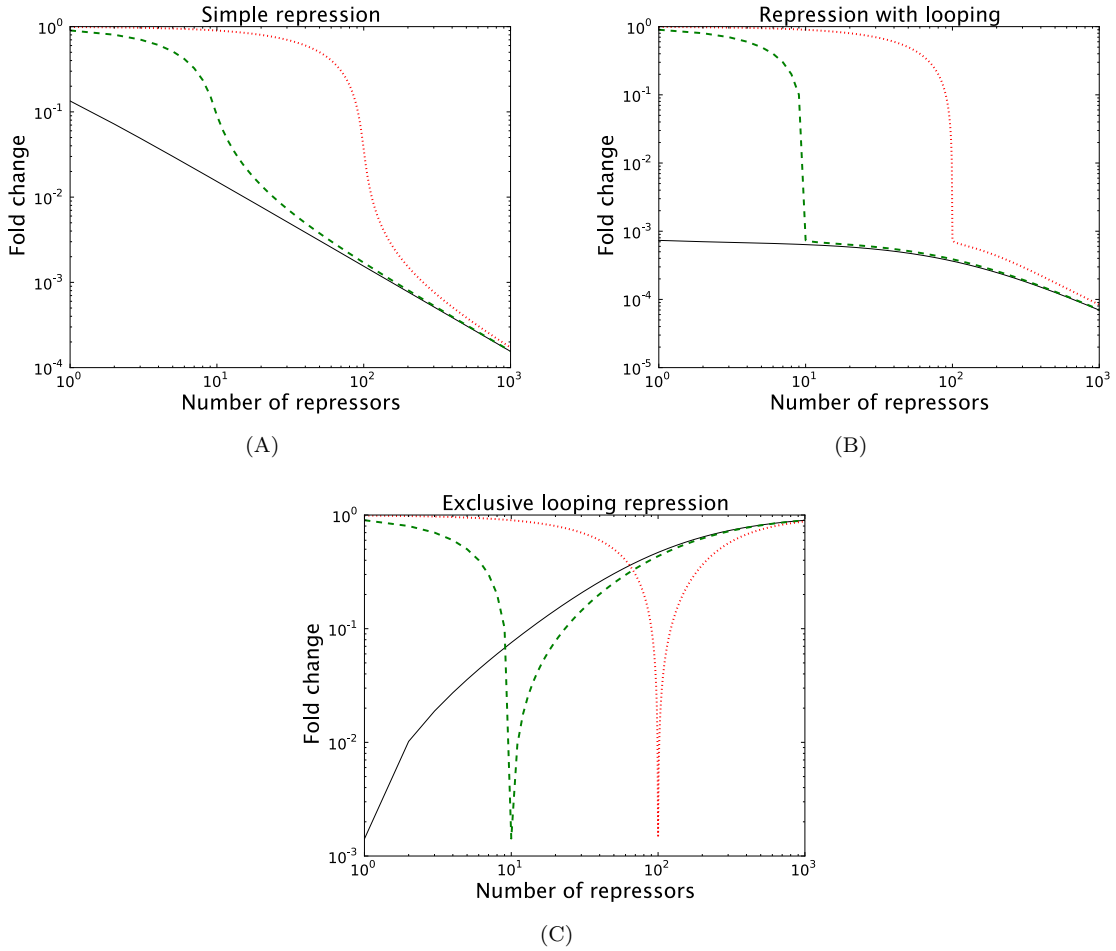


Figure 2.3: Fold change as a function of repressor copy number (R) for gene copy numbers $N = 1$ (solid line), $N = 10$ (dashed), and $N = 100$ (dotted) for three different promoter architectures: (A) simple repression, (B) repression with looping, and (C) exclusive looping repression. For these plots we used operator binding energy $-17.3 k_B T$ (equivalent to the strongest known *lac* operator *O_{id}* [18]), the number of nonspecific sites as the genome length of *E. coli* ($N_{NS} = 5 \times 10^6$), number of RNAP $P = 1000$, and the looping energy $\Delta F_{loop} = 10 k_B T$ [26]. The RNAP promoter binding energy is assumed to be weak ($p \ll 1$) [75, 18].

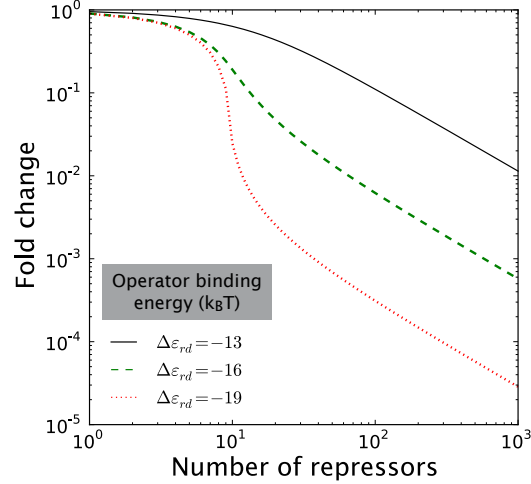


Figure 2.4: Fold change of a simple repressor with gene copy number $N = 10$ for three different TF binding site strengths, with strengths chosen to correspond to the range observed for real repressors. Stronger repressor binding leads to a steeper response in fold change around $R \approx N$. The RNAP promoter binding energy is assumed to be weak ($p \ll 1$), and the number of nonspecific sites $N_{NS} = 5 \times 10^6$.

to the identical promoter case [Eq. (2.33)] we define the fold change $f^{(n)}$ with respect to promoter n as

$$f^{(n)} \equiv \frac{\text{Occupancy for promoter } n (F)}{\text{Occupancy for promoter } n (F = 0)}, \quad (2.35)$$

where the occupancy is given by

$$\text{Occupancy for promoter } n = p^{(n)} \frac{\partial}{\partial p^{(n)}} \ln Z^{tot}, \quad (2.36)$$

with $p^{(n)} \equiv \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{p_n d}}$, and $\Delta \varepsilon_{p_n d}$ the energy difference between specific and nonspecific RNAP binding to promoter n . If one promoter has stronger TF binding sites than the other promoters these binding sites will, in general, be filled first by TFs, but as soon as this happens the other promoter might experience a sudden regulatory response [1]. As an example [3], let us assume we have N_{pl} plasmids, each with one TF binding site of energy $\Delta \varepsilon_{pl}$ as shown in Fig. 2.5(A). The $Z_i^{(1)}$ factors associated with these N_{pl} binding sites are given by

$$Z_i^{(1)} = \binom{N_{pl}}{i} e^{-\beta i \Delta \varepsilon_{pl}}, \quad (2.37)$$

corresponding to $\binom{N_{pl}}{i}$ ways of distributing i repressors on N_{pl} plasmids, each with one binding site. Furthermore, let the same TF act as an inhibitor (see Sec. 2.3.1) for a single simply repressed gene located on the chromosome. We already know the $Z_i^{(2)}$ factors for this promoter architecture from

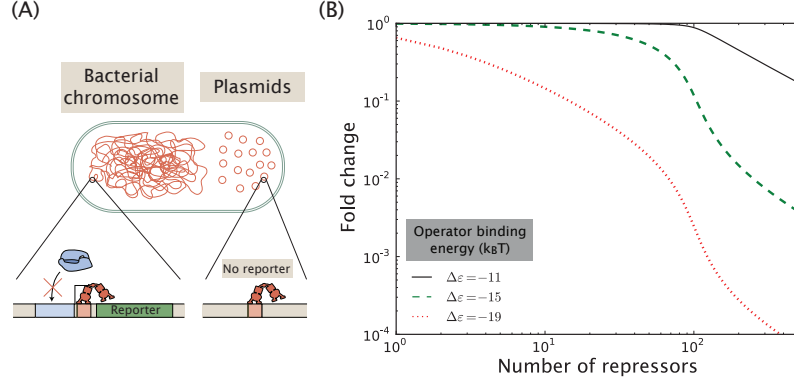


Figure 2.5: Effect of TF sequestration on fold change. (A) A repressor can bind to a reporter construct located in a single copy on the chromosome or to a binding site on a multi-copy plasmid which leads to no gene expression. (B) Fold change of a simple repressor for different repressor binding site strengths, where the TF is subject to sequestration from 100 nonfunctional binding sites ($\Delta\varepsilon_{pl} = -15 k_B T$). If the sequestration sites are much stronger than the simple repressor operator, the fold change remains constant until these sites have been filled. The RNAP promoter binding energy is assumed to be weak ($p \ll 1$), and the number of nonspecific sites $N_{NS} = 5 \times 10^6$.

Sec. 2.4.3, namely $Z_0^{(2)} = 1 + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon}$ and $Z_1^{(2)} = e^{-\beta\Delta\varepsilon}$. Using Eq. (2.20) we find the total partition function of the system

$$\begin{aligned}
 Z^{tot} &= \sum_{i_1=0}^{\min(N_{pl}, R)} \sum_{i_2=0}^{\min(1, R-i_1)} \frac{R!}{N_{NS}^{i_1+i_2} (R-i_1-i_2)!} Z_{i_1}^{(1)} Z_{i_2}^{(2)} \\
 &= (1+p) \sum_{i_1=0}^{\min(N_{pl}, R)} \frac{R!}{N_{NS}^{i_1} (R-i_1)!} \binom{N_{pl}}{i_1} e^{-\beta i_1 \Delta\varepsilon_{pl}} \\
 &+ \sum_{i_1=0}^{\min(N_{pl}, R-1)} \frac{R!}{N_{NS}^{i_1} (R-i_1-1)!} \binom{N_{pl}}{i_1} e^{-\beta i_1 \Delta\varepsilon_{pl}} e^{-\beta\Delta\varepsilon}. \tag{2.38}
 \end{aligned}$$

In Fig. 2.5(B) we show the fold change of the simple repressor on the chromosome for three choices of operator strength $\Delta\varepsilon < \Delta\varepsilon_{pl}$, $\Delta\varepsilon = \Delta\varepsilon_{pl}$, and $\Delta\varepsilon > \Delta\varepsilon_{pl}$. As expected when the plasmid binding sites are very strong, we do not get a response in fold change of the simple repressor until all these sites have been filled. However, if the simple repressor binding site is stronger than the plasmid binding sites, this is no longer the case and we see an immediate decline in fold change when repressors are added. Even on a logarithmic plot the fold change shows a rich structure, which makes it an ideal candidate for experimental verification since we expect that this functional form can be easily detected above the intrinsic experimental noise in making such gene expression measurements.

Finally, for *independent* identical promoters, for example when the TFs are in great excess with respect to number of gene copies, the fold change for the set of promoters reduces to the fold change of an individual promoter. This intuitive result can be directly shown from Eq. (2.32) + (2.33),

using the fact that for N independent promoters, each with partition function Z , the total partition function is given by $Z^{tot} = Z^N$. Let f_Z denote the fold change of a single promoter and $f_{Z^{tot}}$ denote the fold change for N promoter copies, then

$$\begin{aligned} f_{Z^{tot}} &= \frac{\frac{p}{Z^{tot}} \frac{\partial}{\partial p} Z^{tot}}{\frac{p}{Z_{F=0}^{tot}} \frac{\partial}{\partial p} Z_{F=0}^{tot}} = \frac{(Z_{F=0})^N \frac{\partial}{\partial p} Z^N}{Z^N \frac{\partial}{\partial p} (Z_{F=0})^N} \\ &= \frac{\frac{p}{Z} \frac{\partial}{\partial p} Z}{\frac{p}{Z_{F=0}} \frac{\partial}{\partial p} Z_{F=0}} = f_Z. \end{aligned} \quad (2.39)$$

This equality, $f_{Z^{tot}} = f_Z$, greatly simplifies calculating the fold change of the promoters.

2.6 Transcriptional correlation

There are many reasons why expression of different genes might be correlated [76, 77, 78]. One obvious example is if a gene A regulates another gene B , then random intrinsic fluctuations in A will affect the expression of B (with a time delay), resulting in correlated expression of the two genes. For genes without direct regulatory connections, such random fluctuations due to intrinsic noise do not lead to correlated expression. Extrinsic noise, on the other hand, refers to fluctuations which affect the expression of *both* A and B simultaneously; this includes “global noise” such as fluctuating number of RNAP molecules or cell size, which leads to a positive correlation in transcription rates of the two genes. Another example of extrinsic noise, which we study in more depth in Sec. 2.7.2, is fluctuations in TF copy number if A and B are regulated by the same TF.

In addition to these mechanisms we predict that promoter entanglement due to TF titration constitutes another source of correlation in transcription rates for genes regulated by the same TFs. Quantifying this effect is the topic of this section.

2.6.1 Toy model of transcriptional correlation

To develop intuition for the correlation in transcription from different promoters due to promoter entanglement, we first consider a hypothetical system of two unregulated promoters (P_A, P_B), transcribed by a single RNAP molecule ($P = 1$). This system can be found in three different states: no promoter bound by RNAP, P_A bound by RNAP, or P_B bound by RNAP. Since the single RNAP molecule can only bind to one of the promoters at a time, transcription of the two promoters will become anticorrelated.

Let A, B denote the number (0 or 1) of RNAP bound to promoters P_A and P_B respectively.

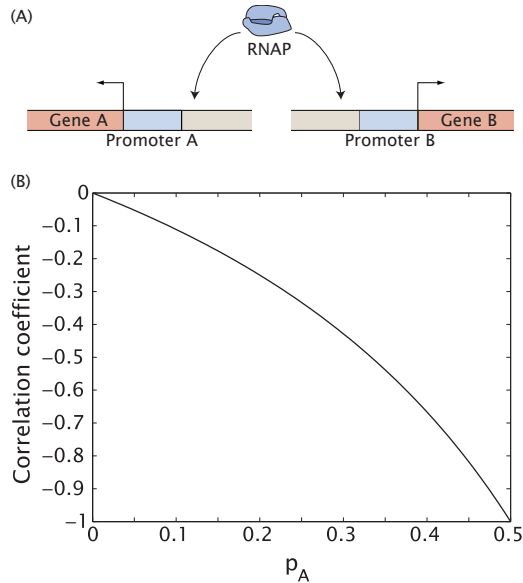


Figure 2.6: Correlation coefficient between transcription rates of two equally strong promoters P_A, P_B for a single RNAP molecule ($P = 1$), as a function of probability $p_A = p_B$ of one of the promoters being bound.

These two random variables are correlated with the Pearson correlation coefficient,

$$\rho_{corr} = \frac{\langle (A - \bar{A})(B - \bar{B}) \rangle}{\sqrt{\langle (A - \bar{A})^2 \rangle \langle (B - \bar{B})^2 \rangle}} \quad (2.40)$$

$$= \frac{\langle (A - \bar{A})(B - \bar{B}) \rangle}{\langle (A - \bar{A})^2 \rangle}. \quad (2.41)$$

For the sake of simplicity we assume that the two promoters P_A, P_B have the same strength, and hence in Eq. (2.41) we set $\langle (A - \bar{A})^2 \rangle = \langle (B - \bar{B})^2 \rangle$. Let p_0 and $p_A = p_B$ denote the probabilities of the three states listed above. In terms of these probabilities the correlation coefficient translates to

$$\rho_{corr} = -\frac{p_A}{1 - p_A}, \quad (2.42)$$

which is plotted as a function of p_A in Fig. 2.6. When the promoters are very strong P_A or P_B will always be bound by RNAP ($p_A = p_B = \frac{1}{2}$); hence, knowledge of the state of one promoter is sufficient to tell the state of the other promoter ($\rho_{corr} = -1$). However, when the promoters are weak, at most times both promoters are empty and the correlation between the promoters will be weak.

These results can be framed in terms of the familiar partition functions used throughout this chapter. We now consider a statistical mechanical model of RNAP binding. The partition function

for the two-promoter system is given by

$$Z = 1 + \frac{1}{N_{NS}} e^{-\beta \Delta \varepsilon_A} + \frac{1}{N_{NS}} e^{-\beta \Delta \varepsilon_B} \quad (2.43)$$

$$= 1 + \frac{2}{N_{NS}} e^{-\beta \Delta \varepsilon}, \quad (2.44)$$

where we again assume that both promoters have the same binding energy $\Delta \varepsilon = \Delta \varepsilon_A = \Delta \varepsilon_B$. The probability p_A for promoter A to be in the bound state is then given by

$$p_A = \frac{\frac{1}{N_{NS}} e^{-\beta \Delta \varepsilon}}{Z}. \quad (2.45)$$

Plugging p_A back into the correlation coefficient [Eq. (2.42)] gives the transcriptional correlation as a function of promoter strength, namely,

$$\rho_{corr} = -\frac{1}{1 + N_{NS} e^{\beta \Delta \varepsilon}}. \quad (2.46)$$

These results are intended to illustrate how the correlations will be computed in the more general case considered next.

2.6.2 General theory

As reported in Sec. 2.5, a state with i specifically bound RNAP molecules to a certain promoter type has a statistical weight of the form $\propto p^i$ with $p = \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}$. This weight generalizes for a set of N different promoter types to the form $\propto p_1^{i_1} \cdots p_N^{i_N}$. Using this observation it is easy to derive the statistical moments for promoter occupancies

$$\langle i_1 \dots i_m \rangle \equiv \frac{1}{Z^{tot}} p_{i_1} \frac{\partial}{\partial p_{i_1}} \cdots p_{i_m} \frac{\partial}{\partial p_{i_m}} Z^{tot}, \quad (1 \leq i_j \leq N, \forall j). \quad (2.47)$$

On the left hand side we use $\langle i_1 \dots i_m \rangle$ as a shorthand notation for the expectation value of the product of number of RNAP simultaneously bound to the promoters specified by the indices i_1, \dots, i_m . For two promoter types ($N = 2$) the Pearson correlation coefficient can be expressed in terms of the partition function as

$$\rho_{i_1 i_2} = \frac{\langle (i_1 - \bar{i}_1)(i_2 - \bar{i}_2) \rangle}{\sqrt{\langle (i_1 - \bar{i}_1)^2 \rangle \langle (i_2 - \bar{i}_2)^2 \rangle}} \quad (2.48)$$

$$= \frac{p_1 p_2 \frac{\partial}{\partial p_1} \frac{\partial}{\partial p_2} \ln Z^{tot}}{\sqrt{\left[\left(p_1 \frac{\partial}{\partial p_1} \right)^2 \ln Z^{tot} \right] \left[\left(p_2 \frac{\partial}{\partial p_2} \right)^2 \ln Z^{tot} \right]}}. \quad (2.49)$$

Here $\bar{i}_{1,2}$ denotes the occupancy ($\langle i_{1,2} \rangle$) for promoters 1, 2, respectively.

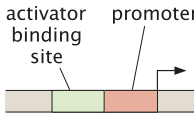
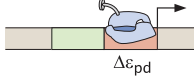
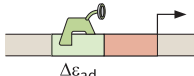
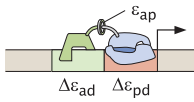
STATE	WEIGHT
	1
	$\frac{P}{N_{NS}} e^{-\Delta\varepsilon_{pd}/k_B T}$
	$\frac{A}{N_{NS}} e^{-\Delta\varepsilon_{ad}/k_B T}$
	$\frac{P}{N_{NS}} \frac{A}{N_{NS}} e^{-(\Delta\varepsilon_{pd} + \Delta\varepsilon_{ad} + \varepsilon_{ap})/k_B T}$

Figure 2.7: States and weights for the simple activation regulatory motif [26].

2.6.3 Two anticorrelated genes

Let us now study the specific example of transcriptional correlation for a system with two genes located together on N_{pl} identical plasmids, where both genes are regulated by the same A activating TFs (*activators*), as shown in Figs. 2.7 and 2.8. The transcription rates for the two genes will be anticorrelated, because when one gene is highly activated there are fewer activator molecules left to also activate the other gene. When there are no activators ($A = 0$) transcription of the two genes is clearly independent, but this is also true if $A \gg N_{pl}$ because the number of activators available for promoter binding will be essentially constant. Hence, we expect anti-correlation of transcription rates between the two genes to have a peak in magnitude when the number of activators is roughly comparable to the number of plasmids ($A \approx N_{pl}$).

There are four different states for a simple activator promoter architecture (see Fig. 2.7): empty state, activator bound, promoter bound (by RNAP), and activator and promoter bound. The last state has a (negative) activator-RNAP interaction energy ε_{ap} , used by the activator to “recruit” RNAP to the promoter. For simplicity we assume that the two genes have the same operator strength and promoter strengths, and hence the same partition function,

$$\begin{aligned}
 Z = 1 + \frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_{pd}} + \frac{A}{N_{NS}} e^{-\beta \Delta\varepsilon_{ad}} \\
 + \frac{A}{N_{NS}} \frac{P}{N_{NS}} e^{-\beta(\Delta\varepsilon_{pd} + \Delta\varepsilon_{ad} + \varepsilon_{ap})}.
 \end{aligned} \tag{2.50}$$

We use Eq. (2.20) to find the partition function for the two genes on one plasmid copy, then Eq. (2.25) to find the partition function for multiple plasmid copies. Once we have the total partition function we can calculate the transcriptional correlation using Eq. (2.49).

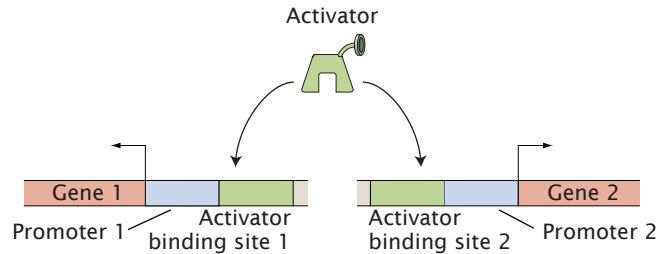


Figure 2.8: Two simple activators regulated by the same TF.

In Fig. 2.9(A) we show the transcriptional correlation of the system as a function of activator copy number for different numbers of plasmids. As expected, the correlation exhibits a peak when the number of activators is similar to the number of plasmids (peak value $\rho \approx -0.8$). As the number of activators outgrows the total number of binding sites ($2N_{pl}$) the correlation dies off rapidly, at least when the activator operators are strong. In Figs. 2.9(B) and 2.9(C) we show how the correlation depends on the RNAP-activator interaction energy ε_{ap} and the binding site strength of the activator $\Delta\varepsilon_{ad}$. As expected, the transcriptional correlation between the two genes increases in magnitude when these interactions are stronger (more negative). In Fig. 2.9(D) we show how the transcriptional correlation depends on the promoter binding strength. Weak promoters only recruit RNAP when bound by activators. With just one single activator molecule this system becomes similar to the toy model of Sec. 2.6.1, and we see a fast response in correlation. Strong promoters can recruit RNAP well even without activators and hence it takes more of them before we see any substantial effect in fold change and correlation.

A necessary condition for the transcriptional correlation effect to be experimentally observable is that TFs stay bound to their binding sites a sufficient amount of time to avoid rapid switching between different promoter states. For example, if mRNA levels are measured at fixed time points (e.g., using FISH), TFs would need to stay bound longer to the operators than the mRNA lifetime. To see this, consider the opposite extreme when the mRNA lifetime is very long (or say infinite), then the observed mRNA expression merely corresponds to an averaged production over every possible promoter state and no effect of transcriptional correlation will be observed. On the other hand, if the mRNA lifetime is much shorter than the TF binding time, the observed mRNAs were likely produced from the same promoter state (or configuration of TFs). This condition is met, e.g., in the case of LacI regulating *lacZ*, where the TF on average stays bound approximately 10 min to the strongest operator in 37 °C [79], whereas the *lacZ* mRNA lifetime is only about 2 min [80]. Even when this condition is not met one might still be able to detect the transcriptional correlation effect by measuring mRNA or protein production during a relatively short time interval from fluorescence time traces, as long as the uncertainty in production (and maturation) time of mRNA or proteins

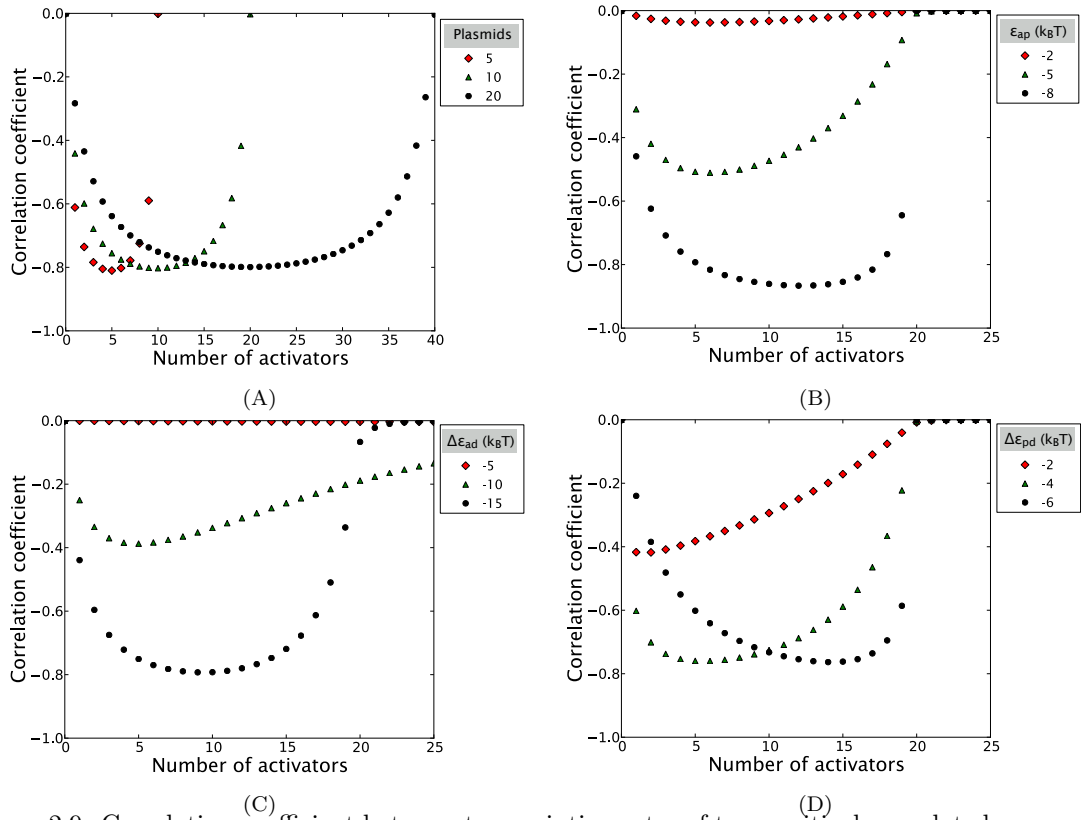


Figure 2.9: Correlation coefficient between transcription rates of two positively regulated genes on a plasmid, as a function of (A) number of plasmids, (B) RNAP-activator interaction energy ϵ_{ap} , (C) activator operator strength $\Delta\epsilon_{ad}$, and (D) promoter strength $\Delta\epsilon_{pd}$. For fixed parameter values we use number of nonspecific sites $N_{NS} = 5 \times 10^6$, 10 plasmids, operator strength $\Delta\epsilon_{ad} = -17.3 k_B T$, promoter strength $\Delta\epsilon_{pd} = -5 k_B T$, and interaction energy between TF and RNAP $\epsilon_{ad} = -7 k_B T$.

is small compared to the TF binding time.

Another condition for the transcriptional correlation effect to be biologically relevant is that extrinsic noise sources, like fluctuations in plasmid or TF copy number, do not have a stronger impact on gene expression than the correlation effect due to TF titration. This matter will be discussed at more length in Sec. 2.7.2.

2.7 Statistically distributed TF and promoter copy numbers

In a cell the number of TFs and promoter copies are, because of inherent stochasticity, not fixed but rather fluctuating according to a statistical distribution. These distributions vary greatly, with examples ranging from the tightly regulated low-copy F-plasmid [81], to the wide distribution of gene copies produced at viral infections [82]. In this section we see how the predicted fold change and transcriptional correlation are affected by fluctuations in promoter copy number and TF copy number. Given the wide range of possible copy number distributions, our goal is not necessarily to model any particular biological system but rather provide a general framework which allows us to compute the fold change and transcriptional correlation for any given such distribution, as well as illustrate this effect on our previously derived results in a few specific cases.

2.7.1 Fold change

In Sec. 2.5 we showed that the fold change of a promoter architecture can depend sensitively on the number of repressors R when this number is comparable to the number of promoter copies N [Fig. 2.3]. We now see how this sensitivity is affected when the number of repressors R or promoter copies N are not fixed but rather fluctuating according to a probability distribution $P(R, N)$. In this case the RNAP occupancy [Eq. (2.32)] to the promoters needs to be replaced by the expectation value with respect to $P(R, N)$,

$$\langle \text{Occupancy} \rangle_{P(R,N)} = \sum_{R,N} P(R, N) \text{Occupancy}(R, N),$$

which we can consequently insert into the definition of fold change,

$$f = \frac{\langle \text{Occupancy} \rangle_{P(R,N)}}{\langle \text{Occupancy}(R=0) \rangle_{P(N)}} \quad (2.51)$$

$$= \frac{\sum_{R,N} P(R, N) \text{Occupancy}(R, N)}{\sum_N P(N) \text{Occupancy}(R=0, N)}. \quad (2.52)$$

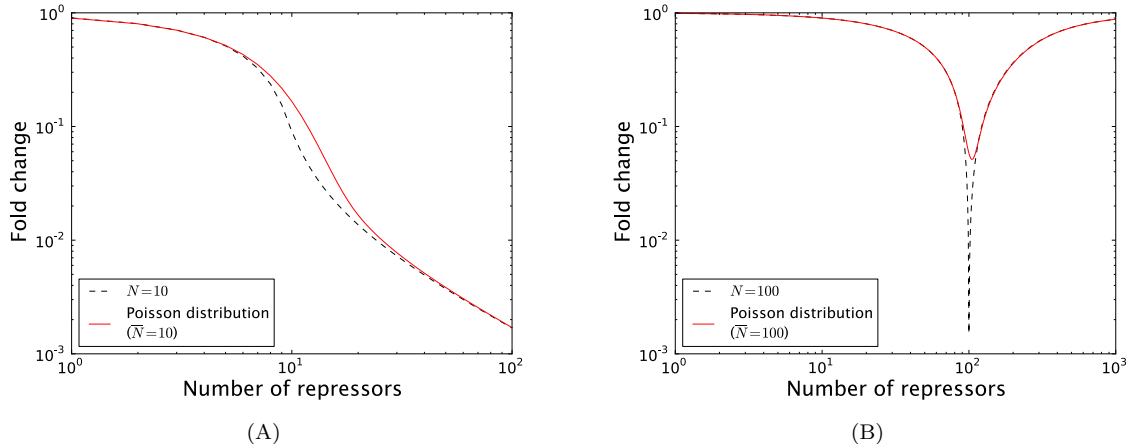


Figure 2.10: (A) Fold change in the simple repression architecture for fixed ($N = 10$) or Poisson distributed (mean $\bar{N} = 10$) promoter copy number. (B) Fold change in the exclusive looping repression architecture for fixed ($N = 100$) or Poisson distributed (mean $\bar{N} = 100$) promoter copy number. For these plots we use operator binding energy $-17.3 k_B T$, looping energy $10 k_B T$, and number of nonspecific sites $N_{NS} = 5 \times 10^6$. The RNAP promoter binding energy is assumed to be weak.

We can simplify the last line [Eq. (2.52)] by noticing that for $R = 0$ the promoters are independent and hence the occupancy must be proportional to N

$$f = \frac{\sum_{R,N} P(R, N) \text{Occupancy}(R, N)}{\langle N \rangle_{P(N)} \text{Occupancy}(R = 0, N = 1)}. \quad (2.53)$$

As an example in Fig. 2.10 we investigate the effect of replacing the promoter copy number with a Poisson distribution in the simple repression (Sec. 2.3.1) and exclusive looping repression (Sec. 2.3.3) architectures. A set of simple repressors will only be effectively repressed when all the promoter copies are inhibited; therefore, the steep decline in fold change around $N \approx R$ will now be shifted up to a higher repressor copy number. For the exclusive looping architecture we note that a trough is still clearly visible but less deep and slightly widened (at half peak depth) compared to the case with fixed promoter copy number. If we in Fig. 2.10 instead were to replace the repressor copy number by a Poisson distribution and keep the promoter copy number fixed, the fold change will look close to identical (result not shown).

2.7.2 Transcriptional correlation

Fluctuations in TF copy number constitute an extrinsic form of noise that affects the transcription rate of all genes regulated by the TF. In this section we show that such fluctuations, when large enough, can hide the effect of transcriptional correlation due to TF titration. To include extrinsic noise into our calculation of transcriptional correlation [Eq. (2.49)] between two genes we compute

the Pearson correlation coefficient using weighted moments,

$$\rho_{i_1 i_2} = \frac{\langle i_1 i_2 \rangle_{P(F)} - \langle i_1 \rangle_{P(F)} \langle i_2 \rangle_{P(F)}}{\sqrt{\langle (i_1 - \langle i_1 \rangle_{P(F)})^2 \rangle_{P(F)} \langle (i_2 - \langle i_2 \rangle_{P(F)})^2 \rangle_{P(F)}}} \quad (2.54)$$

where the expectation value $\langle \cdot \rangle_{P(F)} \equiv \sum_{i=0}^{\infty} P(F=i) \langle \cdot \rangle_{F=i}$ is evaluated over the distribution of TFs. In Fig. 2.11 we use this formula to show how TF fluctuations affect the transcriptional correlation of the particular system of two genes activated by the same TF studied in Sec. 2.6.3. In this case we use, for illustrative purposes, a Gaussian distribution which allows us to vary the distribution width and see what effect it has on transcriptional correlation. Promoter entanglement and extrinsic noise will have opposite effects on transcriptional correlation, and their relative strengths will determine the resulting sign of the correlation coefficient. For $A > 2N_{pl}$ there is no promoter entanglement but a positive correlation due to TF fluctuations remains until the average number of activators is so high that essentially all operators will be occupied.

As the TF copy number increases the TFs will distribute themselves more and more evenly among their targets, and the transcriptional correlation due to TF titration will have a smaller impact on gene expression. We therefore expect transcriptional correlation due to TF titration to be most relevant when the TF copy number is low and extrinsic noise limited. These conditions can be somewhat relaxed due to recent advances in molecular biology; for example cells with TFs labeled by a fluorescent reporter can be sorted by fluorescence to limit the effect of TF fluctuations on transcriptional correlation, hence allowing precision tests of the thermodynamic model.

2.8 Verifying the thermodynamic model of TF titration using Gillespie simulations

To examine the validity of the thermodynamic calculations, we use Gillespie simulations [83] to predict fold change and correlation in transcription rates. Although this is computationally more onerous than the thermodynamic models used throughout this chapter, it has the benefit of simplicity, requiring only knowledge of the gene/TF copy numbers and allowed reactions. Consequently, the intricate details of TF binding combinatorics are given to us “for free.”

To demonstrate the Gillespie algorithm we consider, as an example, free repressors (R) (un)binding to empty gene promoters (G) to form repressor-gene complexes (GR) through the reactions



Here we assume, as in the law of mass action, that the total rate of repressor association is propor-

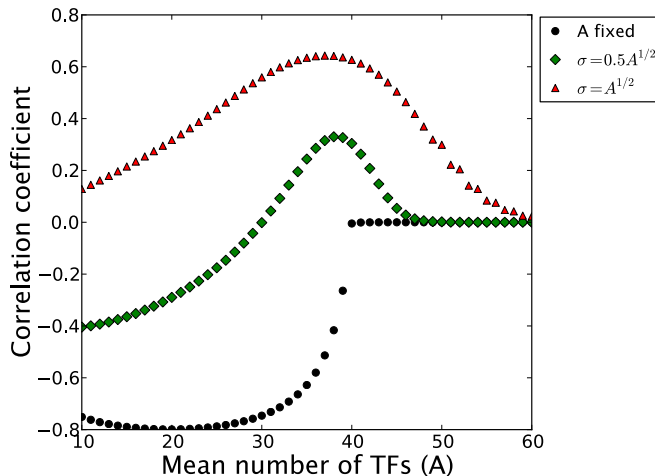


Figure 2.11: Correlation coefficient between transcription rates of two positively regulated genes located on 20 plasmids, as a function of number of TFs. Three different Gaussian TF copy number distributions are considered with standard deviations $\sigma = 0, \frac{1}{2}\sqrt{A}, \sqrt{A}$. TF fluctuations constitute extrinsic noise, affecting expression of both genes, that hides the anti-correlation in transcription rates due to promoter entanglement. As parameter values we choose number of RNAP $P = 1000$, nonspecific sites $N_{NS} = 5 \times 10^6$, 20 plasmids, operator strength $\Delta\varepsilon_{ad} = -17.3 k_B T$, promoter strength $\Delta\varepsilon_{pd} = -5 k_B T$, and interaction energy between TF and RNAP $\varepsilon_{ad} = -7 k_B T$.

tional to both the number of free repressors and empty promoters. The normalized rate parameter k_R^{on} gives number of associations per free repressor, per empty promoter, per time unit. Similarly the normalized disassociation rate parameter k_R^{off} gives number of disassociations per repressor-gene complex, per time unit. These rate parameters will depend on operator strength and number of competing nonspecific binding sites (N_{NS}), but not molecular numbers of the species involved. Notice that since the repressors are assumed to be always bound on DNA we do not consider cell volume, or cytosolic repressor/gene concentration, as parameters of our model. However cell volume will have an indirect effect on above rate parameters through its influence on the nonspecific free energy of binding a repressor to DNA.

In the first step of the Gillespie algorithm we calculate the total accumulated reaction rate, $G \times R \times k_R^{on} + GR \times k_R^{off}$, for both reactions and then draw a random time step at which the next reaction will take place from an exponential distribution, with mean equal to the inverse of this rate. The decision which of the two reactions should be chosen is random but weighted by the accumulated rate for each reaction $G \times R \times k_R^{on}$ vs $GR \times k_R^{off}$. If the repressor binding reaction is chosen we update the corresponding state variables according to $G \rightarrow G - 1$, $R \rightarrow R - 1$, and $GR \rightarrow GR + 1$ (analogously for repressor unbinding). Notice that G, R, GR are *discrete* quantities, not continuous concentrations. By repeating this procedure over and over we acquire time traces for G, R , and GR , which can be used to compute the (time averaged) occupancy of repressors to

genes, fluctuations in G, R and GR etc. To compute fold change, a quantity of central importance throughout this chapter, we use Gillespie's method to find the average number of promoters bound by RNAP, with and without TFs present.

In order to connect the stochastic model with our thermodynamic calculations much effort in this section is dedicated to finding mathematical relations between the stochastic model rate constants and corresponding thermodynamic free energy parameters. This matter is alleviated by the fact that the rate constants are independent of gene copy number, TF copy number, and RNAP copy number, which allows us to determine the rates using stripped-down version of the full promoter architectures.

2.8.1 Simple repression

To determine the rate parameters corresponding to repressor (un)binding in the simple repression architecture (Sec. 2.3.1) we consider a minimal system with a single promoter ($N = 1$) and no RNAP. In this system there are only two states (see Fig. 2.12): repressor bound (state B) and empty promoter (state 0), with dynamics described by the following master equation

$$\frac{dP(B)}{dt} = Rk_R^{on}P(0) - k_R^{off}P(B). \quad (2.56)$$

Here $P(B), P(0)$ correspond to the respective state probabilities [$P(B) + P(0) = 1$]. In equilibrium there is no net probability flux between the two states, or mathematically,

$$\begin{aligned} Rk_R^{on}P(0) &= k_R^{off}P(B) \implies \\ \frac{k_R^{on}}{k_R^{off}} &= \frac{1}{R} \frac{P(B)}{1 - P(B)}. \end{aligned} \quad (2.57)$$

In the thermodynamic model we find the probability $P(B)$ from the partition function computed in Eq. (2.7),

$$P(B) = \frac{\frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}}, \quad (2.58)$$

which gives us a simple expression for the ratio between the rates

$$\frac{k_R^{on}}{k_R^{off}} = \frac{1}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}. \quad (2.59)$$

This argument holds equally well for RNAP and we find

$$\frac{k_{RNAP}^{on}}{k_{RNAP}^{off}} = \frac{1}{N_{NS}} e^{-\beta \Delta \varepsilon_{pd}}. \quad (2.60)$$

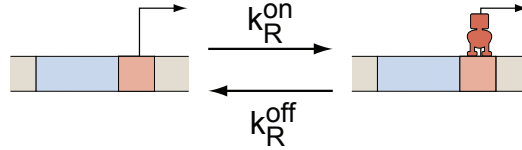


Figure 2.12: States and transition rates in a the simple repression architecture with no RNAP present. The rates correspond to “per molecule” rates, i.e. the total probability flux into the right, repressed state, is given by $Rk_R^{\text{on}}P(0)$.

In equilibrium each reaction will be balanced by its reverse reaction; hence, the final state probabilities can only depend on these ratios, also in the case of multiple gene copies.

We are now ready to apply Gillespie’s method to simulate the full simple repression promoter architecture (see Fig. 2.1(A)), using the following set of reactions:



Here we use the notation: G (empty promoter), R (free repressor), P (free RNAP), GR (promoter bound by repressor), and GP (promoter bound by RNAP). From the resulting simulation time trace we can compute the average number of RNAP-promoter complexes (GP), which we use as a proxy for gene expression. By repeating the simulation with no repressors ($R = 0$) we can then determine the fold change.

Fig. 2.13 shows a precise agreement in fold change between Gillespie simulations and thermodynamic theory, as one would expect.

2.8.2 Repression with looping

In the case of repression by looping [Sec. 2.3.2] we not only need to take the repressor (un)binding rates into account but also the rate of DNA (un)looping between the main and auxiliary binding site. To find the rate constants corresponding to the thermodynamic free energy parameters we consider a simplified system with a single promoter, no RNAP, and only three states: empty promoter (state 0), main operator bound (state M), and looped state (state L). The transitions between these states are illustrated in Fig. 2.14. We consider the state with only the auxiliary operator bound by a repressor to be forbidden. This does not affect the rate constants for repressor (un)binding or DNA loop formation as compared to the full repression with looping architecture, but makes the

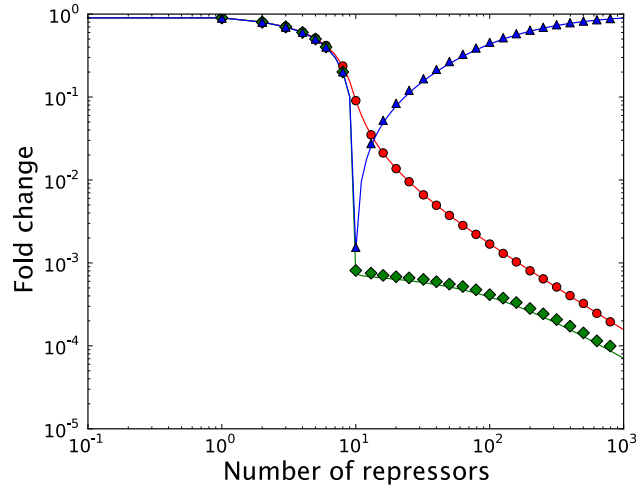


Figure 2.13: Fold change as a function of repressor copy number in the simple repression (\bullet), repression with looping (\blacklozenge), and repression exclusively due to looping (\blacktriangle), promoter architecture, for $N = 10$ promoter copies. Solid lines correspond to thermodynamic model predictions and markers Gillespie simulated data. Here we use the parameters: $k_R^{on} = 1.0$, $k_R^{off} = 0.15$ (simple repression), $k_R^{off} = 0.075$ (looping), $k_{RNAP}^{on} = 3.0 \times 10^{-5}$, $k_{RNAP}^{off} = 1$, $k_{loop} = 1$, and $k_{unloop} = 6.8 \times 10^{-4}$ in arbitrary inverse time units, chosen according to Eqs. (2.59), (2.60), (2.64). The standard deviations, acquired from three separate runs, are smaller than the marker size. Since the rates only enter as *ratios* in the state probabilities we use this freedom to set larger of the two rates to 1. As initial condition we set all promoters to the empty state, $G = 10$, RNAP copy number $P = 1000$, and repressor copy number R indicated by the x axis.

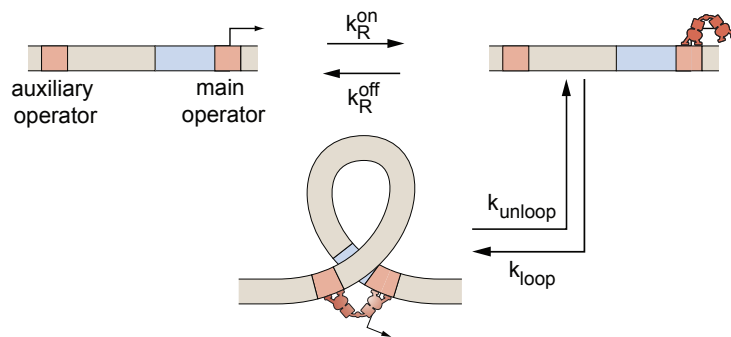


Figure 2.14: States and transition rates in a simplified version of the repression with looping promoter architecture, with no RNAP and where the auxiliary operator is not allowed to be bound individually.

mathematical derivations more straightforward. The detailed balance equations for this system are

$$\begin{aligned}
RP(0)k_R^{on} &= P(M)k_{off}, \\
P(M)k_{loop} &= P(L)k_{unloop}, \\
P(0) + P(M) + P(L) &= 1,
\end{aligned} \tag{2.62}$$

which can be easily solved for $P(0)$, $P(M)$, and $P(L)$.

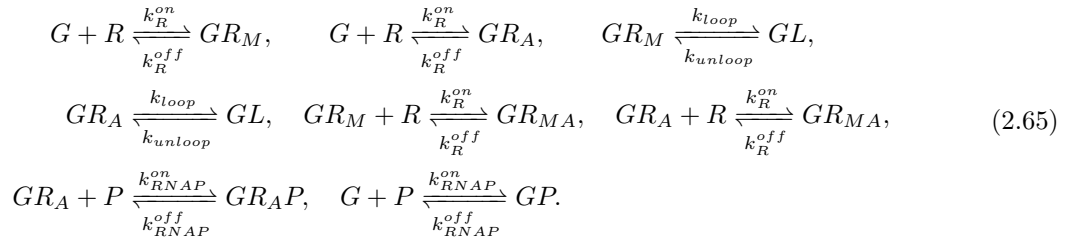
On the other hand the state probabilities for this system can be derived using the statistical mechanical framework, similar to the procedure used in Sec. 2.3.2

$$\begin{aligned}
P(0) &= \frac{1}{1 + \frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}} + \frac{2R}{N_{NS}}e^{-\beta(2\Delta\varepsilon_{rd} + \Delta F_{loop})}}, \\
P(M) &= \frac{\frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}}}{1 + \frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}} + \frac{2R}{N_{NS}}e^{-\beta(2\Delta\varepsilon_{rd} + \Delta F_{loop})}}, \\
P(L) &= \frac{\frac{2R}{N_{NS}}e^{-\beta(2\Delta\varepsilon_{rd} + \Delta F_{loop})}}{1 + \frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}} + \frac{2R}{N_{NS}}e^{-\beta(2\Delta\varepsilon_{rd} + \Delta F_{loop})}}.
\end{aligned} \tag{2.63}$$

Here we assume that the main and auxiliary operators have the same binding energy $\Delta\varepsilon_{rd}$. Equating the state probabilities found in the thermodynamic model with those from Eq. (2.62) allows us to express the (un)binding and (un)looping rates in term of the free energies $\Delta\varepsilon_{rd}$, ΔF_{loop}

$$\begin{aligned}
\frac{k_R^{on}}{k_R^{off}} &= \frac{2}{N_{NS}}e^{-\beta\Delta\varepsilon_{rd}}, \\
\frac{k_{loop}}{k_{unloop}} &= e^{-\beta(\Delta\varepsilon_{rd} + \Delta F_{loop})}.
\end{aligned} \tag{2.64}$$

Notice that, by assuming that the two TF operators have the same binding energy we only need one set of (un)looping rates. We use these rates to apply Gillespie's method on the full repression with looping architecture, where all states in Fig. 2.1 (B) are allowed, using the reaction scheme



where we use the following notation: G (empty promoter), R (free repressor), P (free RNAP), GR_M (main operator bound), GR_A (auxiliary operator bound), GR_{MA} (main and auxiliary operator bound), GL (looped conformation), GR_AP (auxiliary operator bound by TF and promoter by RNAP), GP (promoter bound by RNAP).

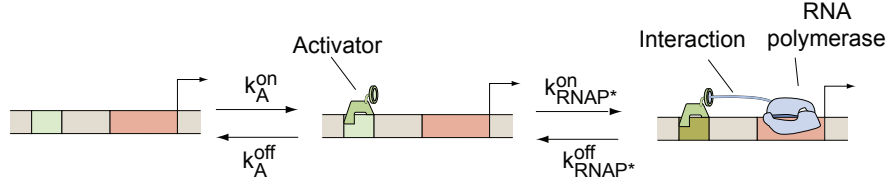
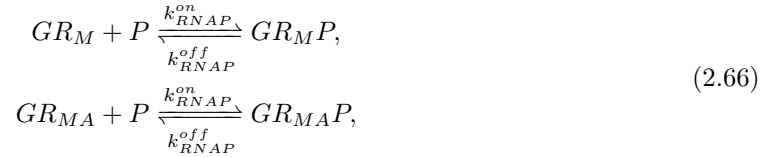


Figure 2.15: Simple activation promoter architecture in the weak promoter approximation, neglecting RNAP binding to the empty promoter.

In Fig. 2.13 we find that our statistical mechanical predictions for fold change are precisely replicated by Gillespie simulations. To achieve the level of precision shown in the figure required around 1h of Gillespie simulations for 30 data points, compared to the analytical framework which allowed us to compute the fold change for 1000 data points in less than 1s.

2.8.3 Repression exclusively due to looping

For repression exclusively due to looping (Sec. 2.3.3) we use the same rate parameters as found in Eq. (2.60)+(2.64), but allow RNAP to bind all states except the looped state (see Fig. 2.1). This means we need to add the following reactions to the scheme in (2.65)



where we use the following notation GR_MP (main operator bound by TF and promoter by RNAP) and GR_{MAP} (main plus auxiliary bound by TF and promoter by RNAP).

In Fig. 2.13 we compare the fold change predicted by the thermodynamic model with Gillespie simulations, and again find them to be in precise agreement.

2.8.4 Transcriptional correlation

In Sec. 2.6 it was shown that under certain conditions the transcription rates of two genes can be correlated and we used the simple activation promoter architecture as a case study. To find the rate constants that correspond to the thermodynamic model free energy parameters for this promoter architecture we solve the detailed balance equations resulting from Fig. 2.15

$$\begin{aligned}
 AP(0)k_A^{on} &= P(A)k_{off}, \\
 P(A)Pk_{RNAP^*}^{on} &= P(AP)k_{RNAP^*}^{off}, \\
 P(0) + P(A) + P(AP) &= 1,
 \end{aligned} \tag{2.67}$$

where we use the notation: empty promoter (state 0), activator bound to promoter (state A), activator and RNAP bound to promoter (state AP), and $k_{RNAP^*}^{off}$, $k_{RNAP^*}^{on}$ refer to RNAP (un)binding rate when the promoter is *already* bound by an activator. For mathematical convenience we invoke the weak promoter approximation and neglect the state with RNAP bound to an empty promoter.

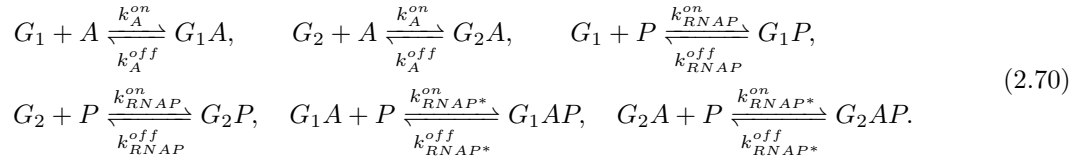
In the thermodynamic model we can write down the corresponding state probabilities (see notation Fig. 2.7)

$$\begin{aligned}
P(0) &= \frac{1}{1 + \frac{A}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}} + \frac{AP}{N_{NS}^2} e^{-\beta(\Delta \varepsilon_{ad} + \varepsilon_{ap})}}, \\
P(A) &= \frac{\frac{A}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}}}{1 + \frac{A}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}} + \frac{AP}{N_{NS}^2} e^{-\beta(\Delta \varepsilon_{ad} + \varepsilon_{ap})}}, \\
P(AP) &= \frac{\frac{AP}{N_{NS}^2} e^{-\beta(\Delta \varepsilon_{ad} + \varepsilon_{ap})}}{1 + \frac{A}{N_{NS}} e^{-\beta \Delta \varepsilon_{rd}} + \frac{AP}{N_{NS}^2} e^{-\beta(\Delta \varepsilon_{ad} + \varepsilon_{ap})}}.
\end{aligned} \tag{2.68}$$

Equating the state probabilities in Eqs. (2.67) and (2.68) allows us to express the TF and RNAP (un)binding rate in terms of the thermodynamic model parameters

$$\begin{aligned}
\frac{k_A^{on}}{k_A^{off}} &= \frac{1}{N_{NS}} e^{-\beta \Delta \varepsilon_{ad}}, \\
\frac{k_{RNAP^*}^{on}}{k_{RNAP^*}^{off}} &= \frac{1}{N_{NS}} e^{-\beta(\Delta \varepsilon_{pd} + \varepsilon_{ap})} = \frac{k_{RNAP}^{on}}{k_{RNAP}^{off}} e^{-\beta \varepsilon_{ap}}.
\end{aligned} \tag{2.69}$$

Using these rates we can apply Gillespie's method to the system of two genes considered in Sec. 2.6.3, described by the reaction scheme



At each time step of the simulation the number of promoters of each type bound by RNAP is recorded, and using the time traces we can compute the correlation coefficient between the two quantities. Fig. 2.16 again shows a precise agreement between our thermodynamic model and Gillespie simulations.

2.9 Conclusion

In this chapter we have developed a general framework based on statistical mechanics to predict gene expression for systems with multiple genes or gene copies regulated by the same TFs. These kinds of systems arise in a multitude of biologically relevant circumstances. In particular, we have

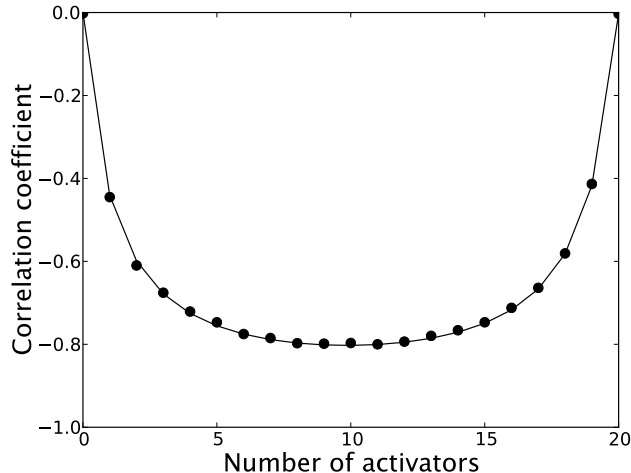


Figure 2.16: Correlation coefficient between transcription rates of two positively regulated genes on a plasmid (copy number $N = 10$) as a function of activator copy number. The solid line corresponds to thermodynamic model prediction, and dots corresponds to Gillespie simulated data. Here we use the parameters: $k_A^{on} = 1.0$, $k_A^{off} = 0.15$, $k_{RNAP}^{on} = 3.0 \times 10^{-5}$, $k_{RNAP}^{off} = 1$, $k_{RNAP^*}^{on} = 0.033$ and $k_{RNAP^*}^{off} = 1$ in arbitrary inverse time units, chosen according to Eq. (2.59), (2.60), (2.69). The standard deviations, acquired from three separate runs, are smaller than the marker size. Since the rates only enter as ratios, we use this freedom to set the larger of the two rates to 1. As initial condition we set all promoters to the empty state, $G_1 = G_2 = 10$, RNAP copy number $P = 1000$, and activator copy number A indicated by the x axis.

shown that when the number of TF binding sites is large enough to titrate the TFs, the predicted gene expression depends in a highly nontrivial way on the relative abundance of promoter and TF copy numbers. New data [4] on protein copy numbers in *E. coli* indicate that such titration might happen more often than previously thought. We have also quantitatively linked the effect of TF titration to correlation between transcription rates of different genes.

An advantage with the presented model is that quantities of interest, e.g. fold change or correlation in transcription rates, can be expressed analytically for a set of promoters explicitly in terms of the individual promoter architectures. This allows us to vary model parameters and TF copy number without the need to run thousands of time-consuming Gillespie simulations.

Recent advances in the field of molecular biology have made it possible to accurately measure and tune protein copy numbers in a cell [4, 62, 84, 18], which provides an excellent opportunity to test the predictions presented here experimentally. This will indeed be the topic of Chapter 3.

Appendix

2.A Partition function for a set of promoters regulated by multiple low-copy TFs

One can easily show that the partition function derived in Eq. (2.20) for a set of promoters regulated by one TF type is valid also when the promoters are regulated by additional TFs, as long as these extra factors are not subject to titration effects and can be summed out together with RNAP in Eq. (2.21). However, in the case of regulation by multiple low-copy TFs the derivation needs to be generalized. To do this let us denote the different TFs by F_1, \dots, F_m and f_{n_j} the number of TFs of type $j \in \{1, \dots, m\}$ bound to promoter $n \in \{1, \dots, N\}$. By analogy to the treatment in Sec. 2.4.1 the total partition function is given by

$$Z^{tot} = \sum_{\substack{f_{n_j}, \forall n, j \\ \sum_n f_{n_j} \leq F_j, \forall j}} \left(\prod_{j=1}^m \frac{F_j!}{N_{NS}^{\sum_n f_{n_j}} (F_j - \sum_n f_{n_j})!} \right) \prod_{n=1}^N Z_{f_{n_1}, \dots, f_{n_m}}^{(n)}, \quad (2.71)$$

where $Z_{g_1, \dots, g_m}^{(n)}$ corresponds to states for promoter n occupied by g_1 number of TFs of type F_1 , g_2 TFs of type F_2 etc. Analogously to Eq. (2.22) the single promoter partition functions with multiple TF types are given by

$$Z^{(n)} = \sum_{g_1, \dots, g_m} \left(\prod_{j=1}^m \frac{F_j!}{N_{NS}^{g_j} (F_j - g_j)!} \right) Z_{g_1, \dots, g_m}^{(n)}. \quad (2.72)$$

For the case when all promoter copies are identical we can also generalize the computationally

more efficient Eq. (2.25) to multiple low-copy TF types F_1, \dots, F_m

$$Z^{tot} = \sum_{\substack{k_{i_1, \dots, i_m}, \forall i_1, \dots, i_m \\ \sum_{i_1, \dots, i_m} k_{i_1, \dots, i_m} = N \\ \sum_{i_1, \dots, i_m} i_j k_{i_1, \dots, i_m} \leq F_j, \forall j}} \binom{N}{\{k_{i_1, \dots, i_m}\}} \left(\prod_{i_1, \dots, i_m} Z_{i_1, \dots, i_m}^{k_{i_1, \dots, i_m}} \right) \quad (2.73)$$

$$\times \prod_{j=1}^m \frac{F_j}{N_{NS}^{\sum_{i_1, \dots, i_m} i_j k_{i_1, \dots, i_m}} (F_j - \sum_{i_1, \dots, i_m} i_j k_{i_1, \dots, i_m})!}.$$

Here k_{i_1, \dots, i_m} is the number of promoters which have i_1 TF of type F_1 bound, i_2 TF of type F_2 bound, etc., Z_{i_1, \dots, i_m} corresponds to states with i_j TFs of type F_j bound, and the notation $\binom{N}{\{k_{i_1, \dots, i_m}\}}$ refers to the multinomial coefficient $N! \prod_{i_1, \dots, i_m} \frac{1}{k_{i_1, \dots, i_m}!}$.

2.B Number of binding sites vs. TF copy number in *E. coli*

For the specific case of *E. coli*, hundreds of TFs and their corresponding vast array of binding sites have been identified [21]. As a result, one can make an educated guess about regulatory architectures where the TF titration effect might play a role by looking for cases where the number of binding sites (N) approaches the number of TF molecules (F) per cell. An attempt to amass such data is shown in Fig. 5.6. The majority of genes belong to a regime where we do not expect strong titration of TFs, however, with a handful of exceptions, especially in the borderline regime $F \approx 2N$ where TFs binding as dimers could experience depletion. As new binding sites are discovered more TFs might fall into this category.

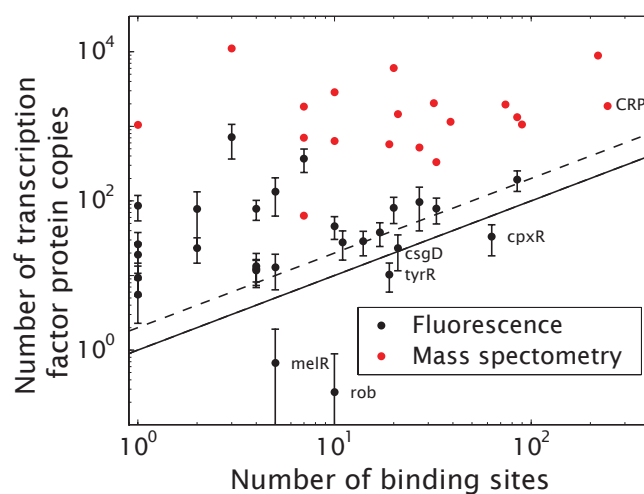


Figure 2.17: Transcription factor (TF) copy number vs. number of binding sites, using two different protein censuses of *E. coli*. Protein copy numbers were determined using mass spectrometry [62] and fluorescence [4]. The number of binding sites was obtained from RegulonDB [21]. The solid line marks the boundary between depletable TFs (more binding sites than TF copies) and nondepletable (more TF copies than binding sites). For TFs forming dimers (e.g. CRP, Fis, GalR), this boundary is replaced by the dashed line. Due to incomplete knowledge about the *E. coli* regulatory system we expect the number of binding sites to be underestimated, and hence more TFs might belong to the depletable category than shown in the figure.

References

- [1] Buchler, N. & Louis, M. Molecular titration and ultrasensitivity in regulatory networks. *J Mol Biol* **384**, 1106–19 (2008).
- [2] Burger, A., Walczak, A. M. & Wolynes, P. G. Abduction and asylum in the lives of transcription factors. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 4016–4021 (2010).
- [3] Lee, T. H. & Maheshri, N. A regulatory role for repeated decoy transcription factor binding sites in target gene expression. *Mol. Syst. Biol.* **8**, 576 (2012).
- [4] Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
- [5] Zhong, C. *et al.* Determination of plasmid copy number reveals the total plasmid DNA amount is greater than the chromosomal DNA amount in *Bacillus thuringiensis* YBT-1520. *PLoS ONE* **6**, e16025 (2011).
- [6] Luria, S. E. & Dulbecco, R. Genetic recombinations leading to production of active bacteriophage from ultraviolet inactivated bacteriophage particles. *Genetics* **34**, 93–125 (1949).
- [7] Hanada, K. *et al.* Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **28**, 377–382 (2011).
- [8] Wang, S., Liu, N., Peng, K. & Zhang, Q. The distribution and copy number of copia-like retrotransposons in rice (*Oryza sativa* L.) and their implications in the organization and evolution of the rice genome. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6824–6828 (1999).
- [9] Navarro-Quezada, A. & Schoen, D. J. Sequence evolution and copy number of Ty1-copia retrotransposons in diverse plant genomes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 268–273 (2002).
- [10] Kentner, E. K., Arnold, M. L. & Wessler, S. R. Characterization of high-copy-number retrotransposons from the large genomes of the Louisiana iris species and their use as molecular markers. *Genetics* **164**, 685–697 (2003).

- [11] Bremer, H. & Dennis, P. P. Modulation of chemical composition and other parameters of the cell by growth rate. In al., N. F. e. (ed.) *Escherichia coli and Salmonella Cellular and Molecular Biology*, 1553–1569 (ASM Press, Washington DC, 1996).
- [12] Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
- [13] Cappuzzo, F. *et al.* Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. *J. Natl. Cancer Inst.* **97**, 643–655 (2005).
- [14] Cook, E. H. & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919–923 (2008).
- [15] Oehler, S., Amouyal, M., Kolkhof, P., von Wilcken-Bergmann, B. & Müller-Hill, B. Quality and position of the three *lac* operators of *E. coli* define efficiency of repression. *EMBO J* **13**, 3348–55 (1994).
- [16] Müller, J., Oehler, S. & Müller-Hill, B. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol* **257**, 21–9 (1996).
- [17] Vilar, J. M. & Leibler, S. DNA looping and physical constraints on transcription regulation. *J Mol Biol* **331**, 981–9 (2003).
- [18] Garcia, H. G. & Phillips, R. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A* **108**, 12173–8 (2011).
- [19] Garcia, H. G. *et al.* Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Rep* **2**, 150–161 (2012).
- [20] Boedicker, J. Q., Garcia, H. G. & Phillips, R. Theoretical and experimental dissection of DNA loop-mediated repression. *Phys. Rev. Lett.* **110**, 018101 (2013).
- [21] Gama-Castro, S. *et al.* RegulonDB version 7.0: Transcriptional regulation of *Escherichia coli* *K-12* integrated within genetic sensory response units (Sensor Units). *Nucleic Acids Res.* **39**, 98–105 (2011).
- [22] Cournac, A. & Plumbridge, J. DNA looping in prokaryotes: experimental and theoretical approaches. *J. Bacteriol.* **195**, 1109–1119 (2013).
- [23] Weickert, M. J. & Adhya, S. The galactose regulon of *Escherichia coli*. *Mol Microbiol* **10**, 245–51 (1993).
- [24] Ackers, G. K., Johnson, A. D. & Shea, M. A. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A* **79**, 1129–33 (1982).

- [25] Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A* **100**, 5136–41 (2003).
- [26] Bintu, L. *et al.* Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev* **15**, 116–24 (2005).
- [27] Bintu, L. *et al.* Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev* **15**, 125–35 (2005).
- [28] Elf, J., Li, G. W. & Xie, X. S. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316**, 1191–4 (2007).
- [29] Winter, R. B., Berg, O. G. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli lac* repressor–operator interaction: Kinetic measurements and conclusions. *Biochemistry* **20**, 6961–77 (1981).
- [30] Müller-Hill, B. *The lac Operon: a short history of a genetic paradigm* (Walter de Gruyter, Berlin; New York, 1996).
- [31] Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–40 (2008).
- [32] Raveh-Sadka, T., Levo, M. & Segal, E. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* **19**, 1480–1496 (2009).
- [33] Gertz, J., Siggia, E. D. & Cohen, B. A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**, 215–8 (2009).
- [34] He, X., Samee, M. A., Blatti, C. & Sinha, S. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol* **6** (2010).
- [35] Kinney, J. B., Murugan, A., C. G. Callan, J. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**, 9158–63 (2010).
- [36] Fakhouri, W. D. *et al.* Deciphering a transcriptional regulatory code: Modeling short-range repression in the *Drosophila* embryo. *Mol Syst Biol* **6**, 341 (2010).
- [37] Sherman, M. S. & Cohen, B. A. Thermodynamic State Ensemble Models of cis-Regulation. *PLoS Comput Biol* **8**, e1002407 (2012).
- [38] Kuhlman, T., Zhang, Z., Saier, J., M. H. & Hwa, T. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc Natl Acad Sci U S A* **104**, 6043–8 (2007).

- [39] Meijnsing, S. H. *et al.* DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* **324**, 407–10 (2009).
- [40] Wall, M. E., Markowitz, D. A., Rosner, J. L. & Martin, R. G. Model of transcriptional activation by MarA in *Escherichia coli*. *PLoS Comput. Biol.* **5**, e1000614 (2009).
- [41] Voss, T. C. *et al.* Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell* **146**, 544–54 (2011).
- [42] Kuhlman, T. E. & Cox, E. C. Gene location and DNA density determine transcription factor distributions in *Escherichia coli*. *Mol Syst Biol* **8**, 610 (2012).
- [43] Santillán, M. On the Use of the Hill Functions in Mathematical Models of Gene Regulatory Networks 2 . Derivation of the Hill function. *Mathematical Modelling of Natural Phenomena* **3**, 85–97 (2008).
- [44] Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–42 (2000).
- [45] Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–8 (2000).
- [46] Cherry, J. L. & Adler, F. R. How to make a biological switch. *J Theor Biol* **203**, 117–33 (2000).
- [47] Bolouri, H. & Davidson, E. H. Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. *Proc Natl Acad Sci U S A* **100**, 9371–6 (2003).
- [48] Suel, G. M., Garcia-Ojalvo, J., Liberman, L. M. & Elowitz, M. B. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* **440**, 545–50 (2006).
- [49] Alon, U. *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).
- [50] Kim, H. D. & O’Shea, E. K. A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* **15**, 1192–1198 (2008).
- [51] Tsai, T. Y. *et al.* Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science* **321**, 126–9 (2008).
- [52] Riley, T., Sontag, E., Chen, P. & Levine, A. Transcriptional control of human p53-regulated genes. *Nat Rev Mol Cell Biol* **9**, 402–12 (2008).

- [53] Cagatay, T., Turcotte, M., Elowitz, M. B., Garcia-Ojalvo, J. & Suel, G. M. Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell* **139**, 512–22 (2009).
- [54] Peter, I. S. & Davidson, E. H. Modularity and design principles in the sea urchin embryo gene regulatory network. *FEBS Lett* **583**, 3948–58 (2009).
- [55] Sprinzak, D. *et al.* Cis-interactions between Notch and Delta generate mutually exclusive signalling states. *Nature* **465**, 86–90 (2010).
- [56] Sprinzak, D., Lakhanpal, A., Lebon, L., Garcia-Ojalvo, J. & Elowitz, M. B. Mutual inactivation of notch receptors and ligands facilitates developmental patterning. *PLoS Comput Biol* **7**, e1002069 (2011).
- [57] Balaskas, N. *et al.* Gene regulatory logic for reading the Sonic Hedgehog signaling gradient in the vertebrate neural tube. *Cell* **148**, 273–84 (2012).
- [58] Warmflash, A. *et al.* Dynamics of TGF-beta signaling reveal adaptive and pulsatile behaviors reflected in the nuclear localization of transcription factor Smad4. *Proc Natl Acad Sci U S A* **109**, E1947–56 (2012).
- [59] Jishage, M. & Ishihama, A. Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of sigma 70 and sigma 38. *J Bacteriol* **177**, 6832–5 (1995).
- [60] Grigorova, I. L., Phleger, N. J., Mutalik, V. K. & Gross, C. A. Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. *Proc Natl Acad Sci U S A* **103**, 5332–7 (2006).
- [61] Klumpp, S. & Hwa, T. 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).
- [62] Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**, 117–24 (2007).
- [63] Kao-Huang, Y. *et al.* 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**, 4228–32 (1977).
- [64] Runzi, W. & Matzura, H. *In vivo* distribution of ribonucleic acid polymerase between cytoplasm and nucleoid in *Escherichia coli*. *J Bacteriol* **125**, 1237–9 (1976).

- [65] Malan, T. P., Kolb, A., Buc, H. & McClure, W. R. Mechanism of CRP-cAMP activation of lac operon transcription initiation activation of the P1 promoter. *J Mol Biol* **180**, 881–909 (1984).
- [66] McClure, W. R. Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.* **54**, 171–204 (1985).
- [67] Reppas, N. B., Wade, J. T., Church, G. M. & Struhl, K. The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol. Cell* **24**, 747–757 (2006).
- [68] Hsu, L. M. Promoter clearance and escape in prokaryotes. *Biochim. Biophys. Acta* **1577**, 191–207 (2002).
- [69] Salgado, H. *et al.* RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* **41**, D203–213 (2013).
- [70] Garcia, H. G. *et al.* Biological consequences of tightly bent DNA: The other life of a macromolecular celebrity. *Biopolymers* **85**, 115–30 (2007).
- [71] Schleif, R. AraC protein: A love-hate relationship. *Bioessays* **25**, 274–82 (2003).
- [72] Tricomi, F. Sulle funzioni ipergeometriche confluenti. *Annali di Matematica Pura ed Applicata* **26**, 141–175 (1947).
- [73] <http://functions.wolfram.com/07.33.03.0040.01>.
- [74] <http://functions.wolfram.com/07.33.20.0005.01>.
- [75] Brewster, R. C., Jones, D. L. & Phillips, R. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput. Biol.* **8**, e1002811 (2012).
- [76] Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12795–12800 (2002).
- [77] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).
- [78] Dunlop, M. J., Cox, R. S., Levine, J. H., Murray, R. M. & Elowitz, M. B. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics* **40**, 1493–1498 (2008).
- [79] Hammar, P. *lac of Time : Transcription Factor Kinetics in Living Cells*. Ph.D. thesis, Uppsala University, Computational and Systems Biology (2013).

- [80] Liang, S. T., Dennis, P. P. & Bremer, H. Expression of lacZ from the promoter of the *Escherichia coli* *spc* operon cloned into vectors carrying the W205 *trp-lac* fusion. *J. Bacteriol.* **180**, 6090–6100 (1998).
- [81] Jones, K. L. & Keasling, J. D. Construction and characterization of F plasmid-based expression vectors. *Biotechnol. Bioeng.* **59**, 659–665 (1998).
- [82] Delbruck, M. The burst size distribution in the growth of bacterial viruses (bacteriophages). *Journal of Bacteriology* **50**, 131–135 (1945).
- [83] Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* **81**, 2340–2361 (1977).
- [84] Martin, R. G., Bartlett, E. S., Rosner, J. L. & Wall, M. E. Activation of the *Escherichia coli* *marA/soxS/rob* regulon in response to transcriptional activator concentration. *J Mol Biol* **380**, 278–84 (2008).

Chapter 3

The transcription factor titration effect dictates level of gene expression

This chapter reproduces the submitted version of [1].

3.1 Introduction

Models of transcription are often built around a picture of RNA polymerase (RNAP) and transcription factors (TFs) acting on a single copy of a promoter. However, most TFs are shared between multiple genes with varying binding affinities. Beyond that, genes often exist at high copy number; in multiple identical copies on the chromosome or on plasmids or viral vectors with copy numbers in the hundreds. Using a thermodynamic model, we characterize the interplay between TF copy number and the demand for that TF. We demonstrate the parameter-free predictive power of this model as a function of the copy number of the TF and the number and affinities of the available specific binding sites; such predictive control is important for the understanding of transcription and the desire to quantitatively design the output of genetic circuits. Finally, we use these experiments to dynamically measure plasmid copy number through the cell cycle.

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 [2, 3, 4, 5, 6]. With increasing regularity, these systems are examined in tandem using both theoretical models with precise “governing equations” and precision measurements whose ambition is to explicitly test the validity of these models. 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 even be performed at the single-cell level [7, 8, 9, 10, 11, 12, 13, 14].

Typically, such models rely on the assumption that the number of TFs is in excess with respect to the number of its binding sites in the cell. There are many situations where this assumption might break down, such as those involving highly replicated viral DNA [15], genes expressed on plasmids [16], genes existing in multiple identical copies on the chromosome [17, 18, 19, 20, 21] or even just genes controlled by “overworked” TFs with many available target genes [22]. Additionally, this interplay between the number of TFs and the number of its binding sites provides yet another tuning parameter with which to test and refine theoretical models of transcriptional regulation as well as precisely control the output of synthetic genetic circuits [23, 24, 25, 26]. In fact, it is common to explore regulatory architecture in the context of multicopy plasmids [16, 27, 28, 12]. As a result, precise knowledge of the role of plasmid copy number on the output levels of gene expression is required. This interdependence of a given gene’s input-output relation with the external environment in which it exists has been termed “retroactivity” [29, 30] and is treated in analogy to impedance in electrical circuits. Some studies have explored this interplay, typically in the context of activation with binding competition stemming from molecular depletants [31] or binding arrays [32].

Here we dissect the interplay between TF copy number and the number of its target binding sites using the simple repression regulatory architecture. Simple repression is a ubiquitous motif in *E. coli* [33, 34] which consists of a promoter with a single proximal repressor binding site such that when a repressor is bound no transcription ensues [35, 36, 37]. In particular, we focus on simple repression by Lac repressor (LacI), which has been extensively studied in the context of theoretical models of transcriptional regulation [38, 8, 10, 39, 11, 13, 14]. Using video fluorescence microscopy, we simultaneously measure both the absolute number of repressors and the rate of expression of a reporter fluorescent protein in single cells as they progress through the cell cycle. This method is used to examine several cases of simple repression in which the TF binding sites are placed in multiple locations, shown schematically in Fig. 3.1. In particular, these include transcription from a plasmid at several distinct copy numbers [Fig. 3.1B], transcription from multiple identical copies integrated in the chromosome [Fig. 3.1C], and transcription from a single chromosomal copy that competes for the repressor with plasmids also containing a specific binding target [Fig. 3.1D].

One major outcome of this study is that, when a TF is shared among many binding sites, either due to multiple identical copies of a gene regulated by that TF or unrelated genes which also independently bind the TF, the correlation in occupancy between the binding sites will lead to a complex dosage response to that TF [32, 40]. At low copy numbers (relative to the number of binding sites), this essentially buffers the transcriptional level to the presence of the TF, and at high copy numbers the response of the fold-change is similar to that seen for a single isolated copy of the gene with no binding competition. The sharpness of the transition between these regimes is predicted to depend explicitly on the relative strength of the specific binding site on the gene of interest compared to the specific sites with which it competes. However, we find that the width of the

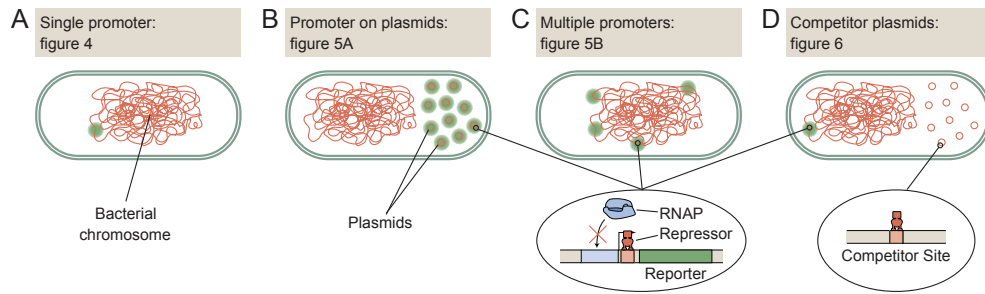


Figure 3.1: **Examples examined in this study of transcriptional regulation with competition for the TF.**

(A) Single chromosomal copy of the gene of interest. (B,C) Competition from multiple, identical genes in the simple repression regulatory architecture when the promoters are (B) placed on a high copy number plasmid or (C) integrated in multiple chromosomal locations. (D) The chromosomal reporter construct competes with competitor plasmids which have binding sites for the repressor, but do not code for the reporter gene. In this particular case the competitor binding sites can have a different affinity than the regulated gene.

plasmid distribution inside the population of cells can also play a role in flattening the transition, and the distribution must be taken into account to accurately predict gene expression when the plasmid distribution itself becomes wider than the transition region in the fold-change curve, which tends to occur for stronger binding operators. Building on the success of the predictive model, we then exploit it as a tool for measuring plasmid copy number throughout the course of the cell cycle. The average number of plasmids per cell increases as the cell cycle progresses with a time-averaged mean value that is consistent with our independent bulk qPCR measurements of the mean copy number.

3.2 Results

3.2.1 Thermodynamic model

Our results are based upon time-lapse fluorescence microscopy [Fig. 3.2] in which we measure the level of gene expression by looking at the rate of production of a fluorescent reporter (i.e. dP/dt , where P is the fluorescent protein number per cell). Specifically, we measure the fold-change given by

$$\text{fold-change} = \frac{\frac{dP}{dt}(R \neq 0)}{\frac{dP}{dt}(R = 0)}, \quad (3.1)$$

which typically measure the steady-state level of the gene product in cell populations. However, we can demonstrate the relationship between the fold-change data from steady-state measurements, where expression is quantified as levels of fluorescence reporter, P , and that obtained using video

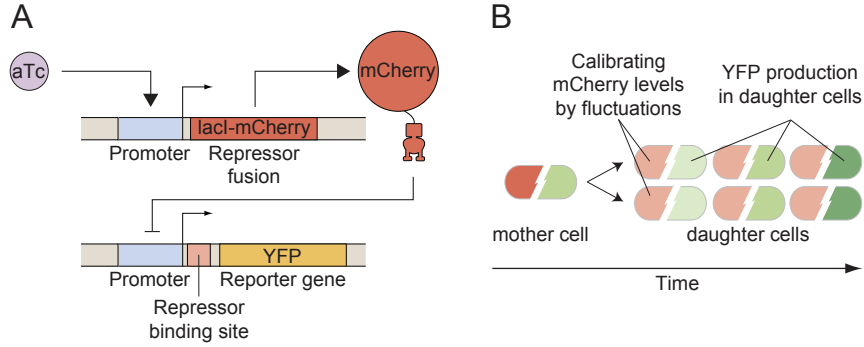


Figure 3.2: **Experimental methods for the single-cell dissection of regulatory architectures.**

(A) Genetic circuit employed in this work. The expression of the LacI-mCherry fusion is induced by the small molecule aTc. The repressor acts on a promoter expressing a YFP reporter gene. (B) Individual cells are observed through a division event. The fluctuations in the partitioning of the LacI-mCherry between the daughters is used to calibrate the signal such that the mCherry fluorescence measurement in each cell can be expressed as an absolute number of repressor molecules. In addition, the rate of YFP production is measured over the cell cycle.

microscopy by observing the rate of production of a fluorescent reporter, dP/dt . In the limit that degradation of the measured product is slow, the equivalence of these methods can be derived (see Appendix section “Equivalence of fold-change in steady-state measurements and video microscopy”),

$$\text{fold-change} = \underbrace{\frac{\frac{dP}{dt}(R \neq 0)}{\frac{dP}{dt}(R = 0)}}_{\text{video microscopy}} = \underbrace{\frac{P(R \neq 0)}{P(R = 0)}}_{\text{steady-state microscopy}} = \underbrace{\frac{p_{\text{bound}}(R \neq 0)}{p_{\text{bound}}(R = 0)}}_{\text{theory}}, \quad (3.2)$$

suggesting that a direct comparison between the bulk measurements and those presented here is admitted, as is the comparison to thermodynamic models.

The basic idea of the thermodynamic model of transcriptional regulation is to enumerate the possible configurations of the molecular players among the available specific and nonspecific binding sites and calculate the probability of finding RNAP bound at the promoter of interest. These models predict the fold-change in gene expression defined as the ratio of the level of gene expression in the presence of TF to the level of expression in its absence. In particular, the fold-change for simple repression in the case where the gene and corresponding TF specifically bind only at the reporter gene [Fig. 3.1A] is [41]

$$\text{fold-change} = \frac{1}{1 + \frac{R}{N_{\text{NS}}} e^{-\Delta\varepsilon/k_B T}}, \quad (3.3)$$

where R is the number of repressors present in the cell, N_{NS} is the size of the nonspecific binding reservoir (which we take here to be the whole *E. coli* chromosome such that $N_{\text{NS}} = 5 \times 10^6$) and $\Delta\varepsilon$

is the binding energy of repressor to its operator. In reference [40], this model has been extended to the case of simple repression from multiple identical copies of the gene, schematically shown in Fig. 3.1B and C. In this case, the fold-change is predicted to have the form,

$$\text{fold-change} = \frac{\sum_{m=0}^{\min(N,R)} \frac{R!}{(N_{NS})^m (R-m)!} \binom{N}{m} e^{-\beta m \Delta \epsilon} (N-m)}{N \sum_{m=0}^{\min(N,R)} \frac{R!}{(N_{NS})^m (R-m)!} \binom{N}{m} e^{-\beta m \Delta \epsilon}}, \quad (3.4)$$

where the only new parameter is N , the copy number of the gene. Finally, the model predicts the regulatory outcome of a single gene copy regulated by simple repression with a binding affinity $\Delta \epsilon$ in the presence of competing binding sites with a distinct affinity $\Delta \epsilon_c$ [Fig. 3.1D]. In this more complex case, the fold-change is given by

$$\text{fold-change} = \frac{Z_u}{Z_b + Z_u} \quad (3.5)$$

where Z_b and Z_u are the partition functions for the case where the repressor is bound or unbound to the chromosomal promoter, given by,

$$Z_u = \sum_{k=0}^{\min(N_c, R)} \frac{R!}{N_{NS}^k (R-k)!} \binom{N_c}{k} e^{-\beta k \Delta \epsilon_c}, \quad (3.6)$$

$$Z_b = \sum_{k=0}^{\min(N_c, R-1)} \frac{R!}{N_{NS}^k (R-k-1)!} \binom{N_c}{k} e^{-\beta(k \Delta \epsilon_c + \Delta \epsilon)}, \quad (3.7)$$

where N_c is the copy number of the plasmid containing the competing binding site and no reporter gene. The extension of this model to N copies of the gene with N_c competitors is detailed in the Appendix section ‘‘Accounting for chromosome replication in competitor theory’’. One feature of the theoretical predictions in Eqs. (3.4) and (3.5), is that in the limit that $R \gg N$ (Eq. (3.4)) or $R \gg N_c$ (Eq. (3.5)) these expressions immediately simplify to Eq. (3.3) (see Appendix section ‘‘Thermodynamic model in the limit $R \gg N$ ’’ for details), meaning that the multiple promoters are independent in this limit. Between all of these situations there are relatively few parameters: the number of TFs (R), the size of the nonspecific reservoir (N_{NS}), the strength of binding sites ($\Delta \epsilon$, $\Delta \epsilon_c$), and the copy number of the gene (N) or of the competing binding site plasmid (N_c). Interestingly, many of the same parameters arise within each of the different scenarios we are considering and a critical test of the theoretical understanding is the self-consistency of those results. Once these quantities are determined, the theory generates falsifiable predictions without any free parameters for all remaining experiments. In the following paragraphs we discuss how these parameters were determined from independent measurements with the ultimate objective of performing a stringent test of the thermodynamic models, in general, and of the impact of gene copy number on regulation, in particular.

3.2.2 Fluorescent measurements of gene expression and absolute TF copy number

We consider a number of distinct regulatory landscapes [Fig. 3.1], all of which involve a rich interplay between the gene copy number and the copy number of the transcription factor controlling that gene. To test the expressions for fold-change given in Eqs. (3.3) - (3.5), we need to simultaneously measure both the rate of gene expression and the absolute number of TFs. To that end, as shown in Fig. 3.2B, our cells harbor two important fluorescent proteins, one to mark the TF and one to mark the gene product.

We use the partitioning statistics of the repressor TF, an mCherry-LacI fusion, during cell division to determine the absolute TF copy number from the arbitrary mCherry fluorescence intensity in a given cell [42, 43, 44]. We find our maximum induction level is ~ 1000 repressors per cell [Fig. 3.9C] and our lower resolution limit is 3 – 5 repressors per cell [Fig. 3.9E]. See Appendix section “Calibrating LacI-mCherry intensity to absolute copy number” and Fig. 3.9 for details on this method. Simultaneously, we determine the level of gene expression by measuring the rate of YFP production.

3.2.3 Gene copy number measured by qPCR

We determine gene copy number using qPCR to measure the average number of plasmids in a cell. In this study we use plasmids based off the ColE1 Δ Rom origin of replication from Lutz and Bujard [45, 46, 47]. We also have made a version where the Rom protein, responsible for regulating the plasmid copy number, is inserted back into the ColE1 Δ Rom origin to arrive at an origin functionally similar to the wild-type ColE1 origin [48, 49, 50, 45]. While previous measurements locate the copy number of ColE1 Δ Rom plasmid in the range of 50 \sim 70 ([45]), the addition of the Rom protein should result in a reduced average plasmid copy number [48]. We find the ColE1 plasmid has an average copy number of 52 ± 5 while the ColE1 Δ Rom plasmid has a copy number of 64 ± 11 (error bars are standard deviation from triplicates). These values for the copy number show up as either N or N_c in the predictions generated by Eqs. (3.4) and (3.5), respectively. One obvious naive aspect to this approach is that the plasmid copy number is treated as a single static value. In any population of cells, the copy number is subject to cell-to-cell variability and thus the copy number is more accurately represented as a distribution rather than a single value [51]. Additionally, plasmid copy numbers are bound to increase as the cell progresses through its cycle under steady state conditions [52]. We will examine the consequences of these simplifications in a later section.

3.2.4 Determining sequence dependent TF binding energies

Finally, the affinities $\Delta\epsilon$ and $\Delta\epsilon_c$ of Lac repressor to its specific binding sites (Oid, O1, O2 and O3 from strongest to weakest) have been previously determined using bulk measurements [38, 8, 14]. Thus, we know all the parameters in Eqs. (3.3), (3.4) and (3.5) necessary to predict the fold-change in gene expression for every one of the regulatory cases considered in this chapter [Fig. 3.1]. Effectively, this means that we can predict the fold-change as a function of the number of repressors without any free parameters at all.

3.2.5 Simple thermodynamic model predicts expression level of single integrated gene copy

Our approach has several facets that require deeper examination. One possible confounding factor in the comparison to other measurements on the same architecture is that the fusion of LacI to a fluorescent protein might affect its function as a TF, thus changing its binding properties with DNA. A second point is that it's not immediately clear that a comparison of expression rate from cells grown under a microscope on a flat surface is comparable to steady-state measurements grown in bulk media [38, 11, 14].

To assess these issues, we compare our video microscopy method against the outcome of previous bulk steady-state results performed using wild-type Lac repressor [38, 14, 53]. In Fig. 3.3 we show the result of measuring fold-change in expression of a single chromosomal copy of our simple repression construct as a function of the number of repressors per cell for different binding sites (filled symbols) using the dilution method and video microscopy advocated here. The limits of our measurement both at low repressor numbers and at low fold-change (where repressed YFP production becomes small) are discussed in the Appendix. The lines are the theory predictions from Eq. (3.3) for each operator without any fit parameters with a shaded region representing the uncertainty in $\Delta\epsilon$. One assumption in this simple theory of Eq. (3.3) is that the copy number of the gene is exactly one. In reality, the copy number of our single integrated copy varies between one and two over the course of the cell cycle [17]. However, the predicted expression for two chromosomal copies, Eq. (3.4) with $N = 2$, is identical to Eq. (3.3) when $R \gg N$. Thus, the promoters will express independently and we can ignore this small correction [Fig. 3.11]. The data from reference [14] is shown as open symbols in the figure. These results lead to the interesting conclusion that single-cell measurement of the expression rates agree precisely with previous bulk measurements of steady-state expression.

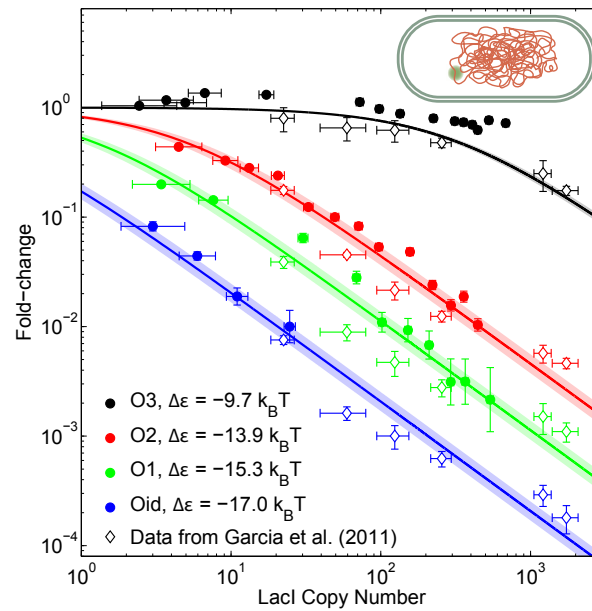


Figure 3.3: **Simple repression of a single chromosomal construct.**

Fold-change of simple repression construct located on the chromosome as a function of Lac repressor copy number. The solid lines correspond to Eq. (3.3) with values for $\Delta\epsilon$ from steady-state measurements of expression. The data from steady-state measurements [14] is shown as open symbols. The data from our experiments (filled symbols) is both consistent with the model with no free parameters (curves) and with expression data obtained from the same construct in steady-state measurements. The shaded regions on the curves represent the uncertainty from the errors in the measurement of the binding energies.

3.2.6 Predicting expression levels from plasmid constructs as a function of gene copy number

We now wish to compare the predictions of the thermodynamic theory against the more complicated cases involving TF binding. In this section, we compare the predictions of Eq. (3.4) to measurements of expression from plasmids as illustrated in Fig. 3.1B.

To begin, we measure the expression of an O1 simple repression construct placed on either the ColE1 or ColE1 Δ Rom plasmids, akin to Fig. 3.1B. The fold-change in gene expression as a function of Lac repressor copy number is shown for both plasmids in Fig. 3.4A. The data shown here is taken from the chronological middle of the cell cycle; the effect of the evolution of the copy number throughout the cell cycle on expression will be addressed later. The solid lines in the figure are plots of the predictions from Eq. (3.4) with no adjustable parameters. The shaded region accounts for the standard deviation in N from our qPCR measurements of the average copy number and the uncertainty in the binding energy $\Delta\epsilon$. For reference, the green points and line are the chromosomal data and theory for the O1 operator from Fig. 3.3. The theory predicts the fold-change in expression and captures the major features observed in our data. When the repressor copy number is much larger than the gene copy number, the fold-change is relatively unchanged with respect to the single copy chromosomal case as predicted for the case $R \gg N$. However, when the repressor copy number is less than the gene copy number the effect of the repressors is largely buffered away and the repressors have little effect on the fold-change. Between these two regions, the transition is sharp and switch-like. An alternative way to look at the data and its agreement with the theoretical predictions is to plot the fold-change as a function of the promoter copy number for a defined number of repressors. The inset to Fig. 3.4A shows this data for three distinct repressor copy numbers (8 black, 64 violet, 256 cyan). The distinct values for promoter copy number, N , are obtained by taking data from the simple repression O1 construct taken at the end of the cell cycle ($N = 2$) in addition to the plasmid data with and without Rom ($N=52$ and 64 respectively).

3.2.7 Simple thermodynamic model predicts expression levels from multiple integrated chromosomal gene copies

Plasmids can differ from the chromosome in their relative distribution throughout the cell, the accessibility to TFs and their segregation mechanisms [54, 55, 56]. However, the effect of the interplay between repressor and gene copy number is not exclusive to constructs located on plasmids. The same regulatory features can be seen at low gene copy numbers from multiple copies of the gene located on the chromosome. We measure gene expression using a strain which has the Oid simple repression construct integrated into the chromosome in five different locations, as schematically shown in Fig. 3.1C. To avoid uncertainty in the copy number of the gene, we examine expression near the

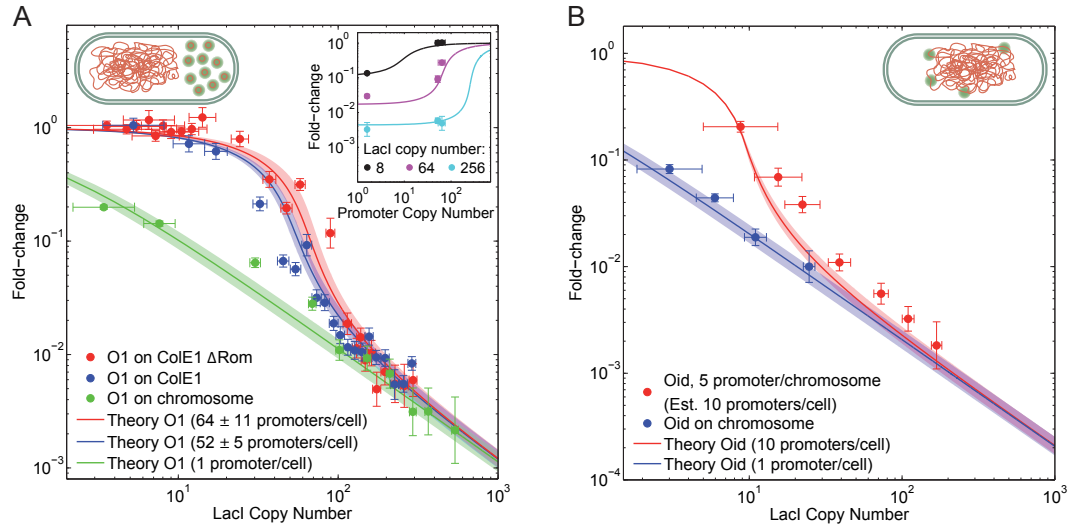


Figure 3.4: **Fold-change of multiple identical gene copies.**

(A) Fold-change as a function of Lac repressor copy number for two distinct plasmids with the O1 simple repression motif on a high copy number (ColE1) plasmid with (blue) and without (red) the Rom protein. Measurements are performed at the middle of the cell cycle. The blue and red solid lines are the theory from Eq. (3.4) using the average copy number measured by qPCR and known binding energies from earlier steady-state measurements as in Fig. 3.3. The shaded regions represent the combined uncertainty in the copy number measurement and the binding energy measurement. For reference, the green symbols and line are the data and theory prediction from Fig. 3.3 for simple repression with the O1 binding site for a single chromosomal copy. The inset shows the predicted scaling (lines) and measured fold-change (points) for three distinct repressor copy numbers as the number of promoter copies is varied. (B) Fold-change as a function of concentration of Lac repressor for multiple gene copies on the chromosome. The red symbols are measurements of the fold-change in expression at the end of the cell cycle of a strain with the Oid simple repression motif integrated into 5 unique sites on the chromosome. We expect 10 copies of the gene at the end of the cell cycle. The red line is the theory prediction for multiple identical gene copies with $N = 10$ from Eq. (3.4). The shaded region represents the uncertainty from the measured value of $\Delta\epsilon$. The blue symbols and line are the data and theory prediction for simple repression with the Oid binding site from Fig. 3.3. In both cases, the fold-change is approximately 1 when the copy number of the repressor is less than the copy number of the gene. At high repressor copy number, the curve coincides with simple repression from the chromosome with a sharp transition between these two regimes.

end of the cell cycle (the last 15 minutes before division); we expect that each of the 5 chromosomal copies will have fully replicated (the D period of the cell cycle, the time between replication completion and division, is roughly 27 minutes at our growth rate [17]) and the copy number of the gene should be 10 resulting from two sets of 5 copies, one on each completed chromosome [17, 57]. The resulting fold-change data for this construct is shown in Fig. 3.4B as red symbols. The red curve is the prediction from Eq. (3.4) for $N = 10$ and the Oid binding energy with, once again, no fit parameters. The shaded region in the fit comes from the uncertainty in $\Delta\epsilon$ for the Oid binding site. For reference, the blue curve and symbols are the theory and data (from Fig. 3.3) for the Oid construct integrated at a single copy in the chromosome. The observed behavior of this construct is qualitatively comparable to that observed for genes on plasmids [Fig. 3.4A]. Additionally, the same theory predicts its quantitative features without any free parameters. Once again, there is a sharp drop in fold-change when the number of repressors equals the gene copy number before rejoining the predictions (and measurements) for fold-change from a single gene copy. The genetic locations of integration and a discussion of the distribution and uncertainty in the measurement of gene copy number can be found in Appendix section “The copy number of multiple chromosomal integrations strain” and Fig. 3.10.

3.2.8 Predicting expression levels in complex TF binding landscapes

Finally, a common situation which results in competition for TFs occurs when different genes share the same TF. For example, in the regulatory databases of RegulonDB for *E. coli* [34] three quarters of TFs are listed as having specific interactions with more than one operon. In fact, many of these TFs target dozens of operons (and it is worth noting that these databases are far from complete and represent only a partial list of binding interactions, implying that these numbers will continue to grow). In this case, the competing specific binding sites which do not modulate the gene of interest may out-compete the gene copy when TFs are limiting.

To examine this competition scenario, we measure the expression from a single copy of the O1 simple repression construct integrated on the chromosome (identical to the O1 construct from Fig. 3.3) in a cellular context containing competitor binding sites. These binding sites are carried in a high copy number ColE1 Δ Rom plasmid that does not express a gene product, illustrated schematically in Fig. 3.1D. In Fig. 3.5A we show the measured fold-change of the chromosomal O1 construct in the case where the competing plasmid has a weaker O2 binding site (green symbols), equal strength O1 binding site (red symbols) or a stronger Oid binding site (black symbols). The theory curves stemming from Eq. (3.5) are shown in the corresponding color with the shaded region corresponding to the uncertainty in the theory stemming from the uncertainty in N_c , $\Delta\epsilon$ and $\Delta\epsilon_c$ (see Appendix section “Determining errors in theoretical predictions”). The stronger binding sites

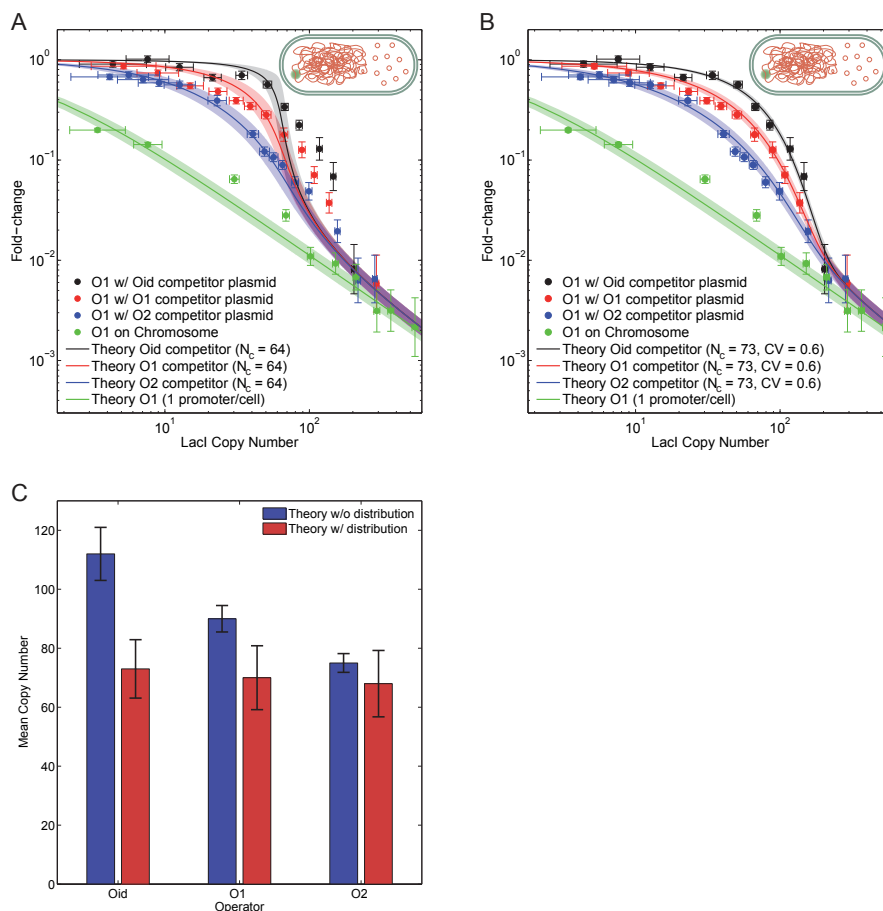


Figure 3.5: **Effects of repressor competition on expression.**

(A) Fold-change as a function of concentration of Lac repressor for the O1 simple repression construct integrated into the chromosome in competition with a ColE1 Δ Rom plasmid containing a stronger (Oid; black symbols), equal (O1; red symbols) or weaker (O2; blue symbols) Lac repressor binding site. For reference, the green symbols and line are the data and theory prediction, from Fig. 3.3, for simple repression with the O1 binding site without the competitor plasmid. (B) The same data used in (A) but now the solid lines represent the plasmid distribution theory assuming a normal distribution. The parameters are found by fitting the Oid (black) data for N , the average copy number, and the CV and these parameters are plotted for each binding energy, *i.e.* the red and blue curves are parameter free. (C) Fitted means of all three plasmid copy numbers for both the theory in Eq. (3.5), which assumes a single static copy number for plasmid (blue bars, $N_c^{\text{Oid}} = 112$, $N_c^{\text{O1}} = 90$, $N_c^{\text{O2}} = 75$), and the same theory where the plasmid distribution is normal with CV as determined from fitting the Oid data (red bars, $N_c^{\text{Oid}} = 73$, $N_c^{\text{O1}} = 70$, $N_c^{\text{O2}} = 68$). All three plasmids in this case have the same origin of replication and differ only by a few bases which alter Lac repressor affinity to their binding sites. As such we expect all three plasmids to have identical copy numbers. This is observed for the theory with plasmid distribution (red bars), however the theory without a plasmid distribution systematically overestimates the copy number for stronger binding (blue bars).

shift and sharpen the transition of the gene of interest with respect to LacI copy number. Once the repressor copy number exceeds the number of competitors, the gene finally gains access to the repressor and becomes regulated. In contrast, when the competitors are weaker the position of the transition is shifted towards lower LacI numbers. This simple example illustrates how the regulatory behavior of a gene can be indirectly controlled through identical competitor plasmids. This effect is, however, more general as in wild-type genes the competition will come from a spectrum of binding sites each controlling a specific gene. The thermodynamic model can produce predictions for any such specific arrangement of binding sites with the theoretical infrastructure demonstrated here.

3.2.9 The influence of plasmid distribution on the repressor titration curves

One unsatisfactory feature of Fig. 3.5A is that the observed transition for strong binding sites (Oid and to a lesser extent, O1) is not as sharp as predicted by the theory. Furthermore, the blue bars in Fig. 3.5C show the resulting fit values for the mean competitor plasmid copy number if Eq. (3.5) is fit to the data in Fig. 3.5A. All three of these measurements correspond to plasmids which differ only in the strength of their LacI binding site (which in this case is isolated and not connected to a promoter) and thus we would expect the mean copy number to be unchanged in the Oid, O1 or O2 strain. However, it is clear that stronger binding sites systematically predict a higher copy number showing that in this case the model lacks internal consistency.

These discrepancies, both in the fit to the sharpness of the transition region and in the measured copy number between similar plasmids, likely stem from the theory not accounting for the fact that there is a distribution of plasmids in the population of cells [16, 58, 51]. This distribution will result in a less sharp transition in the fold-change curve. To see intuitively how a distribution of plasmid copy numbers alters the fold-change curve, we imagine the situation of a simple distribution where cells have a single chromosomal YFP gene with half the cells having N and the other half having $3N$ competing plasmids. Further, assume the binding site on these competing plasmids has an extremely high affinity. The average plasmid copy number is $2N$ and it is at this value that the thermodynamic model predicts a sharp transition in the fold-change curve. However, when the number of repressors is $R = 2N$ all the cells with N plasmids are repressed by free repressors and produce very little YFP because all the competing plasmids are saturated. On the other hand, the cells with $3N$ plasmids will still not be repressed because the competing plasmid can buffer the $2N$ available repressors. Hence, for the entire population, the fold-change is no less than $1/2$. Only when the repressor copy number reaches $3N$ will repression in every cell ensue and begin to show a steep drop on a log scale. This simple argument provides the intuition for why a distribution of plasmids is required in the theory when thinking about the strong operator limit on the competitor plasmids.

Generically, a distribution of plasmid copy number in a population of cells will move the location of the switch-like transition to repressor numbers above the mean plasmid copy number and the transition will be softened. This effect is stronger as the copy number distribution becomes wider than the width of the transition region in the fold-change curve. Therefore we expect that the stronger the plasmid binding site, the worse the simple single copy number theory will fit. This is observed for the data in Fig. 3.5A where O2 fits well to the simple theory, O1 fits worse and Oid fits even worse.

It is relatively simple to account for a distribution of plasmids in the thermodynamic theory, however the derivation is left to the Appendix. In the case of identical copies of the gene, Eq. (3.4) must be modified such that the fold-change in the presence of a distribution, $\text{fold-change}_{\text{dist}}$, is related to the fold-change of the fixed copy number fold-change by,

$$\text{fold-change}_{\text{dist}} = \sum_{n=0}^{\infty} p(n) \frac{n}{\langle n \rangle} \text{fold-change}(n), \quad (3.8)$$

where $p(n)$ is the probability that any cell in the population has n plasmids, with $\langle n \rangle = \sum_{n=0}^{\infty} np(n)$ the average number of plasmids in the population, and $\text{fold-change}(n)$ is given by Eq. (3.4) for a particular value of n . The theory for competitor binding sites on a plasmid can be adjusted for a plasmid distribution in a similar fashion. In this case we find,

$$\text{fold-change}_{\text{dist}} = \sum_{n=0}^{\infty} p(n) \text{fold-change}(n), \quad (3.9)$$

where now $\text{fold-change}(n)$ refers to Eq. (3.5) with n plasmids.

To exploit these ideas in the context of our data, we propose a simple phenomenological distribution for the plasmid copy number, a normal distribution with a fixed coefficient of variation (standard deviation over mean). We have chosen a normal distribution for simplicity as we do not expect the exact details of the distribution to have a major effect, as a plasmid distribution will always dull the sharp transition of the fold-change repressor titration curve. We fit the Oid competitor data from Fig. 3.5A to the distribution treating the mean copy number (N_c) and coefficient of variation (CV), σ/N_c , of the distribution as fit parameters. We find the best fit mean is $N_c = 72$ with $\text{CV} = 0.6$; the resulting fit is shown in Fig. 3.5B as the solid black line. We also plot the data from the O2 and O1 plasmid with the same values for the mean and standard deviation.

The theory, which was fit to the Oid data, now describes the data from all three operator sites very well. This is an important sanity check, as we don't expect the strength of a LacI binding site far from the origin of replication to affect the plasmid copy number or its distribution. This is further demonstrated in Fig. 3.5C where we plot the mean plasmid copy number measured by fitting either the simple no distribution model to our data or fitting the mean copy number while

holding the width of the distribution fixed with $CV = 0.6$. The point is illustrated in these bar graphs. When the transition is not sharp, as in the O2 data (and to a lesser extent the O1 data), the fixed N_c single parameter theory fits well and with a copy number consistent with what we expect. However, as the competitor binding gets strong and the fold-change response curve is expected to get sharp compared to the distribution of plasmids, the single parameter fixed N_c theory fits poorly and N_c goes from being descriptive of the actual copy number to merely a phenomenological fit parameter. However, the fit to the mean copy number for the three operators remains consistent when the distribution of plasmids is accounted for in the model.

3.2.10 Cell cycle dependence of the plasmid copy number and the resulting expression

To this point we have varied the copy number of competing binding sites or gene copy number by comparing the fold-change of different constructs at similar points in the cell cycle. However, it is clear that over the course of the cell cycle all genetic material in the cell must double. As a result, we examine the time dependence of the copy number by binning the data according to when in the cell cycle each measurement is made. In this metric, the time of birth of the cell is represented as 0 and the time of its subsequent division is 1. In Fig. 3.6a, an example of the fold-change curves obtained are shown for the O1 simple repression chromosomal integration with the Oid competitor plasmid. In this case, each time bin is fit to Eq. (3.9) for the copy number N_c , keeping the coefficient of variation fixed; the resulting copy number for that point in the cell cycle is written in the legend. As expected, the measured average plasmid copy number increases as the cell cycle progresses. In addition, the copy on the chromosome will double. The operator associated with this copy will affect our measurement of N_c , however the addition of one extra operator in the presence of dozens of copies on plasmid results in a only a very small change to the predicted fold-change. The exact details of the size of this effect are discussed in the Appendix section “Accounting for chromosome replication in competitor theory”. Repeating this process for all plasmids with the proper theory equations (Eq. (3.8) for identical plasmids expressing YFP, Eq. (3.9) for the competitor plasmid data), we plot in Fig. 3.6B the measured plasmid copy number versus fractional cell cycle for each. The horizontal dashed lines and corresponding shaded region represent our qPCR measurements of the average copy number for ColE1 and ColE1 Δ Rom. It should be noted that the cell cycle parameter relates to when the measurement itself was made, but due to fluorophore maturation the actual measurement may represent a time period earlier in the cell cycle.

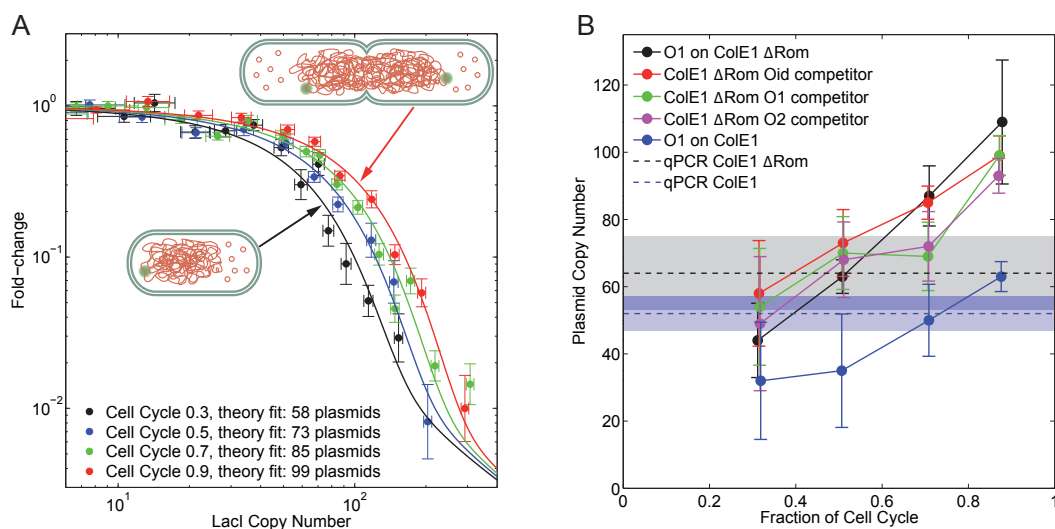


Figure 3.6: **Variation of plasmid copy number throughout the cell cycle.**

(A) Fold-change as a function of LacI copy number for the O1 simple repression construct integrated into the chromosome in the presence of a competing ColE1 Δ Rom plasmid bearing an Oid site for different time points in the cell cycle. The “cell cycle” parameter is the average fraction of the total cell lifetime from which the binned data is taken with 0 representing birth and 1 representing the cell division. The plasmid copy number is fit to Eq. (3.9) at every time point, keeping $CV = 0.6$ fixed, and the resulting value for N is listed in the legend. (B) Plasmid copy number versus the cell cycle. The plasmid copy number is measured by fitting the copy number parameter in the theory to the data from all our experiments binned by time in the cell cycle. The horizontal dashed lines and matching shaded regions are our qPCR measurements for average copy number of ColE1 (blue dashed line) and ColE1 Δ Rom (black dashed line).

3.3 Discussion

Recent experimental and theoretical efforts have focused on understanding the role of regulatory architecture in transcriptional decisions. In these studies, the details of isolated regulatory architectures (number, location and strength of binding sites and TF copy number) are varied systematically and the transcriptional output is compared to corresponding theoretical predictions [7, 41, 39]. However, it is rarely the case that TFs act on only one promoter. As a result, the study of transcriptional decisions at individual promoters, without taking into account the rest of the regulatory network, might be insufficient. In particular, the presence of multiple targets for the same TF can result in a competition that reduces the available free TFs for the gene of interest [40].

In this chapter we explored the interplay between the binding sites for a transcription factor at a gene of interest and competing binding sites regulating other genes in the context of the well studied simple repression architecture [14]. We show that the presence of competing binding sites not only changes the effective amount of available TF, it can also affect the input-output relation by introducing a sharp transition. This transition separates the regime where the repressor is depleted compared to the number of available binding sites and the regime where the repressor is in excess of the number of binding sites. The width of this transition is controlled by the strength of the binding sites. The theory also predicts that the width of this transition, in a population of cells, depends on the size of cell-to-cell fluctuations in binding site copy number; larger fluctuations in the number of available competitor binding sites (or number of identical genes) tend to flatten this transition. We find that when a very strong transition is predicted, the measured transition is considerably dulled which we attribute to the copy number variability in the population.

The quantitative consequences of binding site competition can be predicted using thermodynamic models without any free fitting parameters. Previously, these models had been successful in predicting transcription output for the simple repression architecture in the absence of binding site competition [14]. By measuring binding site copy numbers using qPCR we show that an extended version of these models accounting for the presence of multiple binding sites [40] describes our data precisely with no fit parameters for a wide range of binding site copy numbers and strengths.

Building on the success of the theory in quantitatively predicting the regulatory outcome of the various architectures considered here, we fit the theory to the fold-change curve at different points in the cell cycle as a way to measure the time evolution of the plasmid copy number during the course of the cell cycle. One noteworthy feature of this method is that the reporter fluorescence molecule need not be expressed by the measured plasmid, it only requires that a TF be shared between an unrelated chromosomal copy and the plasmid for the copy number of the competitor to be measured. This can be of benefit as it requires only the insertion of a binding site on the plasmid of interest. Additionally, this approach prevents possible changes in cell physiology due to

starvation or phototoxicity from overexpression and measurement of a fluorescent reporter protein expressed from a high copy plasmid.

The ability to measure absolute numbers of both binding site and input TF copy number is key for contrasting our experimental data with the theoretical predictions stemming from thermodynamic models of transcriptional regulation [41, 39]. In this work we have made use of a recently introduced fluctuation method for taking the repressor census and thereby checking the governing equation for the simple repression regulatory motif in a wide array of situations where the TF was in demand from multiple sources [42, 44]. We find that this dilution method, a form of video microscopy, gives quantitatively compatible results to more traditional steady-state snapshots and bulk measurements [14]. However, there are numerous advantages of the dilution method. This method provides a single cell readout of the number of repressors in each cell. The fact that no new repressors are produced ensures that this “input” level of repressors is held constant through the entire measurement. The single cell nature of this method certainly increases our resolution as compared to measuring a bulk sample where the possible distribution of repressors from cell to cell can have a wider distribution than the feature we wish to illuminate; akin to the issues we see when the plasmid distribution is wider than the sharp transitions we wish to study, seen in Fig. 3.5A. Finally, this method allows one to probe particular regions of the titration curve with varying degrees of resolution; a feature that was essential when trying to distinguish sharp features of the repressor titration curve at 10 repressors for the chromosomal case and over 100 repressors in the plasmid case.

As our characterization of cellular decision making becomes more quantitative so must our theoretical description of this process. The quantitative and predictive control of such decisions allows one to probe their molecular details at a level that escapes any qualitative description. In addition to expanding our understanding of regulation, quantitative models give us the ability to control regulatory output by predictive design. In fact, synthetic biology has focused on the development of standardized regulatory units with known input-output functions [23, 24]. The modification of these input-output functions to have, for example, a particular shape usually requires the reengineering of the regulatory architecture at the DNA level [42, 16, 27]. Our work provides a complementary approach to controlling input-output functions as the introduction of competing binding sites for a TF into a cell makes it possible to tune regulatory response in a predictive fashion without the need of any modifications at the DNA of the gene of interest.

3.4 Experimental procedures

3.4.1 Gene expression measurements

Cultures are grown overnight in 2 ml of LB at 37°C and diluted $\sim 1 : 10^4$ in M9 + 0.5% glucose minimal media with antibiotics and 1, 2, 3, 4, 6, 8, 100 ng/mL anhydrotetracycline (aTc) to induce the

production of various levels of LacI-mCherry that cover the full repressor range (for induction curve, see Fig. 3.7). The diluted cultures are grown at 37°C until they reached an OD600 \approx 0.2 – 0.4 and then they are washed twice with fresh, M9 media (without aTc) to remove the inducer. They are then diluted to give several cells per field of view when placed on a 2% low melting point M9+0.5% glucose agar pad. An automated fluorescent microscope, simultaneously records multiple fields of view for each concentration of aTc. In addition, one pad contains cells without the repressor construct, whose expression measurements serve as the denominator of our fold-change measurements. Growth of cells is observed by fluorescence microscopy at 37°C for 2.5 hours while measuring CFP (used for segmentation), YFP and mCherry intensities.

3.4.2 Data analysis

Data analysis was performed using the Matlab code “Schnitzcells” kindly provided by Michael Elowitz [42]. This code segments cells in a movie and tracks their lineages.

3.5 Acknowledgments

We are grateful to James Boedicker, Michael Elowitz, Ron Milo, Nitzan Rosenfeld and Jon Young and members of the Phillips Laboratory for helpful discussions. We are also grateful to the NIH for support through award number DP1 OD000217 (Director’s Pioneer Award), NIH award number R01 GM085286 and La Fondation Pierre Gilles de Gennes (RP).

Appendix

3.A Genetic elements and details of the dilution method

3.A.1 Dilution circuit

The dilution method is used in this work to measure the transcription factor (TF) titration curve for the expression of a gene under the control of that TF [42]. The required genetic elements of the dilution circuit are: the target gene with a fluorescent protein product (YFP) whose promoter has the regulatory architecture to be queried, a TF-fluorescent protein fusion (LacI-mCherry) whose production is tightly regulated and shut off by another repressor (TetR) that can be inactivated by a small molecule inducer (aTc). Finally, a volumetric marker (CFP) is used to easily segment cells in microscopy images. A schematic of the dilution circuit used here is shown in Fig. 3.7A. To measure fold-change, we also measure the rate of expression of a strain which does not contain the LacI-mCherry fusion gene.

3.A.2 The one step dilution method

As originally outlined [42], the dilution method consists of fully inducing a culture to the maximal level of a TF concentration before shutting off production of the TF and observing under a microscope as individual cells grow to form colonies. During this growth process, the quantity of TF in each cell drops and the response of a gene product which is regulated by that TF is measured in successive generations, with each generation diluting the TF by roughly a factor of two. A powerful aspect of this method is that by observing how the TF partitions to the daughters, one can arrive at a “calibration” relating the fluorescence of the TF-fluorescence protein fusion to the actual number of TF molecules present.

In the experiments reported here we alter the strategy slightly. Instead of fully inducing the culture and taking a long dilution movie, we variably induce with 6 to 10 distinct concentrations of inducer such that the entire range of starting TF concentration is covered. Individual cells are followed over one full division cycle with only one mCherry fluorescence measurement (which measures relative LacI concentrations) followed by 75 minutes of gene production measurements (10

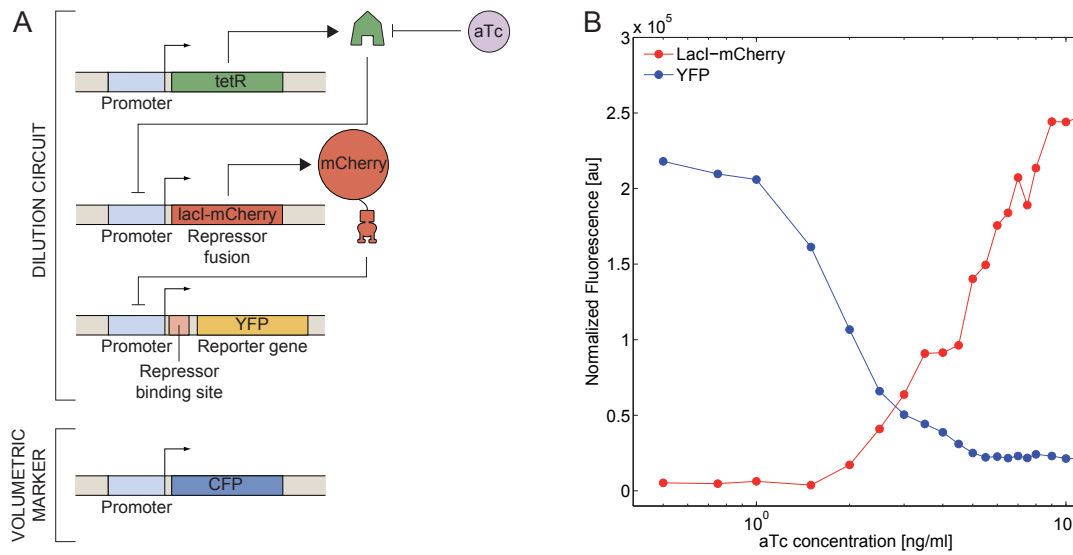


Figure 3.7: **Experimental circuit schematic and response to aTc, related to Fig. 3.2.**

(A) Full schematic of dilution circuit. LacI-mCherry regulates the target gene which expresses YFP. The LacI-mCherry is expressed from a $P_{LtetO-1}$ promoter [45]. When the small inducer molecule aTc is not present, TetR, which is expressed from a P_{N25} promoter, shuts off production of LacI-mCherry. When saturating amounts of aTc are present, the LacI-mCherry level can be induced to roughly 1000 dimers per cell. Finally, CFP is constitutively expressed from the chromosome and is used as a volume marker for segmentation. (B) Induction of the dilution circuit. Steady state fluorescence level per cell of LacI-mCherry and YFP as a function of aTc concentration in the O2 single chromosomal copy strain (used for the green data in Fig. 3.3). Fluorescence is measured in bulk using a plate reader as described in [59].

exposures, once every 7.5 minutes). We argue this method has several advantages which improve the measurement over its original version:

- Data points are acquired in a uniform fashion over the whole range of induction. In the long growth method, the data is exponentially distributed towards lower TF concentrations (for every one cell with N repressors, there are two with $N/2$, 4 with $N/4$, etc.).
- The accuracy of the measurement is uniform over the range of repressor concentrations. In the long growth method, measurements in the high TF copy number regime are less photobleached as compared to cells in the low TF copy number regime, which occurs towards the end of the movie. A bleaching correction needs to be done for the entire movie’s worth of exposures that were previously taken. This results in a statistical averaging of our partitioning events during cell division and amplifies noise, which is the dominant source of error, particularly late in the movie when the signal is low.
- By using one long exposure in the repressor measurement we get an extremely accurate measurement since we don’t need to limit the exposure times in order to minimize photobleaching.
- Colony size is small and independent of TF copy number. In the long growth method, low TF numbers are always correlated with larger colony sizes which can make *very* significant contributions to the background fluorescence from neighboring cells. In particular, once multiple layers of cells begin to grow in the middle of big colonies, we find that the contributions from out of plane fluorescence can be as big as the signal itself.

3.A.3 Physiological effect of repressor induction

To demonstrate that the induction of repressor does not introduce a global physiological change to the cell as a function of induction, in Fig. 3.8 we show that the relative expression of the volume marker from a constitutive *lacUV5* promoter integrated on the chromosome (solid circles) remains unaffected as we change the repressor copy number by almost three orders of magnitude. This promoter is identical to the one responsible for the expression of our reporter gene (data also shown in corresponding color and open circles), except that its repressor binding site has been mutated away (see Appendix section “Constructs and strains”). We conclude that the fold-change of our reporter gene does not change significantly as a result of physiological changes in the cell resulting from varying the intracellular repressor load.

3.A.4 Cell growth and detailed experimental procedure

Three distinct strains are grown for each experiment. First, the rate of YFP expression from the construct of interest is measured in a strain background bearing LacI-mCherry. Second, the rate

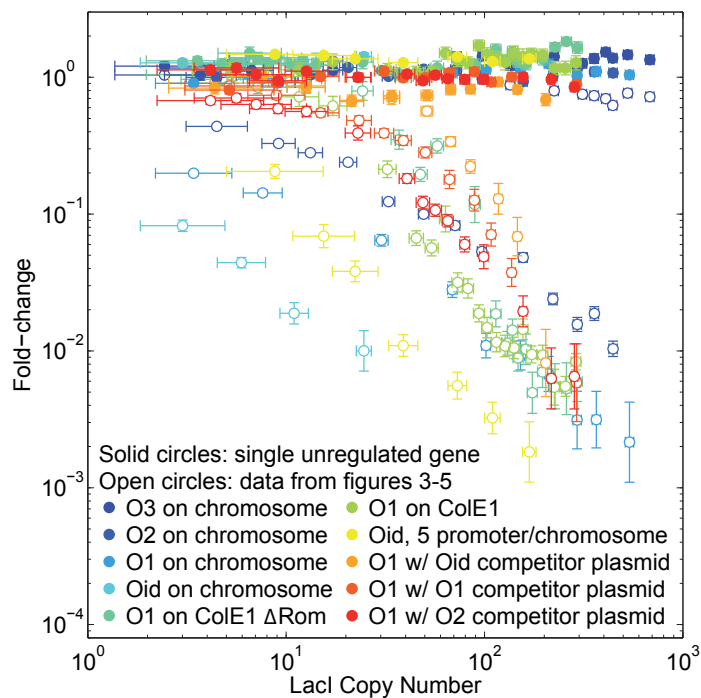


Figure 3.8: **Comparison of fold-change of *lacUV5* promoter with and without repressor binding site, related to Figs. 3.3-3.5.**

Fold-change of unregulated CFP expression from a single integrated gene copy (solid points) compared to the fold-change data from Figs. 3.3-3.5 (open points). Like-colored data is taken simultaneously in the same population of cells and the promoter in both cases is *lacUV5*. The induction of repressor does not have any obvious secondary effects on the global transcription, as demonstrated by the fold-change of the unregulated promoters staying constant at 1 while the same promoter with a repressor binding site exhibits orders of magnitude lower expression.

of expression of the same construct is measured in a strain lacking Lac repressor. Finally, the autofluorescence is determined using a strain which does not have the LacI-mCherry construct or a reporter YFP construct.

Overnight cultures are grown in 2 ml of LB in the presence of the appropriate antibiotic (chloramphenicol is always present for the TetR plasmid and kanamycin for both ColE1 based plasmids or ampicillin for the “competitor” plasmids) at 37°C. They are then diluted $\sim 1 : 10000$ in M9 + 0.5% glucose minimal media with antibiotics and anhydrotetracycline (aTc; Acros Organics cat. num. 233131000) at several different concentrations (1ng/mL, 2ng/mL, 3ng/mL, 4ng/mL, 6ng/mL, 8ng/mL or 100ng/mL) to induce the production of various levels of LacI-mCherry. The induction curve for LacI-mCherry, shown in Fig. 3.7B, is used as a guide for choosing aTc concentrations that cover the full repressor range. These minimal media cultures are grown at 37°C until they reach an $OD_{600} \approx 0.2 - 0.4$ and then they are washed twice with fresh, M9 media (without aTc) to remove the inducer.

The resuspended cultures are then diluted (typically 1:10 in fresh M9+0.5% glucose minimal media) to give several cells per field of view when 2 μ l are placed on a 2% low melting point M9+0.5% glucose agar pad (NuSieve GTG Agarose, Lonza cat. no. 50081). An automated Nikon fluorescent microscope (Nikon Eclipse TI) is controlled by the software Micro-Manager [60], and multiple fields of view (totaling roughly 35 individual fields per experiment) are recorded simultaneously for each concentration of aTc. In addition, one pad contains cells without the repressor construct whose expression measurements serve as the denominator of our fold-change measurements (*i.e.* expression for $R = 0$). Before the growth movie is started, the autofluorescence signal in YFP and mCherry is measured from the autofluorescence strain with 10 positions accounting for roughly 500 individual cells.

Growth of the LacI-mCherry and Δ LacI-mCherry strains is observed by fluorescence microscopy at 37°C over 2.5 hours with CFP exposures every 7.5 minutes for the first 9 frames of growth. This initial period of exposures is used to record the lineage of all cells and identify daughter pairs. In the 10th frame, the CFP exposure is taken along with a single, long exposure of mCherry to determine the LacI concentration in every cell. The last 10 frames consist of both CFP and YFP exposures every 7.5 minutes. The difference in corrected fluorescence of consecutive YFP images (corrections explained below) makes one measurement of expression. By examining only the first division we eliminate colony size as a source of error in our fluorescence measurements; cells in large colonies can have nontrivial contributions to their fluorescence signal from neighboring cells. In addition, only measuring the LacI-mCherry concentration once eliminates the necessity to correct for photobleaching, which necessarily assumes that the bleached fluorophores have been proportionately distributed to the daughters. Furthermore it increases the sensitivity of the LacI-mCherry measurement by allowing for longer exposures without worrying about bleaching. Exposures are chosen to be as long

as possible without impacting the growth rate, compared to a control with no fluorescence exposures.

3.B Image segmentation and analysis

For cell segmentation and lineage identification, we use a modified version of Schnitzcells [61] (kindly provided by the lab of Michael Elowitz, Caltech) designed to segment on a fluorescence marker. We have altered the program slightly such that segmentation and tracking is automated with error checks based on lineage verification (every cell either has a mother or was alive in frame 1, every cell has two daughters or was alive in frame 20) and growth verification (to check that cells do not grow (or shrink) too fast; this usually indicates a tracking error). Failing either of these error checks requires manual intervention, however, most movie positions do not require any intervention. Once all errors are resolved, the program provides a list of all cells, their lineages and the total fluorescence intensity (pixel intensities summed over the segmented pixels of a cell) for every channel and every frame for each cell.

There are essentially two separate data collections going on in the same experiment. One data collection corresponds to gathering pairs of daughter cells whose lineage is known (*i.e.* their common mother cell is known) and have an mCherry measurement (they had divided from their mother already by frame 10). The second data collection corresponds to expression measurements where a cell must have an mCherry measurement to quantify the LacI-mCherry number (*i.e.* the division event that produced the cell must have occurred before frame 10), and must have both been “born” during the movie and divided again sometime later in the movie (it must have an identified mother and daughter set). This prerequisite of two division events allows us to categorize where in the cell cycle the expression measurement occurs. Knowing the location in the cell cycle is important since the copy number of plasmids and chromosomal integrations changes over time. Fluorescence values are corrected for field nonuniformities, chromatic aberration, autofluorescence, photobleaching and crosstalk as described in the following sections. Autofluorescence values for YFP and mCherry are determined as the fluorescence value per pixel from the snapshots of the autofluorescence strain taken immediately preceding each movie.

3.B.1 Flattening fluorescence images

Our illumination is not spatially uniform over an entire field of view. We correct for this fact by taking fluorescence images in the YFP channel of a plastic slide with uniform but bright autofluorescence intensity (Autofluorescent Plastic Slides, Chroma cat. no. 92001) and averaging over 10 – 20 of these images. The resulting image is a map of illumination intensity at any given pixel I_{flat} . The

raw images, I , are then renormalized such that for pixel i, j with raw intensity $I^{(i,j)}$,

$$I_{\text{corrected}}^{(i,j)} = \frac{I^{(i,j)} - I_{\text{dark}}^{(i,j)}}{I_{\text{flat}}^{(i,j)} - I_{\text{dark}}^{(i,j)}} \times \text{mean}(I_{\text{flat}}^{(i,j)} - I_{\text{dark}}^{(i,j)}), \quad (3.10)$$

where I_{dark} corresponds to an image taken with no illumination (mostly these counts are from camera offset).

3.B.2 Chromatic aberration correction

Due to chromatic aberrations in the microscope, the various fluorescence channels are slightly offset from each other. We measure this offset by imaging microspheres (Invitrogen Tetraspeck microspheres, cat. no. T-7281) which fluoresce in all three channels we use (CFP, YFP and mCherry) and rapidly image in all three channels. We then measure the center-to-center distance of the identified sphere in the three images. We find that the YFP image is translated in the x -direction by two pixels and the mCherry is translated by three pixels in the same direction with respect to the CFP image. We find there is no offset in the y -direction. To account for this we translate all YFP images and all mCherry images by the measured offset and trim the edges such that we only look at pixels where there is a measurement in all three channels.

3.B.3 Autofluorescence correction

To calculate the autofluorescence stemming from cells in the YFP and the mCherry channel, we take 8 – 10 snapshots of a strain which is ΔYFP and $\Delta\text{LacI-mCherry}$ and measure the average per pixel intensity of the identified cells in both the YFP channel and the mCherry channel. This average is then subtracted from each pixel of any YFP or mCherry measurement that is made.

3.B.4 Correcting for crosstalk and cross bleaching

We measure the crosstalk between any two channels used in our experiment by determining the difference between the autofluorescence of a strain without a given fluorophore in the presence of the other fluorophore fully induced. So, for instance, we can find the crosstalk of YFP into the mCherry channel by taking exposures of our $\Delta\text{mCherry}$ strain with the appropriate YFP construct (depending on the experiment in question). The ratio of the average per pixel mCherry fluorescence signal to the average per pixel YFP signal (corrected for all the above factors) is the crosstalk. Therefore we correct mCherry measurements for this factor, γ_{cross} , by subtracting from the cells' summed mCherry fluorescence, the summed YFP fluorescence times this crosstalk factor. We do not have to worry about normalizing exposure times because the crosstalk factor between any two channels is measured with the same exposure time used in the experiment. We find that this crosstalk

factor is only relevant in the case of the ColE1 and ColE1 Δ Rom plasmids which express YFP. In this case we measure $\gamma_{\text{cross}} = 0.006$, or 0.6% of the YFP signal can be seen in the mCherry channel. For our other constructs there is 10 – 50 times less YFP (corresponding to 1 – 10 copies of the YFP gene compared to 50 – 70 plasmids) and so we can expect the effect would be correspondingly smaller, though for our experiments it is too small to measure and this correction is not included in those cases.

Conveniently, we do not have to worry about CFP crosstalk into the YFP channel. The first point is that all cells have the same expression of CFP (same constitutively expressed construct), so any potential crosstalk shows up as autofluorescence when we measure YFP; even our autofluorescence strain expresses CFP with the proper construct. Additionally, because our YFP measurement is always measured as a rate of production, which is the difference in production from consecutive frames, most of the autofluorescence and crosstalk corrections cancel out since the correction term is proportional to the size of the cell on both measurements.

We also check for cross bleaching between the fluorescence channels. It is possible that one of our exposures, shaped to excite a particular fluorophore, excites and bleaches a different fluorescent protein species (for instance if the CFP exposure excites and bleaches YFP). Bleaching in the CFP channel does not change our measurements and the mCherry exposure occurs only once before we begin to measure YFP. Since we are only concerned with the rate of YFP production, bleaching YFP molecules before we begin measurements does not change the measured rate of production. Therefore, we only need to worry about how the CFP exposures bleach the YFP or mCherry signal. To account for the CFP exposure, all YFP bleaching curves are measured accounting for both the YFP and CFP exposure; therefore the bleaching from the CFP exposure is rolled into our measurement of the YFP bleaching rate. To check the cross bleach rate of CFP on mCherry, we take an mCherry exposure followed by a long CFP exposure 600 \times longer than that used in experiment, followed by a last mCherry exposure. As a control we also measure the bleach rate of the mCherry exposures alone, without the CFP exposure. We find that the bleaching for this extremely long exposure is roughly 25%, implying that the bleaching from a single exposure is less than 1/10th of a percent.

3.B.5 Correcting for photobleaching

The only channel which must be corrected for photobleaching is YFP. Due to the fact that only one mCherry image is taken per experiment, we do not need to correct it for photobleaching. Before each movie we measure a photobleaching curve of YFP using the highest expression strain available. The characteristic bleaching rate, τ , is extracted by fitting the autofluorescence subtracted bleaching curve to a single exponential decay. Then, all measurements of YFP production, ΔYFP , are corrected

such that,

$$\Delta\text{YFP} = \text{YFP}_{i+1} - \text{YFP}_i(1 - \gamma), \quad (3.11)$$

where $\gamma = e^{(-t_{\text{exp}}/\tau)}$ and t_{exp} is the exposure time for a YFP image.

3.C Calibrating LacI-mCherry intensity to absolute copy number

The absolute number of TFs per cell is usually obtained by cross-calibrating to independent measurements such as immunostaining [38, 62, 14]. In our case, where our signal comes from the LacI-mCherry fusion, the total fluorescence intensity of a cell, I , can be related to the absolute number of TFs N through the calibration factor α such that,

$$I = \alpha N. \quad (3.12)$$

The calibration factor is often determined by measuring the mean intensity of a single copy of the fluorescent molecule [63, 64, 59] or of a bulk solution of purified fluorophore [65, 66, 67, 68, 64]. Here we determine α using a calibration method based on fluctuations in protein partitioning during cell division [42, 43, 44]. By tracking fluorescence partitioning between two daughters after a division, the properties of binomial partitioning state that the average size of fluctuations in the signal of daughter 1 and 2 will be proportional to the total fluorescence signal partitioned. This circumvents the need for a cross-calibration as it allows us to obtain α and simultaneously measure absolute TF copy number, R , in single cells. We expect the distribution of our LacI-mCherry between the two daughter cells should obey the statistics of a fair binomial partitioning [Fig. 3.9A and B, discussed below]. This simple fact alone is enough to determine the calibration factor α . In particular, by observing the fluorescence of the two daughters, captured in the quantities I_1 and I_2 , it can be shown that

$$\langle (I_1 - I_2)^2 \rangle = \alpha(I_1 + I_2), \quad (3.13)$$

where α is the desired calibration factor that links fluorescence intensity and number of fluorophores via $I = \alpha N$. This relation follows from the properties of the binomial distribution as shown in the following subsection. In Fig. 3.9C, we show an example of the calibration data from the experiment in Fig. 3.3. The exact value for the calibration factor is specific for a given acquisition and the current settings of the microscope and is determined for each experiment's unique imaging conditions (exposure times, illumination intensity, etc.).

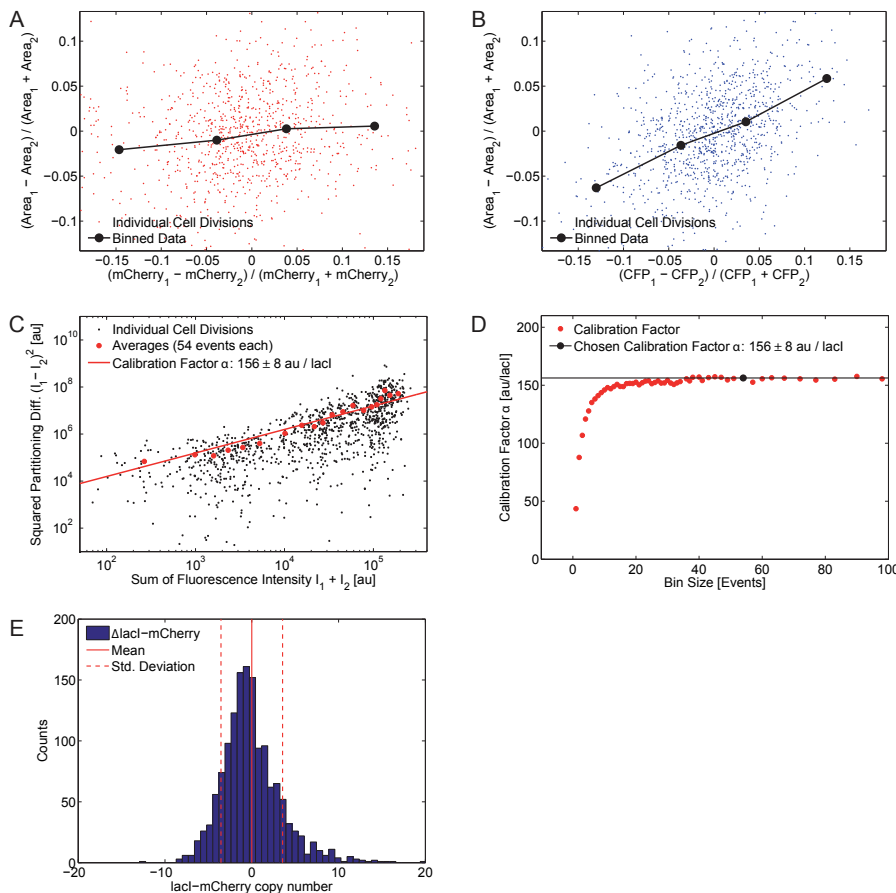


Figure 3.9: **Determination of the calibration factor of LacI, related to Fig. 3.3.**

(A,B) Fraction of area partitioned between a daughter pair on division versus fraction of fluorescence signal partitioned between a pair of daughters for (A) LacI-mCherry and (B) CFP. The partitioning of the cytoplasmic CFP volume marker (B) is strongly influenced by area partitioning differences (more fluorescence partitions to larger cells with more volume). However, the LacI-mCherry (A) is only weakly influenced by volume area partitioning. This is expected since LacI tends to be DNA bound and the chromosomal DNA is partitioned in approximately equal measures independent of differences in cell volume of the daughters. (C) Calibration of fluorescence of LacI-mCherry 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. These points are binned resulting in the red, averaged data points which are fitted to Eq. (3.13) in order to obtain the calibration factor α relating mCherry fluorescence to the absolute number of LacI-mCherry molecules inside the cell. (D) Sensitivity of the calibration factor to data binning. The determined calibration factor as a function of the number of points in each total fluorescence bin. For sufficient averaging (bin sizes ≈ 15 or greater), the calibration factor is relatively insensitive to the binning. (E) Histogram of mCherry measurement of the auto fluorescence strain in units of number of LacI-mCherry. This histogram depicts the inherent limit of detection for LacI-mCherry proteins calculated in an example experiment. The histogram is the mCherry fluorescence of a collection of cells with no LacI-mCherry, but due to autofluorescence fluctuations are typically measured to have nonzero fluorescence. We use the width of this distribution to set the limit of where we can distinguish signal from autofluorescence fluctuations. This is calculated for every experiment and we ignore points in our data which have less than one standard deviation above zero fluorescence.

3.C.1 Fairness of repressor partitioning

The fluctuation-based counting method employed here relies on measuring the asymmetries in partitioning of a TF-fluorescent protein fusion during the cell division process. In this scenario of DNA bound TFs, it is *assumed* that the partitioning between daughter cells is random, mediated by the segregation of the chromosomal DNA (to which the LacI-mCherry are bound) to the daughter cells. This corresponds effectively to each molecule making a coin flip. In Fig. 3.9A and B, we show the partitioning error of fluorescence with area. On the y -axis the percent of the difference in partitioned area of each daughter at division, normalized by the total area of the two daughters, is shown. The x -axis shows the percent of the total difference in mCherry fluorescence between the two daughters divided by the total. If the protein was more likely to partition into bigger cells (because it has more volume), larger cells would have an increased probability of obtaining more protein and the cloud of points would tilt towards the upper right and lower left quadrants of Fig. 3.9. This behavior is seen in Fig. 3.9B for the partitioning of a cytoplasmic protein (CFP). The CFP results are consistent with previous reports where it was shown that for cytoplasmic proteins the error in volume partitioning on division can influence the “fairness” of the distribution [44]. However, the correlation we see for LacI-mCherry is very weak, indicating that volume partitioning fluctuations do not have a strong effect on the fluctuations in the partitioning of LacI-mCherry.

3.C.2 Derivation of calibration factor

It is of interest to have a simple derivation of the relation between fluorescence intensity and repressor number. To do this we exploit a convenient statistical property of binomial partitioning, namely if a mother cell had N_{tot} repressors and divided them randomly between two daughter cells, which now have N_1 and N_2 repressors respectively, then the variance in the total number of repressors in one daughter, N_1 , is $\sigma^2 = N_{\text{tot}}/4$. However, the variance can also be written,

$$\sigma^2 = \langle (N_1 - \langle N_1 \rangle)^2 \rangle, \quad (3.14)$$

$$= \left\langle \left(\frac{N_1 - N_2}{2} \right)^2 \right\rangle, \quad (3.15)$$

using $\langle N_1 \rangle = (N_1 + N_2)/2$. By combining these two expressions for the variance, we arrive at the final expression relating the total number of repressors in the daughters to the difference in that number,

$$\langle (N_1 - N_2)^2 \rangle = N_1 + N_2. \quad (3.16)$$

By assuming that the measured 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, we now find,

$$\langle (N_1 - N_2)^2 \rangle = N_1 + N_2 \Rightarrow \left\langle \left(\frac{I_1}{\alpha} - \frac{I_2}{\alpha} \right)^2 \right\rangle = \frac{I_1 + I_2}{\alpha} \quad (3.17)$$

$$\frac{1}{\alpha^2} \langle (I_1 - I_2)^2 \rangle = \frac{I_1 + I_2}{\alpha} \quad (3.18)$$

$$\Rightarrow \sqrt{\langle (I_1 - I_2)^2 \rangle} = \sqrt{\alpha(I_1 + I_2)}. \quad (3.19)$$

This gives us the relationship between the fluctuations in the difference between the intensities of two daughter cells and the total intensity present between the two daughters, $I_{tot} = I_1 + I_2$. 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 set of daughters plotting $\langle (I_1 - I_2)^2 \rangle$ versus $I_1 + I_2$. A more sophisticated treatment using information from tracking over multiple generations and the introduction of random errors can be found in reference [43].

3.C.3 Interpretation of $\langle (I_1 - I_2)^2 \rangle$

As noted above, the mathematical derivation for the error in partitioning is predicated on the idea that for a given value of $(I_1 + I_2)$, we have many division events to average over to arrive at a well-averaged value for the partitioning error $\langle (I_1 - I_2)^2 \rangle$. However, the data itself in the experimental case does not come in this convenient format. This raises the concern of how data will be binned. In practice, data is binned by fracturing the data into bins of a set number of data points. The data point corresponding to the bin is placed in the geometric center of the data comprising that bin such that data points in the bin fall with equal weight to the left and right of the bin center in log space.

In Fig. 3.9D we show the effect of choice in bin size on the calibration factor by plotting the calibration factor as a function of the number of points in each bin. Over the majority of the range of bin sizes the calibration factor is relatively insensitive to the bin size. However, when bins have few points the calibration factor is strongly affected by the presence of data points where, by chance, $\langle (I_1 - I_2)^2 \rangle \approx 0$, which weighs heavily on a log-log fit. It is interesting that the fit is not changed by making the size of the bins very large, thus averaging data over a larger range of $(I_1 + I_2)$. In this case, where the data should fit a straight line in log-log space, a point located a distance ϵ away from the bin center on either side is expected to contribute equal and opposite weight to the function value and thus should not change the fit.

3.C.4 Photon counting noise

One possible additional source of noise might be simply the Poisson noise corresponding to counting photons. Our camera is a Photometrics CoolSNAP ES² which has a linear full-well count of 13,500 electrons. This means that for this 12-bit output, a count on the camera corresponds to roughly 3.3 photons detected. As seen in Fig. 3.3 of the main text, a LacI typically corresponds to roughly 100 counts on the camera which means 330 photons counted per LacI. The Poissonian standard deviation for a single LacI is then 18 photons or 6% of a LacI. This is small even in the single LacI limit. When we have 10 LacI-mCherry molecules or more, the noise is lower than 2%. This error is smaller than any of the errors related to quantifying our fluorescence levels. As a result, Poisson statistics are not expected to influence the partitioning error significantly.

3.C.5 Limits in LacI-mCherry detection

In order to check our ability to distinguish low repressor copy numbers from cellular auto-fluorescence fluctuations, we examine the mCherry fluorescence signal of a collection of cells from our Δ LacI-mCherry control strain. On average these cells will have zero signal, once they are corrected for autofluorescence (each pixel has the average signal per pixel of the Δ LacI-mCherry strain subtracted) the remaining signal is, on average, 0. However, due to fluctuations the signal is typically not exactly 0 and instead has a distribution. This can be seen for an example experiment in Fig. 3.9E, when we histogram the mCherry signal from a collection of our Δ LacI-mCherry cells. As can be seen, the average is indeed 0, but the distribution has a standard deviation of 3 LacI-mCherry repressors. Therefore, we set our confidence regime for measuring LacI-mCherry signal in this experiment at 3 LacI-mCherry and do not consider cells which are measured as having less signal than this since our measurements show we cannot resolve the difference between 0 and 3. This detection limit is calculated in every experiment and that value is used as a threshold for all data from that experiment; we do not accept cells with an mCherry signal lower than our detection threshold.

3.C.6 Limits in YFP production detection

In a similar fashion to the LacI-mCherry detection threshold, the production measurements also have a lower limit of detection. We account for this in the fold-change vs. LacI number measurements by rejecting binned data points where the standard error is larger than the value of the point itself. In almost all cases, this threshold occurs at a fold-change in the range of 10^{-3} and 10^{-2} . Intuitively, this is the range where the fluctuations of the autofluorescence in YFP become significant. For instance, taking our autofluorescence measurements of YFP (static snapshots), normalized by the YFP of the Δ LacI strain, we find that the standard deviation of the fold-change in YFP of these cells is between 10^{-2} to 10^{-3} . This choice is designed to remove points without significant information

and does not affect the quality of the data. This limitation can be seen in Figs. 3.3-3.6 where the points typically cut off around a fold-change of 10^{-3} .

3.D qPCR measurement of average plasmid copy number

To measure the average copy number of the ColE1 and ColE1 Δ Rom plasmids we performed qPCR measurements. The primers we used target part of the YFP gene and the sequences for these primers are given in Table 3.1. The probe primer is ordered from Integrated DNA Technologies and the /56-FAM/, /ZEN/ and /3IABkFQ/ tags refer to modifications from parts of the ZEN internal quencher system.

A DNA sample to be used as a standard is obtained by Maxiprep (Qiagen Hi Speed Plasmid Maxi Kit) of the ColE1 Δ Rom plasmid which was further concentrated in a PCR purification column (QIAquick PCR Purification Kit). The final concentration of the stock plasmid is ≈ 600 ng/ μ L as measured by a Qubit fluorimeter (Invitrogen Qubit dsDNA HS Assay Kit). As a control to determine the purity of our purified plasmid stock from chromosomal DNA contamination, we also perform the same Maxiprep on a culture without the plasmid and find a final concentration of less than 5% of the measured plasmid concentration. Then, starting with a 16x dilution of the stock, we step down by factors of 4 to generate a standard dilution series; meaning we have 8 standard concentrations ranging from a 16x dilution of the stock down to a 10^6 dilution of the stock separated by factors of 4 in concentration.

For the qPCR measurement we start by growing the ColE1 Δ Rom and ColE1 cells in the same conditions as our cells used for microscopy measurements. However, we chose an aTc concentration (4ng/ μ L) which corresponds to a LacI concentration close to the transitional region of the fold-change curve [Fig. 3.7B]. We also grow a strain with no plasmid or YFP genes which will act as a background for the standard and make 8 samples out of this strain, one for each standard. When the cells are at the proper OD they are spun down, washed twice (exactly as described in Sec. 3.A.4) and finally resuspended in 200 μ L of Qiagen P1 lysis buffer without LyseBlue or RNaseA. We then add 1 μ L of the prepared pre-diluted standards to each control tube, such that the standards will undergo the exact same process as our samples to be measured. The cell mixtures are then set on ice while the cellular density in each sample is measured by hemocytometer chips (InCyto DHC-S01) under 10x phase magnification.

Meanwhile, 25 μ L of cells is then added to 25 μ L of Qiagen buffer P2 to lyse the cells. The cells are allowed to sit for 5 minutes. The cells are then diluted 1:100 into 1x NEB buffer 2 (1 μ L +99 μ L) and 20 μ L of that mixture is added to a thin walled PCR tube with 0.5 μ L HindIII (NEB) restriction enzyme. The mixture digests at 30°C for 30 minutes followed by heat deactivation for 20 minutes. This mixture is then diluted 1:10 in water. The final 20 μ L qPCR reaction consists

of: 4.2 μ L of template, 10 μ L Supermix (PerfeCTa MultiPlex qPCR SuperMix, Quanta BioSciences Cat. no. 95063-200), 0.4 μ L forward primer, 0.4 μ L Rox, 5 μ L water. The number of copies of the YFP gene are determined by comparing the measured CT of each sample and interpolating from the standard. Together with the knowledge of the number of cells in the sample (from the hemocytometer measurements) we arrive at an average copy number of the plasmid in our cells.

3.E The copy number of multiple chromosomal integrations strain

The genetic location (and position on the chromosome in minutes; where 1 minute = 1/100th of the *E. coli* chromosome and *oriC* is located at minute 85) of each specific integration is: *intS* (53 minutes), *yffO* (55 minutes), *intB* (97 minutes), *intE* (26 minutes), and *essQ* (35 minutes). There is some uncertainty in the number of copies of these genes at any given time in the cell cycle. We chose to make measurements at the end of the cell cycle because we know that there are two completed copies of the genome at that point in time [17]. However, the gene copy at *essQ* is directly opposite of the origin of replication, *oriC*, on the chromosome (50 minutes away) and is one of the last parts of the chromosome to be replicated during a round of replication. Therefore, although all of our measurements take place in the D period when the first round of chromosomal replication should be complete, fluorescent protein maturation times may make it such that the extra copy of *essQ* is not fully measurable yet. A second source of uncertainty comes from the fact that at 65 minutes division time, we expect that the next round of chromosome replication to have already begun by the end of the cell cycle. *intB* is a mere 12 minutes (or 600 kbp) away from *oriC*, the origin of chromosomal replication. Therefore it is possible that there are already 4 copies of the *intB* integration when we make our measurements. As a result, we estimate the range of chromosomal construct copy number during our measurements to be between 9 and 12, with 10 being most probable. As such, we expect that there is some cell-to-cell variation in copy number within our measurement. However this small, tight range would not cause a major correction to the predictions of the thermodynamic model. Fig. 3.10 shows the difference in theoretical predictions between assuming exactly 10 copies (red line, as is reported in Fig. 3.4B) and allowing a normal distribution centered on 10.5 copies with a standard deviation of 1.5 copies (black line). While the model of chromosome copy number as a normal distribution is not correct in detail, we intend to show an upper limit on the effect of copy number distribution on our predictions.

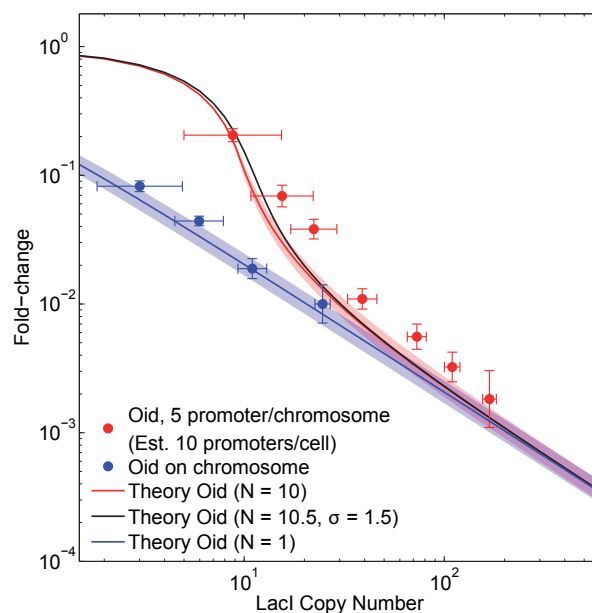


Figure 3.10: **Effect of copy number distribution on multiple chromosomal integration theory, related to Fig. 3.4B.**

Fold-change versus LacI copy number for the multiple integration strain (red points) overlaid with theoretical predictions from the thermodynamic model using a fixed value for the promoter copy number $N = 10$ (red line), as used in the main text, and using a normal distribution for promoter copy number centered around 10.5 with a standard deviation of 1.5 (black line). This choice of distribution is only meant to approximate the effect of chromosomal copy number variation within our expected copy number range of 9 – 12 and demonstrate the relatively small effect of the distribution for this case. For reference, the theory predictions and data for the single integrated copy case are shown as blue.

3.F Additional theoretical details of the thermodynamic model

3.F.1 Equivalence of fold-change in steady-state measurements and video microscopy

In bulk, the fold-change is calculated by comparing the steady state fluorescence, P , of cells with repressor to the fluorescence of those without repressor. To determine this steady-state fluorescence, we consider the rate of production of the fluorescent reporter,

$$\frac{dP}{dt} = rp_{\text{bound}} - \gamma P, \quad (3.20)$$

where r is the rate of production, p_{bound} is the probability that the promoter is occupied by RNAP and γ is the degradation rate. In steady-state, we find

$$P = \frac{rp_{\text{bound}}}{\gamma}, \quad (3.21)$$

which implies that the fold-change in steady-state experiments can be written as

$$\text{fold-change} = \frac{P(R \neq 0)}{P(R = 0)} = \frac{p_{\text{bound}}(R \neq 0)}{p_{\text{bound}}(R = 0)}. \quad (3.22)$$

The right hand side is a quantity that is directly calculable in the thermodynamic framework [7, 8, 41, 39]. However, over the timescales of our experiments, YFP is stable (i.e. $rp_{\text{bound}} \gg \gamma$) [69]. As a result, the rate of fluorescent increase we measure in video microscopy is simply rp_{bound} . This implies that we can write the fold-change in our experiments as

$$\text{fold-change} = \frac{\frac{dP}{dt}(R \neq 0)}{\frac{dP}{dt}(R = 0)} = \frac{p_{\text{bound}}(R \neq 0)}{p_{\text{bound}}(R = 0)}, \quad (3.23)$$

and thus the comparison of fold-change as measured in steady-state experiments should be directly comparable to that measured as a production rate in video microscopy and to the theoretical predictions of the thermodynamic theory which calculated p_{bound} .

3.F.2 Thermodynamic model in the limit $R \gg N$ or $R \gg N_c$

Equations 3.4 and 3.5 from the main text predict the fold-change in expression as a function of the number of binding sites available to the repressor (N or N_c , respectively). However, when the number of repressors is much larger than the number of binding sites available, such that the

approximation $R!/(R - N)! \approx R^N$ is valid, these equations immediately simplify to,

$$\text{fold-change} = \frac{1}{1 + \frac{R}{N_{\text{NS}}} e^{-\Delta\epsilon/k_B T}}, \quad (3.24)$$

identical to the prediction for a single isolated copy of the gene in Eq. (3.3).

3.F.3 Accounting for chromosome replication in competitor theory

In the theoretical predictions of Eq. (3.5) it is assumed that the reporter gene integrated into the chromosome exists at only a single copy. This introduces an error in our calculation of N_c during the portion of the cell cycle where two copies of the reporter gene exist. This error does not come from the presence of an extra copy of the gene producing more of the reporter gene product; measuring fold-change ensures that we are normalizing by cells expressing with the same average copy number of the gene. However, the addition of a new operator site associated with the chromosomal gene copy will change the expression profile by contributing to the demand for repressor and this will be interpreted in our measurement as a larger value for of N_c . The general formula to derive this effect follows from the partition function,

$$Z = \sum_{r_c=0}^{\min(R, N_c)} \sum_{r_{\text{int}}=0}^{\min(R-r_c, N_r)} \frac{R!}{N_{\text{NS}}^{(r_c+r_{\text{int}})} (R-r_c-r_{\text{int}})!} Z_{r_c}^c Z_{r_{\text{int}}}^{\text{int}}, \quad (3.25)$$

with $Z_i^c = \binom{N_c}{i} \exp(-\beta i \Delta\epsilon_c)$ and $Z_i^{\text{int}} = \binom{N_{\text{int}}}{i} \exp(-\beta i \Delta\epsilon_{\text{int}}) (1+p)^{(N_{\text{int}}-i)}$ where N_{int} is the number of integrated copies that exist on the chromosome and N_c is the number of competitor plasmids, $\Delta\epsilon_{\text{int}}$ and $\Delta\epsilon_c$ are the repressor binding energies to the chromosomal operator and plasmid operator, respectively, and finally $p = (n_P/N_{\text{NS}}) \exp(-\Delta\epsilon_p/k_B T)$ where n_P is the number of RNAP, and $\Delta\epsilon_p$ is the energy of polymerase binding to the promoter. The fold-change is then,

$$\text{fold-change} = \frac{\partial_p \ln(Z)}{\partial_p \ln(Z_{R=0})}. \quad (3.26)$$

For our particular experiments, the integrated copy begins at a single copy and doubles over the course of the cell cycle. Fig. 3.11 shows the predicted fold-change for an integrated O1 promoter with $N_{\text{int}} = 1$ (solid lines) or 2 (dashed lines) and a competitor plasmid with $N_c = 64$ and an O1 operator site identical to the chromosomal operator, an O2 operator site weaker than the chromosomal operator, and an Oid operator site stronger than the chromosomal operator. In all cases the predicted change between one and two integrated gene copies is small.

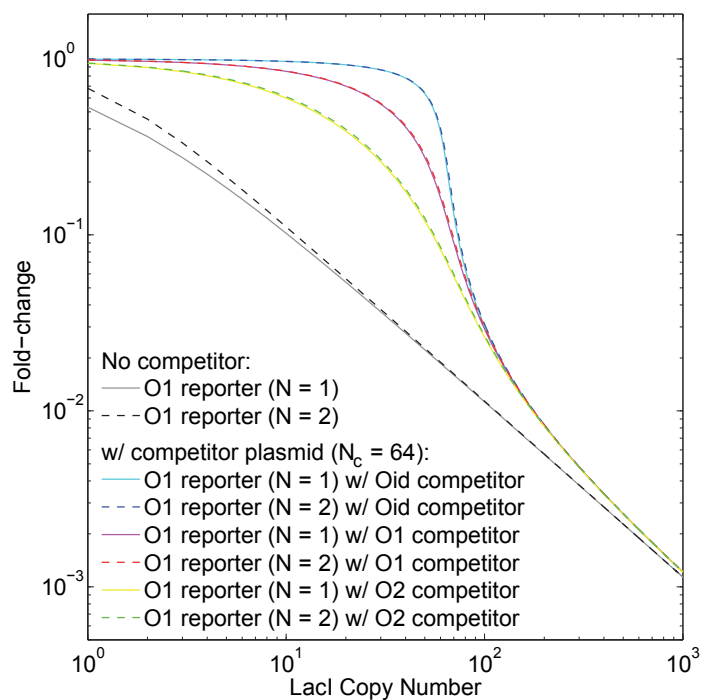


Figure 3.11: **Comparison of fold-change predictions for multiple chromosomal gene copies, related to Figs. 3.3 and 3.6.**

The fold-change predicted from 1 or 2 chromosomal copies with an O1 operator site either without a competing plasmid (grey solid and black dashed line) or competing with a high copy number plasmid $N_c = 64$ and bearing an O1 site (purple solid and red dashed line), O2 site (yellow solid and green dashed line), or Oid site (light blue solid and dark blue dashed line). In this case, simply going from 1 (solid lines) to 2 (dashed lines) copies of the chromosome does not change the expected fold-change significantly.

3.F.4 Thermodynamic model with plasmid distribution

The fold-change predictions in Eqs. (3.4) and (3.5) of the main text are derived by assuming that any given cell has exactly N plasmids. However, our measurements are averaged over many different cells and thus we do not expect the copy number of the plasmid to be exactly the same in every cell. While this static single parameter characterization of the copy number is sufficient to predict the fold-change repression titration curve in most of the cases we examine, we wish to determine how the reality of the plasmid distribution changes our predictions. To begin, we rewrite the fold-change in terms of expression measurements for the case of a static number of plasmids, N ,

$$\text{fold-change} = \frac{\text{expression}(R, N)}{\text{expression}(R = 0, N)}. \quad (3.27)$$

However, if there is a distribution of plasmids $p(n)$ then the expression is the sum of the probability of finding a cell with N plasmids times the expression from a cell with N plasmids such that,

$$\text{fold-change}_{\text{dist}} = \frac{\sum_{n=1}^{\infty} p(n)\text{expression}(R, n)}{\sum_{n=1}^{\infty} p(n)\text{expression}(R = 0, n)}. \quad (3.28)$$

First let's examine this in terms of a single chromosomal copy expressing YFP in the presence of competitor plasmids which do not express the measured gene product (corresponding to Eq. (3.5)). In this case when there is no repressor, the number of plasmids is irrelevant to the predicted expression. Now we can rewrite

$$\text{fold-change}_{\text{dist}} = \frac{\sum_{n=1}^{\infty} p(n)\text{expression}(R, n)}{\text{expression}(R = 0)}, \quad (3.29)$$

which can be rewritten as

$$\text{fold-change}_{\text{dist}} = \sum_{n=1}^{\infty} p(n)\text{fold-change}(R, n), \quad (3.30)$$

where $\text{fold-change}(R, n)$ is the expression for the fold-change of a static, fixed- N plasmid from Eq. (3.5) from the main text, and the above equation is listed in the main text as Eq. (3.9). The situation differs slightly when one considers, instead, identical genes expressing the same measured gene product. Now the expression of the $R = 0$ strain (in the denominator of the fold-change) does depend on the number of plasmids; the expression of a cell with n plasmids is equivalent to n times the production of a cell with just one plasmid. As such we rewrite Eq. (3.28),

$$\text{fold-change}_{\text{dist}} = \frac{\sum_{n=1}^{\infty} p(n)\text{expression}(R, n)}{\langle N \rangle \text{expression}(R = 0, N = 1)}, \quad (3.31)$$

and by breaking up the above sum term by term we see the same equivalence, $\text{expression}(R = 0, N = n) = n \times \text{expression}(R = 0, N = 1)$, allows us to arrive at Eq. (3.8) from the main text,

$$\text{fold-change}_{\text{dist}} = \sum_{n=1}^{\infty} p(n) \frac{n}{\langle n \rangle} \text{fold-change}(R, n), \quad (3.32)$$

where $\langle n \rangle = \sum_{n=1}^{\infty} np(n)$ and $\text{fold-change}(R, n)$ is the fold-change from a static fixed- N distribution from Eq. (3.4) of the main paper.

3.F.5 Determining errors in theoretical predictions

Figures of fold-change vs. repressor copy number often show the standard deviation in theoretical predictions stemming from uncertainty in the parameters of the model such as operator binding energies $\Delta\epsilon$, $\Delta\epsilon_c$, gene copy number N , or competitor binding site copy number N_c while assuming the repressor copy number is fixed. We estimate the standard deviation in fold-change by a first order Taylor expansion around the mean values of these parameters, $\overline{\Delta\epsilon}$, $\overline{\Delta\epsilon_c}$, \overline{N} , $\overline{N_c}$. For instance, calculating the error bars for Fig. 3.5A where the uncertainties in $\Delta\epsilon$, $\Delta\epsilon_c$ and N are all included, the calculation goes as follows,

$$\begin{aligned} \text{fold-change}(R, \Delta\epsilon, \Delta\epsilon_c, N_c) &\approx (\Delta\epsilon - \overline{\Delta\epsilon}) \frac{\partial}{\partial \Delta\epsilon} \text{fold-change}(R, \overline{\Delta\epsilon}, \overline{\Delta\epsilon_c}) \\ &\quad + (\Delta\epsilon_c - \overline{\Delta\epsilon_c}) \frac{\partial}{\partial \Delta\epsilon_c} \text{fold-change}(R, \overline{\Delta\epsilon}, \overline{\Delta\epsilon_c}) \\ &\quad + (N_c - \overline{N_c}) \frac{\partial}{\partial N_c} \text{fold-change}(R, \overline{\Delta\epsilon}, \overline{\Delta\epsilon_c}, \overline{N_c}) \end{aligned} \quad (3.33)$$

which gives us the corresponding estimated variance in fold-change

$$\begin{aligned} V[\text{fold-change}(R, \Delta\epsilon, \Delta\epsilon_c, N_c)] &\approx V[\Delta\epsilon] \left(\frac{\partial}{\partial \Delta\epsilon} \text{fold-change}(R, \overline{\Delta\epsilon}, \overline{\Delta\epsilon_c}) \right)^2 \\ &\quad + V[\Delta\epsilon_c] \left(\frac{\partial}{\partial \Delta\epsilon_c} \text{fold-change}(R, \overline{\Delta\epsilon}, \overline{\Delta\epsilon_c}) \right)^2 \\ &\quad + V[N_c] \left(\frac{\partial}{\partial N_c} \text{fold-change}(R, \overline{\Delta\epsilon}, \overline{\Delta\epsilon_c}, \overline{N_c}) \right)^2, \end{aligned} \quad (3.34)$$

where we used the additional assumption of no correlation between any of the expanded parameters. The derivatives in Eq. (3.34) can be computed either numerically or analytically using standard mathematical software. To be explicit, here we list the relevant figures and the parameters which contribute to the uncertainty. Fig. 3.3 has uncertainty stemming only from uncertainty in the binding energy $\Delta\epsilon$. Fig. 3.4A has uncertainty from both $\Delta\epsilon$ and the copy number of the reporter plasmid N , while in part B of that figure we use only the error from $\Delta\epsilon$. Fig. 3.5A has uncertainty

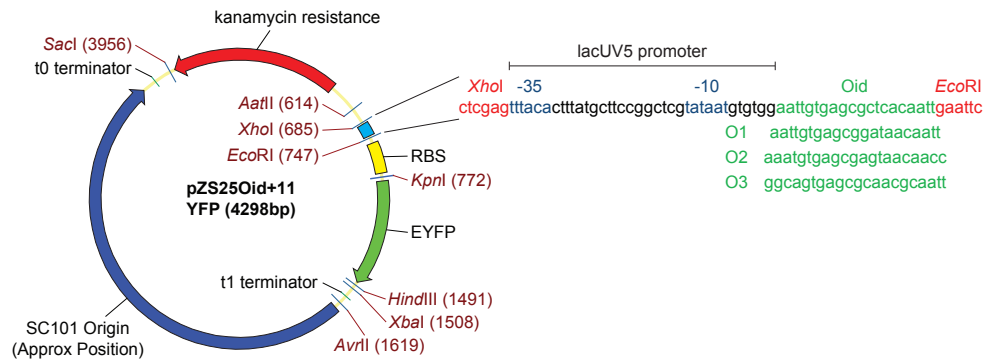


Figure 3.12: **Plasmid diagram and promoter sequence, related to Fig. 3.1.**

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. This plasmid was used as a basis for creating the plasmids and chromosomal integration reporters for this study.

contributions from $\Delta\epsilon$, as well as the binding strength and number of competitor plasmids, $\Delta\epsilon_c$ and N_c . In Fig. 3.5B, the distribution is initially fit to the Oid data and thus the only uncertainty shown there is due to $\Delta\epsilon$ and $\Delta\epsilon_c$.

3.G Constructs and strains

The base strain through this work is HG105, which is MG1655 with a *LacIZYA* deletion [59]. A constitutive CFP marker has been integrated at the *gspI* chromosomal location [70]. The marker is expressed from a *lacUV5* promoter with no Lac repressor binding site. In addition, every strain contains a low copy number plasmid which expresses TetR pZS3P_{N25}-tetR. This is a plasmid with a sc101 origin and a P_{N25} promoter controlling the TetR gene and was obtained by PCR from the chromosome of DH5 α Z1 [45]. Specific details of the construction of the individual strains used in each part of the experiment are now listed below:

- **Single copy chromosomal integration:** This originates from plasmid pZS25O1+11-YFP (map shown in Fig. 3.12). From this plasmid, we have produced, by site-directed mutagenesis, the same plasmid with the Oid, O2 and O3 repressor binding sites in place of O1 (sequences listed in Table 3.12) [59]. These constructs, consisting of the terminators, resistance marker, and EYFP gene are integrated into the chromosomal location of *galK* using recombineering [71, 59] with primers listed in the table below.
- **High copy number plasmids:** The sc101 origin of plasmids pZS25O1+11-YFP was removed by digestion with *SacI* and *AvrII* and ligated to a *ColE1* Δ Rom origin to make pZE25O1+11-YFP

[45]. This procedure was repeated for plasmids with the binding sites Oid, O2 and O3. To create the ColE1 origin, we have added the Rom protein near the origin of the pZE25O1+11-YFP plasmid to make pRE25O1+11-YFP. This is achieved by PCR of the Rom protein from plasmid pBR322 followed by Gibson assembly with plasmid pZE25O1+11-YFP to make our ColE1 plasmid.

- Multiple chromosomal integrations: The plasmids pZS2*5Oid+11-YFP, pZS3*5Oid+11-YFP and pZS4*5Oid+11-YFP contain resistance genes for kanamycin, chloramphenicol and spectinomycin, respectively. These resistance genes are flanked by FLIP recombinase sites. The kanamycin and chloramphenicol cassettes were obtained by PCR from plasmids pKD4 and pKD3, respectively [72] and placed between the SacI and AatII sites of pZS25Oid+11-YFP. FLIP recombinase sites were placed flanking the spectinomycin resistance gene in pZS4*5Oid+11-YFP by site-directed mutagenesis on pZS45Oid+11-YFP using primers 15.15 and 15.16 [Table 3.1]. These constructs were integrated into the chromosomal locations of genes *intS*, *yffO*, *intB*, *intE*, and *essQ* [70, 56] using recombineering [71]. The oligos used to amplify the pZS plasmid to integrate constructs at every chromosomal location are listed in Table 3.1. All resistances are then flipped out by FLP recombinase transiently expressed from plasmid pCP20 [72].
- Competitor plasmids: These plasmids are made from the pZE25O1+11-YFP plasmid digested with AatII and XbaI. An insert containing Oid, O1, or O2 flanked with sticky ends for the same restriction sites (sequence of inserts listed below) are ordered as annealed double stranded oligos (Integrated DNA Technologies) and then ligated into the pZE vector. The result is a plasmid with the ColE1 Δ Rom origin of replication, a resistance marker and a LacI binding site without an active YFP gene or promoter.
- Constitutive marker: The cerulean (CFP in this work) gene was obtained from [73], amplified using primers 15.14 and 15.14R [Table 3.1], and ligated between the KpnI and HindIII sites of pZS4*5O1+11-YFP to create pZS4*5O1+11-CFP. The O1 binding site was deleted using mutagenesis primer 21.3 [Table 3.1] [38] in order to create pZS4*5NoO1-CFP. This construct was integrated into the *gspI* gene.
- TetR plasmid: The *tetR* gene controlled by the pN25 promoter was amplified from the genome of DH5 α Z1 [45] using primers 13.6 and 13.7v2 [Table 3.1]. The PCR product was digested between the XhoI and HindIII of pZS3*1-LacI to create pZS3P_{N25}-tetR.
- LacI-mCherry fusion: A construct bearing mCherry was obtained from [74] and amplified using primers 13.12 and 13.13 [Table 3.1]. The *LacI* gene was amplified from pZS3*1-LacI [59] using primers 13.28 and 13.30. Both of these PCR products were combined and amplified once again

using primers 13.28 and 13.13 [Table 3.1]. The resulting LacI-mCherry PCR product has a KpnI site on its 5' end and a HindIII site on its 3' end. This repressor cannot tetramerize due to the deletion of the last 11 amino acids of its sequence. The fusion was ligated between the KpnI and HindIII sites of pZS3*1-LacI to create pZS3*1-LacI-mCherry. Finally this construct was integrated into the chromosome at the *ybcN* chromosomal location with the *ybcN* primers listed below.

3.H Primers used in this study to create strains

Name	Sequence
Chromosomal integrations	
galK<>res	TTCATATTGTTTCAGCGACAGCTTGCTGTACGGCAGGCACCAGCTCTTCGGGGCTAATGCACCCAGTAAGG
galK<>YFP	GTTTGCGCGCAGTCAGCGATATCCATTTTCGCGAATCCGGAGTGTAAAGAACTAGCAACACCAGAACAGCC
gspI<>res	TGCCAGAACTGGACGTGTTTTCTCGCCGAATGAATCTTGACTGAAGCGGCTAATGCACCCAGTAAGG
gspI<>YFP	TCAAACGCTCGCCAGAGATACCCGCCATGAACAACAATCAGGGATGACACTAGCAACACCAGAACAGCC
intS<>res	ATAGTTGTTAAGTTCGCTCACTCCACCTTCTCATCAAGCCAGTCCGCCAGGCTAATGCACCCAGTAAGG
intS<>YFP	CCGTAGATTTACAGTTCGTCATGGTTCGCTTCAGATCGTTGACAGCCGCACTAGCAACACCAGAACAGCC
yffO<>res	TTTCAAATATTACAGCTTGGCTGCTGCCAGTAGTGCCCTTGCCTTTGCTTGGCTAATGCACCCAGTAAGG
yffO<>YFP	GGTGGAATCATGAAACACGTTTTTAAATATCTTGATTTTGAGAAAGACCCACTAGCAACACCAGAACAGCC
intB<>res	ACGCATATTTACCGTATTCTCACTCATGGGTTTGTGCGAATCGTGATCAGGCTAATGCACCCAGTAAGG
intB<>YFP	TGTCCATCCAATGGTTCTAAGTACTGGCGTTTGCAGTACCGTTATGAGGACTAGCAACACCAGAACAGCC
intE<>res	CAAGCGATCCAGGATGACAGGCTTAAAAGTGGTGATATAAGACTCAACACGGCTAATGCACCCAGTAAGG
intE<>YFP	TCACAACGCTACTTTGCTCCATCCTTTACCTCGATCATCATGATAACGATACTAGCAACACCAGAACAGCC
essQ<>res	TAAGGCTACAGTTACCGTAACTTATCTCAAATACGGACTCCTTTCAGGCTAATGCACCCAGTAAGG
essQ<>YFP	TAAAGGTCCTGCAGCAGCAAATGTCATCTACTGATTAATAATTCATCGCACTAGCAACACCAGAACAGCC
Mutagenesis	
15.15	GGAGTCCAAGCGAGCTCAGTTCCTATTCCGAAGTTCCTATT CTCTAGAAAGTATAGGAACTTCGATATCCGTCGGCTTGAACG
15.16	GGTTCGTGCCTTCATCGAAGTTCCTATTCCGAAGTTCCTATT CTCTAGAAAGTATAGGAACTTCATATCGACGCTAAGAAACC
21.3	CCGGCTCGTATAATGTGTGGGATTGTTAGC GGAGAAGAATTGAATTCATTAAGAGGAG
LacI-mCherry fusion	
13.12	ATTATTGGTACCGCATGGTTTCCAAGGGCGAGGAGG
13.13	ATATCTAAAGCTTATTTGTACAGCTCATCCATGCCACC
13.28	ATTATAGGTACCATATGGTGAATGTGAAACCAGTAAC
13.30	CTCGCCCTTGAAACCATCACCAGTTCAGGCCGCCAGCTGCATTAATGAATCGGCCAA
CFP amplification	
15.14	ATTATTGGTACCGCATGACTAGCAAAAGAAGCAAAGGTG
15.14R	ATAATATAAGCTTTATACAGTTCATCCATGCCACG
TetR cloning	
13.6	ATACAAAAGCTTAAGACCCACTTTCACATTTAAGTTGTT
13.7v2	ATACAAActegagGCGCAACGCAATTAATGTAAGTTAGC
Competitor plasmid	
O1Add-F	CTAGACTCAGCTAATTAAGAATTGTTATCCGCTCACAATTATAATGGTTTCTTAGACGT
O1Add-R	CTAAGAAACCATTATAATTGTGAGCGGATAACAATTCTTAATTAGCTGAGT
O2Add-F	CTAGACTCAGCTAATTAAGGGTTGTTACTCGCTCACATTTATAATGGTTTCTTAGACGT
O2Add-R	CTAAGAAACCATTATAAATGTGAGCGAGTAACAACCCTTAATTAGCTGAGT
OidAdd-F	CTAGACTCAGCTAATTAAGAATTGTGAGCGCTCACAATTATAATGGTTTCTTAGACGT
OidAdd-R	CTAAGAAACCATTATAATTGTGAGCGCTCACAATTCTTAATTAGCTGAGT
qPCR primers	
forward	CAGTGGAGAGGGTGAAGGTG
reverse	GTGTCTTGTAGTTCCCGTCAT
probe	/56-FAM/TCAAGAGTG/ZEN/CCATGCCCGAAGGT/3IABkFQ/

Table 3.1: Primers used in this study to create strains, related to Fig. 3.1.

The names of the chromosomal integration primers are formatted with the gene location followed by the side of the plasmid it binds to (the resistance or the FP reporter; see 3.12) with <> between. The red bases bind to the plasmid to amplify and the black bases are homologous to the integration site on the chromosome.

References

- [1] Brewster, R. C. *et al.* The transcription factor titration effect dictates level of gene expression. *Cell* **156**, 1312–1323 (2014).
- [2] Lim, W. A. The modular logic of signaling proteins: building allosteric switches from simple binding domains. *Curr Opin Struct Biol* **12**, 61–8 (2002).
- [3] Ptashne, M. & Gann, A. *Genes and Signals* (Cold Spring Harbor Laboratory Press, New York, 2002).
- [4] Bhattacharyya, R. P., Remenyi, A., Yeh, B. J. & Lim, W. A. 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).
- [5] Kentner, D. & Sourjik, V. Use of fluorescence microscopy to study intracellular signaling in bacteria. *Annu Rev Microbiol* **64**, 373–90 (2010).
- [6] Garcia, H. G., Sanchez, A., Kuhlman, T., Kondev, J. & Phillips, R. Transcription by the numbers redux: Experiments and calculations that surprise. *Trends Cell Biol* (2010).
- [7] Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A* **100**, 5136–41 (2003).
- [8] Vilar, J. M. & Leibler, S. DNA looping and physical constraints on transcription regulation. *J Mol Biol* **331**, 981–9 (2003).
- [9] Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–92 (2005).
- [10] Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & Van Oudenaarden, A. Multistability in the lactose utilization network of *Escherichia coli*. *Nature* **427**, 737–40 (2004).
- [11] Kuhlman, T., Zhang, Z., Saier, J., M. H. & Hwa, T. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc Natl Acad Sci U S A* **104**, 6043–8 (2007).

- [12] Kinney, J. B., Murugan, A., C. G. Callan, J. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**, 9158–63 (2010).
- [13] Daber, R., Sochor, M. A. & Lewis, M. Thermodynamic analysis of mutant lac repressors. *J. Mol. Biol.* **409**, 76–87 (2011).
- [14] Garcia, H. G. & Phillips, R. Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences of the United States of America* (2011). Under review.
- [15] Luria, S. E. & Dulbecco, R. Genetic recombinations leading to production of active bacteriophage from ultraviolet inactivated bacteriophage particles. *Genetics* **34**, 93–125 (1949).
- [16] Guido, N. J. *et al.* A bottom-up approach to gene regulation. *Nature* **439**, 856–60 (2006).
- [17] Bremer, H. & Dennis, P. P. Modulation of chemical composition and other parameters of the cell by growth rate. In al., N. F. e. (ed.) *Escherichia coli and Salmonella Cellular and Molecular Biology*, 1553–1569 (ASM Press, Washington DC, 1996).
- [18] Wang, S., Liu, N., Peng, K. & Zhang, Q. The distribution and copy number of copia-like retrotransposons in rice (*Oryza sativa L.*) and their implications in the organization and evolution of the rice genome. *Proc. Natl. Acad. Sci.* **96**, 6824–6828 (1999).
- [19] Navarro-Quezada, A. & Schoen, D. J. Sequence evolution and copy number of Ty1-copia retrotransposons in diverse plant genomes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 268–273 (2002).
- [20] Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
- [21] Hanada, K. *et al.* Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **28**, 377–382 (2011).
- [22] Busby, S. & Ebright, R. H. Transcription activation by catabolite activator protein (CAP). *J Mol Biol* **293**, 199–213 (1999).
- [23] Endy, D. Foundations for engineering biology. *Nature* **438**, 449–53 (2005).
- [24] Voigt, C. A. Genetic parts to program bacteria. *Curr Opin Biotechnol* **17**, 548–57 (2006).
- [25] Mukherji, S. & van Oudenaarden, A. Synthetic biology: understanding biological design from synthetic circuits. *Nat Rev Genet* **10**, 859–71 (2009).
- [26] Elowitz, M. & Lim, W. A. Build life to understand it. *Nature* **468**, 889–90 (2010).

- [27] Cox III, R. S., Surette, M. G. & Elowitz, M. B. Programming gene expression with combinatorial promoters. *Mol Syst Biol* **3**, 145 (2007).
- [28] Kaplan, S., Bren, A., Zaslaver, A., Dekel, E. & Alon, U. Diverse two-dimensional input functions control bacterial sugar genes. *Mol Cell* **29**, 786–92 (2008).
- [29] Del Vecchio, D., Ninfa, A. J. & Sontag, E. D. Modular cell biology: Retroactivity and insulation. *Mol. Syst. Biol.* **4**, 161 (2008).
- [30] Kim, K. H. & Sauro, H. M. Measuring retroactivity from noise in gene regulatory networks. *Biophys. J.* **100**, 1167–1177 (2011).
- [31] Ricci, F., Vallee-Belisle, A. & Plaxco, K. W. High-precision, in vitro validation of the sequestration mechanism for generating ultrasensitive dose-response curves in regulatory networks. *PLoS Comput. Biol.* **7**, e1002171 (2011).
- [32] Lee, T. H. & Maheshri, N. A regulatory role for repeated decoy transcription factor binding sites in target gene expression. *Mol. Syst. Biol.* **8**, 576 (2012).
- [33] Babu, M. M. & Teichmann, S. A. Functional determinants of transcription factors in *Escherichia coli*: Protein families and binding sites. *Trends Genet* **19**, 75–9 (2003).
- [34] Gama-Castro, S. *et al.* RegulonDB version 7.0: Transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res* **39**, D98–105 (2011).
- [35] Schlax, P. J., Capp, M. W. & Record, M. T., Jr. Inhibition of transcription initiation by lac repressor. *J Mol Biol* **245**, 331–50 (1995).
- [36] Rojo, F. Mechanisms of transcriptional repression. *Curr Opin Microbiol* **4**, 145–51 (2001).
- [37] Sanchez, A., Osborne, M. L., Friedman, L. J., Kondev, J. & Gelles, J. Mechanism of transcriptional repression at a bacterial promoter by analysis of single molecules. *EMBO J* **30**, 3940–6 (2011).
- [38] Oehler, S., Amouyal, M., Kolkhof, P., von Wilcken-Bergmann, B. & Müller-Hill, B. Quality and position of the three *lac* operators of *E. coli* define efficiency of repression. *EMBO J* **13**, 3348–55 (1994).
- [39] Bintu, L. *et al.* Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev* **15**, 125–35 (2005).

- [40] Rydenfelt, M., Cox, R. S., Garcia, H. & Phillips, R. Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. *Phys. Rev. E* **89**, 012702 (2014). URL <http://link.aps.org/doi/10.1103/PhysRevE.89.012702>.
- [41] Bintu, L. *et al.* Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev* **15**, 116–24 (2005).
- [42] Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962–5 (2005).
- [43] Rosenfeld, N., Perkins, T. J., Alon, U., Elowitz, M. B. & Swain, P. S. A fluctuation method to quantify in vivo fluorescence data. *Biophys J* (2006).
- [44] Teng, S. W. *et al.* Measurement of the copy number of the master quorum-sensing regulator of a bacterial cell. *Biophys J* **98**, 2024–31 (2010).
- [45] Lutz, R. & Bujard, H. 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**, 1203–10 (1997).
- [46] Lee, C., Kim, J., Shin, S. G. & Hwang, S. Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. *J Biotechnol* **123**, 273–80 (2006).
- [47] Lee, C. L., Ow, D. S. & Oh, S. K. Quantitative real-time polymerase chain reaction for determination of plasmid copy number in bacteria. *J Microbiol Methods* **65**, 258–67 (2006).
- [48] Twigg, A. J. & Sherratt, D. Trans-complementable copy-number mutants of plasmid ColE1. *Nature* **283**, 216–218 (1980).
- [49] Cesareni, G., Muesing, M. A. & Polisky, B. Control of ColE1 DNA replication: The rop gene product negatively affects transcription from the replication primer promoter. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6313–6317 (1982).
- [50] Stueber, D. & Bujard, H. Transcription from efficient promoters can interfere with plasmid replication and diminish expression of plasmid specified genes. *Embo J* **1**, 1399–404 (1982).
- [51] Ng, J. W., Chatenay, D., Robert, J. & Poirier, M. G. Plasmid copy number noise in monoclonal populations of bacteria. *Phys Rev E Stat Nonlin Soft Matter Phys* **81**, 011909 (2010).
- [52] Paulsson, J. & Ehrenberg, M. Noise in a minimal regulatory network: Plasmid copy number control. *Q Rev Biophys* **34**, 1–59 (2001).
- [53] Brewster, R., Jones, D. & Phillips, R. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLOS Computational Biology* **8**, e1002811 (2012).

- [54] Ebersbach, G. & Gerdes, K. Plasmid segregation mechanisms. *Annu Rev Genet* **39**, 453–79 (2005).
- [55] Ghosh, S. K., Hajra, S., Paek, A. & Jayaram, M. Mechanisms for chromosome and plasmid segregation. *Annu. Rev. Biochem.* **75**, 211–241 (2006).
- [56] Kuhlman, T. E. & Cox, E. C. Gene location and DNA density determine transcription factor distributions in *Escherichia coli*. *Mol Syst Biol* **8**, 610 (2012).
- [57] Michelsen, O., Teixeira de Mattos, M. J., Jensen, P. R. & Hansen, F. G. Precise determinations of C and D periods by flow cytometry in *Escherichia coli* K-12 and B/r. *Microbiology* **149**, 1001–10 (2003).
- [58] Ghozzi, S., Ng, J. W., Chatenay, D. & Robert, J. Inference of plasmid-copy-number mean and noise from single-cell gene expression data. *Phys Rev E Stat Nonlin Soft Matter Phys* **82**, 051916 (2010).
- [59] Garcia, H. G., Lee, H. J., Boedicker, J. Q. & Phillips, R. The limits and validity of methods of measuring gene expression for the testing of quantitative models. *Biophysical Journal* (2011). Under review.
- [60] Edelstein, A., Amodaj, N., Hoover, K., Vale, R. & Stuurman, N. *Computer Control of Microscopes Using Manager*. Current Protocols in Molecular Biology (John Wiley & Sons, Inc., 2010).
- [61] Young, J. W. *et al.* Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nat Protoc* **7**, 80–8 (2012).
- [62] Martin, R. G., Bartlett, E. S., Rosner, J. L. & Wall, M. E. Activation of the *Escherichia coli* marA/soxS/rob regulon in response to transcriptional activator concentration. *J Mol Biol* **380**, 278–84 (2008).
- [63] Choi, P. J., Cai, L., Frieda, K. & Xie, X. S. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science* **322**, 442–6 (2008).
- [64] Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533 (2010).
- [65] Hirschberg, K. *et al.* Kinetic analysis of secretory protein traffic and characterization of Golgi to plasma membrane transport intermediates in living cells. *J Cell Biol* **143**, 1485–503 (1998).
- [66] Piston, D. W., Patterson, G. H. & Knobel, S. M. Quantitative imaging of the green fluorescent protein (GFP). *Methods Cell Biol* **58**, 31–48 (1999).

- [67] Sourjik, V. & Berg, H. C. 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**, 12669–74 (2002).
- [68] Gregor, T., Tank, D. W., Wieschaus, E. F. & Bialek, W. Probing the limits to positional information. *Cell* **130**, 153–64 (2007).
- [69] Andersen, J. B. *et al.* New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl Environ Microbiol* **64**, 2240–6 (1998).
- [70] Posfai, G. *et al.* Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**, 1044–6 (2006).
- [71] Sharan, S. K., Thomason, L. C., Kuznetsov, S. G. & Court, D. L. Recombineering: a homologous recombination-based method of genetic engineering. *Nat Protoc* **4**, 206–23 (2009).
- [72] Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A* **97**, 6640–5 (2000).
- [73] Dunlop, M., Cox III, R., Levine, J., Murray, R. & Elowitz, M. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics* **40**, 1493–1498 (2007).
- [74] Eldar, A. *et al.* Partial penetrance facilitates developmental evolution in bacteria. *Nature* **460**, 510–4 (2009).

Chapter 4

The transcription factor titration effect in a system of two coregulated genes

*This chapter presents **preliminary results** for an experiment suggested by the theoretical considerations presented in Chapter 2. The experiment was designed and conducted in collaboration with Robert Brewster.*

4.1 Introduction

In Chapter 3 we carefully examined the effect of transcription factor (TF) titration on the expression of a repressed gene that existed in multiple copies. We found that the regulatory function has a sharp transition [1, 2, 3] as the number of TFs grows large enough to simultaneously repress all gene copies. In *E. coli*, most TFs regulate not only a single gene but a family of genes, as is shown in Fig. 4.1 using data from RegulonDB [4]. As an extreme example, the cAMP receptor protein regulates almost 500 genes. A natural generalization of the experiment in Chapter 3 is therefore to study the effect of TF titration in a setting where multiple genes share the same TF [5]. This allows us not only to study fold change in gene expression but also the correlation in expression between different genes [6, 7, 8].

To demonstrate the effect of TF titration on transcriptional correlation, we need to bear in mind that TFs stay bound to their binding sites only for a finite amount of time, around 10 min in the case of LacI binding to O_{id} [9], and hence an expression measurement might reflect only the average production over multiple different TF configurations. By using the method of fluorescence in-situ hybridization [10] (FISH) to directly measure mRNAs (which are inherently short lived [11]) instead of long-lived proteins [12], we maximize our ability to study direct correlation in transcription rates between different genes. The basic idea behind mRNA FISH is to measure mRNA levels through the hybridization of mRNA with a set of complementary base paired probes, each around 20 bp in

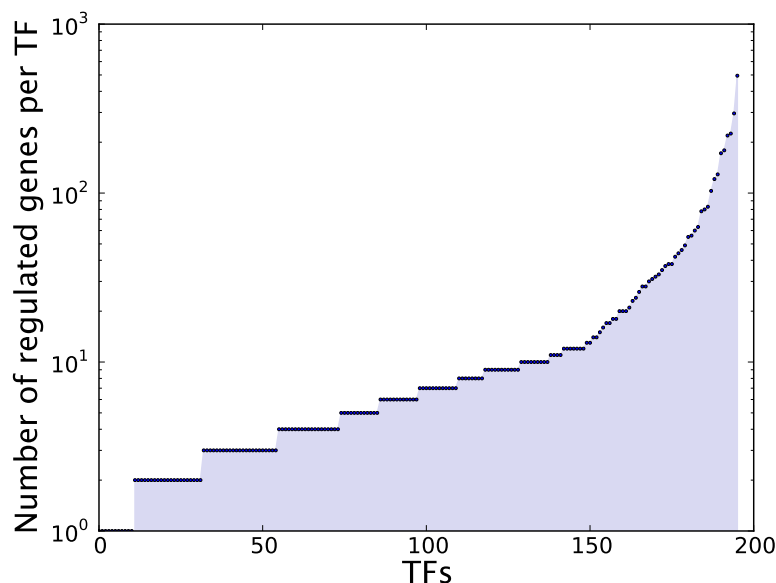


Figure 4.1: Number of genes regulated per TF as reported by RegulonDB 8.5 [4]. Each small dot corresponds to an individual TF. The TFs have been sorted by increasing number of regulated genes. TFs regulating two or more genes are shaded dark.

length, which have been crosslinked to a fluorescent dye. By using multiple dyes that emit light at different wavelengths many different species of mRNA can be observed simultaneously [13].

As proof of principle we investigate the transcription of *two* different genes repressed by the same TF using mRNA FISH. Choosing these reporter genes still requires some care. First of all, the two genes should not be essential to the organism; neither knocking them out nor overexpressing them should have any noticeable impact on the fitness of the cell. Second, since we quantify mRNA levels using FISH it would be desirable, to avoid introducing unnecessary systematic differences between the mRNA signals, if the two genes had approximately the same length (in base pairs), hence binding an equal number probes. For the same reason, it would be desirable if the two mRNAs had similar lifetimes. For convenience we again use *lacZ* as one of the reporter genes and let it be repressed by LacI-mCherry, which allows us to reuse several genetic constructs from Chapter 3. Next we want to find a second gene compatible with the choice of *lacZ*, taking the above considerations into account. In this study we choose *uidA*, which codes for an enzyme that catalyzes the cleavage of β -glucuronides. This gene has been previously used as a reporter gene [14] and fulfills our requirement of being nonessential. The lifetime of the *uidA* transcript has been observed to be very similar to *lacZ* transcripts (around 10 min in M9 at 30 °C [11]), and is, at 1812 bp, closer in length to *lacZ*, which is remarkably long (3075 bp) compared to most other genes in *E. coli*. Finally, *lacZ* and *uidA* have low sequence similarity, which suggests a low off-target hybridization of probes.

To tune the expression of LacI-mCherry, and accordingly the expression of *lacZ* and *uidA*, we

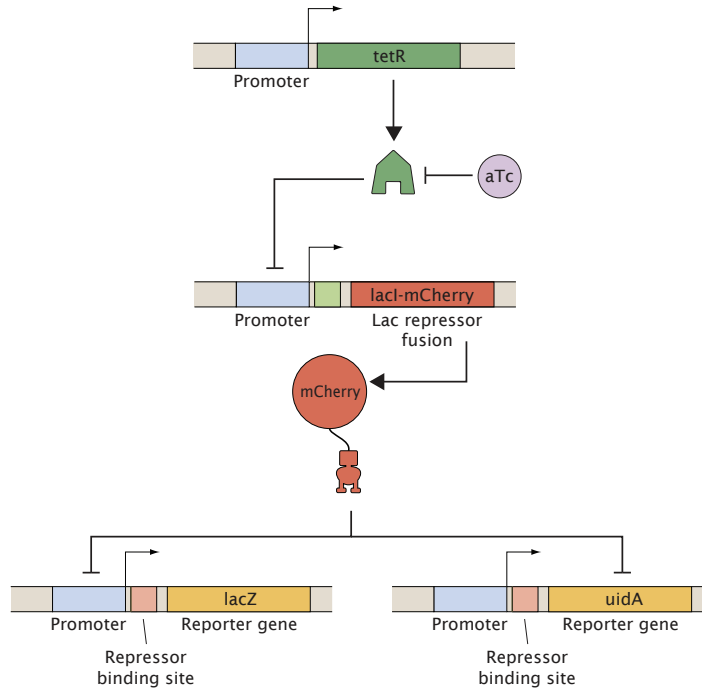


Figure 4.2: Genetic circuit induced by aTc. The two reporter genes *lacZ* and *uidA* are repressed by LacI-mCherry, which in turn is repressed by TetR. Higher aTc concentration leads to inactivation of TetR, and hence higher expression of LacZ and UidA. LacI-mCherry is expressed from the $P_{LtetO-1}$ promoter [15], whereas LacZ and UidA are both expressed from the *lacUV5* promoter [16].

use the same aTc induced *tetR* circuit as in Chapter 3, with the only exception that *tetR* is now integrated into the chromosome. When an active TetR repressor binds *lacI-mCherry* it blocks RNA polymerase (RNAP) from accessing the promoter, thus inhibiting the transcription of the gene. In turn, a lower concentration of LacI-mCherry leads to a *higher* transcription rate of *lacZ* and *uidA*, as these genes are no longer being repressed. The circuit is tuned by aTc, which inactivates TetR [15] and makes it unable to bind the *lacI-mCherry* promoter. By varying the concentration of aTc we can therefore alter the transcription rate of *lacZ* and *uidA*. In particular we can tune the number of LacI-mCherry to be comparable to the number of plasmids, corresponding to the most interesting regime of transcriptional regulation. For calibration we can make an aTc induction curve by measuring, in bulk, the amount of LacI-mCherry produced for different concentrations of aTc using a plate reader [Appendix 4.A]. This helps guide our choices of aTc concentration to ensure our samples are spaced well in expression space.

4.2 Results

4.2.1 Thermodynamic model

With multiple copies of *lacZ* and *uidA* located on plasmids there are a huge number of ways that the genes can be turned off or on, as a function of plasmid and repressor copy number. By using the thermodynamic model presented in Chapters 2-3 we can take all these states into account and predict quantities of experimental interest, such as fold change in gene expression and correlation in transcription rates of the two genes. As a first step we need to compute the statistical weights, or Z_i factors [Eq. (2.22)], associated with specific binding of i repressors to N_p gene copies. We note that there are $\binom{N_p}{i}$ number of ways to choose i genes to turn off in a set of N_p gene copies, and for each such configuration the statistical weight associated with the $N_p - i$ unrepressed promoters is given by $(1 + p)^{N_p - i}$. Here $p = \frac{P}{N_{NS}} e^{-\beta \Delta_{pd}}$ corresponds to either the promoter strength of *lacZ* or *uidA*, which are equal as transcription of both genes is initiated from the *lacUV5* [16] promoter. By using the same strategies as in Chapter 2 one can show that the individual (“unentangled”) partition functions associated with the *lacZ* and *uidA* genes are given by

$$Z^{lacZ} = \sum_{i=0}^{N_p} \frac{R!}{N_{NS}^i (R-i)!} Z_i^{lacZ} \quad \text{with} \quad Z_i^{lacZ} = \binom{N_p}{i} (1 + p_{lacZ})^{N_p - i}, \quad (4.1)$$

$$Z^{uidA} = \sum_{i=0}^{N_p} \frac{R!}{N_{NS}^i (R-i)!} Z_i^{uidA} \quad \text{with} \quad Z_i^{uidA} = \binom{N_p}{i} (1 + p_{uidA})^{N_p - i}. \quad (4.2)$$

Although the promoter strengths are numerically identical, $p_{uidA} = p_{lacZ} = \frac{P}{N_{NS}} e^{-\beta \Delta_{pd}}$, we need to assign them different labels in order to later derive the transcriptional correlation function. To compute the total partition function, including the effect of “entanglement” as to the two genes share a common pool of TFs, we use Eq. (2.20)

$$Z^{tot} = \sum_{f_1=0}^{\min(R, N_p)} \sum_{f_2=0}^{\min(R-f_1, N_p)} \frac{R!}{N_{NS}^{f_1+f_2} (R-f_1-f_2)!} Z_{f_1}^{lacZ} Z_{f_2}^{uidA}. \quad (4.3)$$

The total partition function allows us to derive the fold change [Sec. 2.5] and correlation in transcription rates [Sec. 2.6] between the two genes. In the same way a “competitor plasmid” hosting a TF binding site titrates away TFs and significantly alters the regulatory response of a gene, the two genes *uidA* and *lacZ* will effectively titrate repressors away from each other [see Fig. 3.5]. Hence we expect the two genes to have an impact on each other’s regulatory response function. To compute

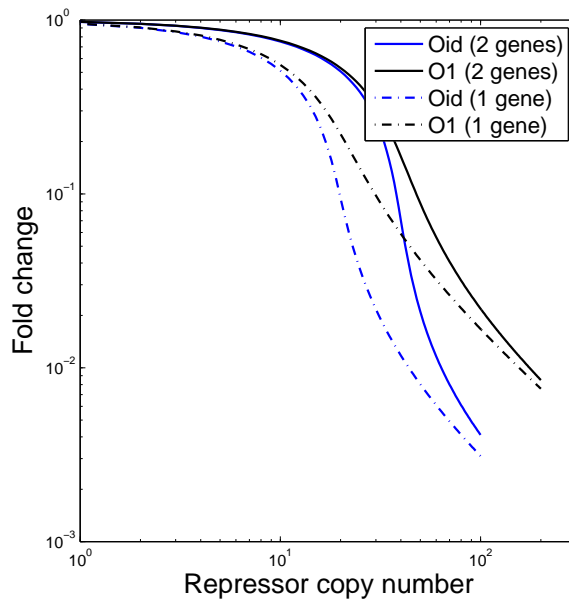


Figure 4.3: (Theory) Fold change as predicted by the statistical mechanical model for two coregulated (“2 genes”) or independently regulated (“1 gene”) genes located on a plasmid. In this plot we use operator binding energies $-17.0 k_B T$ (O_{id}) and $-15.3 k_B T$ (O_1) [17], plasmid copy number $N_p = 20$, number of nonspecific sites as the genome length of *E. coli* ($N_{NS} = 5 \times 10^6$), number of RNAP $P = 1000$, and the RNAP binding energy to the *lacUV5* promoter $-7.0 k_B T$ [18] (same for both genes).

the fold change of the two genes as a function of repressor copy number we use Eq. (2.32)+(2.33)

$$f_{lacZ} = \frac{1 + p_{lacZ}}{N_p} \partial_{p_{lacZ}} \ln Z^{tot}, \quad (4.4)$$

$$f_{uidA} = \frac{1 + p_{uidA}}{N_p} \partial_{p_{uidA}} \ln Z^{tot}. \quad (4.5)$$

The statistical mechanical model predicts [Fig. 4.3], as expected, that it takes twice as many repressors to reach the critical regime where every gene copy can be “turned off”, when the two genes are sharing the same pool of repressors, as compared to when they are independently regulated.

To study the correlation in expression we compute the Pearson correlation coefficient between the number of RNAPs binding the two kinds of promoters using Eq. (2.49)

$$\rho_{lacZ,uidA} = \frac{p_{lacZ} p_{uidA} \frac{\partial}{\partial p_{lacZ}} \frac{\partial}{\partial p_{uidA}} \ln Z^{tot}}{\sqrt{\left[\left(p_{lacZ} \frac{\partial}{\partial p_{lacZ}} \right)^2 \ln Z^{tot} \right] \left[\left(p_{uidA} \frac{\partial}{\partial p_{uidA}} \right)^2 \ln Z^{tot} \right]}}. \quad (4.6)$$

When many repressors are blocking transcription of *lacZ* genes, fewer will be left to block *uidA*, and hence the thermodynamic model predicts [Fig. 4.4] a (weak) *negative* correlation in transcription

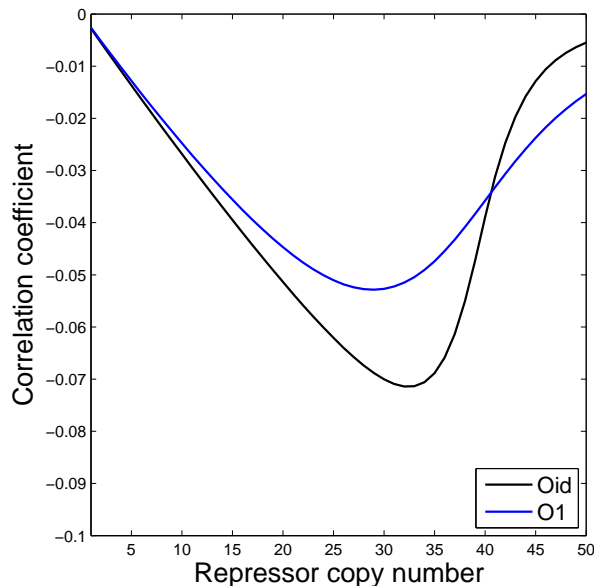


Figure 4.4: (Theory) Transcriptional correlation between two genes repressed by the same TF which binds to either O_{id} or O_1 . In this plot we use the operator binding energies $-17.0 k_B T$ (O_{id}) and $-15.3 k_B T$ (O_1), plasmid copy number $N_p = 20$, number of nonspecific sites as the genome length of *E. coli* ($N_{NS} = 5 \times 10^6$), number of RNAP $P = 1000$, and the RNAP binding energy to the *lacUV5* promoter $-7.0 k_B T$ (some for both genes). For these particular parameter values the transcriptional correlation effect is expected to be small.

of *lacZ* and *uidA*. The correlation will be less prominent as the repressor or RNAP binding energy gets weaker. Our ability to see this correlation, however, is limited. In reality the repressor copy number and plasmid copy number is not fixed, but rather it is fluctuating according to some statistical distribution [19]. Such extrinsic noise leads to a *positive* contribution to the correlation in transcription rates of *lacZ* and *uidA*, as we saw in Sec. 2.7.2. Since the repressors are fluorescently labeled we can reduce the variance in repressor copy number by binning our data, but there is no analogous procedure to reduce the variance in plasmid copy number. We therefore expect plasmid copy number variations to be the dominant source of extrinsic noise among the two.

4.2.2 Stochastic model

Even if extrinsic noise might prevent us from directly observing a negative correlation in the transcription rates of *lacZ* and *uidA*, coupled transcription could still potentially show up in the *relative* mRNA expression

$$Q = \frac{M_{lacZ}}{M_{lacZ} + M_{uidA}}, \quad (4.7)$$

where M_{lacZ} and M_{uidA} denote number of mRNAs in a cell of each kind. By measuring the ratio Q , instead of the absolute numbers M_{lacZ}, M_{uidA} , the effect of extrinsic noise could be greatly reduced. For example, if the number of RNAP were to double, the average transcription rate of both genes would also double, leaving the ratio Q unchanged. The *distribution* of Q over a large set of cells should however depend on whether *lacZ* and *uidA* are transcriptionally entangled or not: If the two genes are transcribed independently of each other, apart from the effect of extrinsic noise, we expect a small variance in Q related to the intrinsic stochasticity of transcription. The width of the Q distribution should, however, increase if a higher transcription rate of one gene leads to a lower transcription rate of the other.

The thermodynamic model is not suitable for predicting the distribution of Q , as it does not take into account the stochastic nature of mRNA production [10, 20, 7, 21]. The promoter used in our experiment, *lacUV5*, produces around 15 transcripts at steady state per cell [18] which follow approximately a Poisson distribution, and the width of this distribution ($\approx \sqrt{15}$) sets a lower bound on the distribution of Q . To qualitatively study how promoter entanglement affects the distribution of Q we resort to stochastic simulations. For a given average repressor copy number \bar{R} , we repeat the following procedure: (A) Choose a random repressor copy number R from a Poisson distribution of mean \bar{R} , and plasmid copy number N_p uniformly from the interval $[20 - 5, 20 + 5]$. The Poisson distribution crudely corresponds to our resolution of binning data by LacI-mCherry copy number. The choice of average plasmid copy number $\bar{N}_p = 20$ will become apparent after next section. To simplify matters we assume that when there are fewer Lac repressors than O_{id} binding sites ($R < 2N_p$), which is the most interesting case for transcriptional entanglement, all repressors are specifically bound, and moreover that repressed promoters are completely shut off. The probability that k out of R repressors will bind to the *lacZ* genes (or *uidA* genes) is then given by

$$P(k; R, N_p) = \frac{1}{Z} \binom{N_p}{k} \binom{N_p}{R-k}, \quad \max(0, R - N_p) \leq k \leq \min(R, N_p), \quad (4.8)$$

where the total number of repressor configurations Z is given by

$$Z = \sum_{k=\max(0, R-N_p)}^{\min(R, N_p)} \binom{N_p}{k} \binom{N_p}{R-k}. \quad (4.9)$$

(B) Draw a random repressor configuration (R_{lacZ}, R_{uidA}) , where R_{lacZ} and R_{uidA} stands for the number of repressors that are bound to *lacZ* genes and *uidA* genes, from the probability distribution determined by Eq. (4.8). Draw one more independent repressor configuration sample (R_{lacZ}^*, R_{uidA}^*) , that will be used as a control. (C) For each unrepressed promoter, sample the mRNA production from a Poisson distribution with a mean burst size of 15 mRNAs per cell to get a total production of (M_{lacZ}, M_{uidA}) and (M_{lacZ}^*, M_{uidA}^*) for the two repressor configurations. (D) Compute the ratio

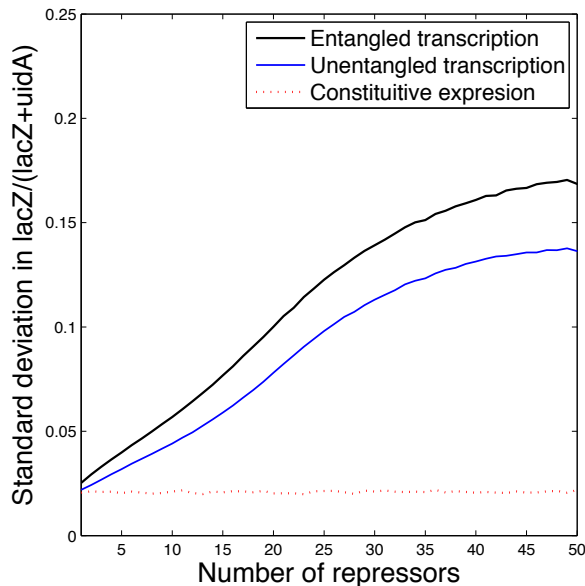


Figure 4.5: (Simulation) Stochastic simulation of the standard deviation σ_Q for the relative expression of two genes repressed by the same TF. The binding site is assumed to be strong enough that all TFs bind specifically, when there are fewer TFs than number of binding sites. The stochastic configuration of repressors on the two genes is determined by the probability distribution in Eq. (4.8). Each unrepressed promoter is assumed to produce an average of 15 mRNAs per cell following a Poisson distribution.

$Q = \frac{M_{lacZ}}{M_{lacZ} + M_{uidA}}$, and the control ratio $Q^* = \frac{M_{lacZ}}{M_{lacZ} + M_{uidA}^*}$ (notice only M_{uidA}^* is starred) guaranteed to lack transcriptional entanglement. To acquire a distribution for Q and Q^* we repeat the above steps 15,000 times for each value of $\bar{R} \in [0, 50]$. In Fig. 4.5 we plot the predicted standard deviation of Q (σ_Q) for the given input distributions of plasmid and repressor copy numbers. As expected σ_Q increases with repressor copy number, as lower transcription leads to higher relative uncertainty in mRNA production, but more interestingly we see that promoter entanglement leads to an increase in σ_Q .

4.2.3 FISH measurements

To demonstrate the effect of TF titration on two coregulated genes we measure the mRNA expression of *lacZ* and *uidA* as a function of LacI-mCherry copy number either when the two genes are located on the same plasmid or when only one of the genes is located on the plasmid. The difference in fold change between these two cases shows the effect of TF titration. In Fig. 4.6 we show the measured fold change when LacI-mCherry binds to the strong O_{id} operator. Our data clearly shows that *lacZ* and *uidA* compete for LacI-mCherry, as it takes approximately twice as many repressors to reach the critical point where transcription of all gene copies can be turned off. With only one regulated

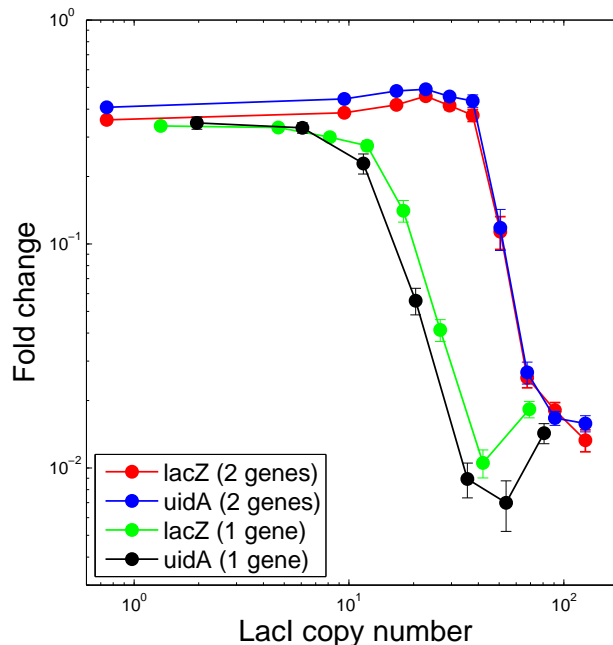


Figure 4.6: (Data) Fold change in transcription of *lacZ* and *uidA* when both genes are located on the same plasmid (“2 genes”) or when only one gene is located on the plasmid (“1 gene”). As transcription of both genes is repressed by LacI-mCherry (binding to O_{id}) it takes twice the number of repressors to inhibit transcription when both genes are located the plasmid. The absolute number of repressors is estimated by calibrating the fluorescence signal to a strain with known absolute expression of LacI-mCherry, as described in Sec. 4.4.5.

gene the critical point occurs at $R \approx 20$, which also suggests [Chapter 3] a plasmid copy number around $N_p \approx 20$, or somewhat less, as a distribution in plasmid copy number tends to shift the estimated N_p upwards [Sec. 3.2.9]. In [15] the plasmid copy number was measured to 10-12, i.e. somewhat lower than our estimate. With both genes on the plasmid the critical point is shifted to around $R \approx 40$, consistent with the idea that all binding sites ($\approx 2 \cdot 20$) on the plasmid needs to be filled before transcription of both genes can be effectively repressed.

Fig. 4.6 raises a couple of concerns. First of all the measured fold change does not approach $f \rightarrow 1$ as $R \rightarrow 0$. An explanation for this, consistent with previous observations in Chapter 3, is that the genetic circuit of Fig. 4.2 produces (“leaks”) a small number of LacI-mCherry even when the *tetR* is fully induced. Even though this number is smaller than can be distinguished from background fluorescence, it is still large enough to repress transcription by roughly a factor of two. If this interpretation is correct we should see a smaller effect of these leaking repressors if we exchange O_{id} with O_1 [Fig. 4.7]. Indeed, our measurements show that when using O_1 the fold change goes up from $f \rightarrow 0.4$ to roughly $f \rightarrow 0.7$ as $R \rightarrow 0$. The leakage production should have less of an impact on fold change when both *lacZ* and *uidA* are located on a plasmid, as there are twice as

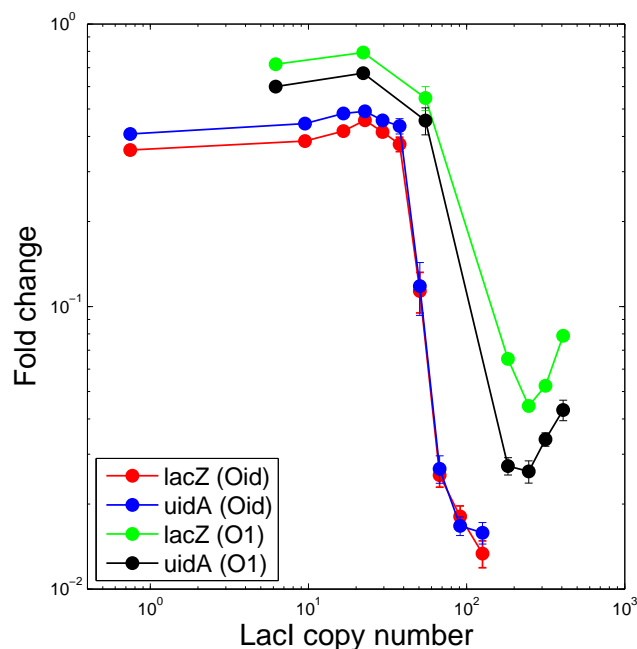


Figure 4.7: (Data) Fold change in transcription for two genes, *lacZ* and *uidA*, repressed by LacI-mCherry binding to either O_{id} or O_1 .

many binding sites for LacI-mCherry to bind. In Fig. 4.6 we do see less repression when the two genes are coregulated, but a much smaller difference ($\approx 10\%$) than the expected factor of two. One speculative explanation to the observation is that the resolution limit [22], i.e. the smallest number of LacI-mCherry molecules that can be distinguished from background, might have been different in the two experiments. As an example, if one of the microscopy dishes had more fluorescent “dirt” attached to it, the ability to resolve low numbers LacI-mCherry would be reduced, causing more cells to get randomly assigned to the lower bins of Fig. 4.6, which would lead to a drop in fold change.

The measured fold change also behaves unexpectedly in the limit when the number of repressors grows large, where fold change seems to plateau or even increase. In this regime, however, our signal is almost identical to the background and we deduce that our experiment lacks sensitivity to measure expression of mRNA in this regime. The ability to measure small numbers of mRNA could be greatly improved by using lasers [23] to excite the FISH probes, rather than a regular fluorescence lamp. One might also consider optimizing the choice of dyes for the given FISH probes. For example, in our experiment the ATTO425 dye gave much higher background signal variance than did ATTO514.

Our measurement of fold change inarguably shows that TF titration can have a strong impact on the regulatory response of coregulated genes. This means that, in principle, the transcription rate

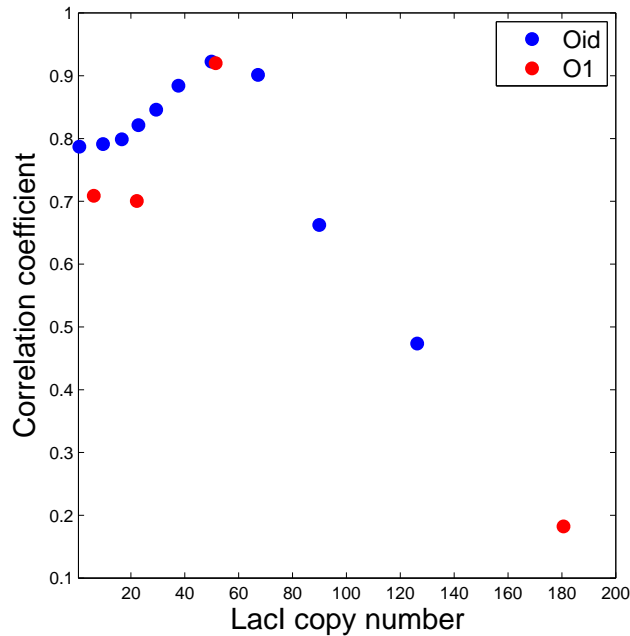


Figure 4.8: (Data) Correlation in mRNA levels for two genes, *lacZ* and *uidA*, repressed by LacI-mCherry binding to either O_{id} or O_1 .

of *lacZ* should affect the transcription rate of *uidA* and vice versa. To see if we can directly observe transcriptional coupling we first look at the correlation in mRNAs levels as a function of repressor copy number [Fig. 4.8]. Due to extrinsic noise from e.g. plasmid copy number and repressor copy number variations the correlation is, as argued previously and predicted by the thermodynamic model [Fig. 2.11], distinctly positive.

A better way to observe transcriptional coupling between *lacZ* and *uidA* might be to look at the *relative* levels of mRNAs of each kind, to reduce the effect of extrinsic noise. We showed in Sec. 4.2.2 that coupled transcription of two genes can produce a widening of the distribution of Q , the relative number of *lacZ* (or *uidA*) mRNAs. Using the dataset from Fig. 4.6 we can plot the standard deviation of Q as a function of repressor copy number. The data and simulations show similar increasing trends in σ_Q for higher number of repressors. In general, the data falls between the two curves corresponding to coupled or uncoupled transcription of *lacZ* and *uidA*. Without carefully studying how experimental noise affects the relative expression of the two genes, as well as the theoretical assumptions (plasmid copy number distribution etc.) which went into the stochastic simulations it is, however, difficult to conclusively say if *lacZ* and *uidA* are transcriptionally coupled or not.

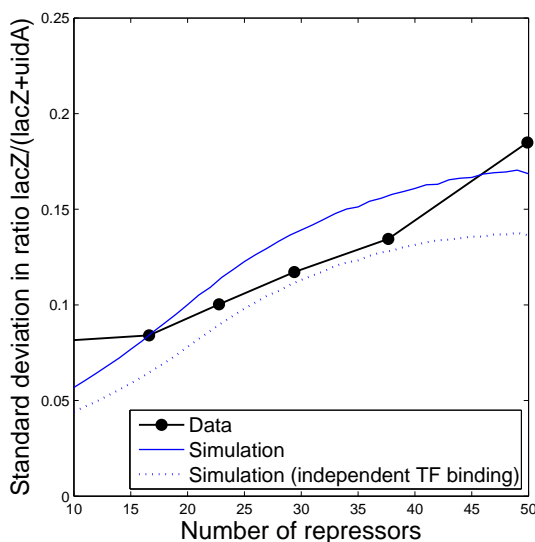


Figure 4.9: (Data) Standard deviation σ_Q in the relative mRNA expression levels for two genes, *lacZ* and *uidA*, which are repressed by LacI-mCherry binding to O_{id} . The fluorescence signals have been normalized to make the observed number of mRNAs of each kind equal, which one would expect for genes transcribed from the same promoter (*lacUV5*). For each cell at least one of the signals must be two standard deviations above background, or else the cell is disregarded. Signals which are not distinguishable from background are set identically to zero, to reduce unwanted fluctuations.

4.3 Discussion

We have shown that TF titration can play an important role in the regulation of genes that share a common pool of TFs. As coregulated genes compete for TFs, it takes higher TF copy number to reach the same level of regulatory response had the genes been independently regulated. In real biological regulatory networks this effect should be accounted for, leading us to believe that TFs with a vast number of binding sites should also have higher TF copy number. We have demonstrated the TF titration effect for the regulation of two genes located on a multi-copy plasmid (copy number $\approx 15 - 20$), but TF titration will most certainly be of similar relevance in settings with a larger number of different genes. In such cases, the affinities of TF binding sites can be different and hence provide a richer set of possible regulatory outcomes [Fig. 2.5, Fig. 3.5].

The TF titration effect shows how one gene can influence the regulatory response of another, which also suggests that one might be able to observe a direct correlation in the transcription rates of two genes. If two genes share a limited pool of TFs, random fluctuations in the TF binding configuration could lead to a temporary enhancement of the transcription rate of one gene at the expense of the transcription of the other. However, our experiments show that if such transcriptional coupling exists it gets drowned by extrinsic noise from e.g. plasmid copy number fluctuations. A rather labor intensive way of reducing this source of extrinsic noise could be to integrate several

copies of *lacZ* and *uidA* on the chromosome [24], and sort cells by size to make sure that all cells are at the same stage in the cell cycle and hence have the same number of gene copies. Even if transcriptional coupling of coregulated genes might be of limited biological significance, as our difficulties in observing it suggest, the effect might still be well-suited for future precision tests of theoretical models of transcriptional regulation.

4.4 Methods

4.4.1 Strains

The cloning procedure consists of two parts. First we create a plasmid with *lacZ* and *uidA*, as well as two binding sites for LacI-mCherry (a fluorescent fusion) to repress each gene. Second we create a host strain for this plasmid with the native *lacIYZA* and *uidA* genes deleted, and the aTc induced circuit of Fig. 4.2 integrated.

As starting point for the plasmid we use pZS25Oid+11-lacZ [25], which contains the *lacZ* part of the desired construct, and replace the coding part of *lacZ* with *uidA* to create pZS25Oid+11-uidA. We do this by PCR amplifying pZS25Oid+11-lacZ minus *lacZ* as well as PCR amplifying *uidA* from the chromosome with DNA overhangs matching the plasmid. We join the two fragments together using Gibson assembly [26]. Next we create a single plasmid that hosts both genes by PCR amplifying *lacZ* (including the promoter) and then (again using Gibson assembly [26]) insert it into pZS25Oid+11-uidA. Untransformed pZS25Oid+11-lacZ plasmids are digested after the *lacZ* PCR amplification using Dpn1 as well as NarI, KasI and SfoI, which cleaves DNA at sites outside the *lacZ* fragment. By growing cells transformed with the new plasmid on agar plates spread with X-gal [27] we can identify colonies of cells containing *lacZ*, which turn blue, hence increasing our ability to select colonies successfully transformed with the desired product. To verify the construct we PCR amplify segments supposed to contain both genes and confirm the presence of *lacZ* and *uidA* through electrophoresis and sequencing.

As starting point for the host strain we use wild-type *E. coli* (MG1655) with *lacIZYA* deleted [17]. A $\Delta uidA$ strain (JW1609-1) was kindly provided by Yale Coli Genetic Stock Center. Using P1 phage transduction we insert $\Delta uidA$ into our host strain. Additional P1 transductions are used to insert *lacI-mCherry* into the locus of *ybcN*, to regulate production of LacZ/UidA, and *tetR* into the locus of *intS*, to regulate the production of LacI-mCherry.

Finally we transform our plasmid into the host strain and into a version lacking *lacI-mCherry*, which allows us to measure unrepressed mRNA production from *lacZ* and *uidA*. The whole procedure above is performed twice to make constructs carrying either the LacI-mCherry repressor binding site O_{id} or O_1 .

4.4.2 Growth

Cultures are grown to saturation in LB and then diluted 1:6000 into 30 mL of M9 supplemented with 0.5% glucose and grown. After growing overnight for around 10 hours, under aluminum foil to avoid degradation of aTc by light, cells are harvested at OD 0.3-0.5.

4.4.3 Single cell mRNA FISH

Following the protocol of [10] we centrifuge cells at 4500 g for 5 min at 4 °C. Harvested cells are fixed by resuspending in 1 mL of 3.7% formaldehyde. After 30 min of fixation cells were washed twice in 1 mL 1x PBS buffer, and then permeabilized in 1 mL of 70 % ethanol, to allow the FISH probes to enter through the cell membrane. After 1 h of fixation cells are gently centrifuged and resuspended in 20 % wash solution (200 μ L formamide, 100 μ L 20x SSC, 700 μ L water). The cells are then resuspended in 50 mL of hybridization solution (0.1 g dextran sulfate, 0.2 mL formamide, 1 mg *E. coli* tRNA, 0.1 mL 20x SSC, 0.2 mg BSA, 10 μ L 200mM ribonucleoside vanadyl complex) and 1.5 mL of probes and crosslinked dye (ATTO425 or ATTO514), and left overnight at 30 °C. One fifth (10 μ L) of the hybridization product is resuspended and centrifuged (600 g, 3.5 min) three times in wash solution with 1 h of incubation at 30 °C between each wash, before finally resuspending in 200-1000 μ L of 2x SSC.

4.4.4 FISH data acquisition

We plate 2 μ L of cells on 1.5% agarose pads made from 1x PBS buffer, and image the pads using a Nikon Eclipse Ti equipped with a CoolSNAP ES² camera. Exposure times for the fluorescence images were set to 7 s for FITC (*uidA* probes), 5 s for CFP (*lacZ* probes) and 5 s mCherry (LacI-mCherry). The camera offset I_{dark} is subtracted from the signal, and a fluorescent calibration plate I_{flat} is used to “flatten” the image and correct for a slightly nonuniform fluorescence exposure over the field of view, yielding a corrected signal $I_{\text{corrected}}^{(i,j)}$ at pixel (i, j)

$$I_{\text{corrected}}^{(i,j)} = \left(I^{(i,j)} - I_{\text{dark}}^{(i,j)} \right) \times \left(\frac{\max_{i,j} (I_{\text{flat}}^{(i,j)} - I_{\text{dark}}^{(i,j)})}{I_{\text{flat}}^{(i,j)} - I_{\text{dark}}^{(i,j)}} \right). \quad (4.10)$$

Cells are segmented on bright-field phase images using the Schnitzcells program kindly provided by the Elowitz lab. In a semi-automated fashion cells which are about to divide or which are located too close to other cells are removed. In total around 1000 cells from each sample are kept for subsequent analysis. Autofluorescence is measured in the host strain minus *lacI-mCherry* and without plasmids expressing LacZ or UidA. The background fluorescence is subsequently subtracted from the measured signal.

4.4.5 Determining the absolute number of LacI-mCherry molecules

A useful application of the TF titration effect, explored in Chapter 3, is that it allows us to measure the copy number of a plasmid that hosts a strong TF binding site. However, this is only possible when one knows the *absolute* number of repressors in the cell, which is typically not the case. In Chapter 3 the calibration factor between fluorescence intensity and the absolute number of Lac repressors was determined by asserting random partitioning of molecules at cell division and measuring the difference in fluorescence intensity between daughter cells. Unfortunately this approach cannot be used in mRNA FISH experiments, where cells are fixed with formaldehyde and, most certainly, no longer dividing. But by reusing a strain from Chapter 3 with a known average absolute number of Lac repressors, we can again relate fluorescence intensity to absolute number of repressors. The strain corresponding to the red points in Fig. 3.3 has, at full induction, an average of 423 repressors after around 105 minutes of additional growth at agar pads, after being taken out of M9. During these 105 minutes the cells divide and dilute the LacI-mCherry molecules. To estimate the average number of LacI-mCherry molecules per cell immediately after growth in M9, N_0 , we assume that the cells are in exponential phase with a division time of 65 min

$$423 = N_0 \times 2^{-105/65},$$

$$\rightarrow N_0 = 1270.$$

To find the calibration factor we grow the strain exactly as described in Sec. 4.4.2, measure the average mCherry signal and divide by $N_0 = 1270$ to find the estimated fluorescence intensity of a single LacI-mCherry molecule.

4.5 Acknowledgements

We wish to thank Daniel Jones for help with setting up the cell segmentation software, and Franz Weinert for useful discussions and help interpreting our data.

Appendix

4.A aTc induction curve

We determine the response of LacI-mCherry as a function of aTc concentration in bulk using a plate reader [Fig 4.10]. To reach the critical point where the number of LacI-mCherry exceeds the number of pZS25Oid+11-lacZ plasmids, it takes only a very small amount of aTc. The used inducer concentrations for the data shown in Fig. 4.6 are: 0.05, 0.10, 0.25, 0.40 ng/ml. The inducer concentrations for the O₁ data shown in Fig. 4.7 are: 0.10, 0.20, 0.40, 0.60 ng/ml. Finally, the calibration strain used to determine the absolute number of LacI-mCherry molecules is maximally induced at 100 ng/ml.

4.B Plasmid structure

The two genes *lacZ* and *uidA* are transcribed from the same strong promoter *lacUV5* and oriented as shown in Fig. 4.11.

The sequence at the junction between *lacZ* and *uidA* is given by:

```

...GGTCGCTACCATTACCAGTTGGTCTGGTGTCAAAAATA]lacZ AAAGCTTAATTAGCT
GAGTCTAGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTT
TATCTGTTGTTTGTTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGCCCTAGACC
TAGGGTACGGGTTTTGCTGCCCGCAAACGGGCTGTTCTGGTGTGCTAGTTTGTATC
AGAATCGCAGATCCGGCTTCAGCCGTTTGCCGGCTGAAAGCGCCTAAGAAACCATTA
TTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCACCT
CGAG [TTTACACTTTATGCTTCCGGCTCGTATAATGTGTGG]lacUV5 [AATTGTGAGCGC
TCACAATT]Oid GAATTCATTAAAGAGGAGAAAGGTACC uidA[ATGTTACGTCTCTGTAGA
AACCCCAACCCGTGAAATCAAAAACCTCGACGGCCT...

```

For O₁ data shown in Fig. 4.7 the O_{id} operator sequence is replaced by the O₁ sequence [AAT TGTGAGCGGATAACAATT]^{O₁}.

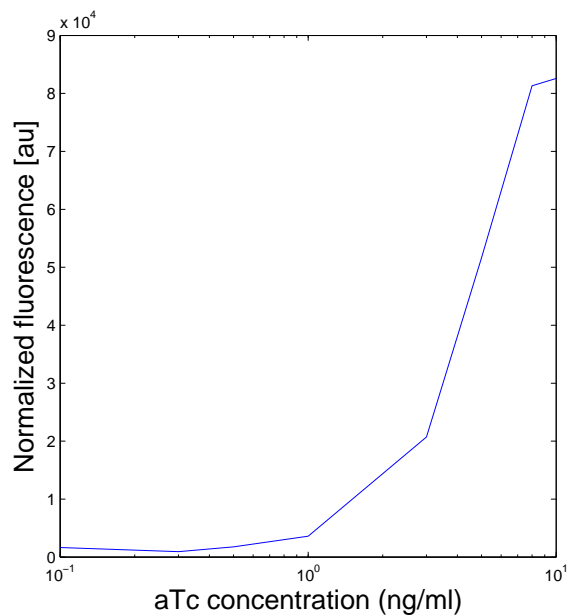


Figure 4.10: (Data) Induction curve of LacI-mCherry by aTc. Fluorescence is measured in bulk using a plate reader and normalized by the optical density (OD) of the sample.

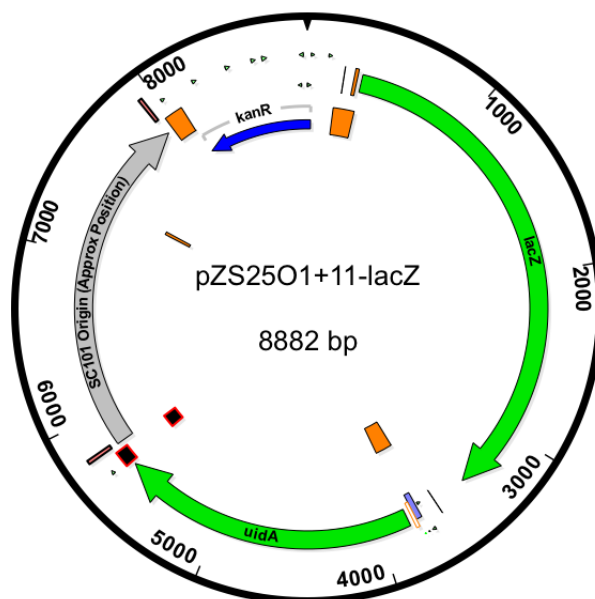


Figure 4.11: Plasmid diagram showing the main features of the pZS25Oid+11-lacZ-uidA/pZS25O1+11-lacZ-uidA plasmid.

4.C PCR primers

To amplify the pZS25Oid+11-lacZ/pZS25O1+11-lacZ plasmid minus *lacZ* we use:

Forward 5'-GGTACCTTTCTCCTCTTTAATGAATTC-3'

Reverse 5'-AAGCTTAATTAGCTGAGTCTAGAGGC-3'

To amplify *uidA* from the chromosome including overhangs (lower case) that match the pZS25Oid+11-lacZ/pZS25O1+11-lacZ plasmid we use:

Forward 5'-cattaaagaggagaaaggtaccATGTTACGTCCTGTAGAAACCCC-3'

Reverse 5'-gactcagctaattaagcttTCATTGTTGCCTCCCTGC-3'

To amplify *lacZ* (with promoter) including overhangs to match the pZS25Oid+11-uidA/pZS25O1+11-uidA plasmid we use:

Forward 5'-ctggcaattccgacgtATTATTATCATGACATTAACCTATAAAAAATAGGCGTATCAC-3'

Reverse 5'-gataataatggttttcttagGCGCTTCAGCCGGCAAAA-3'

To sequence the final plasmid and verify that it has both *lacZ* and *uidA* we use:

Forward 5'-ATTGGTGGCGAGACTCC-3'

Reverse 5'-ACCAACGCTGATCAATTCC-3'

To verify that *lacIZYA* is deleted from the host strain we run PCR with the following primers and verify with electrophoresis that the amplified segment is no more than 300 bp :

Forward 5'-GTGTCTCTTATCAGACCGTTTCCC-3'

Reverse 5'-TGATAAGCGCAGCGTATCAGG-3'

To verify that *uidA* is deleted from the host strain we run PCR with the following primers and verify with electrophoresis that the amplified segment is no more than 300 bp :

Forward 5'-GACGATGGTGCGCCAGGA-3'

Reverse 5'-CAGATTAAGGTTGACCAGTATTATTATCTTAATGAGGAGT-3'

To verify that *tetR* is inserted to the host strain we run PCR with the following primers and verify with electrophoresis that the amplified segment is 1400 bp instead of wild-type 1177 bp :

Forward 5'-CGCAAAATCCCCTGAATATC-3'

Reverse 5'-TTGCACTGGATTGCAAGACT-3'

To verify that *lacI-mCherry* is inserted to the host strain we run PCR with the following primers and verify with electrophoresis that the amplified segment is several thousand base pairs instead a few hundred for wild-type :

Forward 5'-AGCGTTTGACCTCTGCGGA-3'

Reverse 5'-GCTCAGGTTTACGCTTACGACG-3'

References

- [1] Buchler, N. & Louis, M. Molecular titration and ultrasensitivity in regulatory networks. *J Mol Biol* **384**, 1106–19 (2008).
- [2] Burger, A., Walczak, A. M. & Wolynes, P. G. Abduction and asylum in the lives of transcription factors. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 4016–4021 (2010).
- [3] Lee, T. H. & Maheshri, N. A regulatory role for repeated decoy transcription factor binding sites in target gene expression. *Mol. Syst. Biol.* **8**, 576 (2012).
- [4] Salgado, H. *et al.* RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* **41**, D203–213 (2013).
- [5] Allocco, D. J., Kohane, I. S. & Butte, A. J. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **5**, 18 (2004).
- [6] Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12795–12800 (2002).
- [7] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).
- [8] Dunlop, M. J., Cox, R. S., Levine, J. H., Murray, R. M. & Elowitz, M. B. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics* **40**, 1493–1498 (2008).
- [9] Hammar, P. *et al.* Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. *Nature Genetics* (2014).
- [10] So, L. H. *et al.* General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.* **43**, 554–560 (2011).
- [11] Bernstein, J. A., Khodursky, A. B., Lin, P. H., Lin-Chao, S. & Cohen, S. N. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9697–9702 (2002).

- [12] Nath, K. & Koch, A. L. Protein degradation in *Escherichia coli* I. Measurement of rapidly and slowly decaying components. *Journal of Biological Chemistry* **245**, 2889–2900 (1970).
- [13] Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9**, 743–748 (2012).
- [14] Jefferson, R. A., Burgess, S. M. & Hirsh, D. beta-Glucuronidase from *Escherichia coli* as a gene-fusion marker. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8447–8451 (1986).
- [15] Lutz, R. & Bujard, H. 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**, 1203–10 (1997).
- [16] Silverstone, A. E., Arditti, R. R. & Magasanik, B. Catabolite-insensitive revertants of *lac* promoter mutants. *Proc. Natl. Acad. Sci. U.S.A.* **66**, 773–779 (1970).
- [17] Garcia, H. G. & Phillips, R. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A* **108**, 12173–8 (2011).
- [18] Brewster, R. C., Jones, D. L. & Phillips, R. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput. Biol.* **8**, e1002811 (2012).
- [19] Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences* **105**, 17256–17261 (2008).
- [20] Paulsson, J. Models of stochastic gene expression. *Physics of Life Reviews* **2**, 157–175 (2005).
- [21] McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences* **94**, 814–819 (1997).
- [22] Niswender, K., Blackman, S., Rohde, L., Magnuson, M. & Piston, D. Quantitative imaging of green fluorescent protein in cultured cells: Comparison of microscopic techniques, use in fusion proteins and detection limits. *Journal of Microscopy* **180**, 109–116 (1995).
- [23] Pawley, J. *Handbook of Biological Confocal Microscopy* (Springer, 2010).
- [24] Yu, D. *et al.* An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proceedings of the National Academy of Sciences* **97**, 5978–5983 (2000).
- [25] Garcia, H. G., Lee, H. J., Boedicker, J. Q. & Phillips, R. Comparison and calibration of different reporters for quantitative analysis of gene expression. *Biophysical Journal* **101**, 535–544 (2011).
- [26] Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**, 343–345 (2009).

- [27] Horwitz, J. P. *et al.* Substrates for cytochemical demonstration of enzyme activity. I. Some substituted 3-indolyl- β -D-glycopyranosides1a. *Journal of Medicinal Chemistry* **7**, 574–575 (1964).

Chapter 5

The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*

This chapter presents results which are about to be submitted for publication.

5.1 Introduction

One of the most impressive accomplishments in molecular biology over the past half-century has been the mapping of thousands of gene interactions to create genetic networks for a broad collection of different organisms. Such maps have made it possible to qualitatively understand how groups of genes can together provide important functionality. Still, the genetic network descriptions leave us with a picture of the regulatory landscape that is not quantitatively predictive. Although impressive, genetic networks do not provide information necessary to make concrete predictions, such as the number of proteins produced of a given kind under particular environmental conditions, not the least because the notions of ‘activation’ and ‘repression’ are inherently verbal rather than quantitative. The amount of activation or repression achieved by transcription factors (TFs) can vary by many orders of magnitude, depending on how tightly TFs bind to the promoter of interest [1] and many other factors. Moreover the resulting response curves depend on *promoter architecture*, i.e. the particular configuration of TF binding sites. For example, a repressor that blocks a promoter through DNA looping (e.g. LacI) has been shown to have a steeper response curve than its unlooped counterpart [2]. Furthermore, genetic networks do not tell if the TF supposed to regulate a gene is actually present in the cell at all, which might not be the case if it is inactivated through nucleosomal organization or by chromatin remodeling complexes [3, 4].

For genetic networks to be predictive tools in biology they need to be augmented with quan-

titative descriptions of the census of regulatory players. Our goal with the present chapter is to take a step in this direction by studying the role of promoter architectures in transcriptional regulation, from a genome-wide point of view. Few organisms offer a better opportunity to do so than *E. coli*, which after more than half a century of intense study demonstrates arguably one of the most well understood regulatory networks. Through ambitious efforts many cold and hard facts about transcriptional regulation in *E. coli* have been collected and made easily accessible in databases like RegulonDB [5] and EcoCyc [6]. These contain information including, but not limited to, which TFs regulate different operons, where they bind to promoters, and their regulatory effect (activation or repression). All of these features play an important role in transcriptional regulation. A TF which binds cooperatively to multiple binding sites, either through direct contact or DNA looping, provides a steeper regulatory response than TFs binding just a single site [7], which is typically reported by Hill coefficients. The position of binding sites play an equally important role. In experiments where a single repressor binding site has been systematically moved along the promoter region [8, 9, 10, 11], the effect of repression has shown a clear dependence on position, interestingly featuring a 10-11 bp modulation following the periodicity of the DNA helix. The most dramatic position effect occurs when a TF binds near the promoter and allosterically blocks RNAP from accessing the promoter, hence turning off transcription of the gene.

In this chapter we study both the positions and multiplicities of TF binding sites in *E. coli*, for the 2500 or so known TF-DNA interactions in RegulonDB 8.5. A challenge inherent in using RegulonDB, EcoCyc, or any other biological database as primary information source is that the data is inevitably incomplete. More than half of the genes in *E. coli* still lack any regulatory annotation, including important genes such as those responsible for mechanosensation. We must therefore be cautious when interpreting our results. Whereas there is no obvious reason that, for example, binding site positions are biased, the absolute number of binding sites is almost certainly underestimated. This assertion is supported by the fact that the rate of newly discovered TF binding sites does not show any sign of slowing down, thanks to the advent of powerful techniques such as ChIP-seq [12] and Sort-Seq [13]. A healthy skepticism from the reader is thereby encouraged and the results should be viewed as provisional until more of the underlying regulatory facts are in hand. Although incomplete, we can use RegulonDB to construct expectations for what promoter architectures should look like and identify overrepresented promoter architectures motifs deviating from this expectation, in the same way as overrepresented genetic network motifs have been identified from an incomplete genetic network [14].

We view the work presented here as a step towards using promoter architectures to give a more detailed understanding of transcriptional regulation than can be given by a genetic network map alone. Hopefully these findings can also provide valuable input for the theoretical dissection of transcription regulation, which has shown increasing capability to make distinct predictions for the

response function of different promoter architectures [15, 16, 17, 18]. Perhaps most importantly, the analysis presented here shows how far short the current factual understanding of regulatory architectures and measured expression levels falls from serving as a predictive framework, and thus should be seen as a call for higher predictive expectations and a more rigorous treatment of the relation between regulatory architecture and input-output functions.

5.2 Models

5.2.1 Random promoter architecture model

In the study of genetic networks, *network motifs* refer to recurring patterns that are overrepresented in biological networks as compared to random networks [14]. A well-studied network motif is the feed-forward loop, where a single gene is regulated by two TFs, and in addition one of the TFs regulates the other. Motifs are presumed promoted in biological systems because of functionality they provide, for example robustness against concentration fluctuations of regulatory molecules. In this section we will apply the idea of network motifs to the configuration of TF binding sites. For this we need to introduce a *random promoter architecture model*, analogously to a random genetic network, to be used as reference for identifying overrepresented reported promoter architectures.

TF binding sites can be both lost and gained due to the steady pace of mutations across the genome. These mutations are thought to occur randomly across the genome, and hence in the absence of selection any specific distribution of a given number of TF binding sites over a set of operons would be as probable as any other. If a certain class of promoter architectures occurs more frequently in real regulatory networks than in this null model, we expect them to encode biological functions which are advantageous. The simple approach we will adopt to implement a random promoter architecture is therefore to imagine all binding sites reported in RegulonDB as being “sprinkled” over all operons with *uniform* probability. The mathematical implications of this simple postulate will be developed here, saving for the Results section the task of identifying promoter architecture motifs.

As a first application of the random promoter architecture model we will look at the distribution of number of binding sites per operon. The problem of independently assigning N_{bs} binding sites (2871 in RegulonDB 8.5) to N_{op} operons (2642 in RegulonDB 8.5) can be formulated as the flipping of a biased coin, where each binding site has a probability $1/N_{op}$ to be assigned to a particular operon, in each of N_{bs} repeated coin tosses. Here we neglect the complicating fact that a small number of binding sites can regulate multiple operons (9 % in RegulonDB 8.5). Hence the probability distribution $P_{bs}(m; N_{bs})$ for an operon to end up with m binding sites is given by a simple binomial

distribution

$$P_{bs}(m; N_{bs}) = \binom{N_{bs}}{m} \left(\frac{1}{N_{op}} \right)^m \left(1 - \frac{1}{N_{op}} \right)^{N_{bs}-m}, \quad m = 0, 1, 2, \dots \quad (5.1)$$

Since the probability is small for a binding site to be assigned a particular operon, namely $1/N_{op}$, the binomial distribution can be approximated by a Poisson distribution with mean $\lambda = \frac{N_{bs}}{N_{op}}$. For the numbers of N_{bs} and N_{op} given by RegulonDB 8.5 the Poisson approximation is valid to within 1 % for $m \in \{0 \dots 10\}$, which covers 98 % of the operons in the dataset. However, for very highly regulated operons, $m \gtrsim 20$, the Poisson approximation should not be used.

We can generalize the distribution in Eq. (5.1) to incorporate several types of binding sites, say activators or repressors, or different particular TFs. Since all binding sites are assumed to be independently distributed, the probability distribution $P_{2bs}(m_1, m_2; N_{bs}^{(1)}, N_{bs}^{(2)})$ for an operon to end up with m_1 binding sites (from a total of $N_{bs}^{(1)}$) of one type and m_2 binding sites (from a total of $N_{bs}^{(2)}$) of a second type will be given as an independent product of two binomial distributions

$$P_{2bs}(m_1, m_2; N_{bs}^{(1)}, N_{bs}^{(2)}) = P_{bs}(m_1; N_{bs}^{(1)}) P_{bs}(m_2; N_{bs}^{(2)}) \quad (5.2)$$

$$= \binom{N_{bs}^{(1)}}{m_1} \binom{N_{bs}^{(2)}}{m_2} \left(\frac{1}{N_{op}} \right)^{m_1+m_2} \left(1 - \frac{1}{N_{op}} \right)^{N_{bs}^{(1)}+N_{bs}^{(2)}-m_1-m_2}. \quad (5.3)$$

Several TFs in *E. coli* preferentially bind to multiple binding sites at a given promoter, for example NarP binds to two sites or more at 10 out of 11 regulated operons according to RegulonDB 8.5. This can hardly be a coincidence, but how do we rigorously define the level of “self-cooperativity”? Simply looking at the absolute number of operons with multiple binding sites is not a good measure of self cooperativity of a TF, as it does not take into account the total number of binding sites available. A *global TF* [19, 20] with hundreds of binding sites will likely bind at multiple sites at several promoters simply by chance.

Instead we will use the random promoter architecture model to derive the probability $P_{co}(M)$ for M operons to be regulated by at least two binding sites for the same TF, as a function of total number of TF binding sites N_{bs} ($N_{bs} > 2M$). To find the number of ways N_{bs} binding sites can be distributed over N_{op} operons, with two or more sites at M of these, we first choose M operons, in any of $\binom{N_{op}}{M}$ ways, and assign two binding sites to each of them. See illustration Fig. 5.1. Next we put k of the remaining $N_{bs} - 2M$ binding sites into k of the remaining $N_{op} - M$ operons (i.e. one binding site per operon), which we can choose in $\binom{N_{op}-M}{k}$ ways. Finally we put the remaining $N_{bs} - 2M - k$ binding sites into the M operons, which already have two binding sites, however we want. The number of ways this can be done equals the number of nonnegative integer solutions to the equation $x_1 + x_2 + \dots + x_M = N_{bs} - 2M - k$, a famous problem from combinatorics with $\binom{N_{bs}-M-k-1}{M-1}$ solutions. To find the probability $P(M)$ we sum over k and divide by the total number

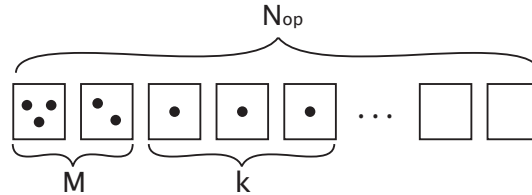


Figure 5.1: Binding sites are distributed into M operons with at least two binding sites, and k operons with exactly one binding site. The remaining $N_{op} - M - k$ operons are empty.

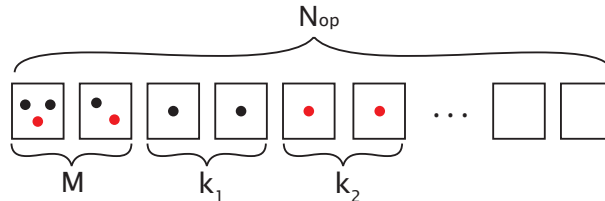


Figure 5.2: Two different kinds of binding sites (black and red) are distributed into: M operons with at least one binding site of each kind, k_1 operons with at least one binding site of first kind but none of the second, and k_2 operons with at least one binding site of the second kind but none of the first. The remaining $N_{op} - M - k_1 - k_2$ operons are empty.

of ways to distribute N_{bs} binding sites over N_{op} operons, which according to the same argument as above is given by $\binom{N_{bs}+N_{op}-1}{N_{bs}}$.

$$P_{co}(M; N_{bs}) = \frac{1}{\binom{N_{bs}+N_{op}-1}{N_{bs}}} \sum_{0 \leq k \leq \min(N_{op}-M, N_{bs}-2M)} \binom{N_{op}}{M} \binom{N_{op}-M}{k} \binom{N_{bs}-M-k-1}{M-1} \quad (5.4)$$

$$M \leq \min(N_{op}, \lfloor N_{bs}/2 \rfloor)$$

By using the general definition of binomial coefficient $\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)}$ where $\Gamma(x)$ is the gamma function [21], Eq. (5.4) gives us the right probability also for $M = 0$, namely $P_{co}(0) = \binom{N_{op}}{N_{bs}} / \binom{N_{bs}+N_{op}-1}{N_{bs}}$ (for $N_{op} \geq N_{bs}$).

We can generalize this problem to cooperativity between TF pairs [7]. TF pairs which coregulate operons more frequently than suggested by the random promoter architecture model are more likely to have related biological function. Let $N_{bs}^{(1)}$ and $N_{bs}^{(2)}$ be the number of binding sites for two different types of TFs. As above we start by choosing M operons where we put one binding site of each kind. Next we put the remaining $N_{bs}^{(1)} - M$ binding sites of first type into the M “shared” operons plus an additional of k_1 operons, which we can choose in $\binom{N_{op}-M}{k_1}$ ways, with at least one binding site in each. See Fig. 5.2. The number of ways this can be done equals the number of integer solutions to

the equation below with the given constraints

$$\underbrace{x_1 + \dots + x_M}_{x_i \geq 0} + \underbrace{x_{M+1} + \dots + x_{M+k_1}}_{x_i \geq 1} = N_{bs}^{(1)} - M. \quad (5.5)$$

After subtracting k_1 from both sides one realizes that the number of solutions to this equation equals the number of solutions to the simpler equation

$$\underbrace{x_1 + \dots + x_M + \tilde{x}_{M+1} + \dots + \tilde{x}_{M+k_1}}_{x_i \geq 0} = N_{bs}^{(1)} - M - k_1, \quad (5.6)$$

which is given by $\binom{N_{bs}^{(1)}-1}{M+k_1-1}$. Next we use the same argument to distribute the remaining binding sites of the second kind onto the M shared operons plus an additional k_2 operons, which we can now choose in $\binom{N_{op}-M-k_1}{k_2}$ ways. We find the probability $P_{2\ co}(M)$ for M operons to be regulated by at least one binding site of each type by summing over k_1, k_2 and dividing by the total number of binding site arrangements, which is given by $\binom{N_{op}+N_{bs}^{(1)}-1}{N_{bs}^{(1)}} \binom{N_{op}+N_{bs}^{(2)}-1}{N_{bs}^{(2)}}$;

$$\begin{aligned} P_{2\ co}(M; N_{bs}^{(1)}, N_{bs}^{(2)}) &= \frac{1}{\binom{N_{op}+N_{bs}^{(1)}-1}{N_{bs}^{(1)}} \binom{N_{op}+N_{bs}^{(2)}-1}{N_{bs}^{(2)}}} \sum_{k_1=0}^{\min(N_{bs}^{(1)}-M, N_{op}-M)} \sum_{k_2=0}^{\min(N_{bs}^{(2)}-M, N_{op}-M-k_1)} \\ &\times \binom{N_{op}}{M} \binom{N_{op}-M}{k_1} \binom{N_{op}-M-k_1}{k_2} \binom{N_{bs}^{(1)}-1}{M+k_1-1} \binom{N_{bs}^{(2)}-1}{M+k_2-1}. \end{aligned} \quad (5.7)$$

As a sanity check we use MATHEMATICA to see that the probabilities add up to one. We can also compare with [22], which solved essentially the same problem but under the assumption that binding sites, even of the same kind, are distinguishable. However, as long as the probability is small that two binding sites regulate the same operon the two methods will give similar results, just like Fermi-Dirac statistics approaches Boltzmann statistics in dilute systems [23].

A different method to identify TF cooperativity based on mutual information from ChIP data was used in [24]. The advantage with the random promoter architecture model is that it takes, for example, biasing due to differences in number of TF binding sites into account, and it allows us not only to determine the expected number of coregulated operons but also the associated p-value with any given observation, i.e. the probability of an equal or more extreme outcome with respect to the random promoter architecture model. This will become useful in the Results section where we want to identify TF binding motifs in the reported distributions from RegulonDB.

5.2.2 Linear energy model of RNAP-DNA binding

The binding affinity of RNAP to the promoter of a gene is determined by the nucleotide sequence of the promoter and has a strong influence on the transcription rate of the gene [18]. The more effectively a promoter can recruit RNAP and initiate open complex formation, the higher the transcription rate of the gene will be. Creating a predictive map between DNA sequence and RNAP binding affinity is a problem which has received much attention [13, 25, 26, 27, 28]. One of the simplest but yet most successful approaches to the modeling of RNAP-DNA (or TF-DNA) interactions is to assume *independent* energy contributions from each individual nucleotide in the binding sequence. Under this linear assumption the total binding energy of a sequence S can be expressed as a simple matrix trace

$$E(S) = \sum_{i=1}^L \sum_{j=A,C,T,G} M_{i,j} S_{j,i} = \text{Tr}(MS), \quad (5.8)$$

where $S_{A/C/T/G,i} = 1$ if the identity of the base at nucleotide position i in the sequence is given by $A/C/T/G$ and otherwise $S_{A/C/T/G,i} = 0$, $M_{i,A/C/T/G}$ represents the energy contribution at position i for base $A/C/G/T$ respectively, and L is the length (in base pairs) of the binding sequence. The RNAP σ^{70} complex has two binding domains which interact with the promoter at -35 and -10 base pairs upstream of the transcription start site [29]. In this study we compute the binding energy from these two boxes separately using the energy matrix $M_{i,j}$ of Brewster et al. [25] and allow a spacer region between 16 – 18 bp that does not influence the total binding energy. Further we allow the -10 box to deviate one base pair up or down from its consensus position. The RNAP binding energy is taken as the minimum binding energy of the $3 \times 3 = 9$ possible binding configurations.

To relate the RNAP binding affinity and transcription rate a commonly used assumption, the *occupancy hypothesis* [15], states a linear relationship between the transcription rate of a gene and the probability of its promoter being occupied by RNAP. This probability is, according to the Boltzmann distribution, proportional to $e^{-E(S)/k_B T}$ for systems in (quasi)equilibrium, an approximation which can be made if RNAP homogenize throughout the cell at a much higher rate than that at which they are being produced. The occupancy hypothesis has despite its simplicity (for example ignoring details of open complex formation and promoter escape rate) proved surprisingly successful in many different settings [25, 1].

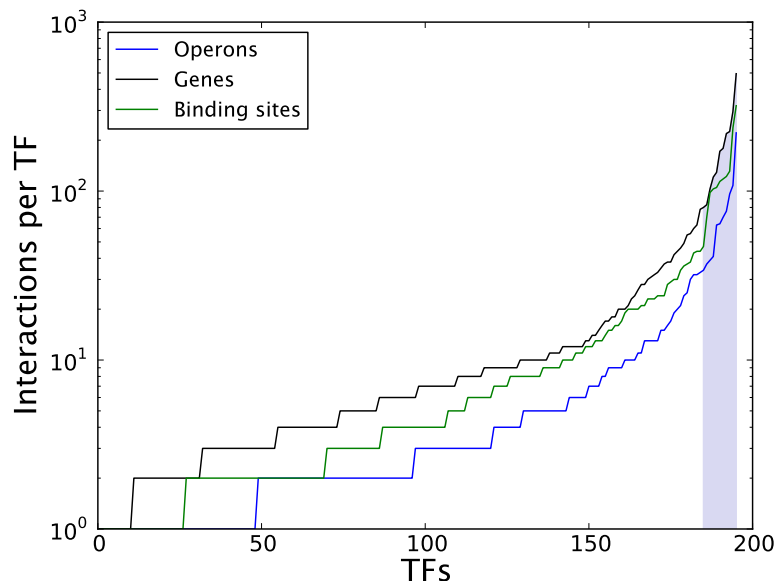


Figure 5.3: Number of operons, genes and binding sites regulated per TF as reported by RegulonDB 8.5. The TFs have been sorted by increasing number of interactions, and the dark shaded area highlights the TFs responsible for 50% of all regulatory interactions in *E. coli*. The median number of operons regulated per TF is 3.0.

5.3 Results

5.3.1 How many genes do TFs regulate?

In regulatory networks, genes of correlated biological function are often coregulated. For example a flagellum in *E. coli* consists of roughly thirty different proteins [30] present at exact ratios. Not all the flagellar genes can however fit the same operon, since an RNA polymerase (RNAP) would likely fall off or stall before reaching the end, and instead these thirty genes are being transcribed from roughly ten different operons [6]. To express the flagellar proteins at correct ratios it is important that these operons are coregulated, a task which in *E. coli* is handled by the TFs FlhC and FlhD. However, other biological functions correlate with the production of flagella. For example production of sugar receptors in the cell membrane such as MglBAC, necessary for chemotaxis, are also regulated by FlhC and FlhD. In general we expect “correlated genes” to form clusters that are regulated by the same TFs, and the question we will address in this section is: What is the typical size of such clusters?

In Fig. 5.3 we show the number of genes and operons that are coregulated by TFs as well as the total number of binding sites for these TFs, reported by RegulonDB 8.5. The numbers provide a lower estimate of the actual *E. coli* regulatory network, acknowledging the fact that not all binding sites have yet been discovered. The figure reveals two almost separate groups of TFs: a large number of *specific TFs* which regulate only a few operons, and a mere handful of *global TFs* [19, 20]

Transcription factor	Operons	Genes	Binding sites
CRP	221	495	320
FNR	108	296	131
Fis	96	225	237
IHF	76	219	114
H-NS	70	179	105
ArcA	64	172	118
Fur	63	129	122
Lrp	41	103	103
⋮	⋮	⋮	⋮

Table 5.1: Global TFs and their associated number of binding sites (RegulonDB 8.5).

regulating up to a hundred operons [see Table 5.1]. Half of all TFs regulate two operons or less, suggesting that many processes in *E. coli* are not strongly correlated with other processes and can be carried out independently, unlike the construction of flagella. For example, in response to varying levels of copper in the cytoplasm ComR reportedly regulates only one single gene *bhsA*, which alters the outer cell membrane permeability for copper [31]. It would not have been far-fetched to believe that other genes would also be regulated by this TF. Global TFs, which regulate core activities in the cell, for example metabolic pathways (e.g. CRP) or the translational machinery (e.g. Fis), are the exceptions to this rule. Despite the small number of global TFs, these are involved in roughly half of all reported regulatory interactions.

Instead of considering the number of operons regulated per TF, we can, conversely, consider the number of TFs regulating each operon. This serves as a simple proxy for the complexity of the regulatory function at a promoter. The more TFs regulating an operon, the more specific its response can be to various cellular conditions. In Fig. 5.4 we show the number of TF interactions and number of different TF types regulating operons as reported by RegulonDB 8.5. The average number of TF binding sites per operon is only 1.1, but climbs to 3.5 when neglecting operons without known regulatory interactions. This observation suggests that data in RegulonDB is, to some extent, collected “one operon at a time”, i.e. the binding site locations for one operon are carefully examined before moving to the next operon. There is an approximately exponential decrease (see fit) in reported number of operons as the number of regulatory interactions increases. It is perhaps surprising that even for such a well studied organism as *E. coli* more than half of the genes still lack any regulatory annotation. Among these unannotated genes we find important examples such as the genes responsible for mechanosensation *mscS*, *mscL*, *mscK*, *ynal*, *ybio* and *ybdG*. Preliminary results from our lab based on the method of Sort-seq [13] show that at least some of these genes might in fact be regulated. Other notable genes lacking regulatory annotation include: *lpp*, a lipoprotein believed to be one of the most abundant proteins in *E. coli* [32]; *rep*, a helicase required for genomic replication [33]; *kdpD* and *nhaB*, genes related to regulation of potassium [34] and sodium [35]

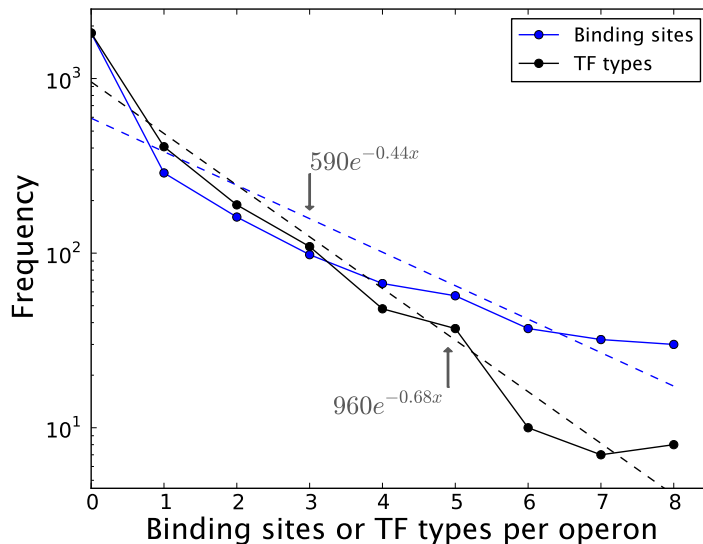


Figure 5.4: Number of binding sites and TF types regulating each operon (RegulonDB 8.5). The mean number of binding sites per operon is 1.1 (3.5 for operons with at least one known binding site).

levels in the cell. Nevertheless, it is still clear that genes in *E. coli* do not strictly depend on TFs to be transcribed. This is in contrast with eukaryotic transcription where TFs are necessary for the transcription initiation process. Moreover, in many eukaryotes, TFs frequently bind in clusters [36], of which we find little evidence in Fig. 5.4.

We can compare the observed distribution of number of TF interactions per operon with the random promoter architecture model [see Models]. Looking at Fig. 5.5 we see some notable differences between the random promoter architecture model and the observed distribution. A larger number of operons are reported as unregulated in RegulonDB 8.5 than expected from the random promoter architecture model. One possible explanation could be that some TFs tend to bind cooperatively at multiple sites at a promoter region rather than in isolation, which would lead to a higher number of unregulated operons for a given number of binding sites. We will address this interesting question in more detail below. Another explanation could simply be that RegulonDB 8.5 is inherently biased and reports a higher fraction of unregulated operons than the actual value. The logic behind this hypothesis is that those operons for which there are known binding sites correspond in general to those that have been studied carefully. To take the latter possibility into account we subtract unregulated operons to consider the distribution of binding sites for only operons known to be regulated. In this case we update the prediction of the random promoter architecture model [Eq (5.1)] by first assigning one binding site to each of $N_{op}^{(reg)}$ regulated operons. Then we randomly distribute the

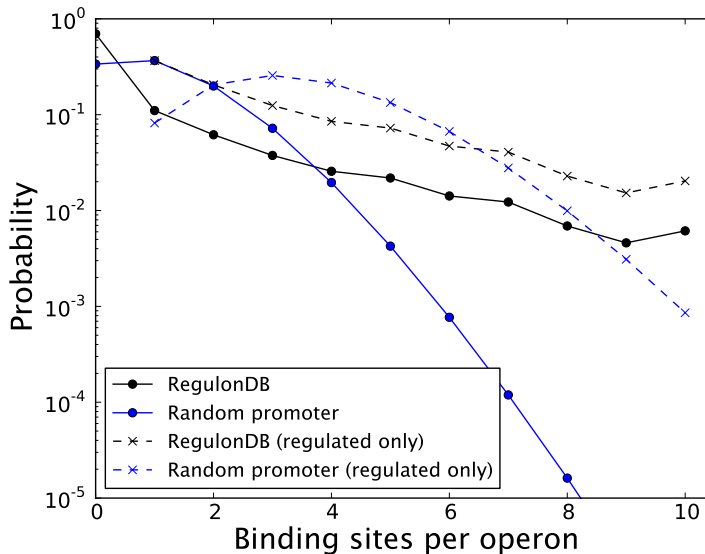


Figure 5.5: Distribution of number of TF binding sites per operon in RegulonDB 8.5 and the random promoter architecture model. Shown separately are distributions after neglecting unregulated operons (“regulated only”).

remaining $N_{bs} - N_{op}^{(reg)}$ binding sites on the $N_{op}^{(reg)}$ operons, as in Eq (5.1), leading to

$$P_{bs}(m; N_{bs}) = \binom{N_{bs} - N_{op}^{(reg)}}{m - 1} \left(\frac{1}{N_{op}^{(reg)}} \right)^{m-1} \left(1 - \frac{1}{N_{op}^{(reg)}} \right)^{N_{bs} - N_{op}^{(reg)} - m + 1}, \quad m = 1, 2, 3, \dots \quad (5.9)$$

In Fig. 5.5 we now observe an overrepresentation of operons regulated from a single binding site, compared to the random promoter architecture model (compare black and blue dashed lines). This supports the idea that *E. coli* generally favors simple regulatory strategies when possible. On the contrary there is also a small group of highly regulated operons. For example *gadAXW*, coding for genes in the acid resistance system[37], is regulated by 35 TF binding sites. The operon *csgDEFG*, coding for genes that regulate the assembly and transport of extracellular amyloid fibres (known as Curli) [38], is regulated by 33 TF binding sites, and the operon *glpTQ*, coding for genes responsible for the uptake of glycerol-3-phosphate [39], is regulated by 21 TF binding sites. These promoter architectures could virtually never ($P \approx 10^{-40} - 10^{-20}$, Eq. (5.1)) occur in the random promoter architecture model, and might as such be of interest for further study.

We can also use the random promoter architecture model to study the number of TF interactions per operon for particular TFs. We expect this number to be higher than suggested by the random promoter architecture model since a TF can regulate an operon cooperatively from multiple sites. As an example the well-studied Lac repressor has three known binding sites in *E. coli* [2], all regulating the same operon (*lacZYA*). Had these three sites been randomly distributed over all operons, it would

TF	Total number of binding sites	Operons regulated by multiple binding sites (RegulonDB)	Operons regulated by multiple binding sites (Random promoter)	p-value
OxyR	44	19	0.69	1.9×10^{-31}
ArgR	34	15	0.41	3.4×10^{-27}
NarP	21	10	0.16	7.7×10^{-22}
NarL	98	25	3.3	1.4×10^{-19}
Fis	237	52	18	2.7×10^{-17}
TyrR	19	8	0.13	2.0×10^{-16}
FlhDC	30	10	0.32	1.0×10^{-15}
IHF	114	25	4.5	1.5×10^{-15}
CRP	320	67	31	3.5×10^{-14}
CytR	23	8	0.19	4.8×10^{-14}
NagC	23	8	0.19	4.8×10^{-14}

Table 5.2: TFs regulating operons at multiple binding sites that differ significantly from the random promoter architecture model. The p-value for data in RegulonDB 8.5 is given by the probability of an equal or more extreme outcome in the random promoter architecture model.

have been an unlikely outcome for them all to regulate the same operon. In Table 5.2 we show the number of operons regulated at multiple binding sites for a given TF, both in RegulonDB 8.5 and as predicted by the random promoter architecture model [Eq. (5.4)]. Many of these TFs differ very significantly from the random promoter architecture model, which could be indicative of multiple TF binding domains (e.g. OxyR [40], ArgR [41]), cooperative binding (e.g. TyrR [42]), TFs which repress operons by DNA looping (e.g. NagC [43]), or chromosomal restructuring through repeated TF binding (e.g. Fis [44]). Interesting exceptions include Rob and MarA, which despite being common regulators do not bind to multiple binding sites at a single operon. Thus the random promoter architecture model allows to us identify TFs of special interest.

With a large number of targets we expect global TFs to be more abundantly expressed in the cell, to avoid running the risk of depleting the reservoir of TFs and hence the TF losing its ability to function effectively [45, 46]. In Fig. 5.6, we explore the relationship between TF copy number and corresponding number of binding sites, using two different genome-wide protein copy number censuses based on fluorescence measurements [47] and mass spectrometry [48]. The data shows a positive but not statistically significant linear relationship in the fluorescence data set (log-log slope= 0.14 ± 0.18), but not in the mass spectrometry data set (log-log slope= 0.01 ± 0.17), where we estimated the uncertainty in the linear fit parameter using the method of bootstrapping [49]. Large systematic deviations in the protein censuses [50, p. 43] makes them difficult to use as means for model testing.

Naively one could also imagine highly expressed genes to be subject to more regulation, because expressing too many of these would be energetically costly and expressing too few could have serious consequences to fitness of the cell. By combining binding site multiplicities from RegulonDB 8.5 with

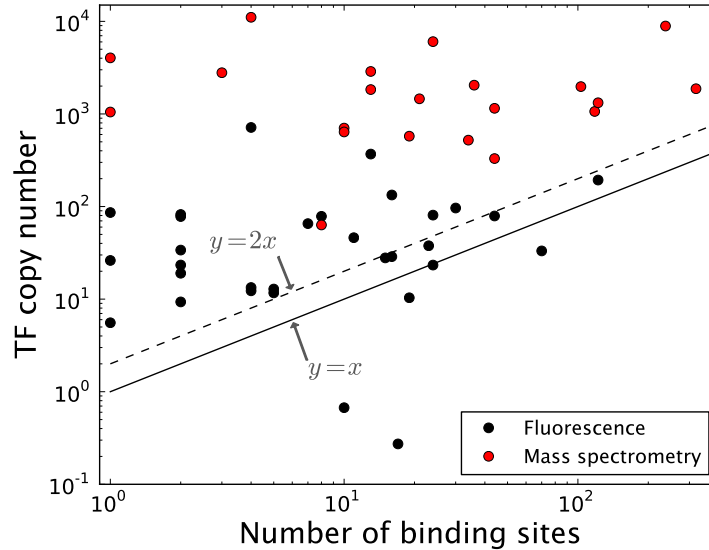


Figure 5.6: Number of TF binding sites (RegulonDB 8.5) vs. number of TF copy number as measured in [48, 47]. The two lines mark the critical boundary where the number of binding sites is large enough to deplete TFs binding as monomers (solid) or dimers (dashed). Updated version of figure published in [45].

the same protein copy number censuses [47, 48] we can explore the possible relationship between these two quantities. In Fig. 5.7 we show number of TF binding sites (RegulonDB 8.5) vs. number of TF copy number. Again we find a positive but not statistically significant relationship using the fluorescence based protein census (log-log slope= 0.20 ± 0.13), but not in the mass spectrometry protein census (log-log slope= 0.02 ± 0.08). Thus, it appears that highly expressed genes are not in general subject to more regulation than others.

5.3.2 How are activator and repressor binding sites configured?

Many genes need to be expressed only under conditions satisfying some “combinatorial rule”. For example the β -galactosidase enzyme LacZ in *E. coli*, which cleaves lactose into glucose, is only expressed if lactose is present and glucose, the more favored energy source, is not present [51]. To achieve combinatorial control TFs must be able to both activate and repress genes, and the configuration of the two types of interactions determines the regulatory response. In this section we will study promoter architectures in more detail and their influence on gene expression.

To classify promoter architectures we adapt the notation (A, R) for a promoter regulated by A activator and R repressor binding sites. Using RegulonDB we can easily find the distribution $P(A, R)$ for (A, R) with respect to all known regulatory interactions in *E. coli*. We show the most dominant promoter architectures in Fig. 5.8, along with their expected frequency in the “two-TF” random promoter architecture model described in Models [Eq. (5.3)]. We see an almost equal

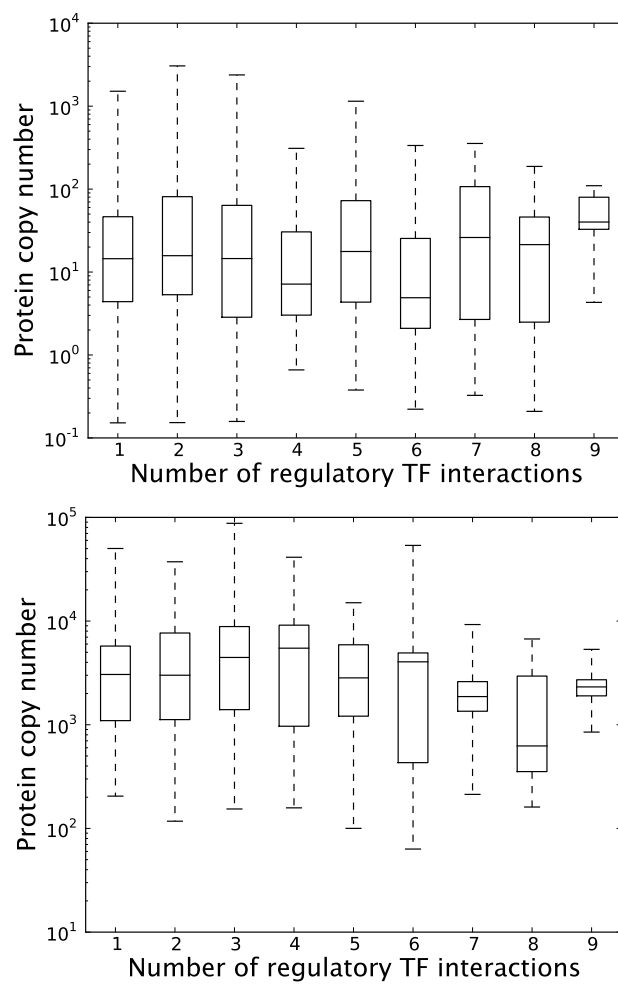


Figure 5.7: Measured protein copy number vs. number of TF binding sites regulating the transcription of the protein. The boxes show median, upper and lower quartiles, and the dashed lines show the range of the data. (Top) Protein data based on fluorescence measurements [47]. (Bottom) mass spectrometry [48].

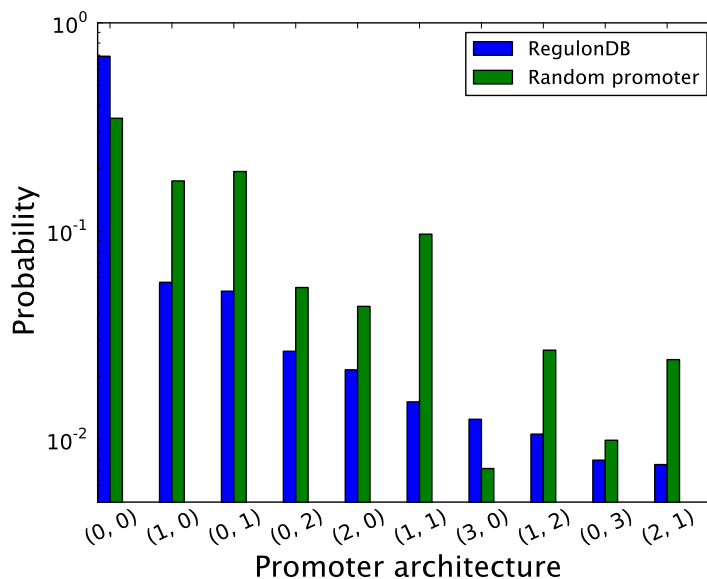


Figure 5.8: Frequency of the most dominant promoter architectures listed in RegulonDB 8.5 and their corresponding frequency in the random promoter architecture model. Binding site configurations with A activator and R repressor binding sites are denoted by (A, R) .

use of repressors and activators, 53 % vs. 47 % interactions, and for each promoter architecture ($A = a, R = r$) shown in Fig. 5.8 its symmetric counterpart ($A = r, R = a$) is almost equally present, both in absolute numbers and compared to the random promoter architecture model. Using the random promoter architecture model we can identify TF pairs which coregulate operons more frequently than one would expect by chance, a possible sign of TF-TF interactions or two TFs with otherwise related biological function [24]. In Table 5.3 we list the ten most such overrepresented TF pairs. The top pairs MarA, SoxS and Rob are all paralogous proteins, having around 45% identical amino acid sequence at their N-terminals [52], responsible for regulating various stress responses. FNR and ArcA are both global regulators responding to the availability of oxygen [53] in the cellular environment. NarL and NarP are homologous proteins responding to availability of nitrate and nitrite [54], and have been shown to act (anti)cooperatively with FNR [55, 56]. Fur and IHF are also global regulators, whose interplay with FNR has been investigated in [57, 58, 59]. GalR-GalS are homologous proteins responding to galactose [60], and GadX-GadW are homologous proteins responding to variations in pH level [37]. Even though TF pairs like Fis-CRP are more frequent coregulators (at 38 operons) in absolute numbers than any of the TF pairs listed in Table 5.3, this pair is still not particularly overrepresented when compared to the random promoter architecture model (p-value “only” 10^{-3}), and their frequent coregulation can simply be attributed to the large number of CRP and Fis binding sites. Hence the random promoter architecture model allows us to find the most interesting TF pairs.

TF 1	TF 2	Total binding sites (TF 1)	Total binding sites (TF 2)	Coregulated operons (RegulonDB)	Coregulated operons (Random promoter)	p-value
MarA	SoxS	24	29	18	0.26	5.5×10^{-34}
MarA	Rob	24	17	14	0.15	1.2×10^{-28}
SoxS	Rob	29	17	14	0.18	4.6×10^{-27}
FNR	ArcA	131	118	30	5.3	6.4×10^{-16}
FNR	NarL	131	98	27	4.5	3.0×10^{-15}
NarP	NarL	21	98	11	0.75	1.7×10^{-11}
FNR	Fur	131	122	24	5.5	2.8×10^{-10}
FNR	IHF	131	114	23	5.2	4.3×10^{-10}
GalR	GalS	12	12	5	0.054	5.5×10^{-10}
GadX	GadW	37	20	6	0.27	1.5×10^{-7}

Table 5.3: TF pairs that frequently coregulate operons. The p-value for data in RegulonDB 8.5 is given by the probability of an equal or more extreme outcome in the random promoter architecture model.

Having identified the most common promoter architectures we are curious to find out how these relate to gene expression. For example, are activated genes more abundantly expressed than repressed genes? To answer this question we identify all genes corresponding to a certain promoter architecture (A, R) in RegulonDB and acquire the protein copy number distribution of these genes from the two *E. coli* protein censuses [47, 48] [see Fig. 5.9]. Perhaps surprisingly, we find no systematic correlation between the number of activator and repressor binding sites, and gene expression in the two sets of 600-1000 genes. The only exception is the promoter architecture with one activator and one repressor binding site each (1,1), whose median expression level is higher than the upper quartile of the other five studied promoter architectures, indicating that genes with this architecture might be more abundantly expressed. The figure shows that even for a given promoter architecture there is a vast spread in protein copy number, spanning up to three orders of magnitude. It seems likely that all promoter architectures in Fig. 5.9 would be capable of producing proteins across the full range of biologically relevant concentrations. The main purpose of activators appears not to be increasing the maximum possible expression of a gene but rather, together with repressors, modulating it around a certain mean level. Such modulation can be achieved through other mechanisms, such as the ribosomal binding sequence (RBS) or promoter strength, which we will discuss in a later section.

5.3.3 Where are TF binding sites located?

There are many different ways in which TFs can regulate the transcription rate of a gene. Perhaps most intuitively TFs can facilitate or block RNAP from interacting with a promoter of interest, to either activate or repress transcription of a gene [15]. However, TFs can modulate basically any step in the chain of events preceding promoter escape [61], or modify the DNA methylation

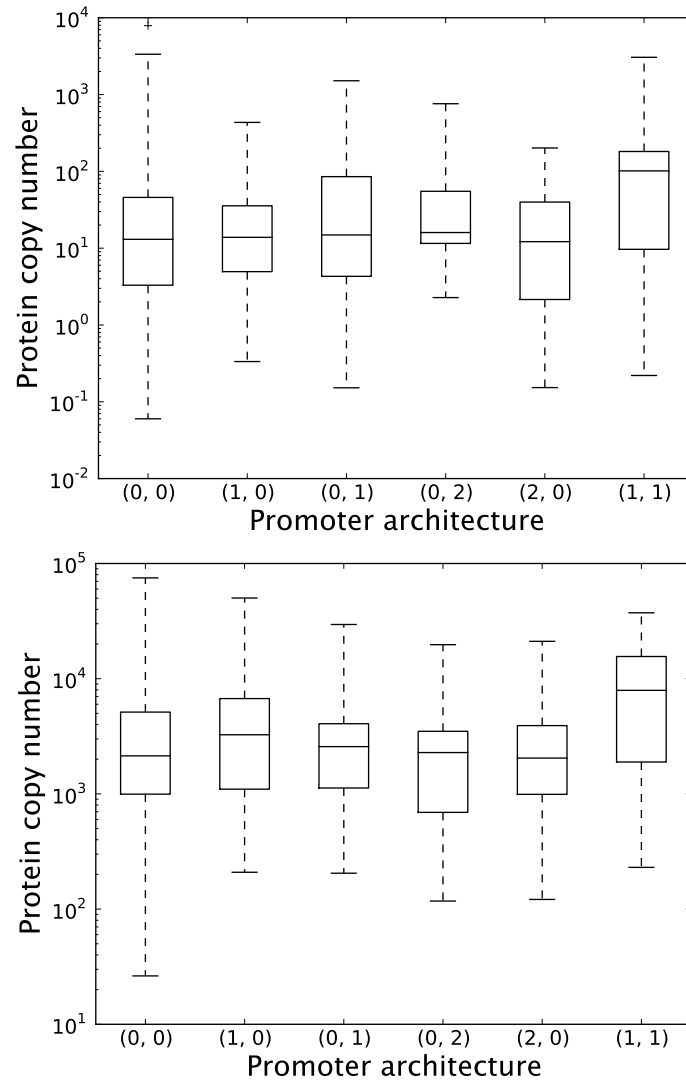


Figure 5.9: Protein copy number as a function of promoter architecture for the most common architectures. The notation (i, j) represents a promoter with i activator and j repressor binding sites. (Top) protein data based on fluorescence measurements [47]. (Bottom) mass spectrometry [48].

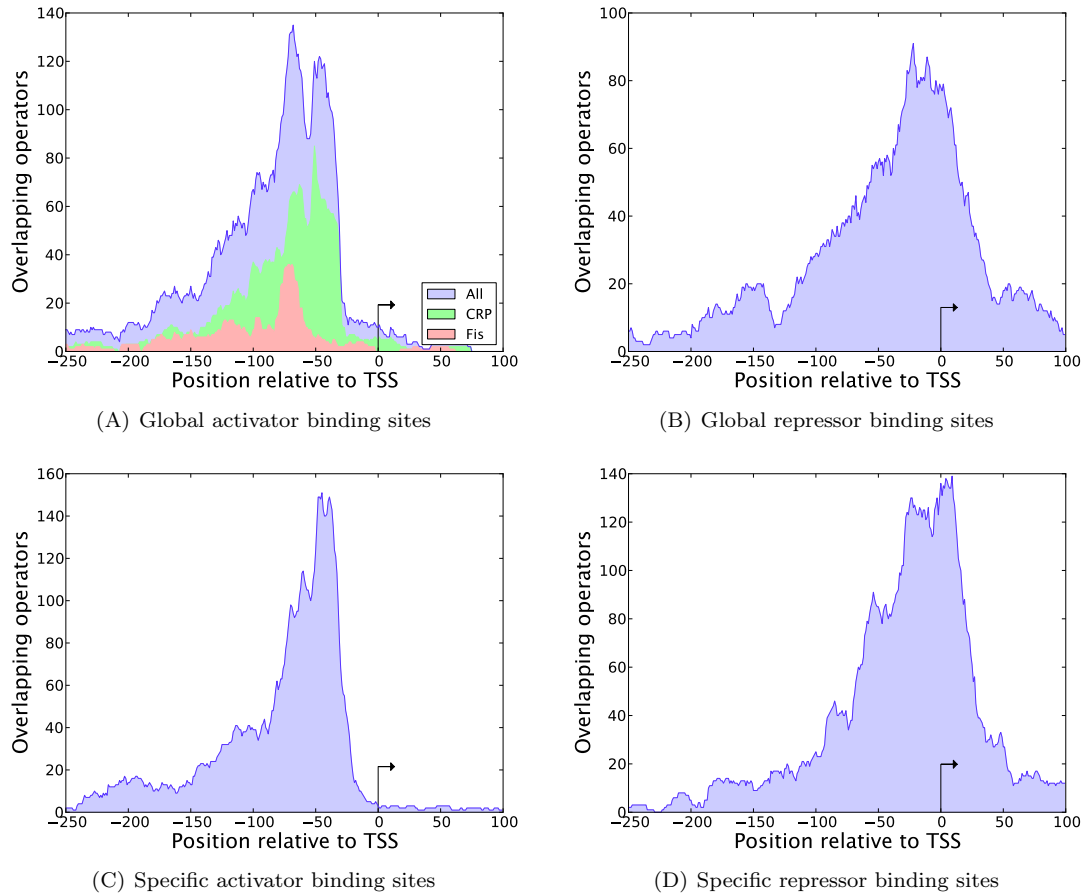


Figure 5.10: Distribution of activating and repressing binding sites bound by global TFs and specific TFs, respectively. The y-axis shows number of binding sites overlapping each nucleotide position, after aligning all promoters with respect to their transcription start site for the different kinds of TFs. Similar figures were reported in [19] using an earlier version of RegulonDB.

or compactification states [62]. In eukaryotes, where the latter regulatory strategies are common, TF binding sites can be located hundreds of thousands of base pairs away from the transcription start site, which means that DNA needs to “loop” to establish a contact between TF and RNAP (if necessary for regulation) [62]. Hence each class of transcriptional regulation will have its own TF binding profile, and in this section we will investigate these profiles in more detail for *E. coli*.

After aligning all known promoters with respect to their transcription start site we can make a histogram [see Fig. 5.10] over the number of binding sites overlapping each nucleotide position. The lack of nucleosomes in prokaryotes leads to a much narrower TF binding profile compared to eukaryotes; in fact 75 % of all reported TF interactions in RegulonDB 8.5 take place within 100 bp of the transcription start site. Activator and repressor binding sites have fundamentally different profiles; whereas repressors overlap the RNAP binding site for maximum repression, activators facilitate transcription initiation from upstream of the -35 region. TFs binding significantly upstream

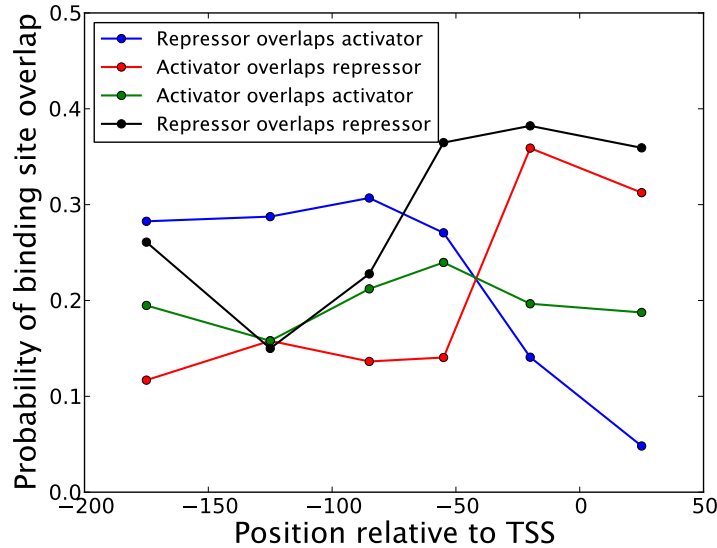


Figure 5.11: Probability of TF binding site overlap. Binding sites are defined as an interval of nucleotides from the 5×10^6 bp *E. coli* genome taken up by a TF upon binding. Two binding sites sharing one or more nucleotides are considered to be overlapping, independently of which strand the TFs bind.

of -35 bp would to a larger extent need to loop DNA to interact directly with RNAP, or regulate expression of genes through other long range mechanisms. An interesting difference between specific activators [Fig. 5.10(B)] and global activators [Fig. 5.10(A)], is that the latter have two separate peaks, located at -70 bp and -45 bp respectively, rather than one. The TFs whose contribution dominates these two peaks, which should correspond to class I and class II activation [63, 61], are CRP and Fis (shown separately in Fig. 5.10(A)). Class I activators interact with the α -CTD domain, whereas class II activators interact directly with the σ factor. Although most repressors function by blocking RNAP from binding the promoter, still around 27 % of the repressors bind upstream of -70 bp, i.e. without the possibility of blocking RNAP. One possible way these upstream repressors could function is by preventing activators from binding the promoter. To test this theory we show in Fig. 5.11 the probability of a repressor binding site to overlap with an activator binding site as a function of position, using the probability for two activators to overlap as a control. The results show that around 30 % of the repressors binding upstream of -70 bp overlap with an activator, compared to 15-20 % for two different activators. This suggests that blocking of activators is an important regulatory strategy for upstream repressors but not the only one, as a large fraction of upstream repressors inhibit transcription through other means. Additional mechanisms through which an upstream repressor could inhibit an activator from accessing its binding site without overlapping it include DNA allostery [10] or DNA bending [64, 65]. In total, almost half of all binding sites reported in RegulonDB 8.5 overlap with other binding sites, which leads us to believe that this constitutes

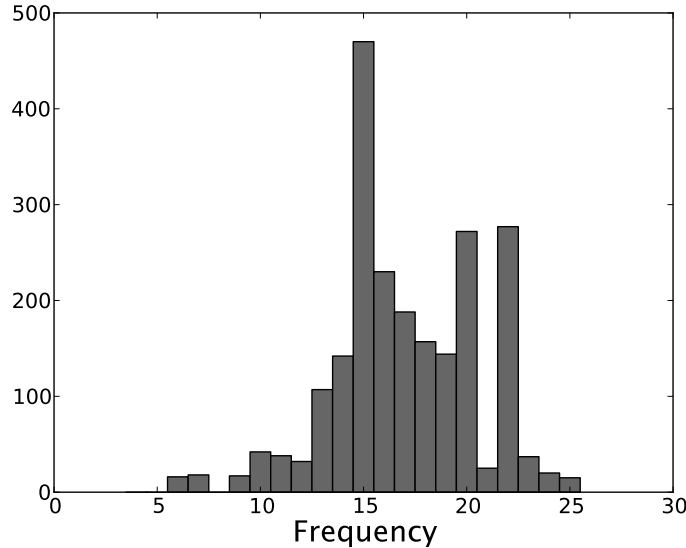


Figure 5.12: Footprint size (in base pairs) for all TF-DNA interactions (RegulonDB 8.5). Mean footprint size: 17.3 bp. Updated version of figure published in [67].

an important regulatory strategy. As more binding sites are discovered, the number of overlapping binding sites will likely increase noticeably, just as the probability of two students in a class having birthday on the same day goes up rapidly with the number of students. Interestingly, TFs often (37 % of the reported overlapping interactions) overlap with themselves. For example, out of the 88 known Fur binding sites, 75% of them overlap with other Fur binding sites [66].

Since the regulatory region of a gene is of limited size, TFs need to compete for space at promoters with other binding sites, in particular TFs which interact directly with RNAP. To study this “real estate” problem we first collect the footprint size of all TF-DNA interaction sites reported in RegulonDB 8.5 [see Fig. 5.12]. A similar figure is reported in [67] using an earlier version of Regulon DB. Some of the notable peaks in Fig. 5.12 correspond mainly to binding of global TFs: Fis (15 bp), ArcA (15 bp) and CRP (22 bp). Most TFs interact with a region of around 15 bp (although outliers exist) which means that one could theoretically fit three nonoverlapping binding sites within 50 bp. Since the majority of operons reportedly have fewer than this number of binding sites [see Fig. 5.4], the size of the regulatory region does in general not seem to be a major constraining factor. However, for promoters with a larger number of binding sites, of which we saw some examples in Fig. 5.4, TFs would either need to bend DNA to access RNAP, or overlap with other TFs. To further study the real estate of the promoter we look at the separation between binding sites [see Fig. 5.13], which shows the the edge-to-edge distance for nonoverlapping adjacent binding sites. The majority of binding sites in this set are separated by less than 15 bp from their neighbors. Hence for an operon with three binding sites the regulatory region would be expected to take up around $3 \times 15 + 2 \times 15 = 75$ bp, around the same as observed in Fig. 5.10.

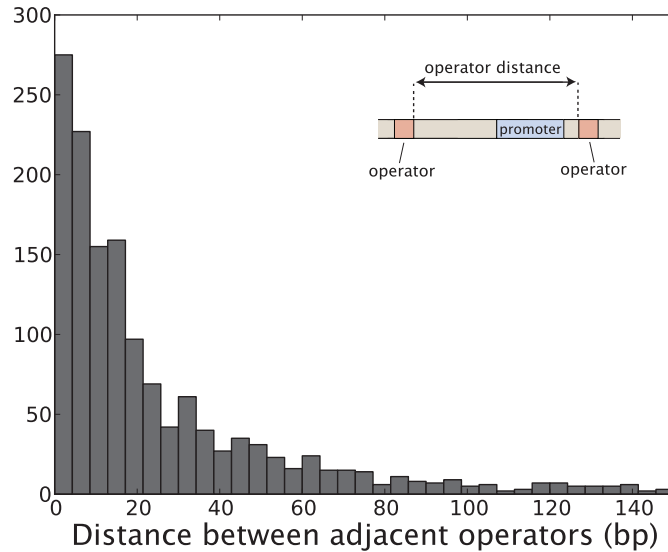


Figure 5.13: Edge to edge distance between adjacent binding sites (RegulonDB 8.5). Figure does not include binding sites separated by more than 150 bp, which would likely correspond to regulation of different operons.

5.3.4 How does promoter architecture relate to promoter strength?

Many prokaryotic genes do not rely on TFs for regulation, and will be constitutively expressed independently of the cellular environment. The production of these genes will, at our current best understanding, only be affected by the global availability of RNAP, sigma factors, ribosomes and the interaction strengths with these different complexes. For proteins which are in constant demand, constitutive expression provides a simple and efficient choice of promoter architecture. Despite its simplicity, constitutive expression allows an impressive dynamic range in protein production, spanning at least three orders of magnitude, as we found evidence for in Fig. 5.9. This demonstrates the power of the basal production machinery, whose transcriptional component we will study further in this section. In particular we will be interested in the relationship between promoter strength and regulation by TFs.

In *E. coli* the transcription rate of a gene can vary by up to three orders of magnitude due to differences in the RNAP affinity alone [25], not taking TFs into account. To illustrate this point we use the linear RNAP-DNA interaction model introduced in Models to predict the binding energy to all known σ^{70} promoters along with the corresponding distribution for nonspecific binding [see Fig. 5.14]. As expected we get two separate distributions, where RNAP binds on average $2.4 k_B T$ more strongly to known promoters than sequences chosen randomly from the *E. coli* genome. The predicted RNAP binding energy distribution spans roughly $8 k_B T$ from the strongest to the weakest promoter, corresponding to a predicted 3000-fold difference in RNAP binding affinity. This difference is similar to that found between the most abundantly expressed proteins (e.g. CRP) and scarcely

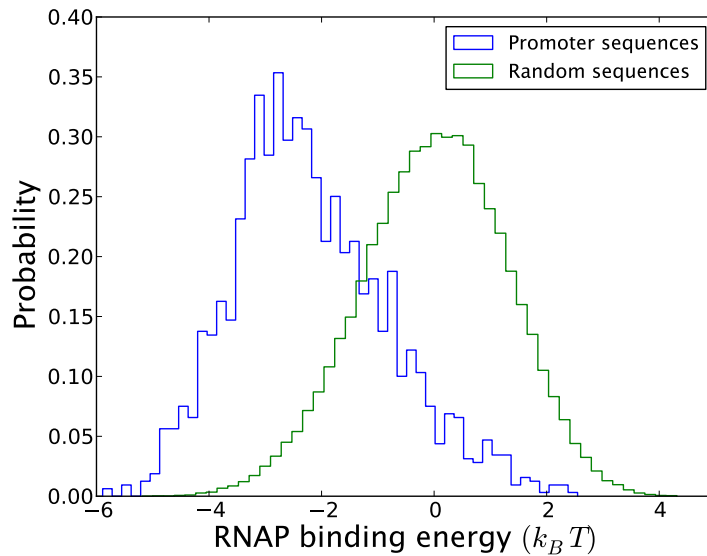


Figure 5.14: Predicted RNAP binding energy [25] for promoters in RegulonDB 8.5 and DNA sequences randomly chosen from the *E. coli* genome. The spacer region is allowed to range from 16-18 bp, and the -10 box is allowed to deviate by one base pair up or down from its consensus position. The RNAP binding energy is taken as the minimum binding energy of the $3 \times 3 = 9$ possible binding configurations.

expressed proteins (e.g. LacI) [68, 69] in *E. coli*, suggesting that promoter strength alone might be a powerful enough tool to set the mean level of gene expression to most biologically relevant values. A disconcerting observation from Fig. 5.14, however, is that 200,000 sites or so in the 5×10^6 bp *E. coli* background interact more strongly with RNAP than the typical promoter. This raises several important questions [70, 71, 72]: Is the linear energy model missing key information, or can all the predicted promoters in principle produce transcripts? Do weak promoters need to be activated by TFs to function? Although trying to solve these important questions falls outside the scope of the current chapter, we note that the paradox might originate from the fact that the promoter sequence encodes detailed information about both RNAP binding affinity, open complex formation rate and promoter escape rate [73, 74] in a way that likely cannot be captured in a simple linear model. Powerful new methods such as RNA-seq [75] could provide further insight into which of the 200,000 predicted promoters are actually transcriptionally active.

In Fig. 5.9 we learned that the number of activator or repressor binding sites did not seem to have any systematic effect on the average gene expression in two sets of 600-1000 genes. Since activators, by definition, increase the expression of a gene and repressors reduce it, the only possible explanation for this observation (if true) is that repressed genes have a higher basal level of expression. This could, for example, be the result if repressed genes have a higher affinity (*promoter strength*) for RNAP to their promoters. Since stronger promoters recruit RNAP more easily they would hence

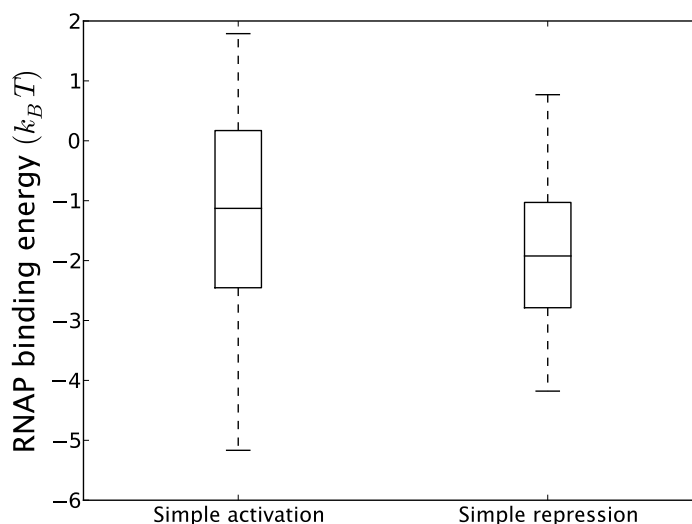


Figure 5.15: Predicted RNAP binding energy to promoters in the simple activation (1,0) and simple repression architecture (0,1). Operons whose transcription is initiated from multiple promoters are excluded.

become transcriptionally more active. To investigate the relationship between promoter strength and promoter architecture we show in Fig. 5.15 the RNAP binding energy distribution for genes which according to RegulonDB 8.5 are regulated from a single activator or repressor binding site. Our data suggest, though not conclusively, that RNAP binds more strongly to the promoters of repressed genes, $E_{\text{repressed}} = -1.9 \pm 0.17$, than promoters of activated genes, $E_{\text{activated}} = -1.1 \pm 0.24$. The reported uncertainty in the median is given by the standard deviation divided by the square root of number of data points (≈ 50). Our results suggest that repressed genes have higher basal rate of transcription, providing a possible explanation as to why we do not see a significant differences in gene expression as compared to activated genes. Conversely, weak promoters are more likely to be activated by TFs, suggesting that these promoters might not work effectively without TF activation.

5.4 Discussion

After more than half a century of intense study *E. coli* remains one of the most important model organisms in biology. In order to make the vast pool of knowledge obtained from these studies publicly available and directly accessible, ambitious initiatives such as RegulonDB have screened thousands of references to collect information relating to TF binding site locations, operon organization, and much more. Although this annotation process is far from complete, as more than half of the *E. coli* genes still lack any known regulation, we now have a better opportunity than ever to study

regulatory interactions in detail.

In this study we have analyzed TF-DNA interactions reported in RegulonDB 8.5, and found distinct differences in binding site locations depending on TF type (activator, repressor, global TF or specific TF). Using an analytic random promoter architecture model, analogous to random genetic networks, we concluded that most promoters in *E. coli* are less heavily regulated than one would expect (with some interesting exceptions). In fact the majority of operons listed in RegulonDB 8.5 have fewer than three associated TF binding sites, and the majority of TFs regulate fewer than three operons, suggesting that many *E. coli* activities can be performed with little “oversight”. The random promoter architecture model further allowed us to identify, with well defined statistical significance, homologous or otherwise related pairs of TFs. Perhaps surprisingly we found no systematic correlation between the number of activating or repressing TF interactions and expression of a gene, as measured by two different genome-wide protein censuses covering 600-1000 genes. A position weight matrix model of RNAP provided some evidence that this might be related to a higher basal rate of transcription for repressed genes compared to activated genes

One of the grand challenges of physical biology is to be able to construct predictive maps between promoter nucleotide sequence and gene expression. Increasingly accurate promoter architecture data, found e.g. using powerful techniques like ChIP-seq, allow predictive maps to be both tested and refined. A difficulty with mapping promoter architecture and gene expression, apart from lacking complete knowledge of the regulatory network, is a substantial disagreement on protein concentrations as measured using different experimental methods and under different experimental conditions. The protein copy numbers measured using mass spectrometry [48] are for example on average at least one order of magnitude higher than for the same proteins measured with fluorescence based techniques [47], though these kinds of effects can be due to different growth conditions for the cells. As TF copy number plays a central role in regulatory function, we believe resolving these discrepancies will be a necessary step for a deeper understanding of several important aspects of gene regulation. Ultimately, predictive maps of gene expression combined with accurate promoter architecture data could give an opportunity to understand *in silico* how many processes in bacteria are regulated, and hence provide important guidelines to experimentalists.

5.5 Acknowledgments

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01 GM085286 and R01 GM085286-01S (M.R., H.G.G., R.P), as well as National Institutes of Health Pioneer award DP1 OD000217 (R.P.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. H.G.G. holds a Career Award at

the Scientific Interface from the Burroughs Wellcome Fund.

References

- [1] Garcia, H. G. & Phillips, R. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A* **108**, 12173–8 (2011).
- [2] Oehler, S., Eismann, E. R., Kramer, H. & Muller-Hill, B. The three operators of the *lac* operon cooperate in repression. *EMBO J* **9**, 973–9 (1990).
- [3] Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: A quantitative approach. *Nat Rev Genet* **10**, 443–56 (2009).
- [4] Zentner, G. E. & Henikoff, S. Regulation of nucleosome dynamics by histone modifications. *Nat. Struct. Mol. Biol.* **20**, 259–266 (2013).
- [5] Salgado, H. *et al.* RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* **41**, D203–213 (2013).
- [6] Keseler, I. M. *et al.* EcoCyc: A comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.* **39**, D583–590 (2011).
- [7] Ackers, G. K., Johnson, A. D. & Shea, M. A. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A* **79**, 1129–33 (1982).
- [8] Müller, J., Oehler, S. & Müller-Hill, B. Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J Mol Biol* **257**, 21–9 (1996).
- [9] Garcia, H. G. *et al.* Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Rep* **2**, 150–161 (2012).
- [10] Kim, S. *et al.* Probing allostery through DNA. *Science* **339**, 816–819 (2013).
- [11] Ryu, S., Fujita, N., Ishihama, A. & Adhya, S. GalR-mediated repression and activation of hybrid lacUV5 promoter: differential contacts with RNA polymerase. *Gene* **223**, 235–45 (1998).
- [12] Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–502 (2007).

- [13] Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9158–9163 (2010).
- [14] Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits (Chapman & Hall/CRC Mathematical & Computational Biology)* (Chapman and Hall/CRC, 2006), First edn.
- [15] Bintu, L. *et al.* Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev* **15**, 116–24 (2005).
- [16] Bintu, L. *et al.* Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev* **15**, 125–35 (2005).
- [17] Sherman, M. S. & Cohen, B. A. Thermodynamic state ensemble models of cis-regulation. *PLoS Computational Biology* **8**, e1002407 (2012).
- [18] Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A* **100**, 5136–41 (2003).
- [19] Madan Babu, M. & Teichmann, S. A. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* **31**, 1234–1244 (2003).
- [20] Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S. & Ishihama, A. Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. *J. Bacteriol.* **181**, 6361–6370 (1999).
- [21] Abramowitz, M. & Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, 1964), Ninth edn.
- [22] Nishimura, K. & Sibuya, M. Occupancy with two types of balls. *Annals of the Institute of Statistical Mathematics* **40**, 77–91 (1988).
- [23] Reif, F. *Fundamentals of statistical and thermal physics* (McGraw-Hill, New York,, 1965).
- [24] Ho Sui, S. J., Fulton, D. L., Arenillas, D. J., Kwon, A. T. & Wasserman, W. W. oPOSSUM: Integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res.* **35**, W245–252 (2007).
- [25] Brewster, R. C., Jones, D. L. & Phillips, R. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput. Biol.* **8**, e1002811 (2012).

- [26] Mulligan, M. E., Hawley, D. K., Entriken, R. & McClure, W. R. *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic Acids Research* **12**, 789–800 (1984).
- [27] Brunner, M. & Bujard, H. Promoter recognition and promoter strength in the *Escherichia coli* system. *Embo J* **6**, 3139–44 (1987).
- [28] Stormo, G. D. DNA binding sites: Representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
- [29] Gross, C. A. *et al.* The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harb. Symp. Quant. Biol.* **63**, 141–155 (1998).
- [30] Macnab, R. M. Genetics and biogenesis of bacterial flagella. *Annu. Rev. Genet.* **26**, 131–158 (1992).
- [31] Mermod, M., Magnani, D., Solioz, M. & Stoyanov, J. V. The copper-inducible ComR (YcfQ) repressor regulates expression of ComC (YcfR), which affects copper permeability of the outer membrane of *Escherichia coli*. *Biometals* **25**, 33–43 (2012).
- [32] Hirashima, A., Wang, S. & Inouye, M. Cell-free synthesis of a specific lipoprotein of the *Escherichia coli* outer membrane directed by purified messenger RNA. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4149–4153 (1974).
- [33] Takahashi, S., Hours, C., Chu, A. & Denhardt, D. T. The rep mutation. VI. Purification and properties of the *Escherichia coli* rep protein, DNA helicase III. *Can. J. Biochem.* **57**, 855–866 (1979).
- [34] Laimins, L. A., Rhoads, D. B. & Epstein, W. Osmotic control of kdp operon expression in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **78**, 464–468 (1981).
- [35] Thelen, P., Tsuchiya, T. & Goldberg, E. B. Characterization and mapping of a major Na⁺/H⁺ antiporter gene of *Escherichia coli*. *J. Bacteriol.* **173**, 6553–6557 (1991).
- [36] Gotea, V. *et al.* Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* **20**, 565–577 (2010).
- [37] Tramonti, A., De Canio, M. & De Biase, D. GadX/GadW-dependent regulation of the *Escherichia coli* acid fitness island: Transcriptional control at the gadY-gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. *Mol. Microbiol.* **70**, 965–982 (2008).

- [38] Hammar, M., Arnqvist, A., Bian, Z., Olsen, A. & Normark, S. Expression of two *csg* operons is required for production of fibronectin- and congo red-binding curli polymers in *Escherichia coli* K-12. *Mol. Microbiol.* **18**, 661–670 (1995).
- [39] Rao, N. N., Roberts, M. F., Torriani, A. & Yashphe, J. Effect of *glpT* and *glpD* mutations on expression of the *phoA* gene in *Escherichia coli*. *J. Bacteriol.* **175**, 74–79 (1993).
- [40] Toledano, M. B. *et al.* Redox-dependent shift of OxyR-DNA contacts along an extended DNA-binding site: A mechanism for differential promoter selection. *Cell* **78**, 897–909 (1994).
- [41] Tian, G., Lim, D., Carey, J. & Maas, W. K. Binding of the arginine repressor of *Escherichia coli* K12 to its operator sites. *J. Mol. Biol.* **226**, 387–397 (1992).
- [42] Bai, Q. & Somerville, R. L. Integration host factor and cyclic AMP receptor protein are required for TyrR-mediated activation of *tpl* in *Citrobacter freundii*. *J. Bacteriol.* **180**, 6173–6186 (1998).
- [43] Plumbridge, J. & Kolb, A. DNA loop formation between Nag repressor molecules bound to its two operator sites is necessary for repression of the *nag* regulon of *Escherichia coli* in vivo. *Mol. Microbiol.* **10**, 973–981 (1993).
- [44] Schneider, R. *et al.* An architectural role of the *Escherichia coli* chromatin protein FIS in organising DNA. *Nucleic Acids Res.* **29**, 5107–5114 (2001).
- [45] Rydenfelt, M., Cox, R. S., Garcia, H. & Phillips, R. Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. *Phys. Rev. E* **89**, 012702 (2014).
- [46] Brewster, R. C. *et al.* The transcription factor titration effect dictates level of gene expression. *Cell* **156**, 1312–1323 (2014).
- [47] Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
- [48] Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**, 117–24 (2007).
- [49] Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Statistics* **7**, 1–26 (1979).
- [50] Phillips, R., Kondev, J., Theriot, J. & Garcia, H. *Physical Biology of the Cell*, 2nd ed. (Garland Science, New York, 2013).
- [51] Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).

- [52] Cohen, S. P., Hachler, H. & Levy, S. B. Genetic and functional analysis of the multiple antibiotic resistance (*mar*) locus in *Escherichia coli*. *J. Bacteriol.* **175**, 1484–1492 (1993).
- [53] Levanon, S. S., San, K. Y. & Bennett, G. N. Effect of oxygen on the *Escherichia coli* ArcA and FNR regulation systems and metabolic responses. *Biotechnol. Bioeng.* **89**, 556–564 (2005).
- [54] Rabin, R. S. & Stewart, V. Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate- and nitrite-regulated gene expression in *Escherichia coli* K-12. *J. Bacteriol.* **175**, 3259–3268 (1993).
- [55] Darwin, A. J., Ziegelhoffer, E. C., Kiley, P. J. & Stewart, V. Fnr, NarP, and NarL regulation of *Escherichia coli* K-12 *napF* (periplasmic nitrate reductase) operon transcription in vitro. *J. Bacteriol.* **180**, 4192–4198 (1998).
- [56] Overton, T. W. *et al.* Microarray analysis of gene regulation by oxygen, nitrate, nitrite, FNR, NarL and NarP during anaerobic growth of *Escherichia coli*: New insights into microbial physiology. *Biochem. Soc. Trans.* **34**, 104–107 (2006).
- [57] Myers, K. S. *et al.* Genome-scale analysis of *Escherichia coli* FNR reveals complex features of transcription factor binding. *PLoS Genet.* **9**, e1003565 (2013).
- [58] Mettert, E. L. & Kiley, P. J. Contributions of [4Fe-4S]-FNR and integration host factor to fnr transcriptional regulation. *J. Bacteriol.* **189**, 3036–3043 (2007).
- [59] Troxell, B., Fink, R. C., Porwollik, S., McClelland, M. & Hassan, H. M. The Fur regulon in anaerobically grown *Salmonella enterica* sv. Typhimurium: Identification of new Fur targets. *BMC Microbiol.* **11**, 236 (2011).
- [60] Weickert, M. J. & Adhya, S. Isorepressor of the gal regulon in *Escherichia coli*. *J. Mol. Biol.* **226**, 69–83 (1992).
- [61] Dove, S. L., Joung, J. K. & Hochschild, A. Activation of prokaryotic transcription through arbitrary protein-protein contacts. *Nature* **386**, 627–630 (1997).
- [62] Ptashne, M. & Gann, A. *Genes and Signals* (Cold Spring Harbor Laboratory Press, New York, 2002).
- [63] Busby, S. & Ebright, R. H. Transcription activation by catabolite activator protein (CAP). *J Mol Biol* **293**, 199–213 (1999).
- [64] Pérez-Martín, J., Rojo, F. & De Lorenzo, V. Promoters responsive to DNA bending: A common theme in prokaryotic gene expression. *Microbiological Reviews* **58**, 268–290 (1994).

- [65] Kim, J., Zwieb, C., Wu, C. & Adhya, S. Bending of DNA by gene-regulatory proteins: Construction and use of a DNA bending vector. *Gene* **85**, 15–23 (1989).
- [66] Chen, Z. *et al.* Discovery of Fur binding site clusters in *Escherichia coli* by information theory models. *Nucleic Acids Res.* **35**, 6762–6777 (2007).
- [67] Ruths, T. & Nakhleh, L. Neutral forces acting on intragenomic variability shape the *Escherichia coli* regulatory network topology. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 7754–7759 (2013).
- [68] Ishihama, Y. *et al.* Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* **9**, 102 (2008).
- [69] Gilbert, W. & Muller-Hill, B. Isolation of the Lac Repressor. *Proc Natl Acad Sci U S A* **56**, 1891–1898 (1966).
- [70] Djordjevic, M. Efficient transcription initiation in bacteria: An interplay of protein-DNA interaction parameters. *Integr Biol (Camb)* **5**, 796–806 (2013).
- [71] Gershenzon, N. I., Stormo, G. D. & Ioshikhes, I. P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res.* **33**, 2290–2301 (2005).
- [72] Djordjevic, M., Sengupta, A. M. & Shraiman, B. I. A biophysical approach to transcription factor binding site discovery. *Genome Res.* **13**, 2381–2390 (2003).
- [73] McClure, W. R. Rate-limiting steps in RNA chain initiation. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5634–5638 (1980).
- [74] McClure, W. R. Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.* **54**, 171–204 (1985).
- [75] Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

Chapter 6

Conclusion

To survive in an unpredictable environment, such as the 20 km thick life-sustaining region near the surface of the earth, living organisms depend critically on their ability to sense and respond to external and internal stimuli. These responses are programmed in the DNA of every cell, which carry instructions for how to produce the right configuration of proteins, at the right (average) concentrations, and at the right time. These actions are not only carried out by proteins, but also regulated by proteins, which are connected into a complex genetic network. Regulatory networks are under immense pressure from mutations to disintegrate, and their ability to stay, largely, intact over time demonstrates the critical role of gene regulation for the survival of an organism.

Regulatory proteins which bind DNA, or transcription factors (TFs), typically target not only a single gene but a multitude of genes, either identical gene copies (for example located on plasmids or viral vectors), or different coregulated genes. Moreover, for each regulated gene a TF can have a varying number of binding sites. This leads to a large number of possible ways a given set of TFs or binding sites can be distributed over a set of genes. Using the classical problem of “balls and bins” from mathematical combinatorics, we here derive such distributions and show that they can provide valuable insights into gene regulation

Interestingly, many TFs are believed to exist at a concentration comparable to their target binding sites. As TFs bind to their targets, fewer remain available to also regulate the other genes, resulting in “entangled” transcription of the coregulated genes. One of the major results presented in this work is a general thermodynamic model of transcriptional regulation that takes the promoter entanglement effect into account, and predicts the fold change in gene expression, as well as the correlation in transcription between different genes. One of the signature predictions of this model is a sharp drop in gene expression when the copy number of a repressor exceeds the number of regulated genes.

To test our theoretical predictions experimentally, we design an experiment where we can systematically tune the various model parameters involved, such as the TF copy number, binding site affinities, or gene copy number. All these parameters are experimentally measured or taken from

literature, which allows us to test the theoretical predictions using no free fit parameters. Our results clearly prove the existence of the TF titration effect, and that its influence on transcriptional regulation can (by biology standards), be accurately predicted by the thermodynamic model. An interesting side effect of TF titration is that it can be exploited to measure the plasmid copy number, as well as its variance. This can otherwise be an expensive and time-consuming operation.

Our experiments show that the thermodynamic model can predict the fold change in gene expression as a function of different model parameters for some specific genetic constructs. However, a grand challenge of quantitative biology is to be able to use theoretical models to predict the expression for *every* gene throughout an organism. To take a small step in this direction, we analyze the thousands of known regulatory interactions in *E. coli*, the model organism which arguably demonstrates the most well-explored regulatory network, to better understand the strategies of transcriptional regulation in a real organism. Specifically, we look at the number of TF binding sites per promoter, binding site positions, and their regulatory action (activation or repression) for every annotated gene. To identify frequently biologically recurring strategies in transcriptional regulation we compare our findings to a “random promoter architecture model”, analogous to random genetic networks used in the study of genetic network motifs, where promoters are constructed by random “sprinkling” of a set of binding sites. Interestingly, many of the mathematical ideas underlying the thermodynamic model can here be reused in a very different context. The random promoter architecture model can also be used to identify TF pairs which coregulate genes more frequently than one would expect by chance, indicating that these TFs might have a related biological function. In fact, many such overrepresented TFs in *E. coli* turn out to be homologous.

Next, we study whether promoter architecture has any systematic effect on the average gene expression in *E. coli*, using two different genome-wide expression censuses. Our preliminary results show, surprisingly, that there seems to be no such strong dependence. Before taking these results for granted one should, however, bear in mind that the regulatory data of *E. coli* is still incomplete, and that genome-wide expression measurements might still not be considered fully reliable.

The goal to understand and accurately predict gene expression in a genome-wide setting, presents a great but exciting challenge in quantitative biology. Such modeling capabilities could open doors to profound applications in other fields where genetic circuits play a role. The work presented here takes a step in this direction, by generalizing a previously successful thermodynamic model of transcriptional regulation, taking into account that TFs in general regulate many different genes or gene copies. Many problems, however, remain to be addressed. The thermodynamic model has little to say about dynamical properties of genetic networks, and fails to describe nonequilibrium processes, such as abortive transcription initiation or regulatory action of TFs which is not only related to the binding occupancy. An even greater challenge, perhaps, is to obtain information about promoter architectures, e.g. binding site identities to feed the quantitative models, for thousands of genes in

an organism at once. Current approaches that explore promoter architectures one at a time would need further development to be suitable for genome-wide studies. However, given the rapid advances in both the experimental and theoretical understanding of regulatory processes in a cell, there are good reasons to be optimistic that many valuable future lessons will be learned about gene regulation at a larger scale.