

DNA Mechanics and Transcriptional Regulation in the *E. coli* *lac* operon

Thesis by

Stephanie Johnson

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2012

(Defended April 18, 2012)

© 2012

Stephanie Johnson

All Rights Reserved

Dedicated to the memory of Jonathan Widom (1955–2011).

Jon was to me a consummate biophysicist, who spoke with equal proficiency the languages of both physics and biology, and whose unique perspective and creativity were evident in all my conversations with him, as well as in the ideas for which he is well known in the broader scientific community, such as the “mechanical code” to genomes that was a main focus of his work. More admirable even than his scientific prowess, though, was his comportment in interactions with other scientists: unfailingly polite and considerate despite being at the center of several heated controversies, conveying equal regard and thoughtfulness for a graduate student’s remarks as any professor’s; in Rob’s words in Jon’s obituary, “He loved life, and everyone around him felt happier and smarter in his presence.”

Acknowledgements

I have loved my time at Caltech and readily acknowledge that this is largely due to the people with whom I have had the privilege of interacting these past six years. First and foremost, my advisor Rob, who has been great fun to work with, and who has been instrumental in both my scientific and personal growth throughout my time at Caltech. Rob's impact on my professional development is almost too extensive for words, but is probably best exemplified by one of my first interactions with him, as a rotation student in his lab during my first year: I proposed an experiment to try to measure the effects of gyrase without a magnetic trap, and was all set to order the necessary components when Rob rocked my world by asking, "That's an interesting idea, but why don't you do a calculation to estimate if the experiment will work or not?" (The calculation gave the experiment a "maybe;" my rotation ended before I could try it.) I have learned an amazing amount from Rob, both about quantitative biology and about bigger life issues (such as, finishing the last 10% is always the hardest!), and I am grateful for how involved he gets with each of his students' lives. I couldn't have asked for a better mentor.

I am also indebted to the late Jon Widom, who inspired and closely guided the projects described in this thesis, and who did a lot to shape my idea of what it means to be a scientist; my committee members, Doug Rees, Zhen-Gang Wang, and Niles Pierce, for their advice and guidance on this project throughout my time at Caltech; and Liz Haswell at Washington University in St. Louis, who mentored me on a project that unfortunately did not become part of my thesis, but who taught me a lot about the next stages in an academic career, particularly as I had the opportunity to spend a month in St. Louis just after she started her lab.

It has been a great pleasure to have spent much of the last six years working with (and hanging out

with!) the other members of Rob’s lab, who have become not just scientific collaborators but good friends as well: Lin Han (my predecessor on this project, who laid an amazing amount of foundation for the work that I’ve done), Paul Grayson, Frosso Seirtaridou, Eric Peterson, Dave Wu (never too busy to drop everything and help me with a code bug!), Dave van Valen (who patiently spent many hours explaining math, physics, and Matlab to me), Hernan Garcia (who taught me molecular biology and helped develop the TPM analysis and the second TPM setup), Heun Jin Lee (who often provided much-needed technical advice, especially about surface chemistry and microscopy), Tristan Ursell, Arbel Tadmor, Sidney Cox, Maja Bialecka, Christoph Haselwandter, James Boedicker, Franz Weinert, Dan Jones, Rob Brewster, Mattias Rydenfelt, my fellow TPM-ers, Geoff Lovely and Yi-Ju Chen, and my coauthor and close collaborator on TPM analysis and theory, Martin Lindén. I also had the opportunity to work with three outstanding undergraduates during my Ph.D. tenure: Kate Craig, who worked with me the summer after my first year when I didn’t know much more than she did, and helped me develop both the TPM technique and the original versions of the “masterscript” analysis code; Kiefer Aguilar, still a good friend, whom I had the pleasure of watching mature from a kid just out of high school to a young professional college graduate, and who came up with several creative improvements to our TPM assay, such as dual-channel slides; and Chao Liu, who made a lot of progress on the initial stages of the polyA project almost entirely on her own, as she worked with me during one of the busiest times in my grad school career, and who was also a joy to TA with for one quarter of Rob’s APh 161. Linda Song and Pradeep Ramesh, though not technically “my” undergrads, still helped me quite a bit with teaching and research projects (and were a lot of fun to have around!). I am also very grateful to our admins Linda Scott and Katie Miller, who not only keep the lab from falling apart but have often provided a much-needed sympathetic ear.

The work presented here would not have been possible without the advice and technical expertise of Kathy Matthews at Rice University and Jia Xu in her lab, who spent a week teaching me the Lac repressor purification (and many hours after that helping me debug it back in California); Dan Grilley in Jon Widom’s lab, who helped Yi-Ju and me with making and verifying the DNAs for the polyA project; John Beausang and Phil Nelson at the University of Pennsylvania, who helped

with some of the earlier stages of TPM analysis; and my roommate (for all six years!) Young In Oh in the Hsieh-Wilson lab, Beth Huey-Tubman in the Bjorkman lab, and the Shan lab, for borrowed equipment and advice (especially Beth) on the Lac repressor purification.

I am grateful for the financial support I received from a Virginia Gilloon Fellowship for Women in Science and Engineering, and from a National Science Foundation graduate fellowship.

And finally, a very heartfelt thank you to my family and my friends both in the LA area and in Northern California. As much as I've enjoyed my time at Caltech, there were certainly rough parts, and I am incommunicably grateful for the support of my family and friends during those times (as well as for all the laughter during the fun times!). A special thank you to my parents, whose support has taken different forms as I've grown up, but has always been overwhelming and unconditional; and to my brother Matt Johnson and my fiancé Luke Breuer, who have provided not only emotional but tangible support as well, in the form of lots of IT help and very patient tutorials in math, code writing, and physics (and in some cases, beautifully written code itself).

Abstract

Many gene regulatory motifs in both prokaryotes and eukaryotes involve physical manipulations of the genetic material, often on length scales short enough that the mechanical properties of the DNA significantly impact gene expression. One class of such manipulations, called “action at a distance”, includes transcription factor-mediated DNA looping, in which a binding site some distance away on the DNA is brought into close proximity with the transcription machinery at the promoter. DNA looping is a key component of several important regulatory systems in bacteria, and is crucial to the combinatorial control that is common at eukaryotic promoters regulated by more transcription factors than can physically bind adjacent to the promoter. Here we use a prototypical DNA looping protein, the Lac repressor from *E. coli*, to explore questions regarding the role of DNA mechanics in DNA looping and combinatorial control, particularly concerning the role of sequence flexibility in short-length-scale looping. We combine a statistical mechanical model of looping by the Lac repressor with a single-molecule technique called tethered particle motion that allows us to quantify this looping, and the systematic tuning of four biologically relevant and experimentally tractable parameters: loop length, loop sequence, repressor-DNA affinity, and repressor concentration. We show that this combination is a powerful approach to measuring repressor-DNA binding affinities and sequence-dependent DNA flexibilities in a way that is orthogonal, and therefore complementary, to conventional ensemble assays. Our results show that the sequence dependence to looping is more complicated than has been observed in other contexts, suggesting that “sequence flexibility” as a general term is misleading, and, we argue, that the measurement of sequence flexibilities depend more strongly than previously appreciated on the shape of the deformation used to make the measurement. Finally, we present preliminary results with a more complicated system that is a case

study for broader issues in combinatorial control, and a new hidden Markov model approach, based on variational Bayesian inference, to analyze these more complicated systems, which we hope will allow more precise dissections of, and more robust extraction of kinetic parameters from, tethered particle motion assays.

Contents

Acknowledgements	iv
Abstract	vii
1 Introduction	1
1.1 The importance of the physical state of the DNA to gene regulation	1
1.2 DNA looping and combinatorial control	4
1.3 The controversial flexibility of DNA at short length scales	7
1.4 The role of sequence flexibility in transcriptional regulation	10
1.5 The <i>lac</i> operon as a case study for measuring DNA flexibility in the context of DNA looping, and for broader questions of combinatorial control	13
1.6 The single-molecule tethered particle motion assay for studying DNA looping and questions of DNA bendability	15
1.7 Structure of the thesis	18
2 A statistical mechanical model of the <i>in vitro</i> looping probability	24
2.1 Tuning the simple titration curve	25
2.2 The case of multiple looped states	32
2.3 Effect of an inactive fraction of repressor	33
2.4 Effect of the presence of dimers in solution	34
2.5 Effect of cooperative binding of repressor heads	39
2.6 Low repressor concentrations	41

2.7	Calculating relative J-factors	43
2.8	Conclusion	44
3	Precision single-molecule measurements of dissociation constants and J-factors	46
3.1	Improvements over previous work	47
3.1.1	Accurate measurements of dissociation constants and J-factors requires protein purified in-house	47
3.1.2	The PUC306 construct exhibits anomalous behavior even in the presence of protein purified in-house	49
3.2	Computational controls: Dimers at low concentration, the active fraction of repressor, and low repressor concentrations	51
3.2.1	The dimer-to-tetramer transition, and the active fraction of repressor	52
3.2.2	Data analysis at low repressor concentrations	54
3.3	Experimental controls: Different bead sizes and nonspecific adsorption to chamber walls	55
3.3.1	Smaller beads result in similar looping probabilities	55
3.3.2	No detectable loss of protein to chamber walls	56
3.4	Conclusion	57
4	The sequence dependence of transcription factor-mediated DNA looping	58
4.1	Effect of repressor concentration and operator strength on the looping probability	59
4.2	Effect of sequence on the looping probability	63
4.3	Effect of loop length on the looping probability	64
4.4	A need to revisit our understanding of sequence flexibility	67
4.5	Preliminary results with additional sequences	70
4.6	The masking of sequence effects <i>in vivo</i> by nonspecific DNA-bending proteins	75
4.7	Conclusion	78
4.A	Appendices to Chapter 4	80

4.A.1	RMS of the unlooped and looped states as a function of concentration and of loop length	80
4.A.2	Compiling looping predictions from several recent theoretical analysis	82
5	A kinetic analysis of looping by the Lac repressor	84
5.1	Kinetics of looping by the Lac repressor by conventional methods	87
5.1.1	State lifetimes and missed events	87
5.1.2	Looping rate constants	92
5.1.3	Direct interconversions between looped states?	94
5.2	Preliminary results with a hidden Markov model analysis	96
5.2.1	Overview of a variational Bayesian hidden Markov model analysis of TPM data	96
5.2.2	Examples of results	101
5.3	Conclusion	104
5.A	Appendices to Chapter 5	106
5.A.1	Obtaining kinetic information from dwell time histograms	106
5.A.2	Calculating the dead time of a filter	109
5.A.3	A physical model for the observable distributions	111
6	The three operators of the wild-type <i>lac</i> system: A case study in combinatorial control	116
6.1	A statistical mechanical model of the wild-type <i>lac</i> system	119
6.2	DNA-bending proteins may be essential elements of the <i>lac</i> regulatory system	124
6.3	Conclusion	128
7	Conclusion	132
	Appendices	140
A	Detailed derivation of the model that includes the dimer-to-tetramer transition	140
A.1	Assumptions	140

A.2	Derivation of $p_{\text{loop}}([R])$, taking into account $T \Leftrightarrow 2D$	142
A.3	Dimers due to damaged protein	148
B	DNAs	150
B.1	Constructs containing E8 and 601TA	150
B.2	Constructs containing poly(dA:dT)	153
B.3	Constructs derived from the naturally occurring <i>lac</i> operon	154
C	Lac repressor purification	157
D	Tethered particle motion: Methods	160
D.1	TPM sample preparation	160
D.1.1	Method summary	160
D.1.2	Detailed protocol	160
D.2	TPM data acquisition and analysis	162
D.2.1	Acquiring data	162
D.2.2	Particle tracking and calculation of the root-mean-squared motion of the bead	163
D.2.3	Determining the looping probability for each trajectory	165
D.2.4	Minimum number of trajectories and minimum observation time	166
D.2.5	Calculating the average looping probability for a set of trajectories	169
D.2.6	Fitting concentration curves	171
D.2.7	Calculating J-factors without concentration curves for each construct	173
E	Representative traces	178
	Bibliography	183

Chapter 1

Introduction

1.1 The importance of the physical state of the DNA to gene regulation

The publication of the first draft of the sequence of the human genome in 2001 [1], a crucial moment in an effort that began with the first complete genomic sequence of a free living organism (that of the bacterium *Haemophilus influenzae*) in 1995 [2], and the publication of numerous other genomes from mice [3] to platypus [4] since then, was in many ways one of the crowning achievements of modern biology. To name two of many revolutionary changes brought about by these fully sequenced genomes, the completion of the human genome ushered in an entirely new era of medical research—for example, by streamlining the process by which disease genes of unknown biochemical function are identified [1, 5]—and offered a clear path towards a not-so-distant future of highly personalized medical treatment [6]. Fully sequenced genomes have also spawned a host of additional bioinformatic databases that contain information related to, but a level above, the sequence of nucleotides in a genome (e.g., RegulonDB [7] and EcoCyc [8] for *E. coli*), such as locations of binding sites for transcriptional regulators.

As important as these advances to our understanding of the content of genomes have been, it has become increasingly clear that genomic-sequence and protein-binding-site databanks do not contain the sum total of the information content of a cell's genome. Rather the *mechanical properties* of the DNA polymer in which the genomic sequence information is encoded, and the *physical state* of the

DNA in a cell, are known from many examples to play crucial roles in the regulation of the genetic information encoded by the sequence. For example, cellular differentiation and tumorigenesis often involve the rearrangement of chromatin (the packaged and organized DNA in eukaryotic nuclei), indicating that the localization of a gene in a eukaryotic nucleus can control the level of its output [9, 10, 11]. And it is now clear that mutations to DNA sequence alone cannot account for all aspects of cellular progression from normal to cancerous, but instead that epigenetic changes—including modifications to the structure and organization of the DNA in the cell—play significant roles in the progression of many types of cancers [12].

Perhaps the most telling indicator that genome structure and the mechanical properties of DNA are tightly controlled by cells is the fact that all domains of life express proteins whose sole function seems to be genomic structuring. Eukaryotic genomes are tightly spooled around protein complexes called histones [13], with the resulting DNA-protein complex, called a nucleosome, being the fundamental unit by which the approximately 3 gigabases of DNA (about 1 meter) are packaged into the roughly $100 \mu\text{m}^3$ nucleus [14, 15, 16]. Nucleosomes play a crucial role in the regulation of transcription as well [14, 15], with genes sequestered into nucleosomes expressed less than genes in the linker DNA that connects adjacent nucleosomes. Bacteria express at least six kinds of “nucleoid-associated proteins” (NAPs), which are thought to package the genome in a similar manner to nucleosomes [17]. Many of these NAPs are DNA-bending proteins—that is, they modify the flexibility of the genomic DNA, not only its organization [17, 18]—and are known to influence gene expression [17, 19, 20]. Mitochondrial genomes (contained in structures called mt-nucleoids) are packaged and organized by nonspecific DNA-bending proteins as well, and there is evidence that the organization of mt-nucleoids changes with cellular metabolic demands [21]. Archea also express at least two kinds of architectural proteins, called chromatin proteins, that compact the genome and probably also influence DNA metabolic processes; one class, called histones, is homologous to eukaryotic histones [22, 23].

Cellular manipulation of the DNA polymer is not restricted to the structuring and packaging of genomes, however. Instead DNA is subjected to a wide variety of physical manipulations in

cellular processes as diverse as the looping events that occur during DNA replication [24, 25], the bending of DNA during recombination [24, 25], and the physical rearrangements of genomic DNA induced by transcription factors [24, 25, 26, 27]. In fact one of the most ubiquitous classes of regulatory architecture found in all domains of life depends upon the physical manipulation of the DNA polymer: so-called “biological action at a distance”, where proteins (often transcription factors) bring two sites separated by some distance on the DNA into close proximity, thus looping the intervening DNA [28, 29, 30].

Interestingly, many of the biological manipulations experienced by DNA, but especially many cases of “action at a distance” in transcriptional regulation, involve bending and twisting the DNA on length scales that are short in comparison with its natural scale of deformation, that is, the persistence length (discussed in more detail below) [27, 31]. Eukaryotic DNA is subjected to enormous deformations when packed in nucleosomes, with 147 bp of DNA (already smaller than the persistence length) wrapped 1 3/4 times around the histone octamer [13, 26]. Similarly, in the context of prokaryotic transcription factor-mediated DNA looping, not only are such lengths the default in naturally occurring transcriptional networks, but the optimal *in vivo* lengths as determined by the maximal regulatory effect are often at loop lengths smaller than 100 bp [27, 32, 33].

Here we examine the role of the mechanical properties of the DNA polymer, and especially the role of sequence-dependent bendability, in the regulation of gene expression at the level of transcription. We will focus on the short-length-scale bending that is so prevalent in cellular processes but that, as will be described in more detail below, remains poorly understood. Although many aspects of gene regulation involve such short-length-scale bending, we will focus on the process of DNA loop formation by a prokaryotic transcription factor, with some reference as well to nucleosome positioning, which impacts transcriptional output in eukaryotes, and to the DNA-bending proteins that structure the genome in prokaryotes.

1.2 DNA looping and combinatorial control

DNA looping, in which two disparate sites on a single DNA molecule are brought together by a single protein or protein complex, is one kind of biological “action at a distance” and occurs in both prokaryotes and eukaryotes, though not necessarily by the same mechanisms in both [25, 26, 28, 29, 30]. The prevalence of loop formation in transcriptional regulation should not be surprising: given the widespread occurrence of combinatorial control in both eukaryotes and prokaryotes (i.e., the fact that more than one transcription factor at a time often influences the regulatory state of a promoter, as shown for a few key examples in Figs. 1.1(C) and 1.3), it is not surprising that regulatory proteins must bind other sites besides those immediately adjacent to the promoter they regulate [25, 34]. There is only space for one or two regulatory proteins to bind and “touch” the transcription apparatus directly. The side effect of such distal binding is that the DNA has to loop in some way to give access to the promoter of interest.

DNA looping was first discovered in the *ara* operon in *E. coli* [35], where it is mediated by a protein called AraC that has two DNA binding domains in the same molecule. When these two domains bind to two sites separated by some distance along the DNA, the intervening DNA is looped out, and the genes of the *ara* operon are repressed [35, 36]. Such looping induced by a two-headed DNA binding protein, with one binding site near the promoter of interest and the other some distance away, has since been shown to play a key role in the regulation of several other operons in *E. coli*, including the *lac*, *deo*, and *gal* operons [26]. It is also a key feature of a well-studied viral protein called the lambda phage repressor, which was the first looping protein whose activity was verified *in vitro* [26, 37]. In the case of these two-headed looping proteins, looping is thought not only to enable combinatorial control, but also to contribute to efficient transcriptional control by increasing the effective concentration of the transcription factor in the vicinity of the promoter [25]. If one head of a DNA looping transcription factor releases from the DNA, it is more likely to rebind and reform the loop than to dissociate entirely from the DNA, as it is tethered near its binding site by the second head. Additional implications for the role of looping in other aspects of fine-tuning control of transcription continue to be suggested [38, 39].

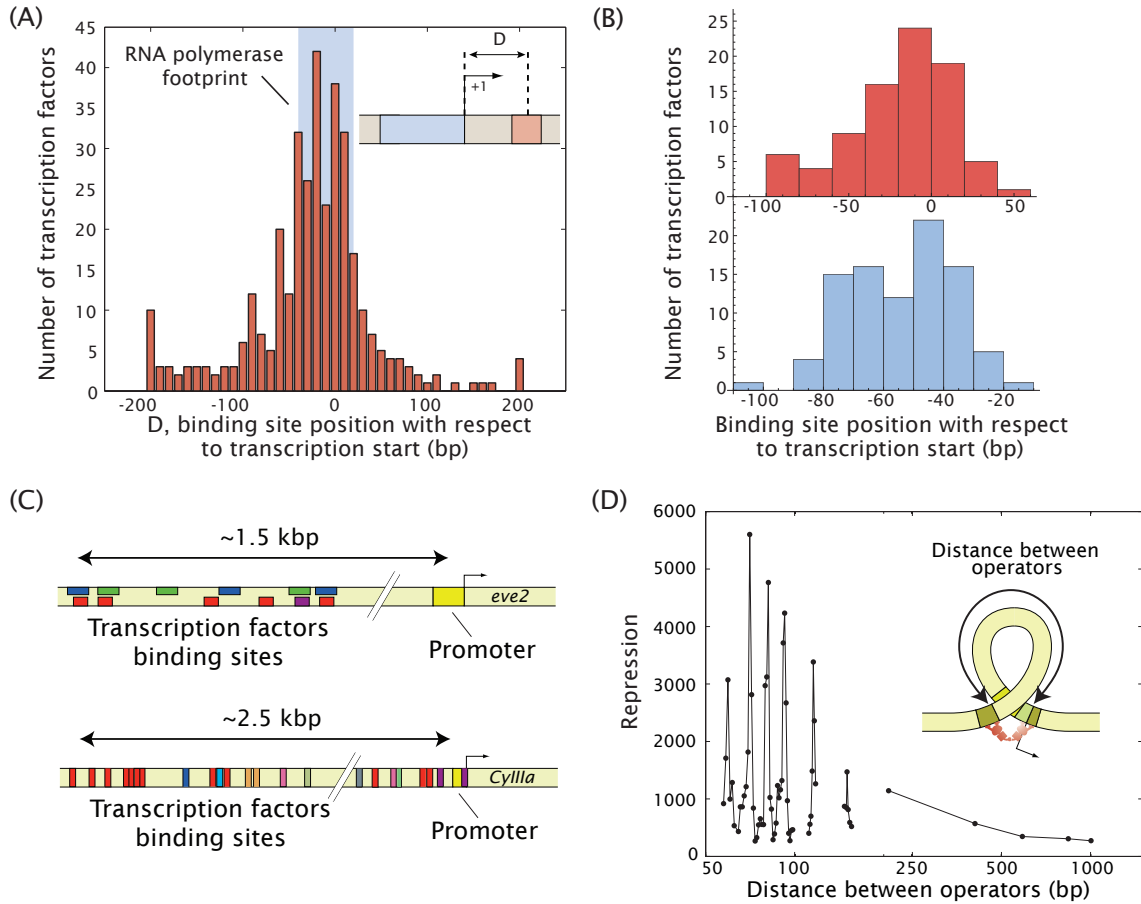


Figure 1.1: Action at a distance and combinatorial control in prokaryotes and eukaryotes. **(A)** Many promoters in *E. coli* are regulated by one or more transcription factors that bind tens or even hundreds of base pairs away from the promoter they regulate. Shown here are all known transcription factor binding sites in *E. coli*; many do bind adjacent to the promoter, indicated by the blue box, but a significant number bind some distance away. Data from RegulonDB; figure courtesy of Hernan Garcia (modified from [40], ©2010 Elsevier Ltd). **(B)** Even at promoters regulated by a repressor (top) or activator (bottom) where that activator or repressor has only one binding site (a subset of the data in (A)), that binding site can be up to 100 bp away from the transcription start site. These cases, even more so than those that make up the rest of the distribution in (A), are suggestive of a key role for loop formation in many regulatory systems. Data from RegulonDB; figure courtesy of Mattias Rydenfelt. **(C)** Combinatorial control in eukaryotes. In these two well-known examples from *Drosophila* (top, adapted from [41]) and sea urchin (bottom, adapted from [42]), not only are many binding sites up to several kilobases away from the promoters, but the inputs from the binding of transcriptional regulators to all of these sites must be integrated to produce the observed output. In the *Drosophila* example, gradients of several different transcription factors combine to produce the famous “eve stripes” that establish the body plan during development. In the sea urchin example, the *CyIIIa* gene encodes a form of cytoskeletal actin, and is tightly regulated both spatially and temporally during development by a set of at least nine transcription factors, including both activators and repressors, and at least one putative DNA looping protein (the SpGCF1 protein, red boxes) [42]. **(D)** In this classic example of the effect of loop length on repression of the *lacZ* gene in *E. coli* (introduced in detail in Fig. 1.3), a significant modulation of gene expression is observed as a function of the spacing of the binding sites (“operators”) that form the boundaries of the DNA loop. The spacings between peaks in repression is roughly 10 bp, the helical repeat of DNA. Such modulation with 10 bp periodicity is a signature of DNA looping. Figure adapted from [27], ©2006 Wiley Periodicals, Inc., by Hernan Garcia, based on data from [32].

DNA looping in transcriptional regulation is not limited to those cases in which a two-headed protein is an activator or repressor of the gene. Many genes in *E. coli* are regulated by transcription factors that bind tens or even hundreds of base pairs away, as shown in the histogram of *E. coli* transcription factor binding sites in Fig. 1.1(A)), and in fact a significant fraction of these genes seem to be regulated by a single activator or repressor with only a single binding site that is not immediately adjacent to the promoter (Fig. 1.1(B)). Some of these belong to a class of promoters in bacteria that are regulated by an enhancer-dependent mechanism similar to that of eukaryotes, in which a loop forms between a distally bound activator and the promoter of the gene of interest, the most well-known example being the nitrogen-assimilation genes regulated by NtrC [43]. The implication with all of these promoters with distantly bound regulating transcription factors is that they must involve some form of DNA looping to bring the regulatory factors in contact with the transcription machinery at the promoter.

In higher eukaryotes non-adjacent binding sites for transcriptional regulators are the rule rather than the exception, with regulatory sites often located kilobases away from the target promoter [44]. Two well-known examples of eukaryotic promoter regions and their control factors are shown in Fig. 1.1(C). It has long been postulated that these cases of truly long-range action-at-a-distance involve some form of DNA loop formation [29], but it was only within the last decade that such loop formation was demonstrated for eukaryotes *in vivo* [30, 45]. It appears that in eukaryotes, DNA looping is most often effected not by two-headed looping proteins, as in bacteria, which have the ability in and of themselves to loop DNA; rather, a looped complex is formed by the transcriptional regulator, RNA Polymerase and linker proteins like Mediator or Cohesin [45]. (There are, however, at least two single-protein, bidentate looping complexes in eukaryotes, both involved in cancer in humans, RXR and p53 [46, 47].)

Loop formation is perhaps one of the clearest examples of the importance of DNA mechanics to gene regulation. One of the classic signatures of looping is a modulation of regulatory activity as the distance between the two binding sites for the activator or repressor is changed [25], as shown in Fig. 1.1(D). Regulatory activity at loop lengths shorter than several hundred base pairs shows

peaks and troughs with a periodicity of about 10–11 bp [32, 48], corresponding to the helical period of double-stranded DNA. The interpretation is that minima of loop formation (in the example of Fig. 1.1(D), indicated by troughs in repression) correspond to loop lengths for which the DNA has to be *twisted*, in addition to being *bent*, in order for a loop to form [49, 32, 48, 50, 25]. That is, how twistable the DNA of the loop is, or at least how much it must be twisted—a parameter that can be modulated simply by the addition or removal of one or a couple base pairs in the loop—can have very large effects on the ability of the transcription factor to either activate or repress the gene of interest.

The example shown in Fig. 1.1(D) highlights an intriguing aspect of loop formation that is not as yet thought to be well understood. As noted above, the default loop lengths in many prokaryotic transcriptional networks, and the optimal *in vivo* loop lengths as determined by maximal regulatory effect, as in the example in Fig. 1.1(D), are often shorter than 150 bp [27, 32, 33]. Even in eukaryotes, where very long loops are more common, the behavior of short DNAs still plays a role in transcriptional regulation, in the wrapping of 147 bp sections of the genome almost two full times around the histone cores of nucleosomes. This prevalence of short loops is surprising, given our canonical understanding of DNA as a semi-flexible polymer with a persistence length, a length over which the DNA tends to be straight, of roughly 150 bp [31]. Generally speaking it should be energetically costly to bend DNA at lengths shorter than 150 bp. In part because of the prevalence of short loops and bends *in vivo*, however, this canonical behavior of DNA is still a highly controversial issue.

1.3 The controversial flexibility of DNA at short length scales

Despite the clear importance of the short-length-scale mechanical properties of DNA in loop formation as well as the many other cellular processes noted above, there remains both uncertainty and controversy about the ease with which such short DNAs can be deformed [51], and also about the role of sequence at these short scales, particularly in the context of protein-mediated bending or looping [51, 52]. The controversy surrounding short-length-scale DNA bending has been reviewed recently [51] but we will summarize the key points here.

The classic conception of DNA as a semi-flexible polymer is usually encapsulated in the worm-like chain (WLC) model [53], which describes DNA as relatively flexible at long length scales, where entropy dominates the energetics of the polymer, but relatively stiff at short length scales, where elasticity dominates [51]. The parameter that determines the length scale in question is the *persistence length*, the length over which the polymer is relatively stiff. More precisely, the persistence length is defined as the length over which the tangent vectors of two points on the molecule become uncorrelated. Under typical conditions the persistence length of double-stranded B-DNA is 50 nm, or 150 bp [31], which, as noted above, raises the question of how the short DNA loops and bends that are so prevalent in biology form.

A number of experimental approaches are available for studying the flexibility of DNA [51], but a particularly common one, especially for studying the behavior of DNAs on the order of one or a couple persistence lengths, is ligase-mediated cyclization [54, 55]. In a cyclization reaction, a linear DNA with complementary single-stranded overhangs on either end (“sticky ends”), is mixed with DNA ligase and allowed to close into circles and to form dimers (or, with lower efficiency, higher-order multimers). The activity of the DNA ligase effectively captures a sampling of the ring closure and dimer formation, which are assumed to be fast compared to the activity of the ligase. The outcome of a cyclization reaction is a parameter called the *cyclization J-factor*, the effective concentration of one end of the DNA in the other, defined as the ratio of the rate of formation of ligated circles to the rate of formation of ligated dimers [54, 56]. A higher J-factor indicates a more flexible DNA.

One reason cyclization assays have found such popularity for the study DNA flexibility is that the molecular conformations of all of the players are known, and extensive theoretical work has been done to predict cyclization J-factors as a function of DNA length based on our current models of DNA flexibility. The Shimada-Yamakawa result [57] is one of the most widely used, and will be the basis of comparison for our looping results in Chapter 4 (see Fig. 4.3). As we will see in that chapter, our currently incomplete knowledge about the conformation of the DNA in a protein-mediated loop, in contrast to the simpler case of a ligated DNA minicircle, prevents us from having a similarly clean result for predicted *looping* J-factors. Nevertheless, the Shimada-Yamakawa result

is the usual starting point for discussions of DNA flexibility; and in particular, it predicts that the sub-persistence length loops that are so common *in vivo* should be so unfavorable as to form only through the assistance of DNA-bending proteins (such as the architectural proteins discussed in the first section above).

One of the first challenges to the applicability of the Shimada-Yamakawa result at short length scales came from a study by Cloutier and Widom of short DNA fragments derived from the nucleosome affinity assays discussed in the next section [58]. Using *in vitro* cyclization studies (and thus in the absence of any of the DNA-bending proteins present *in vivo*), they found the cyclization J-factors for several sub-persistence-length DNAs to be several orders of magnitude higher than predicted by the Shimada-Yamakawa theory. Cloutier and Widom's result was disputed by Du and coworkers for technical reasons [59], but has continued to inspire controversy and additional experimental and theoretical efforts that attempt to explain the results of Cloutier and Widom (and the subsequent work from others that either support or refute their initial results) [51, 60].

So the question of how short DNA loops and bends form *in vivo*, and, now, *in vitro* as well, remains as yet unanswered. Some cases of tightly bent DNA have been solved to a greater extent than the *in vitro* cyclization results discussed here: for example, the favorable interactions between the DNA wrapped in a nucleosome and the histone proteins around which the DNA is wrapped are sufficient to overcome the energy penalty of wrapping a persistence length of DNA one-and-three-quarters times around the histone core [27]. We will return to this question of the bendability of DNA at short lengths in the context of looping by the Lac repressor in Chapter 4, where we argue that the geometry of a protein-mediated loop and/or flexibility of the looping protein can be sufficient to overcome the energy penalty of bending sub-persistence lengths of DNA into transcription factor-mediated loops. However we turn now to a less well-studied aspect of short transcription-factor mediated loops, that of the role of sequence, though as we will see this aspect has already been studied, and generated significant controversy of its own, in the context of the tight bends of eukaryotic histones.

1.4 The role of sequence flexibility in transcriptional regulation

Although it has been known since the 1980s that the sequence of DNA can impact its flexibility and twistability [61], the implications of this sequence dependence to the mechanical properties of DNA on transcriptional regulation have been studied only in a few select cases. It has been shown *in vivo* that flexible or pre-bent sequences in (non-loop-forming) promoter regions can increase transcription [62], and that the inclusion of phased A-tracts that introduce static curves into activation loops (such as those mediated by NtrC) can increase transcription *in vitro* [63, 64] and *in vivo* [65], though activation (and, presumably, loop formation) is surprisingly insensitive to the particular geometry induced by such curved DNAs [63, 66]. In fact an intrinsically curved A-tract DNA is a natural part of the *nifLA* promoter of *Klebsiella pneumoniae*, in a region that is thought to be looped out by NtrC, and this curved DNA is essential for wild-type levels of transcription [65]. Phased A-tracts have been used to examine the effects of intrinsically curved sequences on Lac repressor-mediated DNA loops *in vitro* as well; similarly to NtrC loops, it appears that the Lac repressor can accommodate multiple different loop geometries imposed by static bends, with these static bends inducing the formation of hyperstable complexes that remain looped for days [67, 68, 69, 70]. While these studies have provided valuable insights into the role of static bends and loop geometries in loop formation, they have only brushed the surface of the question of what role sequence plays more generally in DNA looping and transcriptional regulation *in vitro* or *in vivo*.

Though the role of DNA sequence has not been extensively studied in the particular case of transcription-factor mediated looping, it has become a key parameter in the discussion of a different mechanism of transcriptional regulation, that of nucleosome positioning in eukaryotes [14]. Nucleosomes do not have a defined binding site sequence and can form on any DNA of sufficient length; but they do preferentially bind to some sequences over others. A number of sequences with very different nucleosome affinities have been identified, some isolated from natural sources and others from nucleosome affinity assays with synthetic sequences [14]. It has been argued for both classes that their

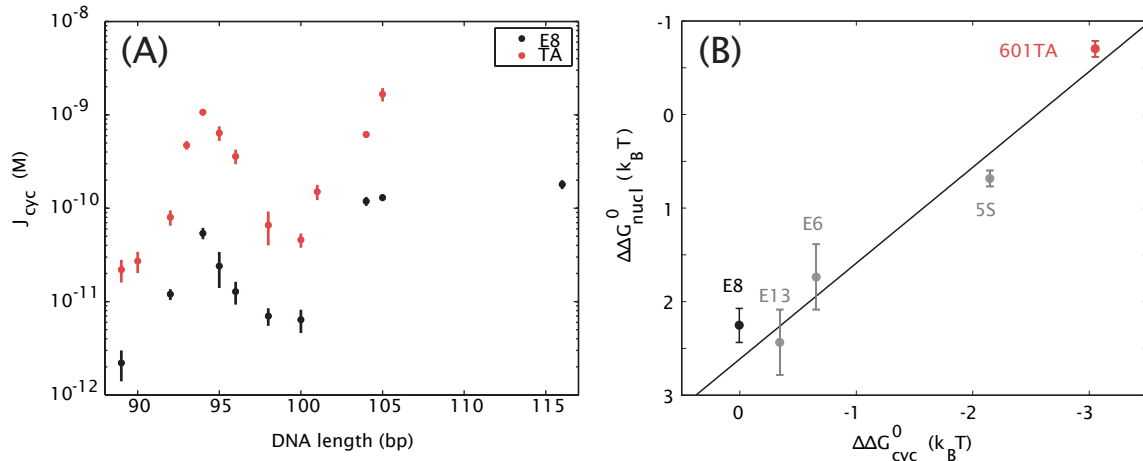


Figure 1.2: Sequences with high nucleosome affinities also have high cyclization J-factors. **(A)** J-factors for two different sequences, a putatively more flexible sequence called 601TA (red, here abbreviated “TA”) that has a high affinity for nucleosomes, and a random sequence called E8 (black) which has a lower nucleosome affinity, as determined by ligase-mediated cyclization assays. At all lengths tested, DNA minicircles composed of the TA sequence form more readily than those of the E8 sequence. Adapted from [85]. **(B)** A sequence’s propensity to be wrapped in a nucleosome (here represented by the energy of forming a nucleosome *in vitro*, $\Delta\Delta G_{nucl}$) correlates with its propensity to form DNA minicircles (here represented by the energy of forming these minicircles, $\Delta\Delta G_{cyc}$). E8, E13, and E6 are all synthetic random sequences; 5S is a strong natural nucleosome positioning sequence from sea urchin. Adapted from [58].

nucleosomal affinities stem from different intrinsic flexibilities, and not in response to some other *in vivo* condition or to a property specific to nucleosome binding [58, 71, 72]: because nucleosomes involve tight bending of short DNAs, sequences with high intrinsic flexibilities are thought to decrease the energy of nucleosome formation, yielding the observed positioning preferences. Though a corollary hypothesis, that sequence flexibility confers preferences for nucleosome positioning and/or occupancy *in vivo*, is quite controversial [15, 73, 74, 75, 76, 77, 78, 79, 80], the original *in vitro* claim has nevertheless led not only to many theoretical and experimental studies on the relationship between sequence and flexibility [51, 52, 81, 82, 83, 84], but also to the elucidation of numerous sequence “rules” that can be used to predict the likelihood that a nucleosome will prefer certain sequences over others [14]. Algorithms that predict nucleosome positions based on these sequence rules are highly predictive *in vitro*, even for sequences from organisms that do not themselves contain nucleosomes [77]. The outstanding question in the field is how predictive these sequence rules, and sequence effects in general, are of nucleosome positions *in vivo*, as compared to other potential nucleosome positioning factors such as chromatin remodeling complexes [15, 78].

More importantly for the purposes of this study, the claim that a sequence’s nucleosome affinity stems from its degree of intrinsic flexibility has also led to the determination of certain sequences that are claimed to be highly flexible in a general sense. For example, Cloutier and Widom characterized a sequence selected from a chemically random pool of sequences, which they called 601TA, and which they showed to have a significantly higher affinity for nucleosomes than a synthetic random sequence called E8 [58, 85, 86]. In fact the 601TA sequence, which we will henceforth abbreviate TA, is the strongest known nucleosome positioning sequence, either synthetic or natural [14, 86], and is often used in *in vitro* assays to ensure the localization of nucleosomes at a precise, desired position (e.g., [87]).

Like other previously-described nucleosome positioning sequences [71, 72], the argument that the TA sequence’s high affinity for nucleosomes stems from a high intrinsic flexibility is based largely on the results of *in vitro* ligase-mediated DNA cyclization assays (described in the previous section). For short DNAs at least, relative to the persistence length of 150 bp, more flexible sequences should cyclize more readily than other sequences, and therefore should have higher J-factors. Cloutier and Widom showed this to be the case for the TA versus E8 sequences, as shown in Fig. 1.2(A), and moreover they correlated the cyclization J-factors for a number of sequences with the energy required to form a nucleosome with these sequences, as shown in Fig. 1.2(B) [58, 85].

It is generally assumed that cyclization assays are a useful tool for measuring sequence flexibility in some general sense and for learning about DNA looping as well [29, 58, 84, 88], with an implication that sequences that appear to be more flexible in cyclization assays might likewise lead to increased loop formation, just as they have been found to increase nucleosome formation. That is, if TA and E8 differ in mechanical bendability in some general sense, then TA should increase *looping* by a bacterial transcription factor just as it increases nucleosome binding and cyclizes more readily than E8. Therefore these sequences, which yield such strong sequence effects in two other *in vitro* assays, should be ideal for addressing the question of how sequence affects DNA looping *in vitro* and, perhaps, *in vivo*. As we will see, the question of a sequence’s flexibility is actually more subtle than these nucleosome formation and cyclization assays reveal.

1.5 The *lac* operon as a case study for measuring DNA flexibility in the context of DNA looping, and for broader questions of combinatorial control

In this work we exploit insights about DNA flexibility garnered from one class of genetic regulation where it has been studied extensively, that of nucleosome formation, to make predictions about how a different class of mechanical deformations in regulatory biology, that of DNA looping by a prokaryotic transcription factor, will be altered by these same sequences. As described in the next section, we test these predictions experimentally with a single-molecule assay in conjunction with ideas from statistical mechanics for the case of one of the most well-known transcriptional regulators in bacteria, that of the Lac repressor, though there are clear implications for other prokaryotic and eukaryotic regulatory motifs as well.

The discovery in 1961 by Jacob and Monod of genes whose products regulated the transcription of other genes [89] led to a restructuring of our understanding of both the content and management of cells' genomes. The *lac* operon in *E. coli* has since become a paradigm of genetic regulation at the level of transcription initiation [90] and continues to be an area of intense research even after more than 40 years (for just two of many examples that illustrate several outstanding questions about this system, see the recent work of [38, 70]).

The *lac* operon, shown in Figure 1.3, encodes a set of three structural genes involved in the uptake and metabolism of the sugar lactose, and one regulatory gene whose product controls transcription initiation at the single promoter for the polycistronic mRNA encoding the three structural genes [91]. The product of the regulatory gene, called the Lac repressor or LacI, is a 154 kDa homotetramer whose binding to a site on the DNA (the O1 operator) overlapping the *lacZYA* promoter (also called Plac) prevents the binding of RNA Polymerase to the promoter, thereby decreasing transcription [91, 95]. However, when lactose is present, a derivative of lactose binds to the repressor and changes its conformation such that its affinity for O1 is significantly decreased. As a result the repressor no longer out-competes the polymerase for binding to the promoter and transcription can readily occur

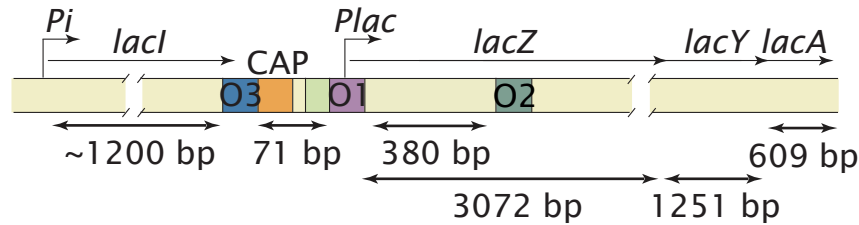


Figure 1.3: Schematic of the *lac* operon. The three structural genes *lacZ*, *lacY*, and *lacA* are transcribed as a polycistronic mRNA and encode the proteins β -galactosidase, lactose permease, and galactoside acetyltransferase. These proteins break lactose into galactose and glucose, transport lactose into the cell, and acetylate galactosides respectively [91] (the natural substrate(s) and specific role of galactoside acetyltransferase in the operon are unknown [92]). The *lacI* gene, located upstream of the three structural genes and under the control of a separate promoter, encodes the Lac repressor, which can bind to any two of the operators O1, O2, and O3 simultaneously (see also Fig. 1.4) [93, 94, 32]. O1 overlaps the promoter for the three structural genes and impedes the ability of RNA polymerase to transcribe the *lacZ*, *lacY*, and *lacA* genes [91]. The operon also contains a binding site for the CAP-cAMP complex, a positive regulator of the operon, between O1 and O3 [91, 95]. Although the protein coding region for the Lac repressor ends before O3, transcription of the *lacI* gene is known to continue into the regulatory region of the *lacZYA* promoter, with the regulatory region of the *lac* promoter possibly serving as the terminator of transcription from P_i [96].

[38, 91]. Thus the function of the repressor is to coordinate the transcription of lactose-metabolizing proteins with the presence of lactose as a carbon source.¹ Additionally, a separate positive regulation mechanism involving cyclic AMP (cAMP) and the CAP protein coordinates transcription of the *lac* operon with the presence or absence of the preferred carbon source, glucose [90, 91].

After this elegant repressor-mediated model of transcriptional regulation was proposed, it was determined that there are actually two additional Lac binding sites in the general region of the *lac* promoter [97, 98] (Fig. 1.3). Originally termed “pseudo-operators” because they did not seem to affect binding of the Lac repressor to DNA *in vitro* [99], it later became clear that both of these auxiliary operators (as they are now known) and the main operator must be present for maximal levels of transcriptional repression *in vivo* [93], a puzzle later solved by the discovery that each dimer of the tetrameric repressor binds a separate site on the DNA, thereby forming a loop in the DNA [49, 93, 94, 100]. Not only, as mentioned above, does such looping offer the advantage of increased local concentration of the repressor, in the case of repression, as with the Lac protein, looping further sequesters the polymerase binding site on a curved DNA fragment, enhancing the repressor’s ability to prevent initiation [25].

The Lac repressor, because it has been so extensively studied, offers one of the best case studies

¹Even when the repressor is fully active transcription of the operon is not completely inhibited, as small amounts of β -galactosidase and lactose permease are necessary to produce the lactose derivative that inactivates the repressor upon first exposure to lactose [91, 38].

for examining the role of DNA mechanics in loop formation and transcriptional regulation. The Lac repressor has been a popular choice in many studies with synthetic looping constructs, both *in vivo* and *in vitro*, where the naturally occurring three-operator architecture is usually replaced by synthetic two-operator versions, and often with completely non-natural sequences comprising the intervening loop DNA (e.g., [32, 49, 50, 101]). Moreover, because the naturally occurring architecture does in fact contain more than two binding sites, with the potential to form multiple loops, plus the binding site for a transcriptional activator (CAP), the Lac repressor and the wild-type regulatory region offer a convenient potential case study in broader studies of combinatorial control.

In this work we will use looping by the Lac repressor in a *in vitro* single-molecule assay, described in the next section, as a tool to probe the role of DNA mechanics in loop formation, specifically the role of the two DNA sequences described in the previous section, and to gain insight into the interplay between sequence flexibility and transcriptional regulation by action at a distance both *in vitro* and *in vivo*. The bulk of this work will make use of the kinds of synthetic two-operator looping constructs that are typically used in studies with the Lac repressor. However, Chapter 6 will discuss the extension of our looping assay to the full, three-operator construct that forms the natural architecture and demonstrate our ability to dissect more complicated architectures as well.

1.6 The single-molecule tethered particle motion assay for studying DNA looping and questions of DNA bendability

Single-molecule biophysics has provided a new generation of insights into the molecular machines that underlie cellular dynamics. One of the most important classes of such experiments has focused on the interaction between DNA and its protein binding partners, as in the looping experiments that we describe here. Many (though not all) of the single-molecule techniques that can be used to monitor loop formation and other deformations in DNA in real time rely on the imaging of microscopic reporter particles or “beads” which are attached through specific, non-covalent small-molecule interactions to the DNA and/or to the protein of interest. These beads, which can be

imaged under a microscope (while the molecular players cannot), act as reporters of the underlying molecular dynamics.

In this work we use the single-molecule tethered particle motion (TPM) technique to study looping by the Lac repressor [102, 103, 104, 105]. As shown schematically in Fig. 1.4(A), a TPM experiment consists of a linear piece of DNA attached at one end to a microscope coverslip and at the other to a microsphere. The dynamics of the microsphere then serve as a readout of the hidden underlying dynamics of the DNA and its partner proteins. In the specific case of looping by the Lac repressor that we are considering, when two operators are present on the DNA tether and repressor is introduced into the sample, the repressor can bind the two operators simultaneously and stabilize a loop in the tether. This loop reduces the effective length of the tether and so reduces the extent of the bead's Brownian motion. As a result, the formation and breakdown of loops can be observed either directly, by measuring the bead's distance from the coverslip [106], or indirectly, as will be done here, by measuring the root-mean-squared motion of the bead in the plane of the coverslip [102, 103]. These measurements result in a telegraph-like signal (see examples in Appendix E) and can be converted into the probability of the system being in the looped state: one useful definition of the looping probability is that it is the total time spent in the looped state divided by total observation time. TPM has been used to examine processes ranging from DNA looping by transcription factors [104, 107, 108, 109], as discussed here, to the dynamics of recombination proteins [110, 111] and restriction enzymes [111], and other processes associated with translation and DNA rearrangement [112]. Looping by the Lac repressor in particular has also been extensively studied by TPM [104, 108, 109, 113, 114, 115, 116].

Here, however, we go beyond previous uses of TPM to study DNA looping by combining this single-molecule experiment with a statistical mechanical model and the systematic variation of four biologically relevant parameters (Fig. 1.4(B)): repressor-operator affinity, loop length, loop sequence, and repressor concentration. In all previous Lac repressor studies with TPM, or other single-molecule techniques such as FRET [67, 68, 69, 70], only one or a couple loop lengths, operators, and repressor concentrations were studied. In many cases, therefore, the repressor-operator dissociation constants

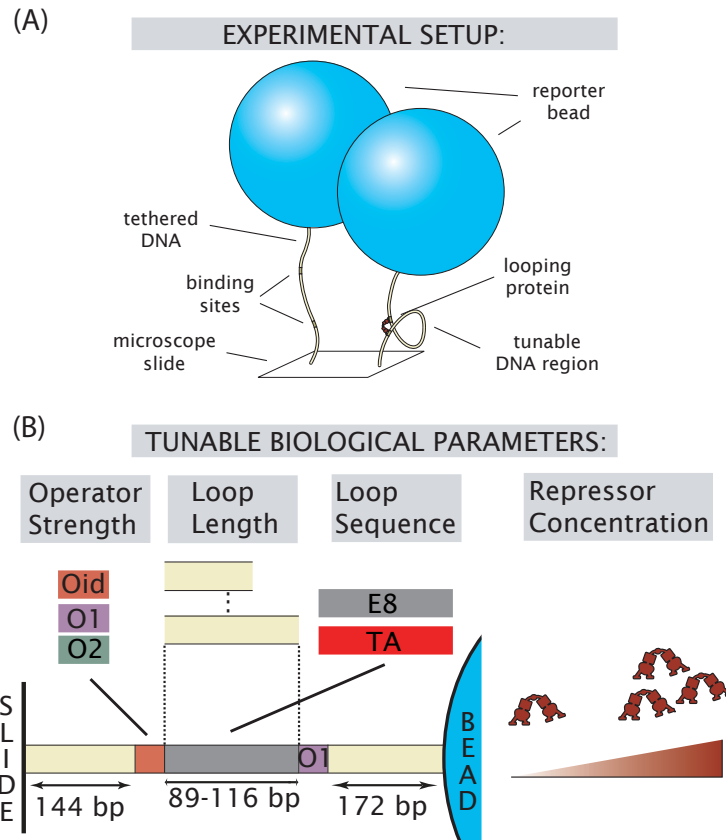


Figure 1.4: Schematic of the tethered particle motion (TPM) assay. **(A)** DNA looping is observed as a result of changes in the Brownian motion of the tethered bead [102, 103, 104, 105]: looping decreases the effective length of the DNA tether, which decreases the bead’s root-mean-squared (RMS) motion. Most of the work here will be concerned with a parameter derived from the RMS motion, the looping probability, which we define as the time spent in the looped state(s) divided by the total observation time. Chapter 5 looks at kinetic parameters that can also be determined from the motion of the bead. **(B)** Four distinct tunable biological parameters: 1. Repressor binding site, or operator. Most of the work here uses the strong, synthetic “Oideal” (O_{id}) operator, the strongest naturally occurring O_1 operator, and the weaker naturally occurring O_2 operator (see Chapter 6 for studies that also involve the weakest naturally occurring operator, O_3). 2. Loop length. The wild-type *lac* operon contains the three operators O_1 , O_2 , and O_3 , which have the potential to generate three loops of different lengths (see also Fig. 1.3): the O_1 - O_2 loop is 380 bp, the O_1 - O_3 loop is 71 bp (shorter than the persistence length of DNA), and the O_2 - O_3 loop is 472 bp. In our synthetic constructs (see Chapters 3 and 4) we use two operators and systematically tune the distance between them as shown in the figure. 3. Loop sequence. Most of the work discussed here will focus on two sequences, “E8” and “TA”. “E8” refers to a synthetic random sequence, “TA” to a synthetic nucleosome positioning sequence (part of the 601TA sequence [86]). The TA sequence has a higher cyclization J-factor than E8 and is wrapped into nucleosomes *in vitro* more readily than E8 [58, 85]. See Section 4.5 for discussion of an additional sequence also related to nucleosome formation, and Chapter 6 for a discussion of the sequences of the wild-type *lac* regulatory region. 4. Lac repressor concentration. One of the key tools we will use in this work is the concentration titration, where the looping probability is measured as a function of the repressor concentration. DNA constructs will be referred to with the operator closest to the microscope slide listed first; operator and loop sequences are given in Appendix B. The promoter-containing DNAs of Fig. 4.2 are identical to those shown here except that the O_1 operator closest to the bead has been replaced by O_2 , 36 bp of the loop closest to this O_2 operator are replaced by the *lacUV5* promoter sequence, and the length of the flanking DNA between O_2 and the bead is 139 bp rather than 172 bp. Fig. B.4 shows the flanking regions of the three-operator constructs of Chapter 6.

were assumed (as opposed to measured) in order for a looping J-factor to be calculated. Here we are describing a new way of measuring both the operator dissociation constants and the relative flexibilities of different DNA sequences as contained in the looping J-factor, by tuning both repressor concentration and operator strengths, with a rigorous comparison between these experiments and the theoretical models we have developed. We will argue here that only through this systematic tuning of parameters and interplay between theory and experiment is it possible to uncover some of the surprises, particularly about sequence-dependent flexibility *in vitro* and *in vivo*, that will be detailed in the following chapters.

The most important of the parameters that we will tune in this study for the purpose of the main goal of this work, that of investigating the role of sequence flexibility in loop formation, is the flexibility of the DNA in the loop, which is captured in a parameter called the looping J-factor. The looping J-factor is analogous to the cyclization J-factor, introduced above, obtained in the ligation-mediated cyclization assays which are commonly used to measure DNA flexibility at short lengths, and can be thought of as the effective concentration of one end of the loop in the vicinity of the other [56, 57]. The J-factor therefore provides a measure of the energetics of bending the DNA into the loop. The approach we have developed here allows us to measure these looping J-factors in a way that provides quantitative insights into how each of the four biologically important parameters we test affects DNA looping and permits us to contrast the role of sequence in DNA cyclization and nucleosome formation with that of looping. As we will see, we find that the two sequences discussed above, E8 and TA, which have significantly different propensities for forming DNA minicircles in *in vitro* cyclization assays or for forming nucleosomes, create a more complicated sequence dependence in the context of DNA loop formation than has been previously appreciated.

1.7 Structure of the thesis

The remainder of this thesis is organized as follows:

In Chapter 2 we develop the theoretical framework that both drives our experimental design and allows us to interpret our experimental results. We begin by analyzing in detail the effects that

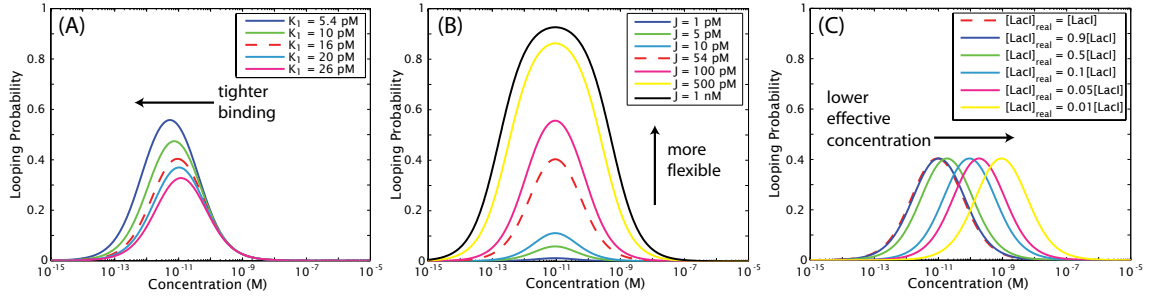


Figure 1.5: Predictions of our statistical mechanical model (Eq. (2.1)) for the effect of intentionally or unintentionally tuning various parameters of the system on the looping probability that we measure with TPM, described in more detail in Chapter 2. One of the key experimental tools we will use in this work is the concentration titration, in which the looping probability is measured as a function of repressor concentration, and so these concentration titrations are the lens through which we view the predictions of the model as well. **(A)** Prediction of the model for the effect of changing the affinity between the repressor and one of the operators, expressed as a change in one of the two repressor-operator dissociation constants (K_1). As will be demonstrated mathematically in Eq. (2.3) and Eq. (2.5), decreasing the K_d of one operator both increases looping and shifts the maximum of looping to lower repressor concentrations. **(B)** Prediction of the model for the effect of changing the J-factor of the loop, that is, changing its flexibility. As derived in Eq. (2.5), increasing the J-factor leaves the maximum of looping unchanged but increases looping at all concentrations. **(C)** In Chapter 2 we consider the effect not only of deliberately tuning the four parameters of Fig. 1.4, as we show here in (A) and (B), but also of unintentional parameter “tuning” caused by potential experimental artifacts. For example, we asked what a concentration titration would look like if there were a discrepancy between the actual concentration of repressor in the TPM chamber, and the concentration we believed we pipetted into the chamber. One example of how such a discrepancy could arise is the loss of repressor from solution by adsorption to the chamber walls. We find that the effect of such a concentration titration is a simple horizontal translation that leaves the *relative* values of the dissociation constants and J-factors unchanged, but does affect our measurement of their *absolute* values. The possible loss of protein to chamber walls is tested explicitly in Chapter 3 and found to be negligible.

the four experimentally tunable parameters of Fig. 1.4 (repressor concentration, loop length, loop flexibility, and operator strength) should have on the looping probability that we observe by TPM. These predictions are summarized in Fig. 1.5(A–B). We then turn to consequences of potential but unintended experimental effects on the looping probability, such as those caused by the adsorption of protein to the TPM chamber walls, as shown in Fig. 1.5(C). The theoretical explorations of this chapter make specific and testable predictions for the changes to the looping probability we might observe through these intentional and unintentional experimental changes.

In Chapter 3 we relate the work presented here to previous work from the Phillips lab, describing the improvements necessary to report accurate dissociation constants and J-factors, as will be done in later chapters. We also present several experimental and computational controls and other verifications of the validity of our combined theory plus TPM approach, most of which are motivated by the considerations of Chapter 2 and are explicit tests of the predictions of that chapter. For example, we designed an experiment to test whether we lose protein to the chamber walls, such that

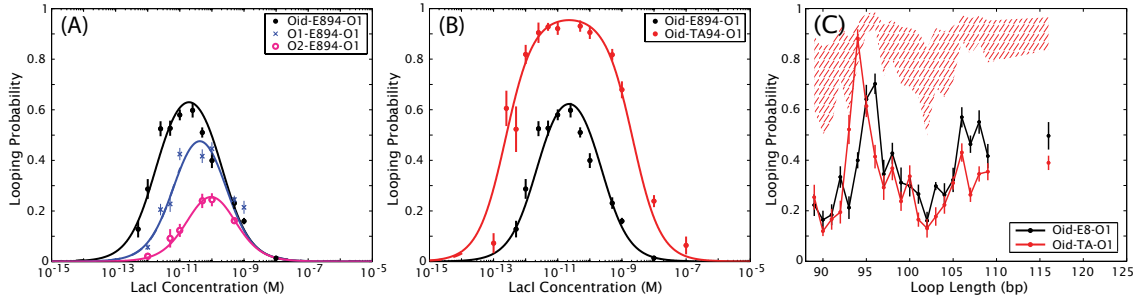


Figure 1.6: The sequence dependence of looping is more complicated than has been observed in cyclization and nucleosome affinity assays, as suggested by these results from Chapter 4. **(A)** We first demonstrate the strength of our combined statistical mechanical model and TPM approach for measuring biologically important parameters such as dissociation constants and J-factors, by showing that the effect of changing the strength of one operator agrees well with the theoretical predictions of Chapter 2, and that the dissociation constants we measure agree well with values obtained from bulk biochemical assays. Shown here is the looping probability as a function of repressor concentration, for 94 bp of the random E8 sequence flanked by three combinations of operators. As predicted in Fig. 1.5(A), increasing the strength of binding to one of the operators increases looping and shifts the maximum of looping to lower repressor concentrations. Curves are fits of Eq. (2.1) to the data, from which we obtain dissociation constants for the operators and J-factors for the DNA in the loop. **(B)** Our model is also robust to changing the J-factor: as predicted by Fig. 1.5(B), changing the sequence of the loop to the putatively more flexible sequence TA does not change the dissociation constants or the location of the maximum, but does increase looping at all concentrations. We find the J-factor for this 94 bp loop of TA to be about 10 times larger than that of a 94 bp loop of E8, qualitatively consistent with our expectations from cyclization and nucleosome affinity assays that TA is more flexible in some general sense than E8. **(C)** However, when we measure the looping probability of these two sequences at fixed repressor concentration but varying loop length, we find a sequence dependence to looping only at the 94 bp used in the concentration titrations, highlighting the importance of systematic experiments tuning several experimental parameters in order to fully capture the behavior of the system. The red hatched region indicates a prediction for where the TA data were to fall if the TA sequence were as much more flexible than E8 as measured in cyclization assays. The sequence dependence to looping is actually more complicated than is captured by the data here: we find that we can restore a sequence dependence to looping by the addition of the *lacUV5* promoter to the loop.

we would need to take into account the modified model shown graphically in Fig. 1.5(C). We verify that our model in Eq. (2.1) is sufficient to account for our data and that the potential experimental artifacts explored in Chapter 2 are not an issue in our experiments.

With a theoretical framework and validation of our approach in hand, in Chapter 4 we turn to experimental results of the systematic tuning of the four parameters listed in Fig. 1.4(B), as well as preliminary results with additional sequences and the relationship of the work presented here to analogous *in vivo* studies. This chapter addresses the main question posed in this work, that of the role of sequence-dependent flexibility in loop formation at short length scales, both *in vivo* and *in vitro*, and touches briefly as well on the flexibility of short DNAs in the context of protein-mediated loops. As summarized in Fig. 1.6(A–B), we find that our combined theory and TPM approach is robust when confronted with the tuning of the four experimental parameters of Fig. 1.4, and in particular we can obtain dissociation constants for three Lac repressor operators that agree

well with literature values obtained using bulk biochemical techniques. However, as summarized in Fig. 1.6(B–C), we find that the dependence of looping on the sequence of the loop is more complicated than suggested by cyclization and nucleosome formation: in some settings, we find that having the TA sequence in the loop leads to an increased looping probability compared to E8, but not always. From a comparison between our results and previous results on cyclization and nucleosome formation with these two sequences, we hypothesize that the *shape* of the deformation of a DNA sequence is more important than has been previously appreciated when determining sequence flexibility. In Section 4.6, we turn to the question of the role of sequence formation not in *in vitro* loop formation, but in *in vivo* gene expression. We find that one of the nucleoid-associated proteins introduced above, the nonspecific DNA-bending protein HU, apparently masks any sequence-dependence to looping *in vivo* that we observe *in vitro*, raising questions about the importance of sequence flexibility to loop formation in its biological context.

The results of Chapters 2–4 focus on one kind of information that can be obtained through TPM experiments, namely the looping probability. In Chapter 5, we discuss preliminary work on obtaining kinetic parameters, instead of equilibrium looping probabilities alone, from TPM traces. We describe two approaches to obtaining these kinetic parameters, schematized in Fig. 1.7(A–B): a half-amplitude thresholding technique that is the standard in the field, and a new hidden Markov model (HMM), based on variational Bayesian inference, as an alternative that we are developing which we hope will provide more information than can be obtained through the standard thresholding method. This HMM approach will be especially important for the analysis of the kinds of systems that we describe in Chapter 6. Chapter 6 contains preliminary work extending our approach to more complicated looping systems, using the wild-type version of the *lac* operon to examine broader questions of combinatorial control and the role of the DNA-bending proteins in the natural functioning of the operon. As shown in Fig. 1.7(C–D), we find a sharp contrast between our *in vitro* results, in which the weakest third operator of the *lac* system has no effect on looping, and *in vivo* results from other groups, in which the weakest operator is as important as the second weakest for wild-type functioning of the system. We present results that suggest that specific or nonspecific DNA-bending

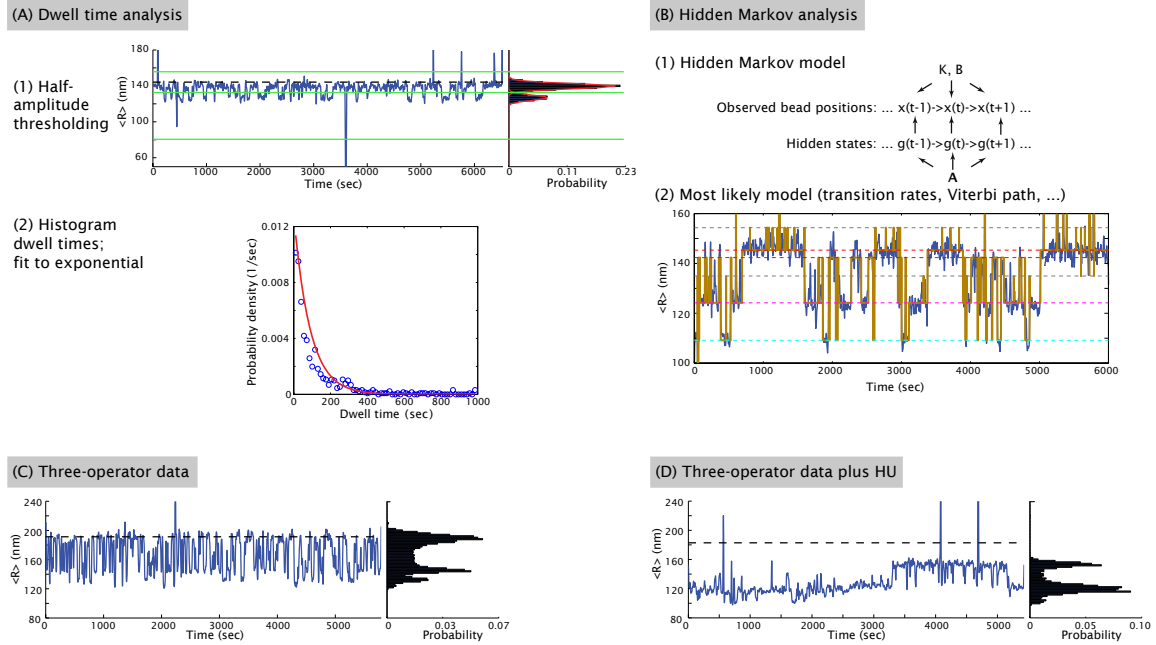


Figure 1.7: A comparison of two methods for performing state identification and obtaining kinetic information from TPM traces, described in more detail in Chapter 5. **(A)** A common technique for obtaining kinetic information from TPM trajectories is based on thresholding each trajectory, after the user has decided how many states to divide the trajectory into, and then collecting the lengths of the dwell times in each state into histograms. Fitting these histograms to exponentials yields state lifetimes and other kinetic information. This method has the advantage of being relatively straightforward and well established; however, it is subject to the temporal resolution of the smoothing filter that is applied to the trajectory, and to user bias in determining state assignments. **(B)** We are developing a new hidden-Markov-model-based approach to analyzing TPM data that overcomes some of the limitations of the method described in (A). In a hidden Markov model, the x and y bead positions that are the raw observable in TPM (the root-mean-squared bead position calculated from these raw data is shown as “ $\langle R \rangle$ ” in (A)) are generated by a series of hidden states according to Markov process, where the hidden states in our case are the tether conformation (looped or unlooped). Our HMM algorithm finds the best sequence of hidden states, as well as the best number of such states, through a statistical analysis that removes some of the user bias in the method in (A). In (2) here we show one of the outputs of the HMM analysis, the Viterbi path, which is the sequence of most likely states, here superimposed on a TPM trace. **(C)** One kind of data for which an HMM analysis like that in (B) will be vital is that generated by looping systems with more than two operators, and therefore multiple looped states, such as those examined in Chapter 6. However, as described in that chapter, we find that the wild-type three-operator system of the *lac* promoter region, diagrammed in Fig. 1.3, behaves identically to an analogous two-operator construct in which the weakest operator has been removed. This is surprising because all three operators are necessary for wild-type gene expression levels *in vivo* [93, 94]. **(D)** We speculated that nonspecific DNA-bending proteins, such as the nucleoid-associated protein HU, could be causing the *in vivo-in vitro* discrepancy regarding the importance of the third operator, and so in Chapter 6 we present preliminary results in which we have added HU to a TPM assay with a three-operator DNA. We find that the addition of HU does change the behavior of the three-operator system, by, for example, enabling long dwells in one or more looped states, as shown here. An HMM approach to analyzing these data, with a statistical method for determining the number of states, is vital, as it is difficult to threshold traces like these by eye.

proteins may be vital for the wild-type behavior of the *lac* system.

Finally, in Chapter 7, we discuss future directions for furthering our understanding of the sequence dependence of loop formation *in vivo* and *in vitro*, and the role of DNA-bending proteins in loop formation and gene expression. Materials and detailed methods can be found in the appendices at the end of this work. The bulk of these appendices and Chapters 2 and 3 will be published as the Supporting Information to [117], with the main text of that paper consisting of the first four sections of Chapter 4 here. Section 4.5 of Chapter 4 will be part of a forthcoming collaborative paper with Yi-Ju Chen of the Phillips lab; Section 4.6 of Chapter 4 appears in [118]; and Chapters 5 and 6 will become part of a forthcoming collaborative paper with Martin Lindén of Stockholm University in [119].

Chapter 2

A statistical mechanical model of the *in vitro* looping probability

In this chapter we sketch a statistical mechanical framework that allows us to see how the looping probability that we measure with TPM depends upon various tunable parameters such as the strength of the repressor binding sites, the concentration of transcription factors and the length and flexibility of the intervening DNA (see Fig. 1.4). The framework presented here builds on earlier work in [115, 120], where a simple model for the looping probability was proposed. Here we move beyond the simple model to examine the effect of tuning the various parameters of the model, and to add complexity to the model that might be needed in order to capture intended or unintended experimental modifications. These theoretical developments serve as an important conceptual framework for making the TPM assay a precise measurement scheme for determining properties of the DNA and of DNA-repressor interactions. However, this framework is neither specific to the Lac protein nor to the TPM technique, but applies equally well to other DNA-protein interactions, and to other single-molecule techniques that can detect DNA bending or looping as a function of protein concentration, such as single-molecule FRET [69], or optical [121] and magnetic [122, 114] tweezers. Further, the concepts presented here may ultimately help understand how looping in cells is controlled by precisely the same parameters since similar statistical mechanical models have been exploited in that setting as well [34, 123, 124, 118].

As we have found repressor concentration titrations, in which a series of TPM experiments are performed at different repressor concentrations and the resulting looping probabilities are measured

as a function of this concentration [115, 117], to be particularly useful tools for understanding DNA-repressor interactions, and have used them extensively in the work detailed in Chapters 3, 4, and 6, these titration curves will serve as the primary windows through which we will view TPM experiments from a theoretical viewpoint. We first give a brief introduction to the statistical mechanical model that characterizes such concentration titrations and examine how the three parameters of the model (effective binding constants and J-factor) affect the looping probability. We then extend this model to examine how the titration curves change when additional, experimentally important complications are added: inactive fractions of repressor, dimers, and low repressor concentrations relative to DNA concentration. We show that some of these effects distort the titration curves in ways that can be recognized. However, there are also effects that only rescale certain parameters without changing the qualitative shape of the curve. The latter is more insidious, as it leads to systematic errors which cannot be detected from within the context of the titration curves themselves. The results of these sections are summarized in Fig. 2.1, which illustrates how the looping titration curves are altered as a result of intentional parameter tuning and unintentional deviations from the ideal case. Throughout this section we also make reference to the experimental results presented in the next two chapters, some of which touch on these intentional and unintentional parameter changes. Finally, we present a method by which J-factors for different constructs can be measured without the need for the full concentration titrations that occupy the rest of this chapter, which will be used extensively in Chapter 4.

2.1 Tuning the simple titration curve

We first analyze the shape of the ideal titration curve in some depth, in order to answer a variety of questions.¹ How do we expect the shape of the looping probability as a function of concentration, $p_{\text{loop}}([R])$, to change if we replace one of the operators with a repressor binding site of a different affinity? What happens if we change the J-factor of the DNA by, for example, changing the distance

¹Thanks to Martin Lindén for the mathematization of the effects of changing dissociation constants or J-factors on the looping probability curve, and for the derivation of the effect in Fig. 2.1(C).

between the operators (and therefore the length of the loop)? Do two identical operators produce different titration curves than a pair of very different strengths?

We begin by summarizing the simple model derived previously in [115]. A repressor titration curve has an intuitive, bell-shaped form. At low protein concentrations, we would expect the probability of forming a loop to be small. Similarly, at high concentrations, the looping probability is also low, because the two operators are each occupied by separate transcription factors. At intermediate concentrations, the looping state has its highest probability. These intuitions can be captured mathematically by statistical mechanical models that take into account all of the different ways that the operators can be decorated with repressors. These models make very strict predictions about the functional form of the looping probability curves as a function of the various biological parameters.

In the specific case we are considering here of a TPM experiment to study looping by a protein such as the Lac repressor, the probability of the looped state can be expressed in terms of the Boltzmann weights of each of the five states available to the system: nothing bound to the DNA, one head of a repressor bound at one operator, one head of a repressor bound at the other operator, two repressors bound with one attached to each operator, or one repressor with the two heads bound to the two operators (the looped state). These states and their corresponding weights as derived in [115] are diagrammed in Fig. 2.2(A). For concreteness here through Section 2.6 we will label the operators O_{id} and O_1 , representing the strong, synthetic “ideal” operator and the strongest naturally occurring operator O_1 . These operators form the example case to which all others are compared in Fig. 2.1 and will be used as such throughout this section, though obviously these results apply equally well to other choices of operators.

For a system that satisfies equilibrium conditions, the looping probability is given by the statistical weight of the looped state divided by the sum of all the states in Fig. 2.2(A), or

$$p_{\text{loop}}([R]) = \frac{\frac{1}{2} \frac{[R]J_{\text{loop}}}{K_1 K_{id}}}{1 + \frac{[R]}{K_1} + \frac{[R]}{K_{id}} + \frac{[R]^2}{K_1 K_{id}} + \frac{1}{2} \frac{[R]J_{\text{loop}}}{K_1 K_{id}}}, \quad (2.1)$$

where $[R]$ is the repressor concentration, K_{id} and K_1 are the dissociation constants for the Lac

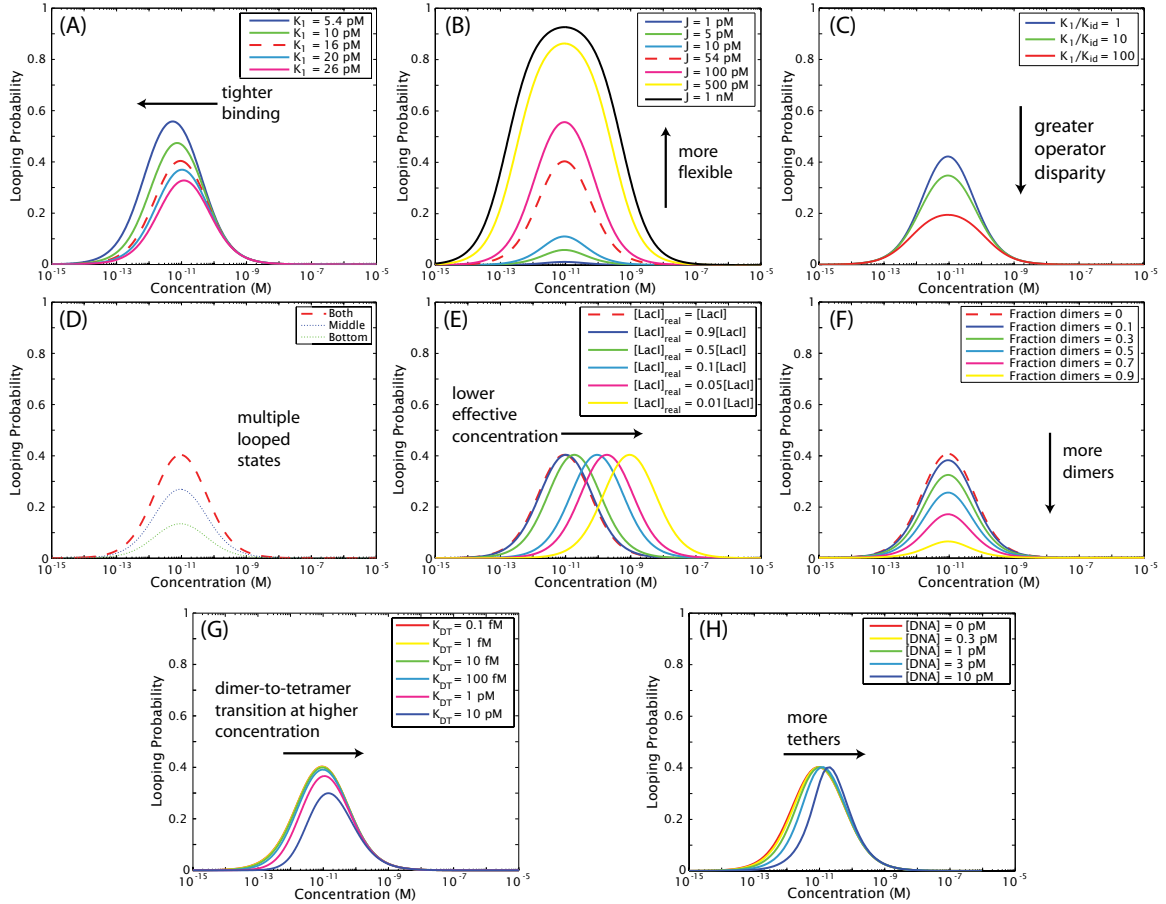


Figure 2.1: Effect of key parameters on the concentration dependence of the looping probability. Unless otherwise indicated, $K_{id} = 5.4$ pM, $K_1 = 16$ pM, and $J_{loop} = 54$ pM. These values for K_{id} and K_1 are comparable to values found in the literature for two of the known binding sites for the Lac repressor [125, 126]. (Note, however, that these are not the values we report in Chapter 4, due to different experimental conditions, such as salt concentration.) Curves with these default parameters are shown as dashed red lines for comparison across panels. **(A)** Effect of changing the strength of one of the operators. **(B)** Effect of changing the flexibility of the DNA in the loop. **(C)** Effect of changing the ratio K_1/K_{id} when the concentration $(K_1 K_{id})^{1/2}$ at which looping is maximal is kept the same. **(D)** Extension of the simple model to the case of two experimentally distinguishable looped states, which we model as having different J-factors. Here the bottom state is one-third that of the default 54 pM, and the middle is two-thirds that of the default (and the dashed red line shows the sum of the probabilities of the two states). **(E)** Effect of a discrepancy between the presumed concentration of repressor and the actual concentration. **(F)** Effect of a constant fraction of repressors that cannot dimerize and therefore cannot loop. **(G)** Effect of taking into account the dimer-to-tetramer dissociation at various low concentrations, for varying values of the tetramer-to-dimer dissociation constant K_{DT} . **(H)** Effect of competition for Lac repressor binding between different tethers at low repressor concentrations. The tether density [DNA] is defined as the number of tethers divided by total volume. Unlike those in (G) and (H), the curves in panels (E) and (F) are indistinguishable in TPM experiments, since the parameters are rescalings of those in the simple model.

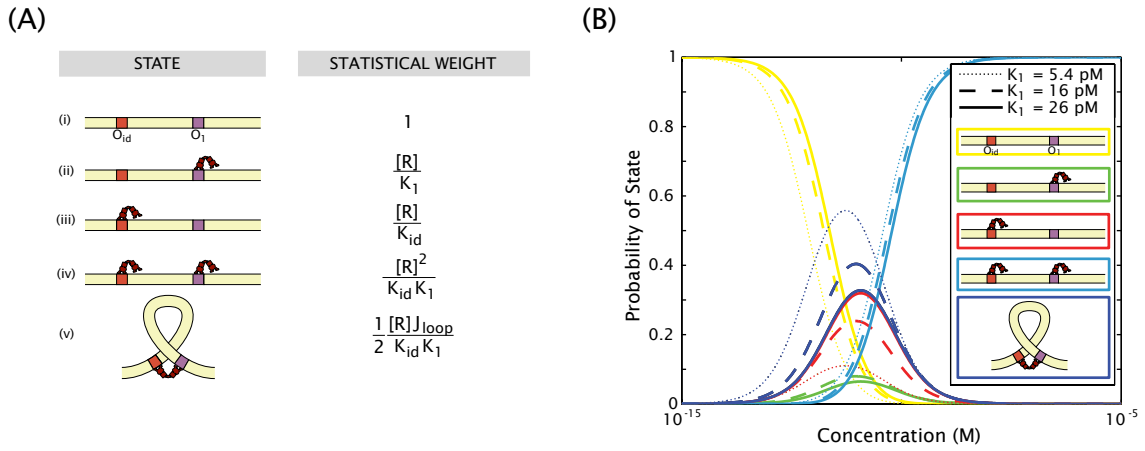


Figure 2.2: States and weights of the simple model. **(A)** Schematized states and Boltzmann weights for the simple model (first derived in [115]). $[R]$ is the repressor concentration, K_1 and K_{id} are the dissociation constants for the repressor binding to operators O_1 or O_{id} , respectively, and J_{loop} is the looping J-factor of the DNA between the operators. The looping probability is given by the weight of state (v) divided by the sum of all five states. **(B)** Probabilities of all the states of the system as K_1 is changed. This figure shows three of the curves of Fig. 2.1(A) but includes not only the looping probability but also the probabilities of the four other states schematized in (A) here. The different colors correspond to the five states as indicated in the legend. Dotted, dashed, or solid lines correspond to $K_1 = 5.4$ pM (in which case $K_1 = K_{id}$), 16 pM, or 26 pM, respectively. For example, the solid dark blue and red curves which overlap indicate that when $K_1 = 26$ pM, there is an almost equal probability at all concentrations that the system will be in the looped configuration (dark blue) or that K_1 will be bound (red); whereas when $K_1 = 5.4$ pM, where the looped state is maximal (dark blue dotted line) the single-operator-bound states have very low probabilities (green and red dotted lines, which here overlap completely because in this case $K_1 = K_{id}$). These curves confirm the intuition that looping is low at low concentrations because the predominant state is the one with no repressors bound at either operator; however at high concentrations looping is also low, because the predominant state is the one in which two repressors are bound to the two operators.

repressor from the O_{id} and O_1 operators, respectively, and J_{loop} is the sum of the individual J-factors of all the possible DNA loop topologies. For the simplest case, the looping J-factor is related to the free energy cost of bending the DNA into a loop through

$$J_{loop} = 1 \text{ M } e^{-\beta \Delta F_{loop}}, \quad (2.2)$$

where ΔF_{loop} is the free energy of forming a loop, and β is $1/k_B T$. The units are concentration: the J-factor can be thought of as the concentration of one binding site in the vicinity of the other [56, 57]. Its value depends on the length, phasing, and flexibility of the DNA, as well as the precise shape of the looped complex, and any energetic contributions from the looping protein [127, 128, 129]. It appears in Eq. (2.1) with a factor $1/2$, which is a combinatorial factor that reflects the symmetry of the Lac protein and the binding sites [115]. In Eq. (2.1), J_{loop} is the sum of the J-factors for each of four possible loop configurations that have different DNA-binding orientations, as well as for any additional loop conformations arising from protein flexibility (see Section 2.2 below). Combining the different loop topologies together in a single state, as we have done here, is appropriate for the situation where they cannot be distinguished experimentally. The generalization to several distinctive looping states is discussed below.

The first observation one can make about the titration curves in Fig. 2.1 is that there is a peak in the looping probability and that the distribution is symmetric (in log scale) around that peak. The concentration at the maximum in the looping probability can be found by differentiating Eq. (2.1) with respect to $[R]$ and results in

$$[R]_{max} = \sqrt{K_{id} K_1}. \quad (2.3)$$

The symmetry of the titration curves around this point can be explained by observing that

$$p_{loop}(10^n [R]_{max}) = p_{loop}(10^{-n} [R]_{max}). \quad (2.4)$$

Note also that the concentration at which the looping probability is maximized does not depend

upon the DNA flexibility as captured in the parameter J_{loop} . The looping probability at $[R]_{\text{max}}$ is given by

$$p_{\text{loop}}([R]_{\text{max}}) = \frac{J_{\text{loop}}/2}{J_{\text{loop}}/2 + (\sqrt{K_{\text{id}}} + \sqrt{K_1})^2}. \quad (2.5)$$

These results explain several features of the titration curves. As shown in Fig. 2.1(A), increasing the binding strength of one of the operators, (i.e., decreasing the value of K_1 or K_{id}) shifts the maximum of the curve to the left and increases its amplitude; that is, stronger operators allow looping at lower concentrations. However, increasing the J-factor (i.e., making the DNA more flexible), as in Fig. 2.1(B), changes only the height of the curve, but not the concentration of repressor at which looping is maximal. These qualitative predictions are borne out in the experimental data presented in Fig. 4.1(D) and (E) in Chapter 4.

Finally, we can consider the behavior of the outer tails of the curves. As shown in Fig. 2.1(A), changing the operator strengths changes the behavior at low concentrations, but not at high concentrations, where curves with different operators (but identical J -factors) fall off in the same fashion. The behavior at high and low concentrations can be read off directly from Eq. (2.1). At high concentrations, the doubly-bound state dominates, which has a weight of $[R]^2/(K_1 K_{\text{id}})$; therefore in the high concentration limit,

$$\lim_{[R] \gg J_{\text{loop}}, K_1, K_{\text{id}}} p_{\text{loop}}([R]) \approx \frac{J_{\text{loop}}}{2[R]}. \quad (2.6)$$

This shows that the binding constants drop out of the equation at high concentrations. In the limit of low concentrations, the state with no repressors bound dominates, which has a weight of 1; that is, in the low-concentration limit,

$$\lim_{[R] \ll J_{\text{loop}}, K_1, K_{\text{id}}} p_{\text{loop}}([R]) \approx \frac{J_{\text{loop}}[R]}{2K_1 K_{\text{id}}} = \frac{J_{\text{loop}}[R]}{2[R]_{\text{max}}^2}. \quad (2.7)$$

These results reflect the different states that compete with the looped state in the two limits. At high concentrations, the looped state is out-competed by the doubly occupied state, and since the

weights of both states have the same dependence on operator strength, the outcome only depends on the J -factor. At low concentrations, on the other hand, the looped state is out-competed by the unoccupied state (with weight unity), and the outcome therefore depends on all parameters. (See also Fig. 2.2(B) above.)

By way of contrast, changing the J -factor moves both tails in a symmetric fashion, as illustrated in Fig. 2.1(B). This behavior is dictated by the symmetry property: since the peak position is independent of J , changes in J have to influence the high- and low-concentration parts of the curve equally. Likewise, both the high-concentration and low-concentration behaviors of the curve in Eqs. (2.6) and (2.7) depend equally on the J -factor. From an experimental perspective this makes data at concentrations near and below $[R]_{\max}$ crucial for measuring dissociation constants. On the other hand, if one is only interested in the J -factor, data at high concentrations is sufficient.

Since the high- and low-concentration limits (and therefore the width of the titration curve), as well as the peak position, depend on the operator strengths only through $[R]_{\max}$, we can ask how changing the relative strengths of the operators in a way that leaves $[R]_{\max}$ unchanged affects the looping probability. As shown in Fig. 2.1(C), if we change the operators in a way that leaves $[R]_{\max}$ unaffected, only the peak height changes, not the peak position or the width of the titration curve. The peak looping probability, given by Eq. (2.5) above, can be rearranged in a way that separates out the dependence on difference in operator strength from the other factors. Specifically, if we define $\alpha = K_1/K_{\text{id}}$, Eq. (2.5) can be written as

$$p_{\text{loop}}([R]_{\max}) = \frac{\frac{J}{2[R]_{\max}}}{\frac{J}{2[R]_{\max}} + (\alpha^{1/4} + \alpha^{-1/4})^2}. \quad (2.8)$$

Since $\alpha^{1/4} + \alpha^{-1/4} \geq 2$, with equality only if $K_1 = K_{\text{id}}$, this tells us that equal operators are “best” for looping, in the sense that for given peak position and J -factor, equal operators maximize the looping probability, as shown in Fig. 2.1(C). Further intuition about this behavior comes from considering the competition of states illustrated in Fig. 2.2(B). The looping probability near the peak is dominated by a competition between the looped state, and the singly occupied state with the repressor bound to

the strongest operator. Changing the relative operator strength while keeping $K_1 K_{\text{id}}$, and therefore $[R]_{\text{max}}$, constant selectively strengthens that singly occupied state, and therefore poisons looping near the peak. However, in the limits of high and low concentrations, the looped state is instead out-competed by the doubly occupied and unoccupied state respectively, whose weights relative to the looped state only depend on the average binding strength $[R]_{\text{max}} = \sqrt{K_1 K_{\text{id}}}$, so these regimes are not affected by changes in the relative operator strengths.

2.2 The case of multiple looped states

Interestingly, looping by the Lac repressor is more subtle than the simple model described so far. Several studies have previously reported two looped states for the Lac repressor in the case of DNAs with two operators [115, 68, 109, 114, 108, 67, 69, 70], and we observe these two states in the work presented here as well (see Figs. 4.1(F), 4.2(B) and (E), and 4.3 in Chapter 4, and Appendix E). These two looped states have been attributed to flexibility in the tetramerization domain of the repressor and/or to superpositions of four DNA loop topologies [68, 109, 108, 127, 128, 129, 130, 70]. Regardless of their underlying physical origin, we and others model the two looped states as having the same dissociation constants but different effective J-factors (see Fig. 4.1(F) in Chapter 4 for the first, to our knowledge, experimental confirmation of this assumption). So the looping probabilities of these two experimentally distinguishable states can be modeled as

$$p_{\text{loop},1} = \frac{\frac{1}{2} \frac{R J_{\text{loop},1}}{K_1 K_{\text{id}}}}{1 + \frac{R}{K_1} + \frac{R}{K_{\text{id}}} + \frac{R^2}{K_1 K_{\text{id}}} + \frac{1}{2} \frac{R J_{\text{loop},1}}{K_1 K_{\text{id}}} + \frac{1}{2} \frac{R J_{\text{loop},2}}{K_1 K_{\text{id}}}} \quad (2.9)$$

$$p_{\text{loop},2} = \frac{\frac{1}{2} \frac{R J_{\text{loop},2}}{K_1 K_{\text{id}}}}{1 + \frac{R}{K_1} + \frac{R}{K_{\text{id}}} + \frac{R^2}{K_1 K_{\text{id}}} + \frac{1}{2} \frac{R J_{\text{loop},1}}{K_1 K_{\text{id}}} + \frac{1}{2} \frac{R J_{\text{loop},2}}{K_1 K_{\text{id}}}}. \quad (2.10)$$

Note that Eqs. (2.9) and (2.10) sum to Eq. (2.1) with $J_{\text{loop}} = J_{\text{loop},1} + J_{\text{loop},2}$. Also note that the ratio of Eqs. (2.9) and (2.10) is just $J_{\text{loop},1}/J_{\text{loop},2}$, which means that the titration curves of the two states should have identical shape up to an overall scaling factor. This scaling can be seen in Fig. 2.1(D), where Eqs. (2.9), (2.10), and (2.1) are plotted with $J_{\text{loop}} = 54$ pM, $J_{\text{loop},1} = J_{\text{loop}}/3$ (labeled the

“bottom” state), and $J_{\text{loop},2} = 2J_{\text{loop}}/3$ (labeled “middle”), as well as in the experimental results of Fig. 4.1(F) in Chapter 4.

2.3 Effect of an inactive fraction of repressor

One question that arises in thinking about actual TPM experiments is how should we expect $p_{\text{loop}}([R])$ to change if the concentration of repressor that is presumed to have been pipetted into the TPM chamber isn’t the “real” concentration of repressor that contributes to the observed looping? The most obvious way in which this could happen is by measurement errors when determining the repressor stock concentration. Another possibility is that some repressor molecules cannot bind DNA, for example due to misfolding that affects both DNA binding sites. Unless they interact in some way with the functional repressor (e.g., via crowding effects at high concentrations) the effect is the same as in the first case: the concentration of active repressors is lowered. A third case is tested experimentally in the next chapter (Section 3.3.2): some protein may bind nonspecifically to the TPM chamber walls and thereby not participate in observable looping. Note that in Section 2.4, we will consider the case in which dimers poison looping, by binding one operator but not forming a loop. Here, inactive repressors cannot bind DNA and so simply contribute to a discrepancy between the real and presumed concentrations.

We can model these cases of inaccurate concentration by a fraction f , such that $[R]$ is the concentration we believe we flow into the chamber, but the concentration of active repressors that contribute to looping is instead $f[R]$. Substituting $[R] \rightarrow f[R]$ in Eq. (2.1) leads to

$$p_{\text{loop, inactive fraction}}([R]) = \frac{\frac{1}{2} \frac{f[R]J_{\text{loop}}}{K_1 K_{\text{id}}}}{1 + \frac{f[R]}{K_1} + \frac{f[R]}{K_{\text{id}}} + \frac{(f[R])^2}{K_1 K_{\text{id}}} + \frac{1}{2} \frac{f[R]J_{\text{loop}}}{K_1 K_{\text{id}}}}, \quad (2.11)$$

which can be rewritten in the form

$$p_{\text{loop, inactive fraction}}([R]) = \frac{\frac{1}{2} \frac{[R](J_{\text{loop}}/f)}{(K_1/f)(K_{\text{id}}/f)}}{1 + \frac{[R]}{(K_1/f)} + \frac{[R]}{(K_{\text{id}}/f)} + \frac{([R])^2}{(K_1/f)(K_{\text{id}}/f)} + \frac{1}{2} \frac{[R](J_{\text{loop}}/f)}{(K_1/f)(K_{\text{id}}/f)}}. \quad (2.12)$$

This means that uncertainty about the overall repressor concentration affects all parameters equally, by literally distorting the basic yardstick of the TPM titration assay. The effect on titration curves is a simple horizontal shift, as shown in Fig. 2.1(E). The experimental implication is that one cannot detect an inactive fraction from any distortion of the titration curve. On the other hand, the parameters are rescaled equally, so ratios of fitted parameters are insensitive to this kind of experimental uncertainty. Section 3.2 discusses the impact of this potential source of error on the fitted parameters for the experimental results presented in this work, and as mentioned above, Section 3.3.2 describes an experimental control to test for one of the potential sources of a concentration discrepancy, namely, loss of protein to the chamber walls.

2.4 Effect of the presence of dimers in solution

The Lac repressor is a dimer of dimers, with each dimer of the wild-type tetramer forming a single DNA-binding domain [95]. Therefore only tetramers can loop DNA, having two DNA binding domains in the same molecule; but dimers can bind individual operators. If a dimer binds one of the operators of a DNA molecule, that DNA cannot form a loop even if a tetramer binds the other site; thus dimers “poison” looping.

There are three conceivable scenarios that would lead to dimers in a solution of otherwise wild-type tetrameric repressors. First, at very low repressor concentrations the tetramer is thought to dissociate into its component dimers [131], a reaction governed by an equilibrium constant that we will call K_{DT} . A second possible scenario is one in which a fraction of repressors is damaged in some way due to the purification, storage, or thawing process, leading to an inability of some repressors to form tetramers. This will lead to a fraction of dimers that is constant with the total repressor concentration. Third, a fraction of monomers could be damaged such that when incorporated into a tetramer they result in a head that is unable to bind DNA. In this case, a fraction of tetramers would be “dimers” in the sense that one head can bind DNA but the protein cannot loop the DNA; however this would also result in a population of tetramers that cannot bind DNA at all. We will first discuss $p_{\text{loop}}([R])$ for the case where we consider the tetramer-to-dimer dissociation at low concentrations,

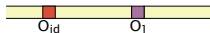
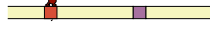
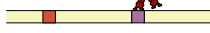
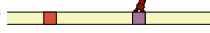





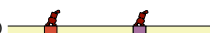
STATE	STATISTICAL WEIGHT	STATE	STATISTICAL WEIGHT
(i) 	1	(vi) 	$\frac{1}{2} \frac{[D]}{K_{id}}$
(ii) 	$\frac{[T]}{K_1}$	(vii) 	$\frac{1}{2} \frac{[D]}{K_1}$
(iii) 	$\frac{[T]}{K_{id} K_1}$	(viii) 	$\frac{[T][D]}{K_{id} K_1}$
(iv) 	$\frac{[T]^2}{K_{id} K_1}$	(ix) 	$\frac{[T][D]}{K_{id} K_1}$
(v) 	$\frac{1}{2} \frac{[T]J_{loop}}{K_{id} K_1}$	(x) 	$\frac{1}{4} \frac{[D]^2}{K_{id} K_1}$

Figure 2.3: Schematized states and Boltzmann weights for a model that includes dimers. Eq. (2.16) is obtained by observing that since $[R] = [T] + [D]/2$, states (ii) and (vii) combine to make the third term of Eq. (2.16), states (iii) and (vi) to make the second term, and states (iv), (viii), (ix), and (x) to make the fourth term. Therefore the presence of dimers affects only the looped state—all other states are insensitive to whether a dimer or a tetramer is bound at each site.

and then comment on the case of a constant fraction of dimers. The third case will not be considered but is well within the scope of scenarios that can be captured by the class of models presented here.

The most recent estimate of K_{DT} for the Lac repressor from biochemical data claims an upper bound in the femtomolar range [131]. This estimate is obtained in part from the fact that no dimers have been observed at any concentrations used in biochemical experiments, which typically do not examine concentrations below about 1 pM (since K_{DT} is the concentration at which half of the repressors in solution are dimers, K_{DT} could be at most in the femtomolar or tens of femtomolar range, in order for the fraction of dimers to be essentially zero at picomolar repressor concentrations). Single-molecule techniques can, however, measure looping at concentrations below 1 pM, and in fact these concentrations are crucial to the determination of the shape of the looping probability versus concentration curve for some choices of operators and J-factors (see Fig. 2.1(A,B)). It is therefore critical to determine the effect of K_{DT} on the looping probability.

In contrast to the simple model, a model that takes into account dimers must have ten states, five the same as those of the simple model and five that allow dimers or combinations of dimers and tetramers to bind to the DNA. These ten states and their associated weights are diagrammed in Fig. 2.3. To calculate the statistical weights of the dimer-containing states, we will assume that dimers and tetramers have the same operator dissociation constants. This assumption is reasonable

given experimental evidence that at least some forms of LacI mutants that cannot tetramerize retain the same dissociation constants as wild-type repressor [132, 133, 134]. On the other hand, this assumption is not critical to the calculation and the more general case is a simple extension of the calculation presented here.

We also assume that the binding of a tetramer or dimer to a DNA tether does not affect the equilibrium between tetramers and dimers in solution. This assumption is similar to that discussed in Section 2.6 below, regarding the independence of different tethers in the same flow chamber. Since it is believed that tetramers dissociate into dimers at low total repressor concentrations, this is not obviously true; at low concentrations, one might indeed expect that single binding and association/dissociation events affect the binding and dimerization equilibria. However, the results of Section 2.6 show that low repressor concentration only affects the simple model for certain values of operator binding strengths and DNA J-factors. Analogously, we expect the calculation in this section to be reasonable for some parameter values, but note that the approach to low concentrations discussed in Section 2.6 could be extended to include dimerization effects as well.

Finally, we define the total repressor concentration $[R]$ such that

$$[R] = \frac{[D]}{2} + [T], \quad (2.13)$$

where $[T]$ and $[D]$ are the concentrations of tetramers and dimers, respectively. This definition arises as follows: experimentally we measure the absorbance of purified repressor at 280 nm, and use the monomer extinction coefficient to obtain a mass concentration of monomers (see Appendix C). We then divide by the molecular weight of a tetramer, which is 4 times the molecular weight of a monomer (since the Lac repressor is a homotetramer), to obtain what we call the concentration of repressor that we flow onto the slide. Therefore we can say that this concentration, $[R]$, is

$$[R] = \frac{\text{num. monomers}}{4 \times \text{vol}}. \quad (2.14)$$

Since a tetramer is four monomers and a dimer two monomers, we have

$$[R] = \frac{2D}{4 \times \text{vol}} + \frac{4T}{4 \times \text{vol}}, \quad (2.15)$$

which simplifies to the equation above.

With these definitions and assumptions, we find that the partition function for the states schematized in Fig. 2.3 can be written as

$$Z_{\text{dimers}} = 1 + \frac{[R]}{K_{\text{id}}} + \frac{[R]}{K_1} + \frac{[R]^2}{K_1 K_{\text{id}}} + \frac{[T]J_{\text{loop}}}{2K_1 K_{\text{id}}}. \quad (2.16)$$

(A detailed derivation is given in Appendix A.) Note that the only state that has changed in the dimers model, compared to the simple model, is the looped state. This makes sense because the looped state is the only one in which it matters if a tetramer or a dimer is bound to the operators.

Finally, we make use of the definition of the tetramer-to-dimer dissociation constant for the reaction $T \Leftrightarrow 2D$, which is

$$K_{DT} = \frac{[D]^2}{[T]}. \quad (2.17)$$

This allows us to eliminate $[T]$ from Eq. (2.16), leaving us with our final result in terms of $[R]$ and K_{DT} :

$$p_{\text{loop, dimers}} = \frac{\frac{[R]J_{\text{loop}}}{2K_1 K_{\text{id}}} \left(1 + \frac{K_{DT}}{8[R]} - \frac{1}{8} \left[\left(\frac{K_{DT}}{[R]} \right)^2 + 16 \frac{K_{DT}}{[R]} \right]^{1/2} \right)}{1 + \frac{[R]}{K_{\text{id}}} + \frac{[R]}{K_1} + \frac{[R]^2}{K_1 K_{\text{id}}} + \frac{[R]J_{\text{loop}}}{2K_1 K_{\text{id}}} \left(1 + \frac{K_{DT}}{8[R]} - \frac{1}{8} \left[\left(\frac{K_{DT}}{[R]} \right)^2 + 16 \frac{K_{DT}}{[R]} \right]^{1/2} \right)}. \quad (2.18)$$

We recover the simple model in the limit that K_{DT} is zero, that is, when tetramers never dissociate into dimers. Fig. 2.1(G) illustrates how the looping probability changes as K_{DT} approaches the K_d 's for the operators. The presence of dimers has two main effects. First, tetramer dissociation breaks the symmetry of the titration curve (since it occurs only at low repressor concentrations), and therefore is an effect that can, at least in principle, be detected in TPM experiments, in contrast to an inactive fraction, which rescales all the parameters. Second, since tetramer dissociation only

occurs at low concentrations, it will contribute mainly an uncertainty factor in the determination of binding constants. The J-factor should be less affected, since it is strongly influenced by the high-concentration data where dimerization is not an issue (see the discussion of Eqs. (2.6) and (2.7) above). In Section 3.2 in the next chapter we discuss the relevance of this extension of the model to our data and use numerical arguments to estimate the value of K_{DT} for wild-type lac repressor.

We turn briefly to the case that involves the presence of a constant fraction of dimers at all concentrations. We note that as in the previous derivation involving K_{DT} , the only state that is affected by the binding of a dimer versus a tetramer is the looped state. The weights of all other states depend only on the total repressor concentration $[R]$. We can therefore start with Eq. (2.16), but then define the concentration of repressors in tetrameric form, $[T]$, to be

$$[T] = (1 - \nu)[R], \quad (2.19)$$

where ν is the fraction that are dimers. Then because $\frac{[D]}{2} + [T] = [R]$, we must define ν as

$$\nu = \frac{[D]}{2[R]}, \quad (2.20)$$

so that $[D]/2 + [T] = 2\nu[R]/2 + (1 - \nu)[R] = [R]$. We can now use this expression for $[T]$ in Eq. (2.16), so that when we form the looping probability we obtain

$$p_{\text{loop, const. dimers}} = \frac{\frac{(1-\nu)[R]J}{2K_1K_{id}}}{1 + \frac{[R]}{K_{id}} + \frac{[R]}{K_1} + \frac{[R]^2}{K_1K_{id}} + \frac{(1-\nu)[R]J}{2K_1K_{id}}}. \quad (2.21)$$

As with the case of an inactive fraction, this model involves only a rescaling of the parameters K_1 , K_{id} , and J_{loop} , by a factor $1/(1 - \nu)$, and therefore cannot be distinguished from the simple model in the absence of additional information. Fig. 2.1(F) above shows how a constant fraction of dimers affects the looping probability as a function of concentration.

2.5 Effect of cooperative binding of repressor heads

In [115], and in all of the models presented here, it is assumed that the binding of the two heads of the Lac repressor is independent, that is, that the binding of one head to DNA does not affect the affinity of the other head for DNA. However, for at least some salt concentrations, the binding of the second head has been found to be anticooperative [135]. We therefore ask what the effect of such (anti)cooperativity would be.

We consider two mathematically equivalent but conceptually distinct definitions of cooperativity. First, we note that we assume non-cooperative binding of the repressor heads in the simple model by asserting that a free head in solution has half the energy of a full tetramer in solution [115]. If, however, we do not make this assumption, but instead maintain the full definition in [115] that the change in energy when a repressor binds to one of the operators is

$$\Delta\epsilon_b = \epsilon_b + \epsilon_t - \epsilon_{sol}, \quad (2.22)$$

where ϵ_t is the energy of a head free in solution, then the looped state acquires an extra energy term $\omega = e^{-\beta(\epsilon_{sol} - 2\epsilon_t)}$ which we will call the “cooperativity factor”. Note that if we re-introduce the assumption that the energy of a head free in solution when the other is bound, ϵ_t , is equal to half the energy of a repressor with both heads free in solution, ϵ_{sol} , then we recover the simple model because this extra factor in the looped term goes to 1. The energy of a head free in solution might change when the other head binds DNA if, for example, the unbound head binds nonspecifically to non-operator DNA. (See also the discussion in Appendix A.)

This additional energy term in the looped state leads to a new looping probability that can be expressed as

$$p_{\text{loop, cooperative}}([R]) = \frac{\frac{1}{2} \frac{[R]J_{\text{loop}}}{K_1 K_{\text{id}}} \omega}{1 + \frac{[R]}{K_1} + \frac{[R]}{K_{\text{id}}} + \frac{[R]^2}{K_1 K_{\text{id}}} + \frac{1}{2} \frac{[R]J_{\text{loop}}}{K_1 K_{\text{id}}} \omega}, \quad (2.23)$$

where again ω measures the degree of cooperativity between heads. Binding is cooperative if ω is greater than 1; binding is anticooperative if ω is less than 1; and binding is independent, as in the

simple model, if $\omega = 1$.

A second and distinct way to capture cooperative or anticooperative binding between the repressor heads is by including an additional “interaction” energy, ϵ_{int} , in the weight of the looped state, to capture changes in the affinity of the second head for operator DNA when the first head is bound. Whereas the conceptual starting point for cooperativity described above focuses on the energetics of a *free head* in solution when the first head is bound, here we focus on the energetics of *binding* of one versus two heads. This second conceptual starting point is essentially that usually used of allosteric interactions, for example in hemoglobin, where the binding of oxygen to one domain of hemoglobin alters the affinity of the other sites for oxygen. In the case of the Lac repressor this could happen if, for example, the repressor-operator dissociation constants are different in the looped state than in the other states, because of strain on the DNA imposed by the loop shape that then affects the affinity of the repressor for the (bent or otherwise strained or distorted) operator DNA. In this second case of cooperativity, the looping probability becomes

$$p_{\text{loop, cooperative}}([R]) = \frac{\frac{1}{2} \frac{[R]J_{\text{loop}}}{K_1 K_{\text{id}}} e^{-\beta\epsilon_{\text{int}}}}{1 + \frac{[R]}{K_1} + \frac{[R]}{K_{\text{id}}} + \frac{[R]^2}{K_1 K_{\text{id}}} + \frac{1}{2} \frac{[R]J_{\text{loop}}}{K_1 K_{\text{id}}} e^{-\beta\epsilon_{\text{int}}}}. \quad (2.24)$$

Despite their different conceptual starting points, Eqs. (2.23) and (2.24) are obviously mathematically equivalent, with $\omega = e^{-\beta\epsilon_{\text{int}}}$.

More importantly, Eqs. (2.23) and (2.24) are also mathematically equivalent to the simple model if we define an “effective J-factor” $J'_{\text{loop}} = J_{\text{loop}}\omega$. As with the case of an inactive fraction of repressor or a constant fraction of dimers, Eqs. (2.23) and (2.24) represent a rescaling of the parameters of the simple model, and therefore concentration titrations cannot detect cooperative binding in the absence of prior knowledge about the J-factor. In particular, concentration titrations of the kind discussed here in fact measure J'_{loop} , not simply J_{loop} , and all of the J-factors discussed in the next chapters should be considered dependent not only on the mechanical properties of the DNA in the loop, but also on parameters related to the looping protein, such as cooperative binding. As we will see in Fig. 4.3 in Chapter 4, such repressor-dependent parameters can have large effects on the

effective J-factors that are measured by looping assays.

2.6 Low repressor concentrations

A major assumption behind the simple titration curves featured throughout this work is that the conformations of different tethers are independent, so that we only have to model one tether. However, a TPM chamber contains many tethers, and one of the ways they might interact is if binding of repressor molecules at some tethers significantly decreases the number of repressors available for binding to other tethers. Intuitively, we expect this to be an issue at low concentrations only, where the total number of repressors is comparable to, or smaller than, the total number of tethers. In that case, the number of repressors available for binding might become significantly less than the total number of repressors, which lowers the looping probability. On the other hand, low concentrations also increase the probability of the unoccupied state. These two trends compete, and in the following, we will use a simple mean field analysis to estimate how these competing effects play out at low repressor concentrations.²

The starting point for this analysis is the observation that the repressor molecules in the test chamber are either bound to an operator, or free in solution and available for binding. (Here, as in [115], we assume the binding of repressors to non-operator DNA is negligible, based on the relative magnitudes of the non-operator DNA concentration in the chamber and the association constant of repressor to non-operator DNA.) If we define a tether concentration $[\text{DNA}]$ as the total number of tethers divided by the total volume, and $\langle n \rangle$ as the average number of bound repressors per tether (that is, n can be 0, 1, or 2, since each tether can have zero, one, or two repressors bound, and $\langle n \rangle$ is the average across all tethers in the sample), we can divide the total repressor concentration into a free and a bound part according to

$$[R] = [R]_{\text{free}} + [\text{DNA}] \langle n \rangle, \quad (2.25)$$

²Thanks to Martin Lindén for this derivation.

where $[R]_{\text{free}}$ is the average concentration of free (unbound) repressors. Next, we make the approximation that the tethers are in equilibrium with the average repressor concentration $[R]_{\text{free}}$. This means that we neglect both temporal and spatial fluctuations in repressor concentration, similar in spirit to simple mean-field theories of spin systems [136], and simply substitute $[R] \rightarrow [R]_{\text{free}}$ in the weights of the simple model of Fig. 2.2(A). We can then use these weights to write an approximate expression for $\langle n \rangle$, by noting that $\langle n \rangle$ is the probability of having no repressors bound times zero, plus the probability of having one repressor bound times one, plus the probability of having two repressors bound times two. That is,

$$\langle n \rangle = \frac{\left(\frac{[R]_{\text{free}}}{K_1} + \frac{[R]_{\text{free}}}{K_{\text{id}}} + \frac{J[R]_{\text{free}}}{2K_1 K_{\text{id}}} \right) + 2 \frac{[R]_{\text{free}}^2}{K_1 K_{\text{id}}}}{1 + \frac{[R]_{\text{free}}}{K_1} + \frac{[R]_{\text{free}}}{K_{\text{id}}} + \frac{J[R]_{\text{free}}}{2K_1 K_{\text{id}}} + \frac{[R]_{\text{free}}^2}{K_1 K_{\text{id}}}}. \quad (2.26)$$

We now have two equations for the two unknowns, $[R]_{\text{free}}$ and $\langle n \rangle$. These can be solved numerically, but it is also instructive to study the behavior at high- and low-repressor concentrations analytically. Since $\langle n \rangle \leq 2$, and $[\text{DNA}]$ is constant, the expected high-concentration limit $[R]_{\text{free}} \approx [R]$, i.e., the simple model, can be read off from Eq. (2.25). The low-concentration limit of Eq. (2.26), which we get by retaining only the linear term in the numerator and the constant term in the denominator, is

$$\langle n \rangle \rightarrow \frac{[R]_{\text{free}}}{c_T}, \quad \text{with} \quad \frac{1}{c_T} = \frac{K_1 + K_{\text{id}} + J/2}{K_1 K_{\text{id}}}. \quad (2.27)$$

If we substitute this back into Eq. (2.25), we can solve for the fraction of free repressors at low concentration,

$$\frac{[R]_{\text{free}}}{[R]} \rightarrow (1 + [\text{DNA}]/c_T)^{-1}. \quad (2.28)$$

At intermediate concentrations, this ratio interpolates smoothly between the high- and low-concentration limits. The low-concentration limit is interesting for two reasons. First, we note that the ratio $[R]_{\text{free}}/[R]$ in Eq. (2.28) becomes independent of $[R]$. This means that the low-concentration part of the titration curve is simply shifted to the right, as illustrated in Fig. 2.1(H). Second, the magnitude of the maximum shift depends on the tether properties, through the characteristic concentration

c_T : only if this concentration is low compared to the tether concentration [DNA] will the simple model fail at low concentrations. A high J-factor, low peak concentration $[R]_{\max} = (K_1 K_{\text{id}})^{1/2}$, and large variation in operator strength (for a given peak concentration) leads to a low c_T . The intuition here is that these features all lead to an increased tendency to have repressors bound at low concentrations, which lowers the number of repressors in solution. Section 3.2.2 describes the applicability of these low-concentration considerations to the data presented in this work.

2.7 Calculating relative J-factors

So far we have been modeling experiments in which we measure the looping probability at several repressor concentrations, and have considered how such concentration titrations may change as we tune the various parameters (e.g., operator strengths or J-factors.) In the following chapters we will fit these concentration titration data to our model to measure J-factors in absolute units (see, for example, Fig. 4.1(D–F) and Table 4.1). However, if we are interested only in the relative flexibilities of two sequences, or if the J-factor for one construct is known and we wish to find the J-factor for a construct with the same operators but a different loop (as in Fig. 4.2(C) and (F) in Chapter 4), our model predicts that we can compute the ratio of the looping J-factors of two sequences based solely on a single pair of looping probabilities, even if we do not know the values of the flanking operator K_d 's.

We can do so by fixing the concentration of repressor and measuring the looping probabilities of the two DNAs, and then computing the ratio

$$\frac{p_{\text{unloop}}/p_{\text{loop}}}{p'_{\text{unloop}}/p'_{\text{loop}}} = \frac{J'_{\text{loop}}}{J_{\text{loop}}}, \quad (2.29)$$

where p_{unloop} for a sequence is $1 - p_{\text{loop}}$. This result is obtained by starting with the probability of being in the *unlooped* state,

$$p_{\text{unloop}} = \frac{1 + \frac{R}{K_i} + \frac{R}{K_{ii}} + \frac{R^2}{K_i K_{ii}}}{1 + \frac{R}{K_i} + \frac{R}{K_{ii}} + \frac{R^2}{K_i K_{ii}} + \frac{1}{2} \frac{R J_{\text{loop}}}{K_i K_{ii}}}. \quad (2.30)$$

We then construct the ratio of unlooping-to-looping probability for a single DNA. The ratio of p_{unloop} to p_{loop} is given by

$$\frac{p_{\text{unloop}}}{p_{\text{loop}}} = \left(\frac{1 + \frac{R}{K_i} + \frac{R}{K_{ii}} + \frac{R^2}{K_i K_{ii}}}{1 + \frac{R}{K_i} + \frac{R}{K_{ii}} + \frac{R^2}{K_i K_{ii}} + \frac{1}{2} \frac{R J_{\text{loop}}}{K_i K_{ii}}} \right) \left(\frac{1 + \frac{R}{K_i} + \frac{R}{K_{ii}} + \frac{R^2}{K_i K_{ii}} + \frac{1}{2} \frac{R J_{\text{loop}}}{K_i K_{ii}}}{\frac{1}{2} \frac{R J_{\text{loop}}}{K_i K_{ii}}} \right) \quad (2.31)$$

which simplifies to

$$\frac{p_{\text{unloop}}}{p_{\text{loop}}} = \frac{1 + \frac{R}{K_i} + \frac{R}{K_{ii}} + \frac{R^2}{K_i K_{ii}}}{\frac{1}{2} \frac{R J_{\text{loop}}}{K_i K_{ii}}}. \quad (2.32)$$

We form the same ratio $p'_{\text{unloop}}/p'_{\text{loop}}$ for a second DNA. When we divide these ratios of unlooped-to-looped probabilities for the two DNAs, we obtain

$$\frac{p_{\text{unloop}}/p_{\text{loop}}}{p'_{\text{unloop}}/p'_{\text{loop}}} = \left(\frac{1 + \frac{R}{K_i} + \frac{R}{K_{ii}} + \frac{R^2}{K_i K_{ii}}}{\frac{1}{2} \frac{R J_{\text{loop}}}{K_i K_{ii}}} \right) \left(\frac{\frac{1}{2} \frac{R J'_{\text{loop}}}{K_i K_{ii}}}{1 + \frac{R}{K_i} + \frac{R}{K_{ii}} + \frac{R^2}{K_i K_{ii}}} \right). \quad (2.33)$$

Note that the concentration dependence (as well as operator dependence) cancels. Fig. D.5(A) in Appendix D.2.7 illustrates this key claim of our model, that the ratio of J-factors computed from a pair of looping probabilities is independent of repressor concentration. We consider this concentration independence to Eq. (2.29) to be an important test of the validity of our model and also a reasonable basis for using single concentrations to measure relative J-factors. (We note, however, that as is clear from Fig. D.5(A), some concentrations result in smaller measurement errors, and would therefore be better choices for measuring relative J-factors.) Appendix D.2.7 discusses how Eq. (2.29) was used in this work to calculate the J-factors presented in Fig. 4.2(C) and (F).

2.8 Conclusion

Many measurements of key biological parameters are reaching the point where they can be carried out reproducibly and with high precision. As a result, it has become possible to expect (and even demand) an interplay between theory and experiment where specific theoretical models can be used as a conceptual foundation for various experimental techniques. Already in the field of single-molecule biophysics, it has become routine to use the well-understood mechanical properties

of DNA as a way to calibrate instruments such as optical and magnetic tweezers. Here we have adopted a similar strategy in which it is shown that a statistical mechanical model of transcription factor binding to DNA can be used as an intellectual filter for both interpreting tethered particle experiments and more importantly, for extracting key parameters of biological interest from these experiments.

In the context of the tethered particle motion assays one of the most useful tools for accessing biologically interesting parameters is the concentration titration curve that explores the relevant protein-DNA binding problem as a function of the transcription factor concentration. The key question addressed here has been an analysis of how these concentration titration curves are altered by various parameters such as the binding strengths of the transcription factor binding sites and the length and flexibility of the DNA substrate, and by unwanted side effects such as an inactive fraction of protein. The centerpiece of the analysis is provided in Fig. 2.1, which shows how both intentional parameter variation and unintended artifacts can result in altering the useful titration curves. Our analysis yields a simple conceptual picture of how both the peak positions and the peak amplitude depend upon the DNA-protein binding constants and on the looping J-factor.

Another important result is that the shape of the titration curve predicted by the basic theory, Eq. (2.1), is very robust. Several of the perturbations we study can be described within the simple theory, as modifications to the parameters rather than as changes to the form of the expression. This shows the robustness of our theory for data analysis, and also clarifies its limitations: some effects simply cannot be detected from “within” a titration curve. Understanding which effects are of this kind is critical, as indications of where further developments are needed, and as an integral part of careful and critical data analysis. In the next chapter we will examine several of these effects and their relevance for the measurements of J-factors and dissociation constants in Chapter 4.

Chapter 3

Precision single-molecule measurements of dissociation constants and J-factors

As introduced in Chapter 1, we have developed a new way of measuring both operator dissociation constants and the relative flexibilities of different DNA sequences as contained in looping J-factors, by tuning the various parameters schematized in Fig. 1.4, with a rigorous comparison between TPM experiments and the theoretical models presented in the previous chapter. Because this is the first use of the TPM assay to make such systematic and quantitative measurements, in this chapter we present a suite of experimental and computational controls that demonstrate the validity of our theoretical framework and that test the impact of potential experimental sources of error on the measurements we report in the next chapter.

The first section describes the relationship of this work to earlier efforts in [115, 120]. Although the theoretical and experimental approaches that we test here were described in this earlier work, which was crucial to laying the foundations for the work reported here, we consider the results presented here to be the first successful test of their applicability to DNA looping experiments and their robustness under numerous experimental variations, and discuss below several improvements to the method that made a rigorous comparison between theory and experiment possible. The second section details experimental controls that support the use of the simple model of Eq. (2.1), and not any of the modifications described in the previous chapter (except the extension to two looped states), in the following chapters.

As described in Chapters 2 and 4, a key tool for making many of the measurements in this work is the concentration titration: by tuning the repressor concentration and measuring the looping probability, we can fit for other parameters that affect the looping probability, namely the operator dissociation constants and, more importantly, the looping J-factors for different DNA sequences and lengths. These concentration titrations will therefore be the lens through which all of the controls presented here are viewed. All of the test titrations in this chapter will make use of one of the DNAs introduced in Chapter 1 and discussed in more detail in the following chapter, Oid-E894-O1, as a case study, though in some instances additional constructs from the next chapter will be referenced. The particular identities of these DNAs are not relevant for the results here and so a more detailed description of these constructs and a discussion of the significance of the results we obtained with them will be deferred to the next chapter.

3.1 Improvements over previous work

The three most significant improvements over the work in [115, 120] that allow us to make the precision measurements discussed in the next chapter are: (1) the use of better quality protein (Section 3.1.1); (2) evidence that one of the main constructs used in [115, 120] is most likely faulty in some way (Section 3.1.2); and (3) the use of global fits to multiple data sets in order to arrive at best-fit dissociation constant parameters (Section 4.1 and Appendix D.2.6). We discuss these first two improvements here, and the last in the next chapter.

3.1.1 Accurate measurements of dissociation constants and J-factors requires protein purified in-house

As in [115, 120], we initially collected data with purified repressor that was kindly shipped to us from the Kathy Matthews lab at Rice University. However, the fitted K_d 's with this protein were not consistent with literature values (see Table 1 of [115] and Table 3.1 below); moreover, we could not obtain consistent results with repressor shipped at different times from different purifications in the

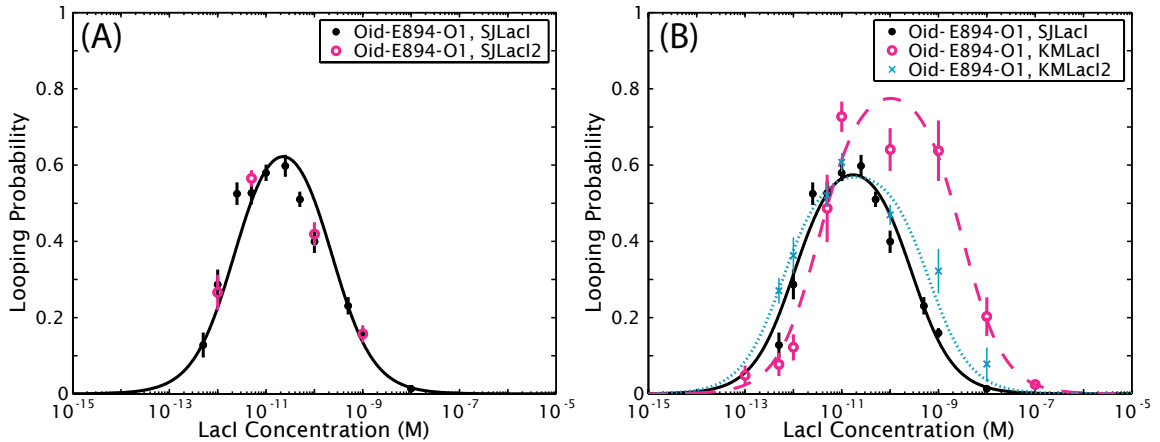


Figure 3.1: Reproducible looping probabilities obtained only with protein purified in-house. **(A)** Concentration titrations with repressor from two different in-house purifications, “SJLacI” (black) and “SJLacI2” (magenta). The black data are Oid-E894-O1 shown in Fig. 4.1(D–E) in Chapter 4; the black curve is the same solid black (global) fit shown in Fig. 4.1(E). Protein from a second purification in magenta gives the same results as the black data. All data in the rest of this work were obtained with SJLacI. **(B)** Concentration titrations with protein from two repressor purifications shipped at different times from the Matthews lab (“KMLacI” and “KMLacI2”). Nonloopers have been subtracted from these data as described in Appendix D.2.5 (as noted in the text there, we find roughly the same proportion of tethers are “nonloopers” across all protein batches), and all data points include at least 20 beads, as described in Appendix D.2.4. Note the significantly larger error bars with the shipped protein despite comparable amounts of data. Data with SJLacI are shown in black for comparison. Unlike in (A), however, the black curve here is the result of the *individual* fit to the SJLacI data alone (see Appendix D.2.6), since that should be a more equivalent comparison for the individual fits shown to the other repressor batches. The fit parameters for KMLacI and KMLacI2 are shown in Table 3.1 (fitting procedures are described in Appendix D.2.6). We were unable to obtain the same results with protein shipped to us as with protein purified in-house.

Matthews lab. Upon a suggestion from Kathy Matthews that shipping the protein on dry ice may damage the protein, we purified two batches of protein in our lab, according to the Matthews lab protocol and after extensive help from their lab (see Appendix C). As shown in Fig. 3.1(A), we were able to obtain consistent results and reasonable parameter values with protein purified in-house, but not with the shipped protein (Fig. 3.1(B)). Except where noted in Fig. 3.1(A), all data presented in the rest of this work were obtained with the “SJLacI” batch.

We also tested the stability of the repressor protein over the course of a day of TPM experiments. We routinely take at least 3 hours of data on a single chamber (in which repressor is at room temperature), and 7 to 9 hours of total data in a day (that is, 2–3 chambers), diluting fresh protein for each new chamber from a stock that remains on ice throughout the course of the experiments. We computed the mean looping probability, with nonloopers subtracted, for a particular data set (Oid-E894-O1, 50 pM LacI) for which we had at least 15–20 beads from each of 2 chambers that represent a day’s worth of data. We find the mean looping probability after 1.5 hours of data taken at

Data, Repressor	K_{id}	K_1	J_{loop}
Oid-E894-O1 KMLacI*	20 (10, 130)	500 (\pm 300)	5000 (\pm 1000)
Oid-E894-O1, KMLacI2	1.5 (\pm 0.7)	300 (\pm 200)	800 (\pm 400)
Oid-E894-O1, SJJacI	3 (\pm 1)	90 (\pm 20)	350 (\pm 40)
Literature values	8.3 \pm 1.7 [125]	37 \pm 5 [137, 138, 139]	-

Table 3.1: Fit parameters, in pM, for the fits to the “KMLacI” and “KMLacI2” repressor batches shown in Fig. 3.1(B), and the “SJJacI” batch purified in-house shown in Fig. 3.1(A). The “SJJacI” fit parameters are the same as in Table 4.1 in Chapter 4. As in that table, 95% confidence intervals are reported where standard deviations of fit parameters would generate negative values. The asterisk indicates that the distributions of fit parameters obtained from bootstrapped data were multimodal. As shown graphically in Fig. 3.1(B), the fit parameters for KMLacI are not within error of the individual fit parameters for “SJJacI”, nor are they within error of the global fit parameters shown in Table 4.1 in Chapter 4, or of values found in the literature using bulk biochemical methods. The fit parameters for KMLacI2 are barely within error of the individual fit parameters for SJJacI, but we still consider this protein batch to be suspect. Mis-measurement of the dissociation constants leads to significant mis-measurement of the J-factor, the parameter which we will attempt to measure as accurately as possible in the next chapter. As will be shown in that chapter, the use of the “SJJacI” batch plus global fits to multiple data sets does result in dissociation constants within error of those found in the literature, and therefore we are confident of the J-factors measured with this protein purified in-house. Fitting procedures are described in Appendix D.2.6.

room temperature, with freshly diluted protein, to be 0.49 ± 0.05 , and after 3 hours to be 0.53 ± 0.14 ; and the mean looping probability after 1.5 hours at room temperature with protein that had been on ice for 3 hours to be 0.53 ± 0.05 , and after 3 hours at room temperature (and 3 hours on ice) to be 0.53 ± 0.08 . The looping probability computed over the entire day was 0.51 ± 0.3 . We therefore detect no loss of protein activity over the course of a day of experiments.

3.1.2 The PUC306 construct exhibits anomalous behavior even in the presence of protein purified in-house

The work in [115] relies heavily on looping data with a 306 bp loop derived from the pUC19 plasmid and flanked by the Oid and O1 operators, in a 901 bp tether. This construct is especially interesting because its two looped states are clearly separated and both well populated at certain concentrations. It was used in proof-of-principle experiments in [115, 120] with what we now consider to be “bad” protein (see the previous section). In Fig. 3.2 we show the looping probabilities for this construct at several repressor concentrations, using repressor purified in-house, and with more data per concentration and at more concentrations than in [115].

As can be seen in that figure and in the fit parameters listed in Table 3.2, despite this additional data and fresh protein, the PUC306 construct is still not well fit by operator dissociation constants that agree with those measured for the other constructs discussed in the next chapter, or with

Data	K_{id}	K_1	$J_{PUC, M}$	$J_{PUC, B}$	$J_{PUC, all}$	K_{DT}
PUC, all*	8 (± 1)	8 (± 1)	-	-	390 (± 80)	-
PUC, B&M global	8 (± 1)	8 (± 1)	240 (± 20)	190 (± 10)	-	-
Global Fit (with E8, TA)	5 (± 2)	36 (± 3)	380 (± 40)	300 (± 40)	-	-
Global Fit, dimers	1 (± 1)	24 (± 4)	360 (± 30)	290 (± 30)	-	7 (1, 30)
Oid-E894-O1 alone	3 (± 1)	90 (± 20)	-	-	-	-
Global Fit, E8 & TA	12 (± 3)	44 (± 3)	-	-	-	-
Literature values	8.3 \pm 1.7 [125]	37 \pm 5 [137, 138, 139]	-	-	-	-

Table 3.2: Fit parameters for the PUC306 construct, in pM. The top row is an individual fit to the total looping probabilities for PUC alone; the second row, a global fit to the bottom and middle looped states simultaneously, according to the model in Section 2.2 (see also Fig. 4.1(F)); the third row is a global fit to the bottom and middle PUC states, plus all three E8 data sets and the TA data set of Fig. 4.1(D–E) in the next chapter. The row labeled “Global Fit, dimers” is the same global fit but where we take into account the dimer-to-tetramer transition at low repressor concentrations, as discussed in Section 2.4. The last two rows are taken from Table 4.1 in the next chapter and are shown for comparison: the second-to-last row is the parameters for the individual fit to the “SJLacI” batch discussed in the previous section, and the last row is representative of what we consider to be the best parameter values for our TPM assays.

literature values based on bulk biochemical assays. Because PUC306 exhibits additional anomalous behavior (for example, the length of the unlooped state as a function of repressor concentration does not decrease monotonically, as with the other DNAs in Fig. 4.6 in the Appendices of the next chapter, but instead decreases and then increases again at high repressor concentration), we argue that the PUC construct contains some aberrant feature that merits its exclusion from further study (for example, perhaps an unidentified pseudo-operator). We note that the disagreement between this 901 bp construct and the 450 bp E8- and TA-containing constructs used elsewhere in this work is probably not due to the difference in total tether length: Chapter 6 discusses two-operator constructs that have total tether lengths of 735 bp, and those constructs are well fit by the K_d ’s derived from the 450 bp constructs. (With only three data points per curve, fitting those 735 bp constructs for the K_d ’s, without reference to the E8- and TA-containing constructs of Chapter 4, does not result in well-constrained parameters. However the K_d ’s from such a fit, $K_1 = 10$ pM and $K_2 = 500$ pM, are not consistent with the trend that PUC shows, which is that longer constructs have K_d ’s that trend to lower values and/or more similar values for two different operators.)

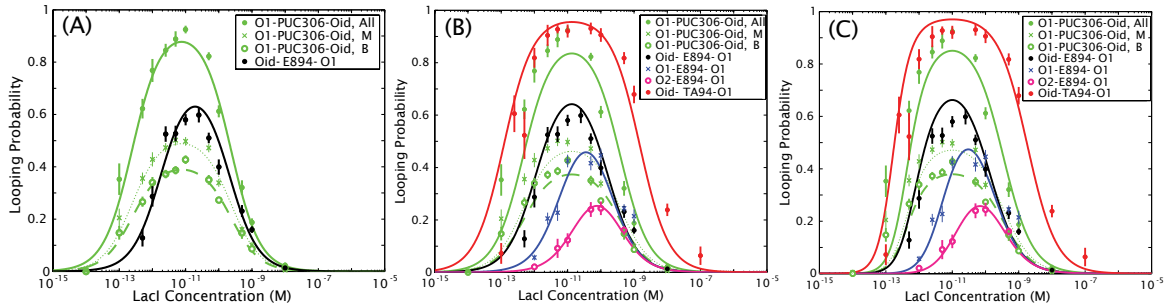


Figure 3.2: Looping probability as a function of concentration for a 901 bp tether with a 306 bp loop, O1-PUC306-Oid (green data), compared to Oid-E894-O1 (black data). **(A)** As with some lengths of the E8 and TA sequences (discussed in the next chapter and in Section 2.2), the 306 bp PUC sequence results in two looped states with distinct average RMS values, the shorter labeled “B” for “bottom” and the longer of the two “M” for “middle”. The green dashed and dotted lines are the results of a global fit to the data for the bottom and middle looped states, where it is assumed the two states have the same operator dissociation constants but different looping J-factors, an assumption that is implicit in the model of Section 2.2 in the previous chapter, and validated in Fig. 4.1(F) of the next chapter. The solid green line, which describes the total looping probability, is then given by the same dissociation constants, and a J-factor that is the sum of the J-factors for the individual loops. The Oid-E894-O1 data and corresponding fit (solid black line) are the same as in Fig. 4.1(E) in the next chapter and represents the best fit parameters we obtained with TPM. The PUC306 construct is supposed to contain the same operators as the E8 construct shown here, such that as in Fig. 2.1(B) the green and black curves should have maximal looping probability at the same concentrations. As can be seen here, however, the maximum of looping for the PUC construct occurs at a lower concentration than for E8, indicating that the two constructs cannot be fit with the same dissociation constants. **(B)** The results of a global fit to the PUC data and the four data sets of Fig. 4.1(D–E). The E8 and TA data sets (described in the next chapter) are all well described by the same parameter set; the PUC construct, however, is not, and inclusion of the PUC data set in the global fits (with the assumption that it contains the O_{id} and O_1 operators) decreases the fidelity of the fits to all of the data sets. **(C)** A global fit that takes into account dimers as in Section 2.4, in the event that the behavior of the PUC construct (and perhaps the TA construct as well—see discussion in the next section) can be explained by the dissociation of repressor tetramers into dimers at low concentrations. As in (B), inclusion of the PUC construct in the global fit, even with dimers allowed, decreases the fidelity of the fit for all constructs. Fit parameters are given in Table 3.2.

3.2 Computational controls: Dimers at low concentration, the active fraction of repressor, and low repressor concentrations

In this section we discuss several of the modifications to the simple model of Eq. (2.1) described in the previous chapter and whether they are necessary for the analysis of the data described in the next chapter. Specifically we address whether it is necessary to take into account the dimer-to-tetramer transition at low repressor concentrations (Section 2.4), a discrepancy between the assumed repressor concentration and the actual concentration (Section 2.3), or an excess of tethers relative to repressors at low repressor concentrations (Section 2.6). Not only do we show here that the simple model of Eq. (2.1) is sufficient to describe our experimental results, by examining the effects of these

K_{DT} or f	$J_{\text{loop, ES}}$	K_{id}	K_1
1 fM	300 pM	9 pM	41 pM
5 fM	300 pM	10 pM	41 pM
10 fM	300 pM	10 pM	40 pM
50 fM	300 pM	11 pM	38 pM
100 fM	300 pM	12 pM	37 pM
500 fM	290 pM	18 pM	27 pM
1 pM	290 pM	24 pM	24 pM
5 pM	300 pM	29 pM	29 pM
10 pM	300 pM	32 pM	32 pM
97%	310 pM	9 pM	43 pM
95%	321 pM	9 pM	44 pM
93%	330 pM	10 pM	45 pM
90%	340 pM	10 pM	47 pM
80%	380 pM	11 pM	50 pM
70%	430 pM	13 pM	60 pM
(experimental)	330 (\pm 30) pM	12 (\pm 3) pM	44 (\pm 3) pM

Table 3.3: Fit parameters of the simple model (Eq. (2.1)) to data generated by a model that takes into account the dimer-to-tetramer transition at low concentrations, or a potential inaccuracy in repressor concentration. The top section gives the fit parameters for data generated by Eq. (2.18), with varying values of K_{DT} ; the middle section, for data generated by Eq. (2.12), with varying values of f ; and the last section gives the best-fit parameters to our real data (row six of Table 4.1 in the next chapter), which were used as the inputs to generate the simulated data that was fit to the simple model here. When K_{DT} exceeds 50 to 100 fM, or f is smaller than at least 90%, the fit parameters to the generated data cease to be within error of the fit parameters to real data.

potential sources of error on the values we measure, we can ask if it is possible from our data to determine upper bounds on K_{DT} , the dissociation constant for the tetramer-to-dimer transition, or f , the fraction of repressor that contributes to looping. We note, however, that as derived in Section 2, the dimer-to-tetramer transition and limiting repressors at low concentrations will affect only our measured K_d 's, and not the crucial parameter discussed in the next chapter, the J-factor, since low-concentration data are more important for measuring K_d 's than J_{loop} .

3.2.1 The dimer-to-tetramer transition, and the active fraction of repressor

We chose a numerical approach to address the questions of the dimer-to-tetramer transition at low concentrations, and a potential concentration uncertainty. We first chose a range of reasonable values for K_{DT} and for f , and then inserted these values along with the best-fit values for K_d 's and one of the J-factors measured in the next chapter (row six of Table 4.1) into the modified models of Eq. (2.12) (which includes an inactive fraction of repressor) and Eq. (2.18) (which includes a

dimer-to-tetramer equilibrium at low concentrations) to generate looping probabilities at a range of concentrations comparable to those used in our TPM assays. We then fit the original model of Eq. (2.1) to these simulated data, and asked how closely the fitted parameter values matched to the “true” values that were the inputs to the simulated data. The results are shown in Table 3.3.

In the first case, that of considering the tetramer-to-dimer transition at low repressor concentrations, we find that the fit parameters to the data generated from the model with K_{DT} (Eq. (2.18)), but fit to the model without K_{DT} (Eq. (2.1)), were within error of the fit parameters for real data until K_{DT} exceeded 50 to 100 fM. Therefore given the uncertainty in our experimental data, we can put an upper bound on K_{DT} in the tens of femtomolar range. This is in good agreement with recent estimates of K_{DT} from other techniques (see [131]), which put an upper bound on K_{DT} in the femtomolar range. Again we note that, as concluded in Section 2.4 above, if the true value of K_{DT} is above 50–100 fM but we do not take it into account in our fits, we would obtain systematic errors in the fitted values for the dissociation constants but not the J-factor. Therefore regardless of the actual value of K_{DT} , our measurement of the J-factor remains the same.

On the other hand, in the second case, that of an inaccuracy in the assumed repressor concentration, we find, as predicted in Section 2.3 above, that both the dissociation constants and the J-factor are affected. However, we find that the assumed value of the repressor concentration could vary by up to about 10% and we would obtain fit parameters within error of those we now have (or within error of current literature values for the dissociation constants); and that within this range our measurement of the J-factor would not change within experimental error.

Since the model that includes f is not an independent model but involves a rescaling of the parameters of the original model in Eq. (2.1), it cannot be used to fit a concentration curve unless one of the parameters (f , K_i , K_{ii} , or J_{loop}) is known from another source. In principle, however, we should be able to fit both the model with K_{DT} and the model without K_{DT} to any concentration curve. A fit of the model that includes K_{DT} to the three E8 data sets of Fig. 4.1(D) in the next chapter, plus the TA data set of Fig. 4.1(E), yields $K_{id} = 8 \pm 2$ pM, $K_1 = 37 \pm 6$ pM, $K_2 = 210 \pm 30$ pM, $J_{\text{loop, E8}} = 300 \pm 20$ pM, $J_{\text{loop, TA}} = 4600 \pm 500$ pM, and $K_{DT} = 0.8$ (0.2, 20) pM. Here we

report a 95% confidence interval for the error on K_{DT} because the fit was not well constrained when K_{DT} was included. As noted above, the J-factors do not change appreciably compared to the fit that does not include the dimer-to-tetramer transition; and in fact the fitted K_{DT} is low enough that the dissociation values do not change significantly either. We therefore conclude, as above, that K_{DT} is no larger than 100's of femtomolar and is low enough to be irrelevant to our experiments.

3.2.2 Data analysis at low repressor concentrations

As noted in Section 2.6, the statistical mechanical model to which we fit concentration curves depends on the assumption that tethers are independent, that is, that the binding or unbinding of a repressor from one tether does not affect the binding or unbinding of repressors on other tethers. This assumption rests in turn on the assumption that repressors are always in excess of the number of tethers, so that the removal of one repressor from the solution when it binds to an operator does not change the effective concentration of repressors that the other tethers “see”. This assumption is valid at most concentrations that we use; however, it could be called into question at very low repressor concentrations. To estimate when the assumption of an excess of repressors over tethers fails, in this section we estimate the number of tethers per chamber and then compare to the numbers of repressors per chamber as a function of concentration.

Our hand-made TPM chambers have volumes of about 40 μL , which means that at the lowest repressor concentrations we use, there are on the order of 240,000 (at 10 fM) to 24 million (at 1 pM) repressors per chamber. To estimate a typical number of tethers per chamber, we note that we usually see fewer than 50 tethers in a field of view, each of which is about $3 \times 10^9 \text{ nm}^2$ in area, corresponding to roughly 0.3 nL in volume, given the double-sided tape's thickness of 100 μm . This means that in a 40 μL chamber, there will be on the order of 7 million tethers. Even if the estimate of tether density is an overestimate (given that 50 tethers per field of view is very high), 1 pM repressor still seems to be the lower bound on concentrations we can use with our model for some choices of operators and J-factors, below which the assumption of an excess of repressors over tethers breaks down. In particular, according to Eq. (2.27) and the parameter values given in Table 4.1, the

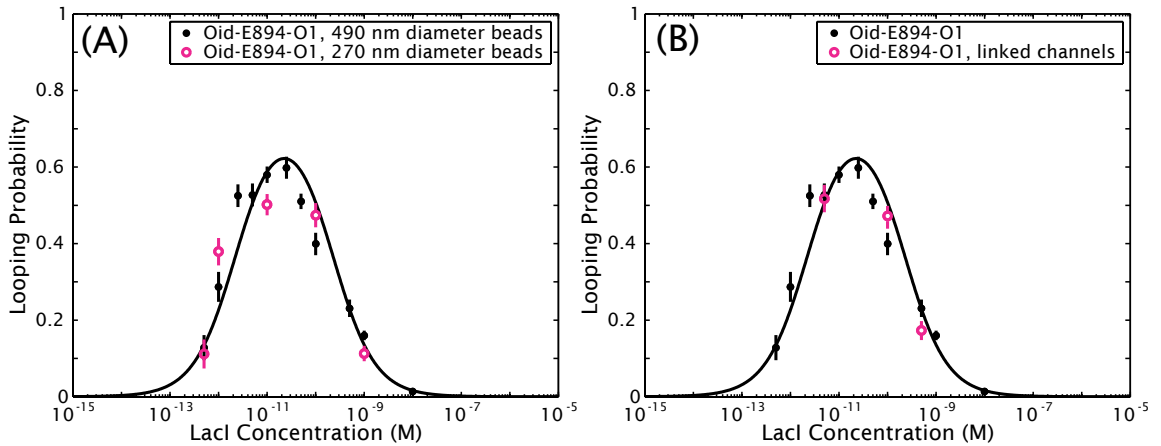


Figure 3.3: Effect of bead size and chamber surfaces on the measured looping probability. Halving the diameter of the reporter bead (A) or linking two chambers to double chamber surface area (B) does not affect the looping probability of Oid-E894-O1 in a way consistent with theoretical predictions of bead-size effects or loss of protein to channel walls (see text for details). In both figures, the black data are the Oid-E894-O1 data shown in Fig. 4.1(D) in the next chapter, and the black curve is the solid black (global) fit of Fig. 4.1(E).

Oid-TA94-O1 construct will be most sensitive to low-repressor-concentration effects. This perhaps contributes to the large spread in looping probabilities at 250 fM and 500 fM for Oid-TA94-O1, as evidenced by the much larger error bars on these data points than others.

3.3 Experimental controls: Different bead sizes and nonspecific adsorption to chamber walls

3.3.1 Smaller beads result in similar looping probabilities

A common concern with single-molecule experiments such as TPM that use large particles as reporters of molecular dynamics is that the reporters affect the observed dynamics. In particular in our TPM experiments, the 490 nm diameter bead is attached to a surface by a roughly 450 bp, or 150 nm, DNA tether, so it is reasonable to ask what the impact of excluded volume effects from the bead may be on the observed looping. Segall and coworkers [140] explored these effects from a theoretical standpoint and found in the regime applicable to our experiments that halving the bead diameter would halve the force experienced by the tethered bead. We expect this force to affect

primarily the measured J-factor: a smaller force would allow more looping and thereby increase the measured J-factor.

We measured the looping probability of the Oid-E894-O1 construct with 270 nm diameter beads in addition to the 490 nm diameter beads used in the rest of this work, and found that the measured looping probability was not within error of the 490 nm points at three out of five measured concentrations (Fig. 3.3(A)). However the trend in these discrepant points are not consistent with an increased effective J-factor (see Fig. 2.1(B)). The 270 nm beads are difficult to image and track with the brightfield microscopy employed in this work, and we attribute discrepancies between the looping probabilities with 270 nm beads compared to 490 nm beads to tracking inaccuracies with the smaller beads, and not to an effect of bead size on the measured looping probability. (See also recent work by Milstein and coworkers [116], who found that looping and unlooping rates varied by only a factor of 2 between 800 nm and 50 nm reporter beads.)

3.3.2 No detectable loss of protein to chamber walls

Our ability to use our simple model (Eq. (2.1)) to obtain K_d 's and J-factors from concentration titrations depends on our knowing the absolute concentration of repressor in the TPM chamber. As described in Section 2.3 in the previous chapter and Section 3.2 in this chapter, if the actual concentration of repressor in our sample is less than we assume, the parameters we measure will be scaled by some constant relative to their true values. In particular, we will measure effective K_d 's that are weaker than they should be, and J-factors that appear larger. We therefore asked if loss of protein to the walls of the TPM chamber could be affecting our measurement, as one potential source of a discrepancy between the assumed and actual concentration of repressor in our experiments. To do so, we made a chamber as described in Appendix D.1, except no DNA was added to the 250 μ L 3P introduced into the chamber after the anti-digoxigenin was washed out, and no beads were added. We then attached the output of this empty chamber to the input of a chamber prepared as usual. Repressor was introduced into the chamber with DNA via the empty chamber, and data were taken on the DNA-containing chamber.

The results of this “linked-channel” experiment, in which the surface area to which repressor could adsorb is effectively doubled, are shown in Fig. 3.3(B). If protein is adsorbing to channel walls, when the surface area is increased, data at lower concentrations than the maximum of looping (e.g., the 5 pM data in Fig. 3.3(B)) should have a lower looping probability than in the normal single-channel experiment, and data at higher concentrations than the maximum (e.g., 100 pM and 500 pM) should have higher looping probabilities. This is not what we observe, and so we cannot conclude that increasing the surface area of the sample leads to a detectable change in looping probability. It should be noted, however, that due to the error on each measured looping probability, the effective repressor concentration in the linked chamber experiment would have to be significantly reduced, compared to the usual single-channel concentration, to be detectable (see Fig. 2.1(E) and Section 3.2 above).

3.4 Conclusion

In this chapter we have shown the robustness of our combined theory plus TPM assay approach when confronted with potential experimental complications such as different repressor purifications (as long as the purification is done in-house), low repressor concentrations, discrepancies between the assumed and actual repressor concentration, and the size of the reporter bead. In all cases we found the simple model of Eq. (2.1) to be sufficient to describe our TPM data, and so that model will be the main workhorse of the following chapters. In addition we were able to validate the combined theory plus TPM experiment approach on which all of the work presented here is based, beyond what was possible in earlier work [115, 120], setting the stage for the application of this approach to the question of sequence flexibility and looping in the next chapter, and of multiple operator systems in Chapter 6.

Chapter 4

The sequence dependence of transcription factor-mediated DNA looping

With the theoretical framework of Chapter 2 and the validation of our approach in Chapter 3 in hand, we now turn to the central question of this work, that of the role of sequence flexibility in DNA loop formation by a prokaryotic transcription factor. The results presented here rely heavily on the systematic tuning of the four biologically relevant parameters introduced in Fig. 1.4: repressor concentration, operator binding strength, loop length, and loop sequence.

In the first two sections, where the roles of the first three of these parameters are examined, we make use of the concentration titrations that figure prominently in the previous two chapters to report new, single-molecule measurements of the dissociation constants for three of the known binding sites for the Lac repressor, and the J-factors of 94 bp loops that contain either the random E8 sequence or the putatively more flexible, strong nucleosome positioning TA sequence (see Chapter 1). We make explicit comparisons between theory and experiment that go beyond those already made in the previous chapter, confirming that we are able to use TPM, in conjunction with our statistical mechanical model, to obtain dissociation constants that are comparable to those measured in bulk biochemical assays. Moreover, this explicit comparison to theory allows the extraction of the looping J-factors for 94 bp loops with the E8 and TA sequences, which becomes especially important when comparing to *in vivo* data as in Section 4.6. We then turn to the fourth tunable parameter introduced in Fig. 1.4, that of loop length, and find that the looping probability depends on sequence and length

in a more complicated way than would be assumed from the concentration titrations of Sections 4.1 and 4.2, or indeed from the cyclization and nucleosome positioning studies that inspired this work. Tuning the length of the loop also yields insights into the two looped conformations that we observe, and allows us to comment on the difference between short transcription factor-mediated loops versus ligated DNA minicircles. Finally, we comment briefly on parallel *in vivo* experiments with the E8 and TA sequences, and the extension of the results presented here to additional sequences.

4.1 Effect of repressor concentration and operator strength on the looping probability

We first explore from an experimental perspective how the Lac repressor concentration and its affinity for several known binding sites alter the looping probability, and compare these experimental results to the theoretical predictions of Chapter 2, finding good agreement between theory and experiment. We also discuss in more detail than in the previous chapters how tuning the repressor concentration and the operator strength may be used to extract the looping J-factor of the DNA, as well as the repressor-operator dissociation constants. In fact we find that the most accurate and logically consistent way of measuring both the J-factors and operator dissociation constants involves not only a fit of our model to a particular concentration curve, but instead a *global* fit of our model to multiple data sets with different combinations of operators simultaneously (see Appendix D.2.6 for procedural details). As noted in the previous chapter, we consider the results presented here to be the first rigorous and successful validation of our combined TPM plus statistical mechanical model, a success which depends in part upon this operator tuning and global fitting procedure.

As described in more detail in Chapters 2 and 3, we can use the tools of statistical mechanics to relate J-factors, operator dissociation constants, and transcription factor concentrations to the experimentally observable looping probability through the expression

$$p_{\text{loop}}([R]) = \frac{\frac{1}{2} \frac{[R]J_{\text{loop}}}{K_i K_{ii}}}{1 + \frac{[R]}{K_i} + \frac{[R]}{K_{ii}} + \frac{[R]^2}{K_i K_{ii}} + \frac{1}{2} \frac{[R]J_{\text{loop}}}{K_i K_{ii}}}, \quad (4.1)$$

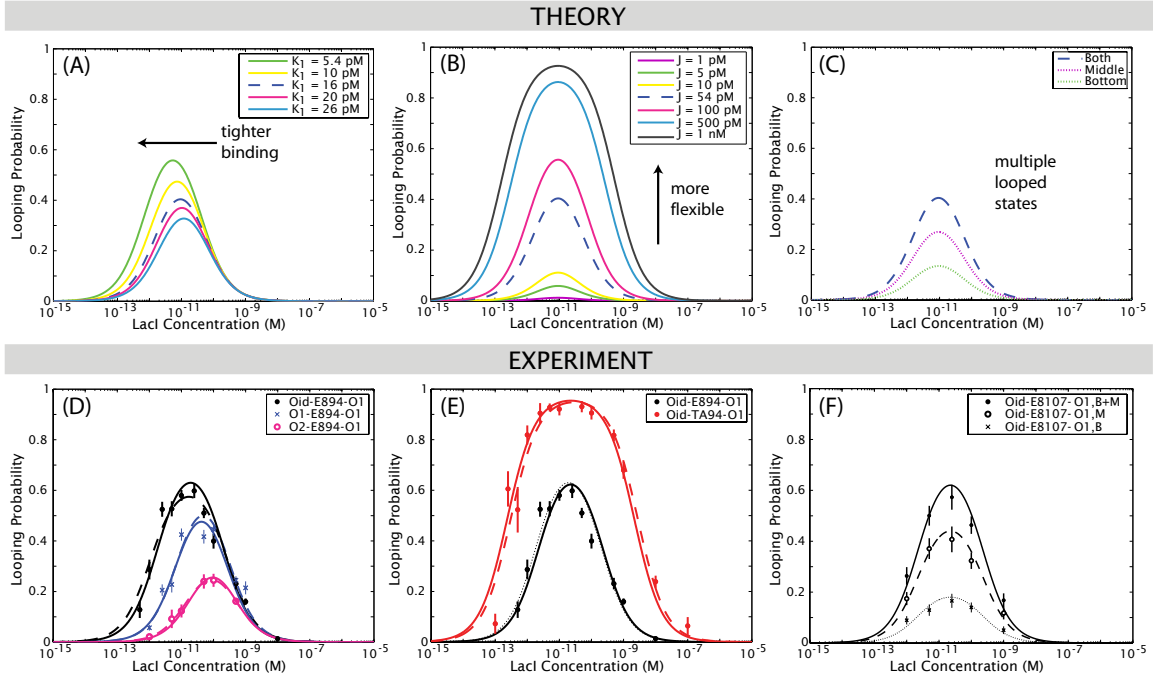


Figure 4.1: Theoretical (A–C) and experimental (D–F) results for the looping probability as a function of operator strength, loop sequence, and repressor concentration. In the theoretical predictions of (A–C) (which are the same as panels (A), (B), and (D) of Fig. 2.1), $K_{id} = 5.4$ pM, $K_1 = 16$ pM, and $J_{loop} = 54$ pM unless otherwise indicated; curves with these default parameters are shown as dashed blue lines for comparison across panels. In the experimental results of (D–F), errors are standard errors on the mean, and in these panels, unlike in (A–C), curves are fits to Eq. (4.1), not predictions. **(A)** Theoretical prediction for the effect of changing the strength of one of the operators on the looping probability as a function of repressor concentration. **(B)** Theoretical prediction for the effect of changing the flexibility of the DNA in the loop. **(C)** Extension of the simple model to the case of two experimentally distinguishable looped states (see Sections 2.2 and 4.3), which we model as having different J-factors. The two looped states are labeled “middle” (“M”) and “bottom” (“B”) in reference to their relative tether lengths: the RMS of the middle state is between that of the unlooped and the bottom state, and the RMS of the bottom state is such that it is the shortest observed (non-sticking) state. Here the J-factor of the bottom state, J_B , is one-third that of the default 54 pM, and J_M is two-thirds that of the default. The dashed blue line shows the sum of the probabilities of the two states, which we refer to as the total looping probability. **(D)** Measured looping probabilities for 94 bp of the random E8 sequence, flanked by three different combinations of operators. Dashed lines indicate individual fits to each data set as described in Appendix D.2.6; solid lines indicate a global fit to all three data sets simultaneously. The global fit, which enforces identical values of the J-factor and O_1 dissociation constant in all three data sets, describes the data as well as the individual fits, demonstrating the consistency of the model when the operators are changed. **(E)** Looping probabilities for the E8 (black) and TA (red) sequences as function of concentration. The Oid-E894-O1 data are the same as in (D); the dotted black line is the result of the global fit shown in that panel as well. The dashed red line represents an individual fit to the Oid-TA94-O1 data; the solid red and black lines are from a global fit to all three E8 data sets in (D) plus this TA data. (The results of this global fit that includes the TA data for the O1-E894-O1 and O2-E894-O1 data sets are shown in Appendix D.2.6.) The TA data can be fit with the same K_d values as the E8 data, but have a significantly larger J-factor, or a more flexible sequence. Fit parameters for (D) and (E) are listed in Table 4.1. **(F)** Looping probabilities for a DNA with two looped states, Oid-E8107-O1. As in (C), “B” refers to the looped state with the shorter tether length, and “M” to the looped state with the longer tether length. Data marked “B+M” are total looping probabilities, that is, the sum of the probabilities of the bottom and middle states. Curves represent a simultaneous fit of the “B” and “M” data to Eqs. (2.9) and (2.10), using the values of K_{id} and K_1 from the global fit to all three E8 data sets in (D) and the TA data in (E). The procedure for determining the errors on the fit follows the bootstrapping scheme used throughout this work and is described in Appendix D.2.6. We find that the two looped states differ only in J-factor, as we and others [128, 129] assume in our models; that is, that the binding affinity of the repressor for operator DNA does not change with the different loop and/or repressor conformations that generate the two observed loop states. For this 107 bp loop, the “bottom” state has a J-factor of 100 ± 40 pM, and the “middle” has a J-factor of 330 ± 40 pM. Note that the total J-factor of 280 pM obtained from this concentration curve is within error of the J-factor of 280 ± 40 pM determined from only the 100 pM data point shown in Fig. 4.2(C). Likewise the J-factors for the two looped states are within error of those determined from the 100 pM data alone ($J_B = 80 \pm 20$ pM, $J_M = 190 \pm 40$ pM), using the method of relative J-factors described in Section 2.7 and plotted for the two looped states in Fig. D.5(B) in Appendix D.2.7.

which is Eq. (2.1) of Chapter 2, reprinted here for convenience. Although Chapter 2 discusses a number of potential modifications to this model, we have found (here and in the previous chapter) this expression for the looping probability to be sufficient to describe all of the experimental results presented here.

As in Chapters 2 and 3, the main workhorse of our approach to testing this statistical mechanical description of the looping probability is the repressor concentration curve, where we measure the looping probability at different repressor concentrations, and then fit Eq. (4.1) to obtain dissociation constants and J-factors. As derived in more detail in Chapter 2, Eq. (4.1) makes very specific and falsifiable predictions for how these repressor concentration curves should change as the model parameters change (Fig. 2.1(A–C)). Figure 4.1 shows a subset of these previously untested predictions, as well as the comparison of these predictions to experiment, which will be examined in more detail below.

Figure 4.1(A) shows the prediction of our model for how the concentration curves should change as the dissociation constant for one of the operators is varied: changing the strength of one of the operators should change both the concentration at which looping is maximal, and the amount of looping at that maximum, but the curves should overlap at high repressor concentrations. These observations can be formalized by appealing to Eq. (4.1), as is done in more detail in Section 2.1. To summarize the results of that section, the concentration at the maximum in the looping probability

Data	K_{id}	K_1	K_2	$J_{\text{loop, E8}}$	$J_{\text{loop, TA}}$
Oid-E894-O1	3 (\pm 1)	90 (\pm 20)	–	350 (\pm 40)	–
O1-E894-O1	–	47 (\pm 4)	–	380 (\pm 30)	–
O2-E894-O1	–	26 (11, 125)	300 (\pm 200)	320 (\pm 90)	–
Oid-TA94-O1	10 (5, 46)	80 (\pm 40)	–	–	5500 (\pm 600)
Global Fit, E8	9 (\pm 1)	42 (\pm 3)	210 (\pm 40)	300 (\pm 20)	–
Global Fit, E8 & TA	12 (\pm 3)	44 (\pm 3)	240 (\pm 50)	330 (\pm 30)	4200 (\pm 600)
Literature values	8.3 \pm 1.7 [125]	37 \pm 5 [137, 138, 139]	350 \pm 130 [137]	–	–

Table 4.1: Measured dissociation constants and looping J-factors, in pM, obtained by fitting Eq. (4.1) to the data shown in Figs. 4.1(D) and (E). In most cases the best fit parameter, plus or minus the standard deviation of the distribution of fit parameters from bootstrapped data, is reported; however in cases where the standard deviation includes negative parameter values, a 95% confidence interval is reported in parentheses instead. The first four rows are individual fits to the indicated data sets; the fifth row is a global fit to all three of the E8-containing data sets in Fig. 4.1(D); and the sixth row is a global fit to these three E8 data sets and the TA data set in Fig. 4.1(E). Fitting procedures are discussed in Appendix D.2.6.

can be found by differentiating Eq. (4.1) with respect to $[R]$ and results in

$$[R]_{\max} = \sqrt{K_i K_{ii}}. \quad (4.2)$$

Note that the concentration at which the looping probability is maximized does not depend upon the DNA flexibility as captured in the parameter J_{loop} . The looping probability at this maximum, however, does depend on J_{loop} , according to

$$p_{\text{loop}}([R]_{\max}) = \frac{J_{\text{loop}}/2}{J_{\text{loop}}/2 + (\sqrt{K_i} + \sqrt{K_{ii}})^2}, \quad (4.3)$$

and will therefore be discussed in more detail in the next section where our measurements of the J-factors of two different sequences are directly addressed. Finally, we note that at high concentrations, Eq. (4.1) approaches the limit $J_{\text{loop}}/(2[R])$, which is independent of operator strength, explaining why the curves in Fig. 4.1(A) overlap at high concentrations. As an experimental consequence, data at low concentrations are essential for determining operator strengths, whereas high-concentration data are sufficient for determining J-factors.

Figure 4.1(D) shows experimental results for a loop containing 94 bp of the synthetic random sequence E8 [58, 85], flanked by three different combinations of the operators O_{id} , O_1 , and O_2 , which are known to have distinct affinities for the Lac repressor. (See Appendix B for the sequences of the loop regions and operators used in this chapter.) As predicted by our model, increasing the binding strength of one of the operators (i.e., decreasing the value of one K_d) shifts the maximum of the curve to the left and increases its amplitude: that is, stronger operators allow more looping at lower concentrations. Similarly, since the J-factor is a property of the DNA loop length and sequence, we would expect all three curves to be fit by the same J-factor, and for the fits to reflect the reality that they share O_1 as one of the operators. This is indeed what we find, as shown in the fit parameters listed in Table 4.1: fits to the individual data sets (dashed lines in Fig. 4.1(D)) and a global fit to all three data sets simultaneously (solid lines), where we have enforced the constraint that all three data sets share the same J-factor and dissociation constant of the O_1 operator, are comparable in

their fidelity. We find that the fitted values for the K_d 's agree well with values in the literature obtained through bulk biochemical techniques (see references cited in Table 4.1), as well as for the most part agreeing between individual fits to different data sets; and that the fitted J-factor also agrees well between data sets, with a value of about 300 ± 20 pM. We are therefore confident that this combined concentration titration plus statistical mechanical model approach provides us with reasonable parameter values for both dissociation constants and J-factors, and that the global fit supplies the most reliable parameter estimates.

The looping J-factor for E894 is higher than the corresponding cyclization J-factor of 54 pM reported in earlier work [85], and significantly higher than cyclization J-factors for other sequences of similar lengths [59]. However, since the looped geometry imposes less stringent constraints on the DNA than does cyclization (discussed in more detail below), we would expect the looping J-factor to be larger than the cyclization J-factor.

4.2 Effect of sequence on the looping probability

As discussed in Chapter 1, we turned to the field of nucleosome positioning for inspiration for sequences that might alter the behavior of transcription factor-mediated loops, because it has been argued that at least *in vitro*, a sequence's nucleosomal affinity stems from its intrinsic flexibility, and not from a property specific to nucleosome binding [14]. We here discuss results with the strongest known nucleosome positioning sequence, 601TA (abbreviated "TA" here and elsewhere in this work), which has a significantly higher affinity for nucleosomes and a J-factor for cyclization 5 to 30 times greater than the random E8 sequence described in the previous section, depending on the phasing discussed in the next section [58, 85, 86]. If TA and E8 differ in mechanical bendability in some general sense, as is assumed from nucleosome affinity and cyclization assays, then TA should increase looping by a bacterial transcription factor just as it increases nucleosome binding and cyclizes more readily than E8.

As derived in Eqs. (4.2) and (4.3) and shown graphically in Fig. 4.1(B), if the TA and E8 sequences have different J-factors, then the concentration at which looping is maximal should be the same for

both sequences, but looping should increase at all concentrations with the more flexible sequence. This is indeed what we find experimentally in Fig. 4.1(E), which shows results for the looping probability as a function of repressor concentration for a loop with 94 bp of a sequence derived from TA, flanked by the O_{id} and O_1 operators. In analogy with the case of different operators discussed in the previous section, the agreement between the individual fit to the TA data (red dashed line) and the global fit to both the E8 and TA data (solid lines) demonstrates that the two data sets can be fit by the same operator dissociation constants but different J-factors (see Table 4.1). The outcome of this measurement is a looping J-factor of 4.2 ± 0.6 nM for the TA sequence, about 10 times higher than the random E8 sequence. This is again higher than the cyclization J-factors in [85] and [59] in terms of absolute magnitude, and significantly so: if we use Eq. (4.3) and the cyclization J-factors of [85] to predict maximal looping probabilities, we would expect the maximal looping probability for Oid-E894-O1 to be 0.25 ± 0.3 (compared to the experimentally observed 0.62 ± 0.01), for Oid-TA94-O1 to be 0.87 ± 0.2 (compared to 0.95 ± 0.01), and the O2-E894-O1 construct to show essentially no looping at all. The looping J-factor we measure for the TA sequence is not, however, as much higher than E8 as the 30-fold difference measured in cyclization [85], hinting that the constraints imposed on the DNA in cyclization versus loop formation may lead to a different dependence on sequence, as indeed we find below.

4.3 Effect of loop length on the looping probability

One of the signatures of looping by transcription factors both *in vitro* and *in vivo* is a significant modulation of transcription factor activity as the distance between the transcription factor binding sites is varied [25, 32, 49, 50]. A similar phasing effect has been observed in cyclization data with the E8 and TA sequences [85]. Our experiments, in conjunction with our model that allows us to extract J-factors, permit us to explore this phasing behavior for both of the sequences discussed in the previous sections and to compare to several recent theoretical predictions of the looping J-factor.

In the spirit of the kinds of theoretical predictions of Fig. 4.1(B), we can use the cyclization results of [85], which looked at the differences between E8 and TA across multiple DNA lengths,

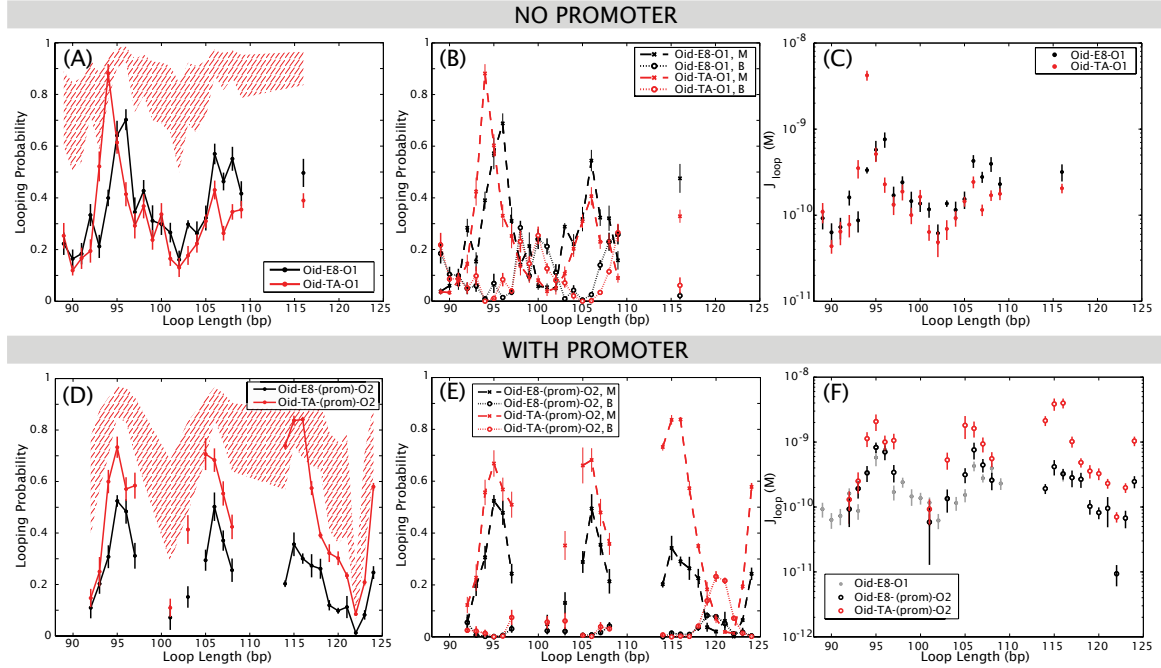


Figure 4.2: Looping probability as a function of loop length at constant repressor concentration. Surprisingly, the sequence dependence of Fig. 4.1(E) for the 94 bp construct is absent at other loop lengths. However, the bottom panels show data for constructs where 36 bp of either E8 or TA nearest O1 has been replaced with the *lacUV5* promoter sequence (and for technical reasons O1 has been replaced with O2, which should not affect our measurements of J-factors as demonstrated by the data in Fig. 4.1(D)). The presence of this promoter *restores* a sequence dependence to looping across several helical periods. **(A)** Total looping probabilities for the constructs Oid-E8-O1 and Oid-TA-O1, at 100 pM repressor. The red hatched region represents a prediction for where the TA data should fall, assuming the TA sequence has a J-factor anywhere from 5 to 30 times larger than the J-factor for the E8 sequence (a range based on the cyclization J-factors of [85]). The lengths used in earlier cyclization assays [85] are a subset of those shown in this figure. **(B)** Looping probabilities for the two looped states separately for the constructs in (A). The two states alternate in likelihood: the bottom state predominates around 89 bp and 100 bp, but the middle state around 94 bp and 106 bp. It is more clear in this panel than in (A) that E8 and TA are in phase with each other, with a period close to the canonical period of 10 bp, everywhere except near 94 bp, where TA has a maximum that is instead at 95–96 bp for E8. Therefore a simple offset in phase between the two sequences cannot account for the behavior at 94 bp. **(C)** Looping J-factors for the constructs shown in (A). The J-factors for both E8 and TA span at least an order of magnitude as a function of loop length, and the J-factors for the two looped states (see Fig. 4.3 and Fig. D.5(B)) can also differ by an order of magnitude at a given loop length. However, as shown in Fig. 4.3, this degree of modulation by operator phasing is less than might be predicted, depending on the assumptions made about Lac repressor conformation and flexibility. **(D)** Looping probabilities for constructs where part of the looping sequence of the constructs in (A) has been replaced with the 36 bp *lacUV5* promoter. The red hatched region is the same kind of cyclization-based prediction as in (A). In sharp contrast to the data in (A), with the promoter sequence in the loop, TA loops as much or more than E8 at all lengths measured, as would be expected from cyclization and nucleosome formation assays with the pure E8 and TA sequences. Note that because of the replacement of O1 by O2, the looping probabilities for these constructs may not match those of (A) even when the J-factors for the loops, plotted in (F), are the same (though as shown in Fig. 4.1(D) and derived in Eq. (2.6), at high concentrations curves with different operators begin to overlap, and 100 pM is sufficiently high that the looping probabilities should in fact be similar). **(E)** As in (B), here the two looped states have been separated out for the constructs in (D). With the promoter in the loop, the two sequences have the same phasing even at 94 bp (and in fact share the same phasing as the pure E8 constructs in (A)). Interestingly, the preferred looped state with the promoter is almost exclusively the middle state at all lengths—note, for example, that at 107 bp without the promoter, the two looped states are comparable in likelihood (see also Fig. 4.1(F)), but with the promoter at 107 bp only the middle state contributes to looping (see also Fig. D.5(D) and (E)). **(F)** J-factors for the constructs in (D) (open circles), overlaid on the J-factors for the no-promoter E8 construct shown in (C) (grayed-out closed circles). The addition of the promoter to the loop does not appreciably change the J-factors for E8-containing loops, only those of the TA-containing loops. See Fig. D.5(C) for the J-factors of the two states of (E). As in Fig. 4.1(D–F), errors in (A), (B), (D) and (F) are standard errors on the mean; the calculation of looping J-factors and associated errors is described in Appendix D.2.7. Solid, dashed and dotted lines in (A), (B), (D), and (E) are guides to the eye only, not theoretical predictions or fits. Their purpose is to highlight general trends.

to make a naïve prediction of how we would expect the sequence dependence to looping shown in Fig. 4.1(E) to manifest as the loop length is changed. Such a prediction is shown as a red hatched region in Fig. 4.2(A). However, as shown in that figure, to our surprise our experimental results for the looping probabilities for the two sequences, at a constant repressor concentration of 100 pM, show no sequence dependence to looping, with the exception of one or two lengths around the length shown in Fig. 4.1(B). The modulation of looping due to phasing is observed in both the E8- and TA-containing sequences, and, with the exception of the 94 bp loop length, it appears that this phasing is the same for both sequences. Yet again, surprisingly, not only does the nucleosome positioning sequence not fall within the hatched predicted region, in fact the nucleosome positioning sequence has comparable or smaller looping probabilities compared to the random sequence at most loop lengths.

Even more surprising is that a difference in loopability between the E8 and TA sequences can be restored when the last 36 bp of the loop is replaced with the bacterial *lacUV5* promoter sequence, as shown in Fig. 4.2(D). We were motivated to make this change since in parallel work (see Section 4.6 below) we have measured how this sequence-dependent looping affects gene expression *in vivo* and the presence of the promoter is a natural part of the full regulatory network. Though these loops contain 36 bp of the loop that are identical between the E8 and TA constructs, the TA-containing DNAs now loop more than the E8-containing DNAs, and at some lengths are even as much more flexible than the E8-containing DNAs as predicted based on cyclization assays, as shown by the red hatched region in Fig. 4.2(D). Interestingly, the J-factors for the E8 sequence with and without the promoter are comparable—that is, the inclusion of the promoter increases the flexibility of the TA-containing loops only (Fig. 4.2(F)).

Before discussing the implications of these complex sequence dependencies, we note several additional features of these length data in light of recent theoretical works on the length dependence of Lac repressor-mediated looping, which are plotted in Figure 4.3. As introduced in Section 2.2, we and others observe two looped states with any pair of operators, which have been hypothesized to arise from the four distinct topological states of the looped DNA and/or several distinct repressor

conformations schematized in the legend of Fig. 4.3 [130, 109, 114, 108, 115, 67, 68, 69, 127, 141, 70]. Regardless of their underlying molecular origins, in Fig. 4.1(F) we show that the two looped states we observe can be modeled as differing only in effective J-factor. We find that the J-factors for these two states have opposite phasings, at least without the promoter, as shown in Fig. 4.2(B), and this phasing does not change between sequences except near 94 bp. Such out-of-phase behavior for two different loop structures has been observed for other DNA looping proteins [142], and has been used to explain key features of *in vivo* repression data [143]. However it is not captured by all of the theoretical models in Fig. 4.3 (e.g., the “va” and “e” states of [128]). Intriguingly, the promoter changes the relative probabilities of the two looped states: as shown in Fig. 4.2(E), the promoter-containing constructs result almost exclusively in the middle state, whereas without the promoter, the two looped states alternate in prevalence (Fig. 4.2(B)). As these measurements represent the first single-molecule study on the phasing of these two looped states at single base-pair resolution, over two helical periods of DNA, at the short loop lengths where the models in Fig. 4.3 show the most pronounced differences in J-factors due to repressor and loop conformations, we hope that our data will help shed light on the molecular origins of the two looped states.

4.4 A need to revisit our understanding of sequence flexibility

We have shown here that the looping J-factors for 94 bp of a random sequence and a nucleosome positioning sequence differ by an order of magnitude, with the nucleosome positioning sequence being more flexible than the random sequence, as expected based on previous cyclization and nucleosome formation assays. To our surprise, however, this sequence dependence occurs only at 94 bp, unless a bacterial promoter sequence is added to the loop, in which case a consistent length-independent sequence dependence is restored.

We hypothesize that the sequence-dependent free energy of bending a DNA depends more strongly than has been previously appreciated upon the specific details of how the DNA double

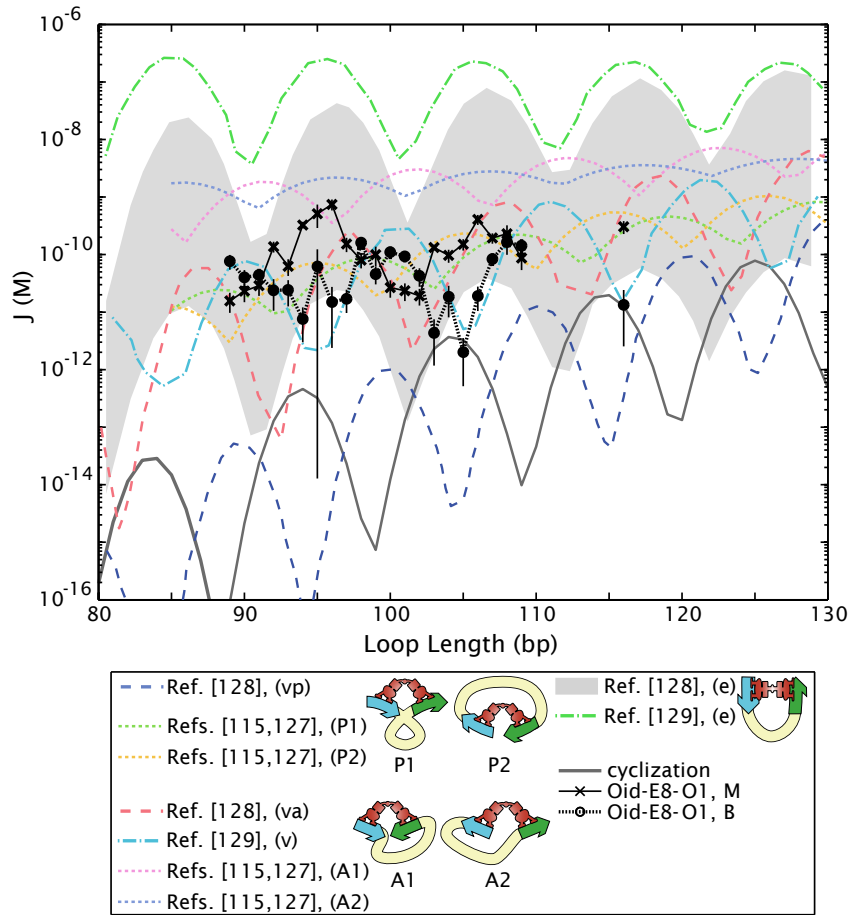


Figure 4.3: Theoretical predictions of the length dependence of the looping J-factor. Elasticity theory with “canonical” values for the stiffness of random DNA sequences, in conjunction with various models of the geometric and mechanical constraints imposed by the Lac repressor tetramer, have been used to compute the looping J-factor [115, 127, 128, 129]. The model of [115, 127] also explicitly includes the boundary conditions of a TPM experiment, with a bead on one end of the DNA and a surface on the other. The assumed constraints can be roughly grouped into V-like repressor conformations, similar to the shape seen in the crystal structure 1LBI [95] (“P1” and “P2”, indistinguishable unless as in TPM there are symmetry-breaking boundary conditions, and therefore collapsed into one state, “vp”, in [128]; and “A1” and “A2”, collapsed into “v” or “va” in [128, 129]); and more extended repressor conformations (“e”), which are favored by the DNA mechanics. These conformations are indicated schematically in the legend; for the case of [115, 127], the blue operator has been chosen to be O_{id} , that is, the operator closest to the surface. The prediction for the extended conformation of [128] is a range of values, reflecting estimated uncertainty in the free energy costs of opening the repressor tetramer. Details of how these curves were obtained are given in Section 4.A.2 below. Our experimental measurements for the two looped states of the no-promoter E8 sequence (“Oid-E8-O1, M” and “Oid-E8-O1, B”), as well as the cyclization result of Shimada and Yamakawa [57, 115] (“cyclization”) have been included for comparison. It is to our looping J-factors for the two looped states separately that we compare the theoretical results, as each of the theoretical results shown here make assumptions about the loop conformation that surely must differ between the two looped states we observe. We caution the reader that a detailed direct comparison between these theoretical predictions and with our data may not be possible for several reasons: (1) assumptions about experimental conditions, such as salt concentrations, differ between references and from the conditions in this work; (2) it is possible, as argued in [115, 127], that the experimentally observed states correspond to superpositions of two or more theoretically predicted states for different loop topologies and/or repressor conformations; and (3) as suggested by FRET data [68, 70] (though see also [130]), TPM with cross-linked repressor [108], and molecular dynamics simulations [144], the protein conformation in both states may involve some degree of rearrangement relative to the V-like conformation observed in the crystal structure (at the least, rotation of the DNA binding domains, as in [144]). In these cases our data would not align with any single theoretical curve presented here. However, we do make some general observations about the relationship of our data to these theoretical models in Section 4.A.2, and hope that these data will help shed light on the origins of the two looped states.

helix is deformed when forming loops versus nucleosomes versus DNA circles. Drew and Travers argued that a DNA minicircle formed by cyclization shares structural similarities with the DNA wrapped around a histone octamer [145], explaining the usefulness of cyclization assays for understanding the sequence preferences of nucleosome formation. Cyclization has often been cited as a model by which to understand looping as well [29, 58, 84, 88]. However, as diagrammed in Fig. 4.3, for DNA loop formation by the Lac repressor, there are multiple looped configurations allowed for a given loop length, most of which are probably quite far from circular as a result of the distinct boundary conditions imposed by repressor binding, and which should have large effects on the associated looping J-factor. We argue that although DNA cyclization may share characteristics with DNA looping such as length-dependent phasing, it apparently does not share other characteristics such as trends in sequence-dependent flexibility, possibly because of this difference in boundary conditions. We also suspect that the strong sequence dependence at 94 bp without the promoter, and with the promoter at all lengths, is due to a change in the preferred loop conformations of these constructs, compared to the majority of the no-promoter constructs. Indeed, the change in the predominant looped state (the “bottom” and “middle” states alternating without the promoter, but the “middle” state predominating at all lengths with the promoter) supports this hypothesis that the promoter alters the preferred conformation of the loop (see also Fig. D.5(D) and (E)). To further unravel these subtleties and to elucidate the sequence rules of loop formation, as has already been done for nucleosome formation [14, 52], we believe a more thorough search of sequence space using the Sort-Seq and high-throughput TPM approaches described in Chapter 7 will be necessary. We also hope that additional theoretical analyses, perhaps involving the observed tether lengths of the looped state with and without the promoter given in Section 4.A.1 below, may shed further light on the conformations of looping for these different sequences.

As discussed in Chapter 1, the mechanics of loop formation at these short loop lengths that are so prevalent in cellular processes is a subject of much debate, regardless of their sequences [27, 51]. However, the question of how flexible we expect short DNAs to be is more complicated to answer in the case of protein-mediated DNA looping than in the case of cyclization. As shown

in Fig. 4.3, varying the boundary conditions of the loop or the assumed protein flexibility can lead to enormous differences in predicted looping J-factors. Some of these predicted J-factors, using canonical assumptions about DNA flexibility, are in fact consistent with the J-factors we measure, so perhaps it should not be surprising that short transcription factor-mediated loops can form readily *in vitro*.

4.5 Preliminary results with additional sequences

A key question raised by the results of the previous sections is how general the conclusions are for other potential looping sequences. Is there no sequence dependence to looping *in vitro*, unless the *lacUV5* promoter is added to the loop, or is this result peculiar to the E8 and TA sequences examined here? Is the ability to restore sequence dependence specific to the *lacUV5* promoter, or do other bacterial promoters yield the same results? What about other looping proteins such as GalR or AraC [26]?

As mentioned in the previous section and described in more detail in Chapter 7, the question of whether any sequences alter loop formation in the absence of the *lacUV5* promoter will most likely require a broad search of sequence space through a combination of the Sort-Seq method and high-throughput TPM. However, we have begun to address this question of the generality of our results by again turning to sequences studied in the context of nucleosome formation, this time to sequences known to *disfavor* nucleosomes, a class of sequences that are rich in long stretches of A and T base pairs.¹

Such poly(dA:dT) sequences are overrepresented in eukaryotic (but not prokaryotic) genomes [146], with beginnings and ends of promoter regions in eukaryotes often especially highly enriched in these sequences [75]. It has been shown both *in vivo* [147, 148, 75, 149] and *in vitro* [150, 151, 152] that poly(dA:dT) sequences disfavor nucleosome formation, and their presence at promoters and the ends of genes has been correlated with increased gene expression levels [153, 147] and

¹This poly(dA:dT) project is an equal collaboration with Yi-Ju Chen in the Phillips lab and was suggested to us by Jon Widom.

decreased transcriptional noise [75]. In fact, it has been argued that poly(dA:dT) tracts are the major determinants of nucleosome positions *in vivo*, rather than nucleosome-preferring sequences such as TA [15].

However, the mechanism by which poly(dA:dT) tracts exclude nucleosomes and influence transcription *in vivo* is as yet unclear [148]. It is known that DNA polymers with 4 or more A nucleotides in a row show unique structural and dynamical properties in a variety of assays [148], and generally it is thought that long stretches of poly(dA:dT) are relatively straight and inflexible *in vitro* [154] (though see [61] for experimental evidence that poly(dA:dT) is *more* flexible, not less; and also it should be noted that the *phased* A-tracts discussed in Chapter 1 are known to induce intrinsic bends into DNA [155], rather than be intrinsically straight). A leading hypothesis for why poly(dA:dT) tracts disfavor nucleosome formation, then, is that their unique structural and dynamic properties lead them to be especially resistant to the deformations that are required for DNA wrapped in a nucleosome [148]. That is, just as the TA sequence favors nucleosome formation because of a high intrinsic flexibility (at least with regards to certain deformations), poly(dA:dT) has a high intrinsic *inflexibility* relative to the deformations involved in nucleosome formation. Again this high inflexibility is thought not to arise from any particular stiffness to AA dinucleotide steps but rather from the special structures known to form when more than two AA steps are found in a row [148]; nevertheless, they should look “stiff” under comparable deformations to those required in a nucleosome.

To test this hypothesis that poly(dA:dT) tracts disfavor nucleosome formation because of a high intrinsic inflexibility in the context of nucleosome-like deformations (regardless of the molecular origin of this stiffness), and also to test the generality of our results with E8 and TA, we chose a poly(dA:dT)-rich promoter region from *S. cerevisiae* that was shown to exclude nucleosomes *in vivo* by microarray analysis [149], and inserted this sequence into both the no-promoter and with-promoter loops described in the previous sections. (See Fig. B.3 and Appendix B.2 for details of these sequences.) If the results from the E8- and TA-containing loops hold more generally, and poly(dA:dT) sequences disfavor nucleosomes due to a high intrinsic inflexibility in the context of nucleosome-like shapes, then we would expect these poly(dA:dT) sequences to yield the same amount

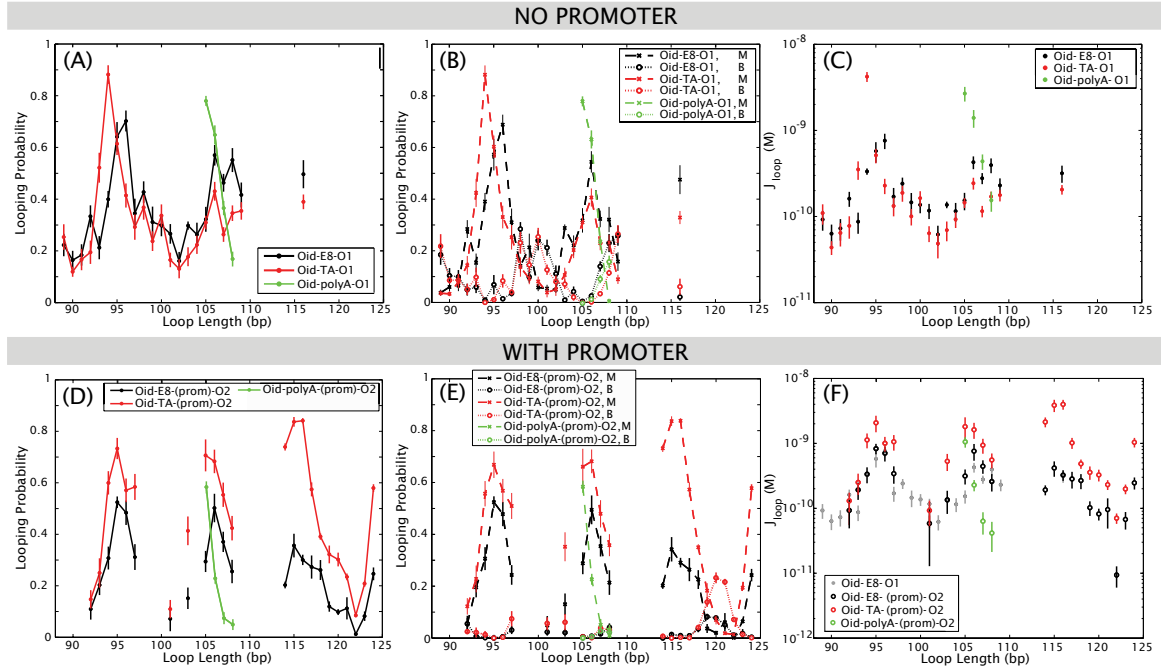


Figure 4.4: Looping probabilities as a function of loop length for a poly(dA:dT)-rich sequence known to exclude nucleosomes in yeast, superimposed on the E8 and TA data of Fig. 4.2, both with and without the promoter. **(A)** Total looping probabilities for the no-promoter constructs, as in Fig. 4.2(A). Data at additional loop lengths with the poly(dA:dT) sequence will be necessary to draw definite conclusions, but it appears that the poly(dA:dT) has a different period than the E8 and TA data, and that, in contrast to the E8 and TA data, the poly(dA:dT) sequence may alter the amount of looping compared to E8, even without the promoter. **(B)** Looping probabilities for the two looped states separately, as in Fig. 4.2(B), for the constructs in (A) here. Interestingly, even though the periods of these three sequences seem to be different, the pattern of which looped state predominates at a given length is consistent between all three sequences: note especially the 107 and 108 bp lengths. **(C)** J-factors corresponding to the total looping probabilities in (A). One of the poly(dA:dT) loops is almost as flexible as TA94, which is surprising given that we expected, based on nucleosome affinity assays, that poly(dA:dT) might be *less* flexible. **(D)** Total looping probabilities for the with-promoter constructs, as in Fig. 4.2(D). As with the data in (A), it appears that the poly(dA:dT) sequence has a different period relative to E8 and TA, so the conclusions we can draw from only four data points are limited. However, it seems that, unlike with E8 and TA, with the promoter in the loop the sequence dependence of poly(dA:dT) may not follow that of nucleosome formation, at least relative to E8: the TA sequence, which nucleosomes preferentially bind to over the random E8 sequence, loops more than E8; but the poly(dA:dT) sequence, known to exclude nucleosomes from a promoter region *in vivo*, loops as much as E8 (though still less than TA) at some lengths. **(E)** Looping probabilities for the two states separately, for the constructs in (A). As was observed for the E8 and TA constructs in Fig 4.2(E), with the promoter in the loop, the middle state predominates at all four lengths, whereas without the promoter in (B), the bottom state is equally or more dominant at 107 and 108 bp. **(F)** J-factors for the constructs in (D). The presence of the promoter decreases the J-factor of the poly(dA:dT) sequence relative to the no-promoter constructs, though not to a value less than that of E8, contrary to what we would expect from nucleosome formation assays.

of looping as the E8 and TA sequences without the promoter, but for the with-promoter loops to follow the same sequence dependence as nucleosome formation, that is, with the poly(dA:dT) sequences looping less than E8 and TA.

As shown in Fig. 4.4, this is not what we find. As our preliminary results include only four loop lengths with and without the promoter, we can at the moment draw only limited conclusions. But the most striking feature of the poly(dA:dT) data is that it appears the period of looping (that is, at what lengths looping is maximized or minimized), is different for the A-tract containing DNAs compared to that of E8 or TA. This is perhaps unsurprising, given that A-tract-containing DNAs are thought to adopt unique structures, and in fact some A-tract-containing DNAs have been shown to have shorter periods than other sequences [148]. What is also striking, however, is how flexible both the no-promoter and with-promoter poly(dA:dT)-containing loops appear to be: the 105 bp loop without the promoter is almost as flexible as the TA94 sequence (Fig. 4.4(C)), and even without the promoter the poly(dA:dT) loop is at least as flexible as E8 loops of comparable (though not identical, due to the period offset) lengths. Indeed, it appears that, in contrast to the results with E8 and TA, the poly(dA:dT) loop does show a sequence dependence in the absence of the promoter, as well as with the promoter, in that its looping probability is different from that of E8 in both cases. One aspect of the data is consistent across all three sequences, though: the relative probabilities of the different looped states. Without the promoter, the two looped states alternate in prevalence, including for the poly(dA:dT) constructs, but with the promoter, the middle looped state predominates.

We argued in the previous section that we suspect that whether or not there is a sequence dependence to looping depends strongly on the shape of the loop, with the promoter altering the preferred conformation such that the promoter-containing loops are more similar in shape to nucleosomes than the no-promoter loops, and therefore the patterns of sequence dependence seen in nucleosome formation hold only for promoter-containing loops, and not the no-promoter loops. We have now shown that that is not generally the case: if the with-promoter DNAs followed the sequence preferences of nucleosomes, then the poly(dA:dT) sequence with the promoter should have looped

less than E8 regardless of the period offset. However, the fact that now neither the with-promoter nor no-promoter constructs follow the sequence dependence trends of nucleosomes underscores even further our argument that “sequence flexibility” is not a general term.

We maintain our original hypothesis that the notion of sequence flexibility needs to be linked to the shape of the deformation induced to measure such flexibility. Because poly(dA:dT) tracts are known to possess unique structural properties, the fact that the poly(dA:dT)-containing loops do not match the sequence-dependence trends of E8 and TA is perhaps further evidence that the shape of the loop plays a large role in the observed flexibility trends. If that were the case, it would also demonstrate that the Lac repressor can accommodate a range of different looped structures, based on the deformation-dependent flexibilities of the loop sequence. Indeed, Haeusler and coworkers have recently shown that the Lac repressor can accommodate a surprisingly large range of designed loop topologies (made with phased A-tracts that introduce static bends in the DNA) [70]. We anticipate that the poly(dA:dT) loops form yet an additional shape, beyond the different shapes we have postulated for with-promoter versus no-promoter E8- and TA-containing loops, because of the unique structural requirements of A-tract DNAs.

As will be described in Chapter 7, rigorous testing of our deformation-dependent hypothesis will require testing a broader region of sequence space than can be accomplished by picking and choosing from sequences studied in the context of nucleosome formation, which addresses only a limited region of shape space (roughly circular) that is probably inaccessible to looping. We will therefore propose a *de novo* search of sequence space to try to identify sequences that are especially good or especially poor looping sequences, to try to build up rules for the sequence dependence to loop formation, as has already been done with nucleosomes [14]. More importantly, as will be seen in the next section, the question of whether sequence flexibility is a “knob” that tunes loop formation *in vivo* remains a very important and outstanding one that cannot, we will argue, necessarily be addressed by *in vitro* techniques alone.

4.6 The masking of sequence effects *in vivo* by nonspecific DNA-bending proteins

The ability of the Lac repressor to form the loops that we have been studying here is crucial for its function as the negative regulator of the *lac* operon *in vivo* [93, 94, 32], with any increases in looping leading, presumably, to an increase in repression of gene expression. As we have shown that sequence can in some cases contribute to significant increases in looping probability *in vitro*, it is important to ask if these increases in looping probability are translated into an increased amount of repression *in vivo*.

Figure 4.5 shows results of *in vivo* repression assays in which the same promoter-containing constructs as were examined *in vitro* in Fig. 4.2(D–F) were integrated into the *E. coli* genome, such that the *lacUV5* promoter drives the expression of a fluorescent reporter gene (YFP) [118].² The activity of the Lac repressor manifests as a decrease in YFP expression: repression is defined as the amount of YFP expression in an *E. coli* strain in which the Lac repressor is not expressed and therefore YFP expression is constitutive and maximal, divided by the amount of YFP expression in the presence of the Lac repressor. That is,

$$\text{Repression} = \frac{\text{YFP}([R] = 0)}{\text{YFP}([R] \neq 0)}. \quad (4.4)$$

Therefore when repression is 1, transcription is unregulated; repression greater than 1 indicates the Lac repressor has decreased YFP expression (and so the denominator of Eq. (4.4) decreases).

The same statistical mechanical approach that was used to derive an expression for the looping probability measured by TPM as a function of key tunable parameters can be used to derive a similar expression for the repression of gene expression measured *in vivo*. In terms of statistical mechanics,

²The *in vivo* model and data in this section are the work of James Boedicker and Hernan Garcia in the Phillips lab.

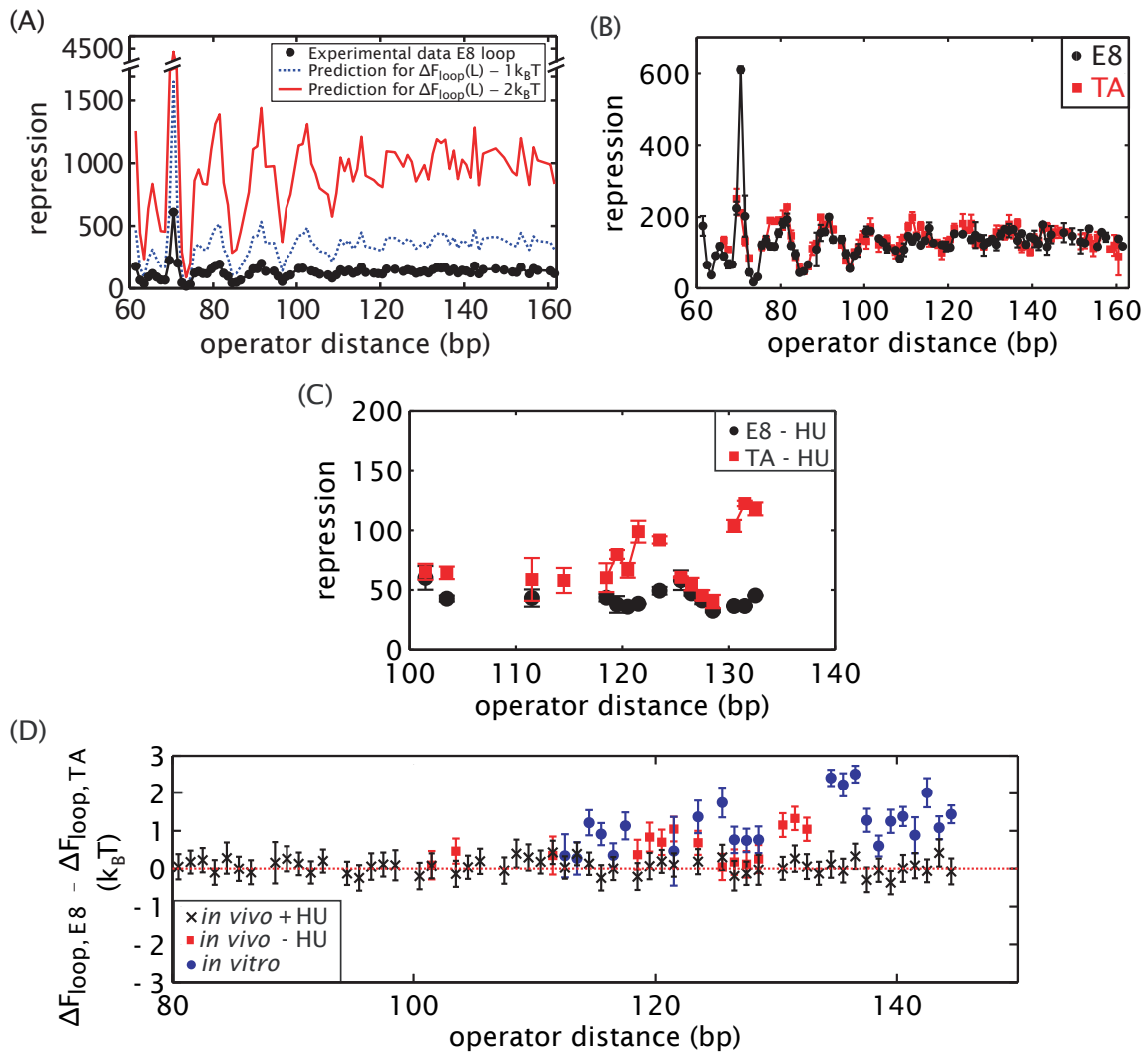


Figure 4.5: The nonspecific DNA-bending protein HU masks sequence-dependent looping *in vivo*. **(A)** Prediction of the effect that having TA instead of E8 in the loop should have on the observed repression, based on Eq. (4.5) and a difference of either 1 (blue) or 2 (red) $k_B T$ in the F_{loop} for E8 versus TA. **(B)** In sharp contrast to the prediction of (A), no sequence dependence to repression is observed *in vivo*: E8- and TA-containing loops yield the same amount of repression. (Note that these are the promoter-containing loops that *do* show a sequence dependence *in vitro* in Fig. 4.2(D–F)). The length-dependent modulation (phasing) that is a signature of looping (see Fig. 1.1(D)) is, however, still observed for both sequences. **(C)** Some sequence dependence to repression is restored when the genes for HU are deleted. Note that the total amount of repression also decreases, which is consistent with previous studies of the effect of HU on looping by the Lac repressor *in vivo* [19], in which HU stabilizes Lac repressor-mediated loops and thereby increases repression. **(D)** Difference in observed free energies between E8 and TA *in vivo* with and without HU, and *in vitro*. A free energy difference of 0 (red dashed line) indicates no sequence dependence to looping. *In vivo*, without HU, E8 and TA look equally flexible (black); however when HU is removed (red), the TA loops form more readily at some lengths, though not to the degree observed *in vitro* (blue). HU is not the only DNA-bending protein in *E. coli* and it is likely that these other bending proteins contribute to the residual discrepancy between the *in vivo* and *in vitro* measurements. Adapted from [118].

then, Eq. (4.4) can be rewritten as

$$\text{Repression} = \frac{1 + \frac{2R}{N_{NS}}(e^{-\beta\Delta\epsilon_{rad}} + e^{-\beta\Delta\epsilon_{rmd}}) + \frac{4R(R-1)}{N_{NS}^2}e^{-\beta(\Delta\epsilon_{rad}+\Delta\epsilon_{rmd})} + \frac{2R}{N_{NS}}e^{-\beta(\Delta\epsilon_{rad}+\Delta\epsilon_{rmd}+\Delta F_{loop})}}{1 + \frac{2R}{N_{NS}}e^{-\beta\Delta\epsilon_{rad}}}, \quad (4.5)$$

where R is the number of repressors in the cell, N_{NS} is the number of nonspecific binding sites (roughly 5×10^6 , the number of base pairs in an *E. coli* genome), $\Delta\epsilon_{rmd}$ is the binding energy of the repressor to the operator overlapping the promoter, $\Delta\epsilon_{rad}$ is the binding energy of the repressor to the other (distal) operator, and β is the reciprocal of the temperature times the Boltzmann constant [118]. Most importantly, the flexibility of the sequence in the loop should be captured by the parameter ΔF_{loop} , related to the J-factors that we measure with TPM according to Eq. (2.2) in Chapter 2; and even small (one or two $k_B T$) changes to ΔF_{loop} should manifest as large effects on the fold-change (Fig. 4.5(A)). The differences in J-factors we measured with TPM for the with-promoter E8 and TA sequences should easily be large enough to be observed as a change in gene expression *in vivo* (Fig. 4.5(D)).

However, to our surprise, we find no sequence dependence to repression *in vivo* with the promoter-containing loops for which there is a strong sequence dependence *in vitro* (Fig. 4.5(B)). The cause of this lack of sequence dependence appears to be the action of one or more nonspecific DNA-bending proteins that organize the *E. coli* genome [17]: it is possible to restore at least some sequence dependence if the genes for one of these proteins, HU, are knocked out (Fig. 4.5(C–D)). Previous experiments on cyclization with intrinsically curved sequences in the presence of DNA-bending proteins, and some limited complementary *in vivo* experiments, are suggestive of a similar effect in eukaryotic cells [156, 157]. It remains to be seen if there are any sequences that can overcome this masking effect of HU in bacteria, or if sequence flexibility is not a parameter that either prokaryotic or eukaryotic cells tune to regulate gene expression.

4.7 Conclusion

Here, we have presented experimental results from a combined single-molecule-plus-modeling approach that allows us to explore how the transcription factor-mediated loops that are a common motif in both bacteria and prokaryotes are influenced by four distinct, tunable biological parameters: transcription factor binding site strength, transcription factor concentration, DNA loop length, and DNA loop sequence. We have demonstrated that this approach explains how the looping probability depends upon the strength of the operator dissociation constants and that our measured K_d 's agree well with values previously obtained by bulk biochemical methods. Further, our model accounts well both quantitatively and qualitatively for the effects of varying the loop flexibility, as well as for details of our single-molecule looping experiments such as the presence of two looped states. Our method provides a way of measuring J-factors that is orthogonal to, and therefore complementary to, current methods in use, which we argue has led to important new insights into the role of sequence in DNA flexibility. In particular we have argued here that the sequence-dependent free energy of bending a DNA must depend more strongly than has been previously appreciated upon the specific details of how the DNA double helix is deformed when forming loops versus nucleosomes versus DNA circles. It is not the case that the TA sequence can be claimed to be more flexible in some general sense, nor the poly(dA:dT) inflexible in some general sense; nor can cyclization and nucleosome affinity assays be used to determine DNA flexibility for all biological contexts, as we have shown here that loop formation does not necessarily follow the same sequence rules as cyclization and nucleosome formation. We hope that the measurement of *looping* J-factors for many more sequences, especially those that are not derived from nucleosome affinity assays but from other biological or synthetic contexts, will begin to elucidate the rules of the sequence dependence to loop formation that we have only begun to glimpse here.

The *in vivo* results of Section 4.6 reveal both the power and limitation of *in vitro* assays such as TPM. On the one hand it is clear that the complex environment of the cell, and particularly the many proteins that structure the genome *in vivo*, creating a context far removed from the naked, linear DNAs of TPM, make it difficult to generalize from TPM results to biological impacts in cells.

In fact, in Chapter 6 we will present another example of a system that may also require architectural proteins to replicate *in vivo* results *in vitro*. On the other hand, without the TPM data we present here, it would not have been possible to claim that HU masks sequence dependence: we would not have demonstrated that the E8 and TA sequences can result in different amounts of looping, as previously they had been studied only in the context of cyclization and nucleosome formation. Furthermore, contact between the *in vivo* and *in vitro* worlds is only possible through our statistical mechanical models, which allow the extraction of looping free energies and J-factors, and allow us to predict if the differences in J-factor we observe in TPM should be sufficient to affect repression detectably *in vivo*. We therefore believe that this three-pronged approach of theory, *in vitro*, and *in vivo* experiments offer the best path towards dissecting the role of DNA mechanics in gene regulation.

4.A Appendices to Chapter 4

4.A.1 RMS of the unlooped and looped states as a function of concentration and of loop length

We hope that the data presented in this work will aid in future attempts to model the interactions of the Lac repressor with DNA, and to that end in this section we comment on the tether lengths of the unlooped and two looped states that we observe with TPM. In this work we have focused on using tether length as an indicator of the state (looped or unlooped) of the system, which in turn enables us to calculate looping probabilities; however we recognize that tether lengths contain additional information about the underlying conformation of the tethered DNA. For example, in Fig. 4.6(A) we discuss an effect that may be indicative of the bending of the operator DNA by bound Lac repressor, seen in the crystal structure of [95].

The tether lengths we observe in populations of otherwise identical tethered DNAs vary noticeably (see the black horizontal dashed lines in Figs. E.1 and E.2(A–B)), which we suspect arises at least in part from variations in bead diameter (the manufacturer reports a coefficient of variation of 1% in the base polystyrene particle, which should correspond roughly to a standard deviation in bead diameter of about 5 nm). Therefore what we report on the y-axes in Fig. 4.6 is the average tether length *relative to the “no lac” length*; that is, the average difference between a bead’s unlooped or looped state(s) and the length of that particular tether recorded before repressor has been introduced into the TPM flow chamber. This allows us to resolve small but detectable changes in tether length in the presence versus the absence of repressor which would otherwise be obscured by the larger bead-to-bead variation in diameter. The improvement in resolution that we obtain by this method is perhaps one reason why we see evidence for operator bending where previous TPM experiments with the Lac repressor has not [109].

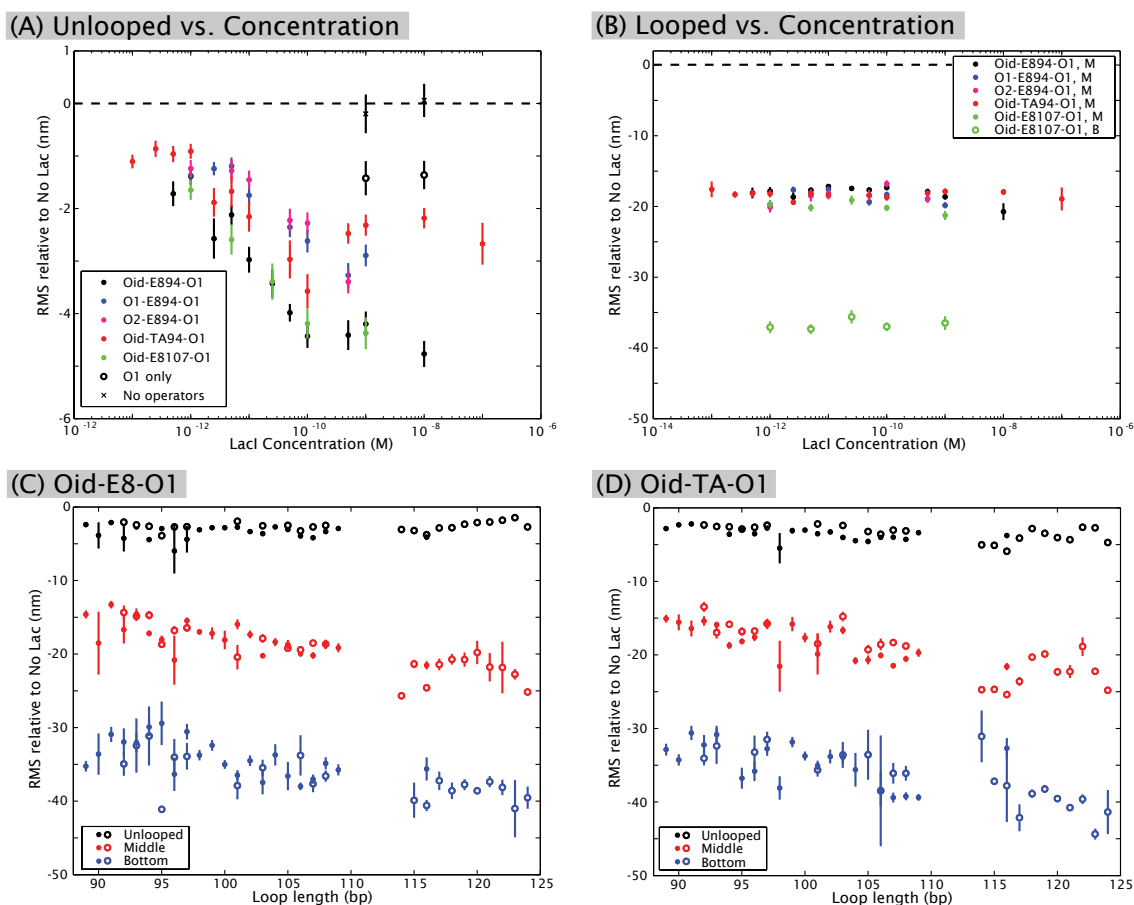


Figure 4.6: The effect of increasing repressor concentration or loop length on observed tether length. As noted in the text, due to initial tether length variability, we here report average RMS motion relative to the RMS in the absence of repressor. The tether length of each state was determined by the thresholding method described in Appendix D.2.3 (even for constructs whose looping probabilities were determined by Gaussian fitting); the average tether length of the state was taken to be a weighted average of RMS values within the threshold limits. The average tether length of the DNAs with 94 bp loops in Fig. 4.1(D–E) in the absence of repressor is about 147 nm; the average tether length of the DNA with a 107 bp loop in Fig. 4.1(F) is comparable (150 nm). **(A)** The average relative RMS of the unlooped state as a function of concentration, for the five DNAs in Fig. 4.1(D–F), plus two DNAs which are missing one (“No Oid”, which is (noOid)-E872-O1) or both (“No operators”, (noOid)-E872-(noO1)) of the operators. These DNAs with missing operators are slightly shorter than the others in this figure, with average tether lengths in the absence of repressor of about 140 nm. For the two-operator DNAs, as the amount of repressor increases, the RMS of the unlooped state decreases. We suspect this shortening is due to the bending of the operator DNA by a bound repressor; in the crystal structure in [95], a repressor bound to the *O_{id}* operator produces a 45° bend. At low repressor concentrations, the unlooped state that we observe in TPM is mostly composed of the state where nothing is bound to the DNA (state (i) of Fig. 2.2(A)), and so little operator bending is observed. However, at high concentrations, the unlooped state is composed mostly of the doubly occupied state (state (iv) of Fig. 2.2(A)), and so the tether length shortens significantly. The addition of repressor has no effect on DNA with no operators at high repressor concentrations, suggesting that the shortening we observe is not due to nonspecific binding of the repressor to non-operator DNA. Note that the observed reduction in tether length for two operators will not necessarily be double the shortening of tether length for a single operator, depending on the phasing of the two bend angles. **(B)** Average relative RMS for the middle and bottom looped states as a function of concentration, for the constructs of Fig. 4.1(D–F). Unlike the unlooped state, the looped states are invariant with concentration. **(C)** Average relative RMS for the unlooped and looped states as a function of increasing loop length, for the E8-containing DNAs whose looping probabilities are shown in Fig. 4.2(A) (no promoter, closed circles) and Fig. 4.2(D) (with promoter, open circles) in Chapter 4. Note the shortening of the unlooped state that we attribute to operator bending, since these data are taken at 100 pM repressor. The contour lengths of the with-promoter DNAs are slightly shorter than the no-promoter DNAs, with an average RMS in the absence of repressor of about 143 bp at a 94 bp loop length. **(D)** Same as (C) but for the TA-containing DNAs.

4.A.2 Compiling looping predictions from several recent theoretical analyses

Cyclization free energies according to the Shimada and Yamakawa approximation [57] were taken from [115, 127].³ Predictions of the looping free energy of various looped conformations from [128, 129] were digitized with Engauge Digitizer (<http://digitizer.sourceforge.net/>). From [128], looping free energies of the V-like conformations P1 and A1 (called “vp” and “va” in Fig. 4.3 in this work) and the extended conformation P1E (“e” in Fig. 4.3) are taken from Fig. 4 of [128]. The extended conformation free energy of [128] contains a contribution that describes the cost of opening the Lac repressor tetramer to the extended conformation. This term was estimated with an interval, reflected by the upper and lower edges of the gray polygon in Figure 4.3 in this work. From [129], looping free energies of the V-like conformation LB and extended conformation SL are taken from Fig 3(B) of that work. 20.5 bp was subtracted from the DNA length values used in that figure in order to convert it to the loop length convention used in [115, 127, 128], and in this work. Ref. [129] also discusses loop conformations similar to that labeled “vp” for [128]; these results were excluded for clarity, as they (and the “vp” conformation for [128] shown in the figure) contribute very little to the total J-factor.

As noted in the caption to Fig. 4.3, detailed comparisons between these theoretical predictions and with our data may not be possible. However, we do make the following observations: first, the most striking feature of our data not consistently captured by all of the models summarized here is that the two looped states we observe can have comparable J-factors at some lengths—that is, the curves for J_B and J_M intersect, at least without the promoter. This is surprising if one postulates that one of the two looped states corresponds to a V-like protein conformation and the other to a more extended protein conformation, as in [109], as computational analyses usually find the extended protein conformation to be so favorable in terms of the DNA mechanics as to generate a J-factor orders of magnitude larger than any V-like conformations at the lengths we examined (e.g., [129]). Second, the “B” and “M” states are out-of-phase with each other, much as the “v” and “e” states of

³Thanks to Martin Lindén for compiling these theoretical results.

[129] or the “A1” and “A2” or “P1” and “P2” states of [115, 127], but not the “va” and “e” states of [128]. Finally, as discussed in more detail in the next chapter, we observe direct interconversion between the two looped states, which would suggest that they differ in protein conformation and not loop topology; however, it is possible, given the 4 second Gaussian filter applied to our data, that we are smoothing out short transitions to the unlooped state. In the next chapter we discuss preliminary attempts to quantify the possibility that the apparent direct interconversions are a result of filtering.

Chapter 5

A kinetic analysis of looping by the Lac repressor

In the preceding chapters we have used the looping probability measured with TPM to gain insight into several important aspects of the short-length-scale mechanics of different DNA sequences in the context of loop formation. However, the RMS-versus-time trajectories that we obtain in TPM contain additional information about looping, namely kinetic information. For example, as shown in Fig. 5.1, two repressor concentrations may lead to comparable looping probabilities for a particular construct, and therefore would look the same in the analyses of the previous chapters; but the dynamics of looping at these concentrations can be very different. Another example of the kinds of insights that are available only from kinetic analyses of TPM data relates to the origins of the two looped states that we and others observe with the Lac repressor (see Section 2.2, Section 4.3, and Fig. 4.3): one of the distinguishing predictions of the two classes of models used to explain these states (two configurations of repressor versus four DNA binding orientations) is whether or not the two states should be able to directly interconvert.

In this chapter we will examine these and similar questions through the use of two different methods to obtain rate constants and state lifetimes from TPM data. The first of these methods, half-amplitude thresholding combined with dwell time histogram analysis, is a classic method first used to analyze single ion channel recordings [158], and is one of the most commonly used methods to obtain kinetic information from TPM [104, 109, 111, 113, 159] as well as from other single-molecule data such as FRET [160]. As described in more detail in the Appendix to this chapter

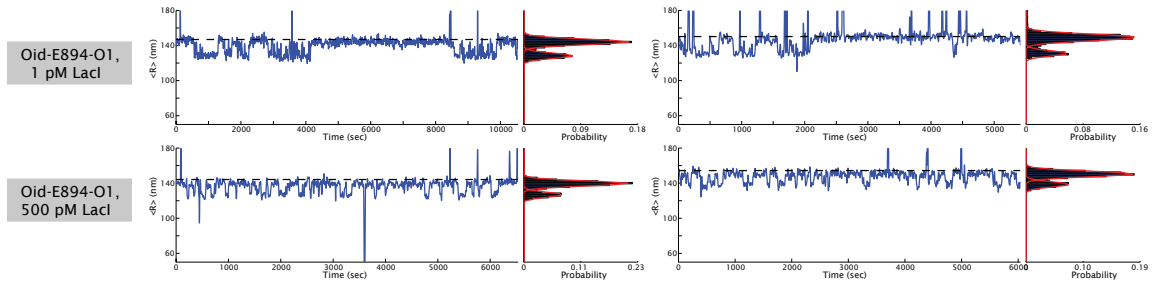


Figure 5.1: An example of the additional information that a kinetic analysis can yield, beyond that provided by an equilibrium analysis of the looping probability. The blue data show the Gaussian-filtered, root-mean-squared motion of 4 different beads tethered by a 450 bp construct with 94 bp of the E8 sequence flanked by the O_{id} and O_1 operators, in the presence of 1 pM (top) or 500 pM (bottom) Lac repressor. The looping probabilities for this DNA at 1 pM and 500 pM are comparable, as can be seen in the histograms of the RMS motion to the right of each trace (looping probability is defined as the area under the looped peak in the histogram, divided by the area under both the looped and unlooped peaks; see also Fig. 4.1(D)); but the lifetimes of the unlooped and looped states at these two repressor concentrations are very different. At 1 pM, bursts of looping are interspersed between long dwells in the unlooped state, whereas at 500 pM, the dwell times in the unlooped state are more uniform. We interpret this difference according to the predictions for the probabilities of each state of the system according to our model, shown in Fig. 2.2(B): at low repressor concentrations, the unlooped state is primarily composed of the state with no repressors bound or a repressor bound at O_{id} ; so the long unlooped dwells may be times when no repressor is bound, and represent the waiting time for another repressor to bind one of the operators. However at high repressor concentrations, the unlooped state is primarily composed of the state with both operators bound by separate repressors, and so the lengths of the unlooped dwells depend on the dissociation of one of the bound repressors so that a loop can form. (These same four traces are shown in Fig. E.1 in Appendix E.)

and shown schematically in Fig. 5.2(A), this approach consists of thresholding the RMS-versus-time trajectories to assign a state (looped or unlooped) to each time point, histogramming the resulting dwell times in each state, and fitting exponentials to these dwell-time histograms. This method has the advantage of being relatively easy to implement, but it is subject to several serious limitations. Since the thresholding must be done on smoothed data (e.g., *via* a Gaussian filter as in this work (Section D.2)), the temporal resolution of this approach is limited by the dead-time of the filter and misses short-lived events. Moreover, the choice of filter width can significantly affect the calculated rate constants (though procedures have been suggested to correct for these filter effects) [113]. To overcome these limitations, hidden Markov models (HMM) [161, 162] and other maximum likelihood approaches such as the “change-point” algorithm [159] have been developed. Here we will introduce a new hidden Markov model-based approach, shown schematically in Fig. 5.2(B), that we believe to be easier to use and more reliable than previously described alternatives, as it does not require training data as does the HMM algorithm of [161, 162], and uses an entire trajectory, not just local information, to compute the most likely sequence of states, in contrast to the change-point algorithm

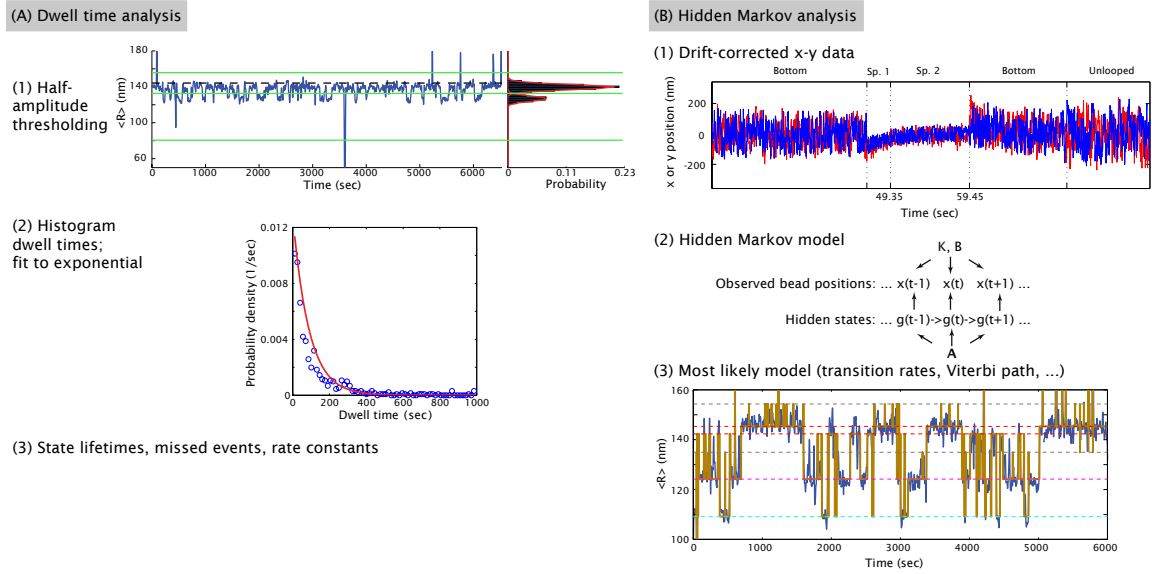


Figure 5.2: Two methods for obtaining kinetic information from TPM data. **(A)** Half-amplitude thresholding and dwell time analysis. In this approach, Gaussian-filtered root-mean-squared (here called $\langle R \rangle$) versus time traces are thresholded (solid green lines) to turn the noisy signal into a step function: in the example here, RMS values between the top threshold and the middle threshold are assigned to the unlooped state, and RMS values between the middle threshold and the bottom one are assigned to the looped state. The thresholds are usually chosen at the minima between two Gaussians fit to the trace histogram (red lines on the right). After every point has been assigned as either unlooped or looped (see Section 5.A.1 below for details on how to deal with spurious events, that is, time points whose RMS is above the top threshold or below the bottom threshold), the lengths of the dwells in each state are histogrammed, as in the bottom panel (blue points, here for the unlooped state). These histograms are made from dwell times aggregated over all the traces in a data set, not on a trace-by-trace basis. Single or double exponentials are fit to this histogram (red line, here a single exponential), which yield a time constant τ (or two time constants for a double exponential fit) that represents the mean lifetime of the state (in this case, the τ from the fit would represent the mean lifetime of the unlooped state). The results of these fits can then be used to calculate missed short-lived events, rate constants, and other information. **(B)** Our hidden Markov (HMM) approach begins not with Gaussian-filtered RMS data but with drift-corrected x and y bead positions (red and blue data in the top panel). The eventual state assignment from the HMM algorithm is given above this panel; note that without the Gaussian filtering step, it is difficult to detect looped versus unlooped states by eye. The removal of the Gaussian filtering step improves the temporal resolution of the data, though the need to drift correct still leads to some filtering artifacts, such as the slow trend towards the origin during the sticking event that occurs near 50 seconds in this trace. This filtering artifact leads the HMM algorithm to assign two distinct spurious states (“Sp. 1” and “Sp. 2”) to the sticking event. The x and y positions in the top panel, the observables, are assumed to arise from an underlying Markovian process represented by the sequence of hidden states in the middle panel. These hidden states correspond to the state of the tether (looped, unlooped, stuck) that we cannot directly observe. The matrix \mathbf{A} contains information about the transition rates between the hidden states; the K and B parameters characterize the distribution of bead parameters that arise from a given hidden state and lead to the observed bead positions. The HMM algorithm finds the most likely number of states in the observed x and y data, as well as the most likely \mathbf{A} and the most likely K and B for each hidden state, and computes both a most likely state for each time point, and a most likely sequence of states (called the Viterbi path) for the entire trajectory. The Viterbi path for a particular trajectory is shown in orange in the bottom panel, superimposed on the $\langle R \rangle$ trace that is the easiest way for us to visualize the trajectory (even though that information is not used in HMM). Colored horizontal lines indicate the $\langle R \rangle$ values calculated from the K and B parameters (see Section 5.A.3) for each genuine state detected; gray lines correspond to spurious events such as the sticking event shown in the top panel.

[159].¹ This HMM analysis will be a key component of the work in the next chapter, as it facilitates the identification of states in trajectories with more than two looped states (especially in cases where it is not known how many states to expect *a priori*).

5.1 Kinetics of looping by the Lac repressor by conventional methods

We begin by calculating state lifetimes by the conventional thresholding method, schematized in Fig. 5.2(A), and discuss what these lifetimes tell us about the number of short-lived events that we miss and the possibility of direct interconversions between looped states. The details of this approach are given in Section 5.A.1. In Section 5.2 we will then consider the hidden Markov model analysis which we believe will be an improvement over the thresholding method.

5.1.1 State lifetimes and missed events

As shown in Fig. 5.2(A), after traces are thresholded, the next step is usually to create dwell time histograms for each state and fit for the time constant τ , which is taken to represent the mean lifetime of the state. However, we have found it informative to also calculate the simple mean value of all the dwell times for a state, as an alternate way of defining the average state lifetime. The difference between these two approaches can be seen by comparing Fig. 5.3(A) versus (C), and (B) versus (D). The mean dwell times for the unlooped states for the E8- and TA-containing DNAs of Fig. 4.1(D–E) are plotted as a function of repressor concentration in Fig. 5.3(A), and the τ parameters from single exponential fits to the corresponding dwell time histograms are shown in Fig. 5.3(B). The values of τ and of the mean dwell time are comparable at high repressor concentrations for most of the DNAs; but the mean dwell time is significantly higher than τ at low repressor concentrations. We believe this is due to the two populations of unlooped states that we see at low repressor concentrations, examples of which are shown in the top traces of Fig. 5.1. The decay constants to the exponential

¹The HMM analysis described here was begun by Martin Lindén while he was a postdoctoral scholar in the Phillips lab, in collaboration with Chris Wiggins' group at Columbia University, and is continuing to be developed by Martin Lindén, now at Stockholm University.

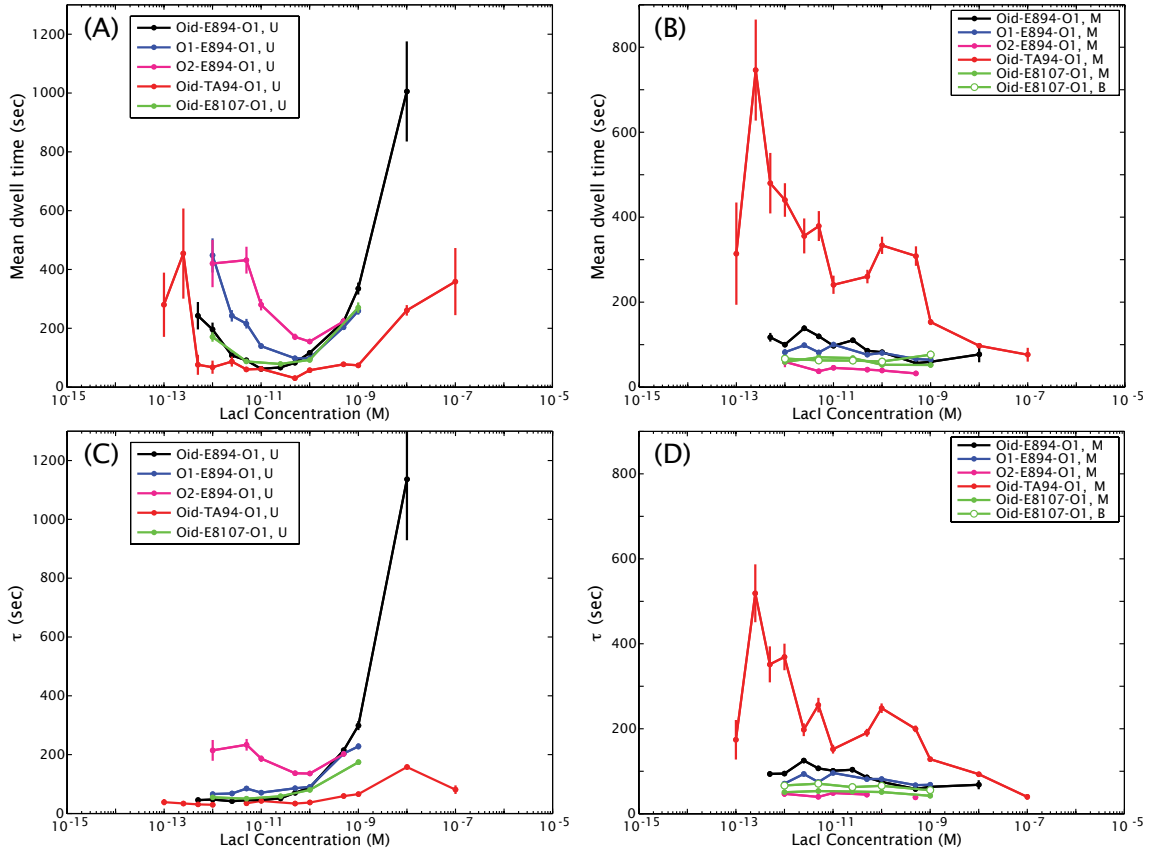


Figure 5.3: Mean state lifetimes and exponential decay constants from dwell time histogram fits for the constructs whose looping probabilities are shown in Fig. 4.1(D–F). **(A)** Mean dwell time of the unlooped state, as a function of concentration, calculated as a simple mean of all the observed unlooped dwell times for the indicated DNAs. The mean dwell time of the unlooped state varies with repressor concentration, as we might expect, because the probability of entering the looped state should depend on how many free repressors are available to bind. **(B)** Mean dwell time of the middle and bottom looped states as a function of concentration. Unlike the unlooped state, the mean dwell times for the looped states are roughly constant with concentration, as might be expected because the rate of exiting the looped state should not depend on the repressor concentration (except for the TA construct, which is discussed in the text). **(C)** Decay constants τ from single exponential fits to dwell time histograms for the unlooped states of the indicated DNAs, another way of defining the average dwell time. At high concentrations, the mean dwell time as calculated in (A) and the decay constants from the fits are comparable; but at low concentrations the mean dwell time is significantly larger than τ . This is because at low concentrations short bursts of looping and unlooping events are interspersed with long dwells in the unlooped state, presumably where no repressors are bound to the DNA (see Fig 5.1), leading to two populations of unlooped dwell times. Exponential fits are sensitive mostly to short events and not the long dwells in the tails of the distribution (see also Fig. 5.8 below). Less variation with concentration is observed for τ for the unlooped state, indicating that during the bursts of looping at low concentration, the kinetics of the unlooped state are roughly the same as those of the unlooped state at high concentrations. **(D)** Decay constants for single exponential fits to the middle and bottom looped states. τ agrees well with the mean dwell time calculated in (B), again with the exception of TA, discussed in the text.

fits describe mostly the short dwell lifetimes (since, as noted in Section 5.A.1 below, we obtained the best fits with single exponentials for all dwell time histograms, including for the unlooped state at low repressor concentrations); but the mean dwell time accounts also for the very long dwells in the unlooped state that occur at low repressor concentrations. The fact that τ is roughly constant with concentration for the unlooped states indicates that the looping dynamics for the short-lived unlooped state at low concentrations is comparable to the dynamics of the unlooped state at high concentrations, which supports our hypothesis that the long-lived unlooped dwells at low repressor concentrations occur when no repressor is bound and a new repressor must diffuse onto the DNA, whereas the short-lived unlooped dwells at low repressor concentrations, with transitions to the looped state interspersed between them, are composed of the state with one operator bound by a repressor. (Another piece of evidence for this hypothesis is given in Section 5.2.2 below.)

On the other hand, both the mean dwell time and τ for the looped states for most of the constructs are roughly constant with repressor concentration, which makes sense as the time spent in the looped state should depend only on the stability of the loop, which is independent of how many repressors are in solution. The data for the Oid-TA94-O1 construct are an obvious exception. We suspect that the behavior of the TA data stems from its large J-factor, which leads to long dwells in the looped state, which is especially problematic at low concentrations because a single hour-and-a-half long trajectory may show only one transition from looped to unlooped or vice-versa (see the examples in Fig. E.1 in Appendix E). This leads to poor statistics on the dwell time lengths, especially since we discard the first and last dwells of any trajectory (since we cannot know their true length, as we do not know when they began or ended aside from the limitations of the observation time). At mid-range concentrations, most of the TA trajectories are entirely in the looped state, which again leads to poor statistics. Probably the mean dwell times and τ 's at high concentrations are the only meaningful ones for this construct.

In addition to information about the average lifetimes of each state, the exponential fits to the dwell time histograms also give us information about how many short-lived events we miss due to the limited time resolution of our Gaussian-filtered data. As described in more detail in Section 5.A.1

below, following the convention of the field we define our temporal resolution as twice the dead time of the Gaussian filter. That means that for the 4 second Gaussian filters we use, we cannot resolve events shorter than 11 seconds. However, we know (or at least, we assume) the dwell times for a given state are exponentially distributed, so we can calculate how many events shorter than 11 seconds we should have observed, given the exponential fit to the rest of the dwell time histogram.

Following [109], we define the fraction of missed events, F , as

$$F = 1 - e^{-t_{min}/\tau}, \quad (5.1)$$

where $t_{min} = 11$ seconds (see Section 5.A.1 for details). The calculated fraction of missed events in the unlooped and looped states is shown in Fig. 5.4(A) and (B) for the constructs whose mean lifetimes were considered in Fig. 5.3. Fig. 5.4(C) shows looping probabilities corrected for these short-lived missed events, compared to the uncorrected looping probabilities that were presented in the previous chapter.

As can be seen in Fig. 5.4(C), the corrected and uncorrected looping probabilities are mostly within error of each other at high repressor concentrations, but deviate significantly at low repressor concentrations, leading to a loss of the inverse-U curve predicted by our model. We do not believe this deviation is due to the fact that we calculate more missed events, at least in the unlooped state, at low concentrations (Fig. 5.4(A)). Rather we believe that the correction scheme that we followed based on [109], which was applied in that work to the high-concentration regime, does not take into account the population of very long-lived dwells in the unlooped state that we see at low concentrations (Fig. 5.1). In Section 5.A.1 below we discuss a modification to the way that corrected looping probabilities are calculated, which takes into account some of these longer dwells. The result of this correction is shown in Fig. 5.4(D), which shows that accounting for some of the longer dwells restores some but not all of the inverse-U shape to the curve, by decreasing the looping probability at low concentrations (while leaving the high-concentration results the same).

As discussed in Section 2.1, data at low concentrations is important for determining dissociation

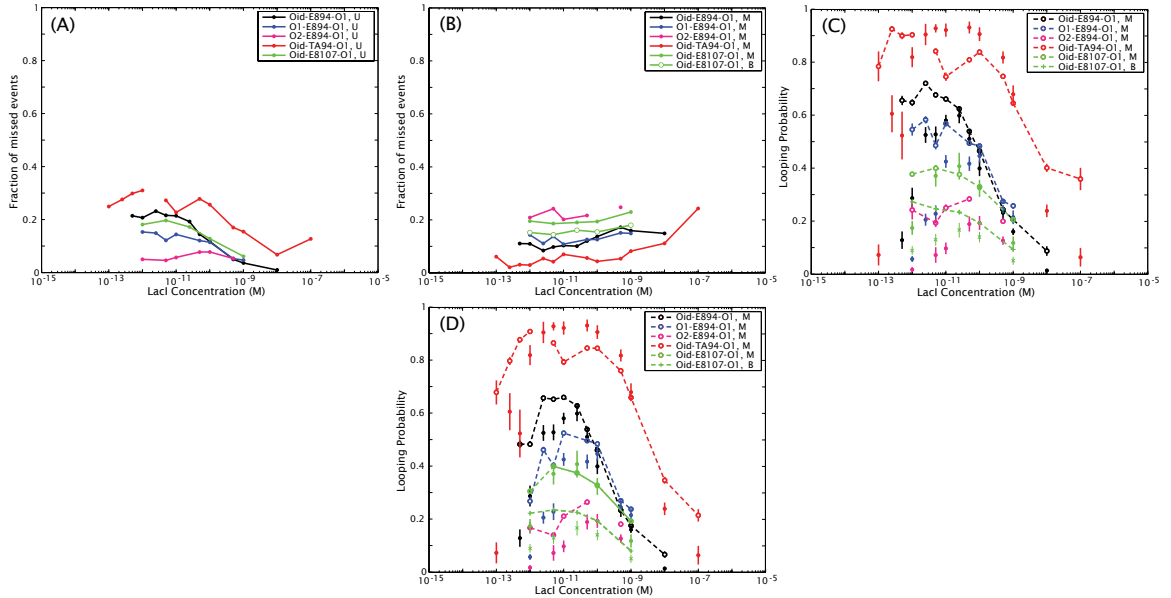


Figure 5.4: Missed events and corrected looping probabilities for the constructs whose (uncorrected) looping probabilities are shown in Fig. 4.1(D–F). **(A)** The fraction of missed short-lived unlooped events. The missing TA point is due to a fit for τ that did not converge. **(B)** The fraction of missed short-lived events for the middle and bottom looped states; again the missing point (for O2-E894-O1) is due to a fit that did not converge. **(C)** Looping probabilities corrected for the missed events in (A) and (B), according to the scheme in [109] (open circles for middle states, and “+”s for the bottom state of Oid-E8107-O1), as compared to the probabilities quoted in Chapter 4 (Fig. 4.1(D–F)) (closed circles for middle states, and “x”s for the bottom state of Oid-E8107-O1). The errors on the corrected looping probabilities are obtained from an error propagation formula as in [109] (rather than being the standard error on the mean of the distribution of looping probabilities for each concentration, as for the uncorrected looping probabilities). Dashed lines are not fits but are to show general trends; for clarity the fits to the uncorrected looping probabilities that are shown in Fig. 4.1(D–F) are not shown here, nor is the total looping probability for Oid-E8107-O1. The corrected looping probabilities are within error of the uncorrected probabilities at high repressor concentrations, but diverge sharply from the uncorrected values at low repressor concentrations. In fact the corrected looping probabilities do not show the inverse-U trend predicted by the model, but instead level out at low repressor concentrations. We believe these corrected looping probabilities at low repressor concentrations to be overestimates, caused by neglecting the long-lived population of dwells that we observe (see Fig. 5.1). **(D)** Corrected looping probabilities that account for more of the long-lived dwells at low repressor concentration. Here missed events were calculated as in (A) and (B), but the time in each state was supplemented by the time spent in the longest dwells (see Section 5.A.1 for details). The corrected looping probabilities at high concentrations remains relatively unchanged, but the looping probabilities at low concentrations decrease, bringing them into better agreement both with the uncorrected values and the inverse-U curves predicted by our model. We suspect that we still have not accounted for all of the long-lived dwells at low concentrations, and so we trust our uncorrected looping probabilities more than the corrected looping probabilities. (See text for details.)

constants from concentration curves, whereas the J-factor is determined by data at high concentrations. The fact that the corrected and uncorrected looping probabilities are comparable at high concentrations means that the J-factors we calculated from high-concentration data are not affected by the temporal resolution of the experiment. Moreover, we have reason to trust the uncorrected looping probabilities at low concentrations more than the corrected probabilities: we show in Chapter 4 that the dissociation constants we measured with the uncorrected data agree well with literature values (determined from ensemble biochemical assays that are not subject to the same temporal limitations as TPM). The “corrected” probabilities in either Fig. 5.4(C) or (D) would not yield values for the dissociation constants that would agree well with literature values. Therefore we conclude that, since in the regime where we trust the results of the missed-events correction the corrected and uncorrected looping probabilities are comparable, the lifetimes of the states generated by the constructs we studied here are sufficiently long enough, compared to our time resolution, to make a correction for short-lived events unnecessary.

5.1.2 Looping rate constants

The τ 's that were calculated in the previous section are related to the rate constants for transitions between states, though not always through a simple expression. Consider a simplified looping reaction in which we allow there to be only one unlooped state and one looped state, with rate constants k_{loop} and k_{unloop} for the transitions between the states. The dwell times for each of the two states will be exponentially distributed, and the decay constants τ_{loop} and τ_{unloop} that we would obtain from fitting single exponentials to histograms of these dwell times would be related to the rate constants by $k_{loop} = 1/\tau_{loop}$ and $k_{unloop} = 1/\tau_{unloop}$.

As shown in Fig. 5.5, however, the looping systems that we have considered in this work are generally more complicated than the simple two-state system described above, particularly because we cannot differentiate between each microstate of the system: we know there can be up to four unlooped microstates that generally are indistinguishable in the TPM assay (as they have the same RMS, aside from the DNA bending in the doubly bound state described in Section 4.A.1), and

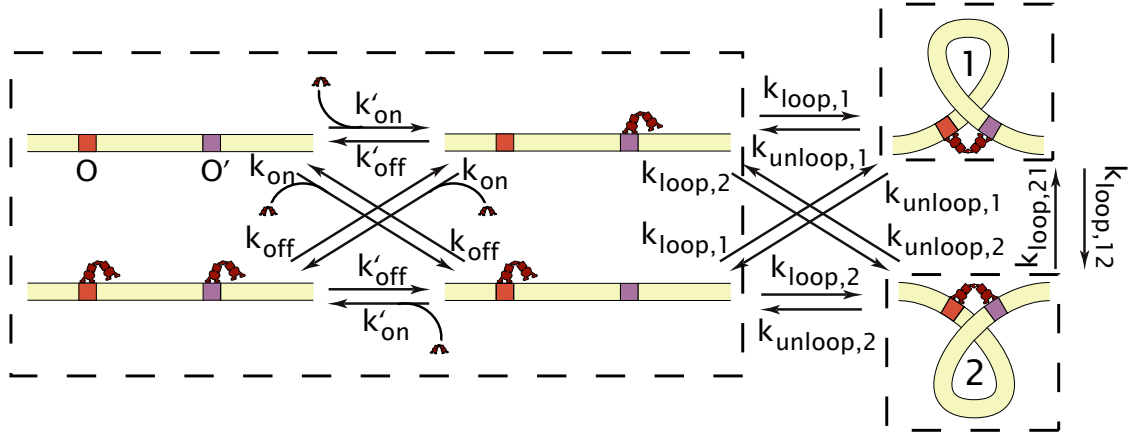


Figure 5.5: A kinetic scheme for looping by the Lac repressor. The six states shown here are the five states in Fig. 2.2(A), but with the two experimentally observed looped states separated out, and with the two operators labeled more generically as O and O' . Though the two looped states are distinguishable in TPM, here they have been drawn schematically the same, as we do not yet know the underlying molecular configurations of these states (see Sections 2.2 and 4.3). The boxes indicate microscopic states that correspond to the three experimentally observed states: for the most part we cannot distinguish the various unlooped states (though see Section 4.A.1 of the previous chapter and Section 5.2.2 below). It is likely that one or both of the looped states contains multiple molecular configurations as well (see the caption to Fig. 4.3). Transition rates between states are labeled as k_{on} and k_{off} , for the association of and dissociation to the O operator respectively, or k'_{on} and k'_{off} for the O' operator; $k_{loop,1}$ and $k_{unloop,1}$ for the looping and unlooping rates associated with loop state 1, and likewise for loop state 2; and finally, $k_{loop,12}$ and $k_{loop,21}$, transition rates between the two looped states. In Section 5.1.3 below we address the question of whether or not a kinetic scheme for the Lac repressor should allow such interconversion between the two looped states.

moreover we suspect there may be at least four loop configurations depending on the directions the operators are bound (see legend to Fig. 4.3) that may or may not correspond to the two looped states we observe. Even though we find single exponentials to fit best to our data (see Section 5.A.1 below), in reality we should find the unlooped state at least, and possibly the looped states as well, to be best fit by a mixture of exponentials characterized by up to four decay constants, if we had perfect temporal and spatial resolution. The decay constants we would obtain (even if we could obtain all four) would no longer be simply the inverses of the rate constants that we wish to find.

Attempts have been made to numerically simulate and/or fit for all of the rate constants in Fig. 5.5, at least for somewhat simpler constructs where $O = O'$ [109]. However most often hidden Markov model-based techniques are used to obtain rate constants for these more complicated schemes [111, 120]. Here we will simply note that there may be cases in our data in which the decay constants from the dwell time histogram fits are the inverses of the rate constants. Those cases are for the O1-E894-O1 construct at high (≥ 100 pM) repressor concentrations, where according to our statistical

mechanical model the unlooped state should consist primarily of the doubly occupied state, and there is only one looped state (the “middle” looped state). (It is possible that we might be able to perform this simplification for the other middle-state-only constructs as well, at sufficiently high concentrations; the convenience of the O1-E894-O1 construct is that both operators are the same, collapsing two of the unlooped states in Fig. 5.5 into one.) If this simplification is valid, then k_{unloop} and k_{loop} for O1-E894-O1 are both about 0.01 per second at 100 pM, and $k_{unloop} = 0.005 \text{ s}^{-1}$ and $k_{loop} = 0.02 \text{ s}^{-1}$ at 500 pM to 1 nM. These rates are in remarkably good agreement with those calculated by Wong and coworkers [109], for a construct that also has two O_1 operators (but a longer loop of a different sequence), at 5.4 nM: they found $k_{loop} \approx 0.02 \text{ s}^{-1}$ and $k_{unloop} \approx 0.007 \text{ s}^{-1}$.

5.1.3 Direct interconversions between looped states?

As noted in the introduction to this chapter, one of the questions we would like to address through a kinetic analysis of looping by the Lac repressor is whether or not the two looped states we see with some constructs directly interconvert, or whether transitions occur only between each looped state and the unlooped state. If we observe direct interconversions between the two looped states, it is more likely that the two looped states correspond to two different conformations of the Lac repressor, rather than different loop topologies, as converting between loop topologies should require at least one repressor head to unbind from an operator and rebind in a different orientation. If a repressor unbinds from an operator, the system will necessarily pass through the unlooped state before entering the other looped state.

One approach to assessing direct interconversions between the looped states is to calculate the partition ratios for the two states [109]. The partition ratio for one of the looped states (say, the middle looped state) is defined as the fraction of transitions from the middle looped state to the other looped state, divided by the fraction of transitions from the middle looped state to the unlooped state. A partition ratio of 1 would indicate equal preference for transitioning to either the other looped state or the unlooped state; a partition ratio of zero would suggest no direct interconversions; and a partition ratio greater than 1 would indicate that direct interconversions between looped states

	1 pM	5 pM	25 pM	100 pM	1 nM
$P_{MtoB/MtoU}$	0.45 ± 0.08	0.32 ± 0.05	0.38 ± 0.06	0.29 ± 0.05	0.19 ± 0.05
$P_{BtoM/BtoU}$	1.4 ± 0.3	1.3 ± 0.3	1.2 ± 0.3	1.0 ± 0.2	0.79 ± 0.3

Table 5.1: Partition ratios for the Oid-E8107-O1 construct, which has both the bottom and middle looped states, as a function of repressor concentration. A partition ratio significantly greater than zero indicates direct interconversions. These partition ratios have been corrected for short-lived missed events in both the unlooped and looped states, according to the scheme of [109]; in all but two cases (at 1 nM) the corrected partition ratios differed from their uncorrected values by less than 7%, and by less than 20% at 1 nM.

are favored.

Following the approach of [109], we calculated partition ratios for the Oid-E8107-O1 construct that has both the middle and bottom looped state, as a function of concentration. We corrected these partition ratios for missed short-lived events in both the unlooped and looped states, because our limited temporal resolution means that even though it looks in the RMS traces like the middle and bottom looped states directly interconvert, we could be missing short excursions to the unlooped state in between these transitions. (Unlike with the looping probabilities, here we must perform the correction for short-lived events; and the long-lived dwells at low concentrations do not come into this calculation and so the method of [109] should work equally well at all the concentrations we studied.)

The resulting corrected partition ratios are given in Table 5.1. Although all of the ratios are larger than zero, none of them are significantly larger than one, suggesting that direct interconversions could be occurring but they are not preferred. Our results are in quantitative disagreement with those of Wong and coworkers in [109], who found that direct interconversions between the looped states were not only possible but preferred. This discrepancy may be due to the fact that the looping probability for our E8107 construct, and especially the probability of the bottom looped state, is much lower than the slightly longer constructs that Wong and coworkers used. A low probability of the bottom looped state means that we do not observe many occurrences of that state, so our statistics are much lower than those of Wong and coworkers. A potentially more informative case study for assessing the potential of direct interconversions between the looped states might be some of the three-operator constructs in the presence of the DNA-bending protein HU described in the next chapter, which not only have much higher looping probabilities but clearly have long dwells that appear to involve direct interconversions between two or more looped states. It will be interesting to

see if missed short-lived events in the unlooped state are sufficient to account for these apparently direct interconversions or not.

5.2 Preliminary results with a hidden Markov model analysis

We are developing a hidden Markov model (HMM) based on variational Bayesian inference for obtaining kinetic information from TPM data, based on the work of (and done in collaboration with) Chris Wiggins and coworkers [163]. Details of our method will be published in Ref. [119]; a main difference between our algorithm and that of Wiggins and coworkers is in how we model the observable of the system (in our case, the motion of the bead), and so we discuss our model in some detail here and in the appendices to this chapter. In the first section below we present an overview of our approach (see also Fig. 5.2(B)) that we hope will serve as a general “users manual” for understanding the graphical user interface (GUI) that we will make available with our HMM algorithm and that we describe in the second section below. That section also presents an example of the kinds of analyses that we hope our approach will facilitate (another is given in Chapter 6).

5.2.1 Overview of a variational Bayesian hidden Markov model analysis of TPM data

The question we wish to address is: given a series of observations (bead positions as a function of time), what is the most likely series of underlying “states” that generated those observations? The states here can be thought of as the conformation of the DNA tether (looped, unlooped, etc) that leads to the observed bead positions.

A useful way to model such time-series data is as a hidden Markov process [164]. The process is “hidden” because we cannot directly observe the state of the DNA tether, only the output bead position; and while bead position and DNA conformation are certainly related, there is not a one-to-one correspondence between the two. We discuss this point in more detail in Appendix 5.A.3, but intuitively, a given state (e.g., a looped state) will generate a distribution of bead positions, not

a single bead position; and the distributions of observed bead positions for different states overlap [105, 140]. The process is Markovian because the hidden state (looped, unlooped, etc) at time t depends *only* on its position at the time $t-1$. (We note that the simplest class of Markovian processes are not actually valid for our system: first, the time series of observed bead positions is dependent not only on the underlying state at time t but also is itself a Markovian process, dependent on the bead’s position at time $t-1$, because the bead can only react so fast to changes to the underlying tether conformation. Therefore we have in fact implemented an *autoregressive* HMM, in which the bead’s position depends not only on the hidden state but on the position at time $t-1$, as can be seen in Eq. (5.2). Second, the hidden state is not independent of the observed bead position: for example, the distribution of potential observed bead positions for the unlooped state includes positions where the tether is maximally extended and from which it is physically impossible for a loop to form at the next time increment. We do not address that critique of TPM as a hidden Markov process here. See [162] for a “diffusive” hidden Markov (dHMM) approach to solving some of these problems.)

HMMs are a class of stochastic processes that can generate data of the kind we are interested in analyzing, namely, TPM time-series data. For the purpose of the discussion here we will consider a *particular* HMM to be of a specific size that corresponds to the number of hidden states it allows, and a set of three parameters listed below. One of the improvements of our HMM algorithm over previous ones [161, 162] is that we need not specify the number of hidden states at the outset; rather our approach will both pick the best number of hidden states, and the best HMM given that number of states. It is therefore a “maximum evidence” approach, as opposed to a “maximum likelihood” one, in which the goal is to select the most likely parameter values given an HMM of a particular size (number of states). Not only does this maximum evidence approach (algorithmically determining the number of states as well as the best parameter values) eliminate some user bias in determining the appropriate number of states, but it also avoids the over-fitting that is common to maximum likelihood [163]: in maximum evidence, overly simplistic models are unlikely because they cannot describe the data well, but overly complex models are also unlikely because there are too many data sets that could come from a too-complex model, and so the probability that *this* particular data were

generated by a particular model decreases. In contrast, the maximum likelihood always improves with more states.

A particular HMM with N hidden states will consist of the following three components, which we will collectively call the parameters of the HMM, $\vec{\theta}$:

(1) A transition matrix \mathbf{A} , which is an N -by- N matrix where element $a_{i,j}$ is the probability of ending in state j if the previous hidden state was state i (that is, $a_{i,j} = p(g_t = j | g_{t-1} = i)$, where g_t is the hidden state at time t). Each row of \mathbf{A} is normalized such that the probability of transitioning to some state at the next time point is one; that is, $\sum_j \mathbf{A}_{i,j} = 1$.

(2) A distribution of initial conditions—that is, the probability that the first hidden state is any one of the N possible states, or $p(g_1 = j)$.

(3) A set of parameters (called *emission parameters*) that describe the probability distribution of observed bead positions given a hidden state g_t . The physical model of the bead’s motion that we use is described in detail in Section 5.A.3 below; here we will mention its main features briefly. We model the motion of the tethered bead with a term related to having a particle diffusing in a harmonic well, and a term that adds Gaussian (white) noise, and parameterize the distribution of the bead’s position with two emission parameters K_j and B_j for a given hidden state j . Specifically, if \vec{x}_t is the bead’s position at time t (consisting, as above, of an x-coordinate and a y-coordinate), then we model the bead’s position as related to the position at time $t - 1$ as

$$\vec{x}_t = K_j \vec{x}_{t-1} + \frac{\vec{w}_t}{\sqrt{2B_j}}, \quad (5.2)$$

where K_j (unitless) and B_j (units of nm^{-2}) relate the bead’s position to $g_t = j$, and \vec{w}_t is a two-dimensional vector whose elements are drawn from a Gaussian distribution with mean 0 and variance 1. The term with K_j corresponds to a particle diffusing in a harmonic well: the values

that K_j adopts are between 0 and 1, such that successive time points are forced closer to the origin (because after each time point, the bead’s position is multiplied by some fractional value, less than 1). The term with B_j is the noise term. As we derive in Section 5.A.3, K_j and B_j are related to the root-mean-squared motion and the correlation time (a measure of the time it takes the tether to explore its configurational space) of the tether by

$$\langle \bar{x}_t^2 \rangle = \frac{1}{B_j(1 - K_j^2)} \quad (5.3)$$

and

$$\tau_{c,j} = \frac{-\delta t}{\ln K_j}, \quad (5.4)$$

respectively. As can be seen from Eq. (5.4), K_j increases with longer tether lengths; the relationship between B_j and tether length is more complicated to derive, but generally B_j decreases with increasing tether length.

If we have an HMM (that is, a transition matrix, initial conditions, and parameters that describe the distribution of observables for every hidden state), we can then obtain the most likely sequence of hidden states given a data set of observed outcomes. This is a process called *inference* (or, in the language of our forthcoming paper on this HMM approach [119], the “VBE step” (variational Bayes-Expectation step)). On the other hand if we were to somehow know the underlying sequence of hidden states for a given data set, it would be a simple matter of standard parameter fitting (called *learning* in this context, or again in the notation of [119], the “VBM step” (variational Bayes-Maximization step)) to obtain the best HMM (the transition matrix and other parameters contained in $\vec{\theta}$) that could generate that sequence of hidden states and corresponding observables.

Of course at the outset we know neither the hidden state sequence nor the HMM, and so we use an iterative process to obtain both given only the data to start with. In our case, we begin by assuming a large number of states, far more than we know should exist—say, 50 hidden states.² We

²In contrast to other approaches of this kind, our algorithm easily handles spurious events such as the transient sticking event shown in Fig. 5.2(B), where the bead has temporarily and nonspecifically adsorbed to the surface, the DNA has adsorbed to the bead, or the DNA has adsorbed to the surface. Our algorithm detects such events and assigns a hidden state to them, as with any genuine state, with some K_j and B_j parameters. As will be discussed in Section 5.2.2, K and B parameters for spurious states tend to be so different from those of genuine states that they

then initialize an HMM with “guesses” for the parameters $\vec{\theta}$, and use inference to obtain the hidden state sequence, given this initial HMM. Next we use an iterative process to learn a new HMM, given that hidden state sequence, then infer a new hidden state sequence, then learn an new HMM, and continue until the parameters of the model change on a new iteration by less than some tolerance (which is called *convergence*). At the end of this iterative process we will have a best guess for a 50 state HMM, along with an estimate of how well it describes the data. We then remove the least populated state from this 50 state model, and begin the iterative process again, until removing additional states does not improve how well we believe we are modeling the data.

For the purpose of introducing some important terminology, we note that the process of finding the best-sized model (the optimal number of states) is done by maximizing the *evidence*, or the probability of the data given a model, $p(\vec{x}_{1:T}|N)$, where $\vec{x}_{1:T}$ represents the entire time trace (that is all \vec{x}_t for $t = 1$ to $t = T$). We express the evidence in terms of the *likelihood* and the *prior*. The likelihood is the probability of the data and the number of hidden states given a particular model and the parameters of this model, $p(\vec{x}_{1:T}, g_{1:T}|\vec{\theta}_N, N)$, where $g_{1:T}$ is the sequence of hidden states for the entire trajectory. The prior contains our previous beliefs about the probability of a particular model and particular parameters $\vec{\theta}$ before we have seen the data, $p_0(\vec{\theta}_N, N|u)$, where u are the *hyperparameters* that characterize the prior distribution (e.g., a mean and a variance if we assume the prior distributions are Gaussian). These hyperparameters are chosen by the user at the start of the process. The evidence that we maximize, in terms of the likelihood and the prior, is

$$evidence = p(\vec{x}_{1:T}|N) = \int d^N \vec{\theta}_N \sum_{g_{1:T}} (likelihood)(prior). \quad (5.5)$$

The process of summing/integrating over all parameter values and all possible trajectories of hidden states is called “marginalization;” it represents the fact that instead of trying to guess values for the parameters and hidden state sequence, which we do not know, we sum over all possible values.

are easily discarded. This saves the user a significant amount of pre-processing time, since spurious events need not be removed beforehand; also, our algorithm can detect very short-lived sticking events that are not readily detectable in the RMS traces and that would be difficult to remove beforehand anyway (but would nonetheless interfere with the detection of genuine states). This does mean, however, that even in cases where we assume there should be only three hidden states—unlooped and two looped states—the number of states that we need to use in our HMM might be larger, to allow for spurious states as well, which can increase the computation time.

The process of determining the most likely parameters for a given HMM with N states is done by finding the *posterior probability distribution*, the probability of a sequence of hidden states and a set of parameters, given the data and a model, or $p(g_{1:T}, \vec{\theta} | \vec{x}_{1:T}, N)$. However the expression for the posterior distribution is generally intractable computationally, and so the result is instead derived by optimizing a *trial distribution* that approximates the “real” distribution [163].

At the end of the iterative process described above, we will have (1) the optimal number of (both genuine and spurious) states to explain the data; (2) for each hidden state, parameters that best describe the distribution of observed bead positions given each hidden state (the K_j and B_j parameters); and (3) a transition matrix for the probability of transitioning from one state to any other state. We will also have the probability of the observed data at any time point being generated by a particular hidden state. From this we can obtain the sequence of most likely states; or, alternatively, the *Viterbi path*, the most likely sequence of states, which tends to be a better way to model the trace (than the sequence of most likely states) because it uses information across the entire trajectory, instead of only at each point. Often the transition rate matrix can be used directly to obtain rate constants; but in cases like the set of multiple unlooped microstates shown in Fig. 5.5 above that generate one distribution of observables, the Viterbi path may be used to make dwell time histograms and obtain lifetimes as we did above with the thresholding method. As we will see in the next section, our HMM approach may be able to distinguish some of the unlooped microstates that we believe should be present.

5.2.2 Examples of results

In addition to the code for our HMM analysis of TPM data, we will be making available a graphical user interface (GUI) that allows the user to visualize the results and to perform some post-processing, such as making or modifying state assignments (e.g., identifying “bottom” versus “middle” looped states, which is not currently automated).³ In this section we describe some of the key features of the GUI and show some examples of results from the analysis of some of the data discussed earlier

³All of the HMM code, including the GUI, was written by Martin Lindén.

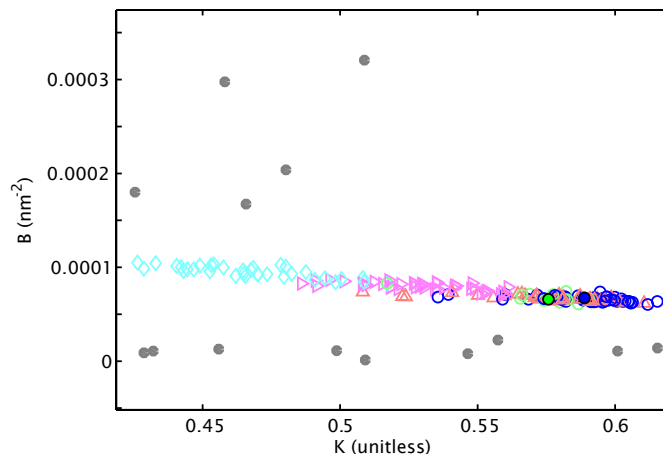


Figure 5.6: Emission parameters K and B for genuine (colored shapes) versus spurious (gray circles) states, for about 50 trajectories. K and B describe the distribution of bead positions that can be observed from a particular hidden state (looped, unlooped, stuck, etc). Though we represent the bead’s motion as being described by these two parameters, there is in fact only one independent parameter, the bead’s distance from its anchoring point that we measure. So it is unsurprising that for genuine states (looped or unlooped), the K and B parameters are linearly related and cluster along a line in K - B space. Spurious states, on the other hand, have different relationships between the K and B parameters than genuine states (for example, because the effective anchoring point may move during a sticking event); and so the spurious states scatter around the line that the genuine states form. (In this example from Oid-E8107-O1 at 25 pM Lac repressor, there are many more spurious states that fall outside this field of view).

in this chapter.

As mentioned above, even when we expect only two or three hidden states (corresponding to the unlooped state and one or two looped states), the best HMM found by our algorithm usually has five to ten states, because of the presence of spurious events in the trajectory, due to, for example, sticking events in which the bead (or the DNA) transiently and nonspecifically adsorbs to the surface (or the DNA to the bead, or to the surface). Fig. 5.2(B) shows a short-lived sticking event, most likely undetectable in the RMS trace, that nevertheless is best described by emission parameters that differ from the genuine states that precede and follow it.

The emission parameters for such spurious states are generally so different from those of genuine states that in fact they can be easily annotated. Fig. 5.6 shows the emission parameters for the genuine states and a subset of spurious states for the Oid-E8107-O1 construct at 25 pM repressor, in which the genuine states fall on a line in K - B space while the spurious states scatter around that line. In fact we have shown with other data that the line that genuine states cluster along is the same line that a set of “calibration” data, bead positions measured for constructs of varying DNA contour length in the absence of any Lac repressor, form in K - B space [119]. This tells us first of all that

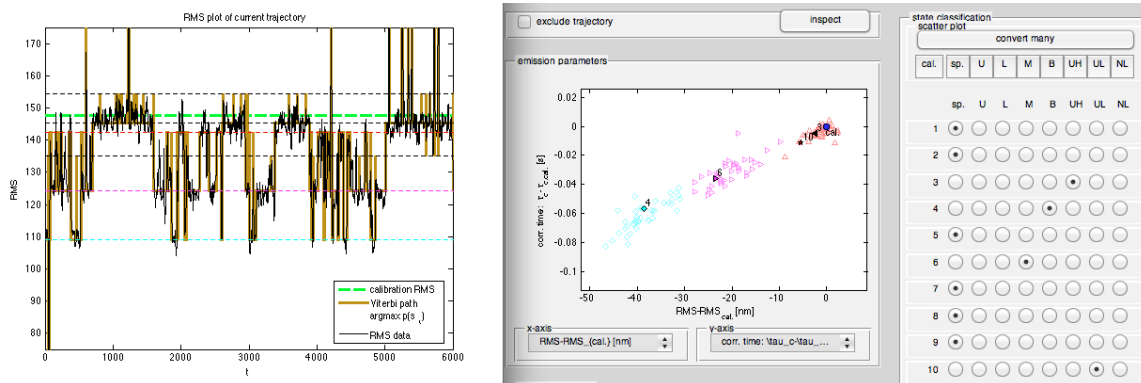


Figure 5.7: An example from the GUI that presents the HMM data. On the left is a particular trajectory from this data set (Oid-E8107-O1 with 1 pM Lac repressor). The RMS-versus-time data (not used in the HMM algorithm but as a visual aide) is shown in black, with the most likely sequence of states (the Viterbi path) superimposed in orange. Dashed horizontal lines indicate the calculated RMS for the hidden states identified by the algorithm: spurious states are in gray, the “calibration” state (obtained from data on this tether without Lac repressor) is shown in a thick green line, the bottom looped state in bright blue, and the middle looped state in magenta. In this case the algorithm identified two unlooped states, shown as red and black horizontal dashed lines, that correspond to the two kinds of unlooped dwells described in Fig. 5.1. The unlooped state that predominates during the long dwells between looping events has a longer RMS than the unlooped state that predominates during the bursts of looping, which we might expect because we see a reduction in RMS in the presence of bound Lac repressor that we attribute to bending of the operators (see Section 4.A.1). On the right is a plot of emission parameters, here the RMS and correlation time for each hidden state calculated from Eqs. (5.3) and (5.4), normalized to the corresponding values for the calibration data, to remove the bead-to-bead tether length variability we observe (see Section 4.A.1). This allows a clearer visualization of the clustering of K - B parameters for the different trajectories: the left-most light blue cluster corresponds to all the bottom looped states (the highlighted and numbered one corresponds to the parameters for the bead whose trajectory is shown to the left), the middle magenta cluster to the middle looped states, and the rightmost orange cluster to the unlooped states. The radio buttons to the right of the emission parameter plot allow the user to change the assigned state for any of the 10 identified hidden states for this trajectory.

looped states can be well described by emission parameters for an equivalent unlooped construct but with a shorter overall tether length. Second of all, it allows us to automate the identification of spurious states, so that when the data are presented to the user in the GUI, the user must only assign “bottom” versus “middle” looped states.

Fig. 5.7 shows how the GUI allows the user to assign states, by selecting radio buttons to the right of a plot of the emission parameters for either a single trajectory or all of the trajectories in a set. This emission parameter plot can show the K and B values for each state, or the RMS and correlation times for each state calculated from Eqs. (5.3) and (5.4). The low-concentration data set shown in Fig. 5.7 is particularly interesting because in some trajectories the HMM algorithm identified two unlooped states, with similar but not identical K and B parameters. One of these states characterizes the long dwells in the unlooped state introduced in Fig. 5.1 at the beginning of this chapter, and the other characterizes the unlooped dwells that intersperse the bursts of looping.

These two unlooped states have different calculated RMS values, with the one during the longer dwells having an RMS closer to that of the tether in the absence of protein; given that we believe we can see operator bending when a repressor is bound (Section 4.A.1), this supports our hypothesis that the long unlooped dwells correspond to periods where no repressor is bound, and that bursts of looping occur when a repressor has bound one operator (and ends when the repressor diffuses away again). The example in Fig. 5.7 also exhibits the two looped states that may or may not interconvert. Not only can we measure, as we did with the thresholding approach, how many (if any) direct interconversions between the two looped states are in the Viterbi path, but we are also working on a modified algorithm that specifically allows or disallows direct interconversions, by altering the prior distributions, so that we can assess how well an HMM with or without direct interconversions describes the data. We are hopeful that our HMM approach will be able to shed additional light on the dynamics of subpopulations of unlooped states, and the question of interconversion between the two looped states.

5.3 Conclusion

In this chapter we have looked at TPM data from a different standpoint than that of the previous chapters, namely from the standpoint of kinetics rather than looping probabilities alone. We presented two methods for obtaining kinetic information from TPM data: a thresholding approach that is commonly used in the field, and a newly developed variational Bayesian hidden Markov model analysis. In both cases we have focused our preliminary work on the two topics raised in the introduction to this chapter, the dynamics of the two populations of unlooped states that we observe at low repressor concentrations, and whether or not the two looped states interconvert. In the case of the former, we find that our HMM algorithm does in some cases identify two unlooped states, and that when it does, the long-lived unlooped state has a longer RMS, consistent with the long-lived state corresponding to having no repressor bound at either operator, whereas the shorter-lived unlooped state has a shorter RMS which is consistent with a repressor bound at an operator and inducing bending. In the case of the latter issue of direct interconversion between looped

states, although the thresholding analysis indicates that direct interconversions could be occurring, the looping probability, especially for the bottom state, is low enough to make gathering enough statistics for a solid conclusion difficult.

In the next chapter we present data on a more complicated, three-operator system, and will argue that an approach like that of HMM not only facilitates the data analysis but is necessary to identify how many looped states we observe in each trajectory. Some of the data in Chapter 6 not only show behavior suggestive of direct interconversions between two looped conformations that are longer-lived (and hence easier to study) than the E8107 states discussed in this chapter, but that are perhaps also suggestive of the direct interconversion among loops formed by different operators.

5.A Appendices to Chapter 5

5.A.1 Obtaining kinetic information from dwell time histograms

Trajectories were thresholded as described in Section D.2.3, which allowed every time point in an RMS trace to be assigned a state: “U”, unlooped; “M”, middle looped state; “B”, bottom looped state; or “Sp”, spurious. Time points were labeled spurious where the RMS value exceeded the highest threshold (and was therefore most likely due to a tracking error), or where it fell beneath the bottom threshold (and was therefore most likely due to a sticking event). If a spurious state was preceded and followed by the same genuine state, then we assumed the underlying genuine state of the system did not change during the spurious event and considered the flanking dwells plus the time spent in the spurious state to be one long dwell time. If the flanking states were not the same, however, we counted half the spurious event’s duration towards the preceding event, and half towards the succeeding event.

Following the convention in the field [104, 109, 113, 165], we ignored any dwells shorter than twice the dead time of the filter (defined in the next section), treating them as we did spurious events: if a transition occurred to a state whose duration lasted shorter than 11 seconds, and the states just before and just after this too-short dwell were the same, we counted the flanking dwells plus the time in the too-short dwell as one long dwell time in the flanking state. If a transition occurred to a third state after the too-short dwell, however, we split the too-short dwell between the preceding and succeeding dwells. Transitions that did not result in dwells longer than twice the dead time of the filter were not counted in the transition count matrix. Excluding too-short dwells was performed before the removal of spurious states (so too-short spurious states, as well as too-short genuine states, were ignored).

The result of this thresholding and dwell time calculation procedure on a given data set was a vector of dwell times for each state, and a transition count matrix, where each row corresponded to a pre-transition state, and each column to a post-transition state. The off-diagonal elements of this matrix, then, were the number of times a transition occurred from the state indexed by the

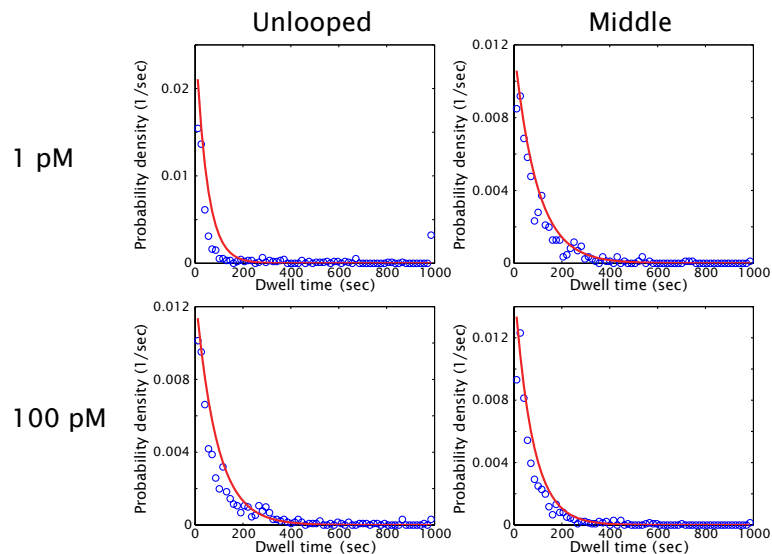


Figure 5.8: Sample dwell time histograms for the construct Oid-E894-O1 at two concentrations. Red curves are single exponential fits. Notice the right-most bin in the top-left histogram (unlooped state for 1 pM) has many more counts than those in the rest of the tail; these are due to the very long dwells in the unlooped state that we observe at low repressor concentrations (see Fig. 5.1).

row to the (different) state indexed by the column; the diagonal elements were simply the sum of the dwell times in each state (because the diagonal elements count transitions from state i at time t to the same state i at time $t + 1$). We normalized each row of this transition count matrix by the sum of the row, and multiplied by 30 frames per second to obtain the number of transitions per second. The off-diagonal transition count matrix elements are then used as the uncorrected $C_{a \rightarrow b}$ (in the notation of [109]) that are used (along with the fraction of missed events, described below) to calculate the partition ratios for the number of transitions from a looped state to another looped state, versus from the looped state to the unlooped state (Section 5.1.3).

We followed the procedure of [109] to plot dwell-time histograms and obtain time constants for exponential fits to these histograms. Unlike in [109], however, our bin sizes were uniformly 15 seconds (instead of variable based on the construct). Also unlike in [109], we obtained the best fits to all distributions with single exponentials; in particular, the distribution of unlooped state dwell times from our data were not well fit by a double exponential. (The poor fit quality was obvious, as the double exponential fit usually resulted in a decay constant similar to that from the single exponential fit, and a second decay constant that was approximately zero and had a 95% confidence

interval of $(-\infty, +\infty)$).

Following [109], we define the fraction of missed events in a given state, F_i , as

$$F_i = 1 - e^{-t_{min}/\tau_i}, \quad (5.6)$$

where $t_{min} = 11$ seconds (twice the dead time of the filter). Wong and coworkers then calculate the corrected time in a given state, D_i as

$$D_i = \frac{\tau_i N_i}{1 - F_i}, \quad (5.7)$$

where N_i is the total number of dwells in state i and τ_i is the decay constant for the exponential fit to the dwell time histogram of state i . We can then calculate the looping probability, corrected for missed short-lived events, as the corrected time in the looped state(s) divided by the corrected time in all states.

However, as shown in Fig. 5.4(C), we believe that these corrected looping probabilities overestimate the probabilities at low repressor concentrations because they neglect the population of long-lived dwells that we observe in the traces. As a simple initial attempt to correct for these long-lived dwells, we added the weight of the dwell time histogram bin with the longest lifetime to the D_i calculated above. That is, we added a term to D_i that included the time spent in the very longest dwells. As shown in Fig. 5.4(D), this modification to the approach of [109] brought the corrected looping probabilities into better (but we think still imperfect) agreement with the inverse-U curve predicted by our model. The weight of the bin with the longest lifetime obviously depends on how many bins of size 15 seconds we have; if we were to include fewer bins, more of the long-lived dwells would be included in the weight of the largest bin, and the added term to D_i would have a larger value. We believe that there should be a better way to handle this double unlooped population than the simplified modification presented here.

5.A.2 Calculating the dead time of a filter

The “dead time” of a filter refers to the duration of an event (looping or unlooping) that gives a half-amplitude response from the filter [158, 159]. The convention in the field is then to assume that the temporal resolution is twice the dead time [104, 109, 113, 165], that is, events shorter than twice the dead time cannot be resolved as true transitions between states instead of noise. In this section we will derive an expression for that dead time for the Gaussian filters that we use in this work (see Section D.2).⁴

In this derivation we will consider the true signal from TPM to be a step function, and neglect the noise that is superimposed on this signal (though that noise also contributes to the temporal resolution of the experiment, it is ignored when calculating the filter dead time). For simplicity consider a two-state system, and let state 1 be at RMS = 0, and state 2 at RMS = A . For an event from state 1 to state 2 back to state 1, where the dwell in state 2 lasts time T and is centered at $t = 0$, we can write the corresponding TPM trace as $A \cdot s_T(t)$, where $s_T(t)$ is 1 between $t = -T/2$ and $t = +T/2$, and zero elsewhere.

If we apply a Gaussian filter $g(t)$ with some standard deviation σ_g to the step-function “trace”, the sharp transitions from states 1 to 2 at $t = -T/2$ and from state 2 to 1 at $t = T/2$ will be smoothed, with the maximum of the filtered signal at $t = 0$, when the filter and underlying trace are aligned. We want that maximum value to become $A/2$ (a half-amplitude response from the filter). So the definition of the dead time of the filter, T_{dead} , becomes the condition that when the length of the dwell $T = T_{\text{dead}}$,

$$\int_{-\infty}^{+\infty} g(\tau) \cdot A \cdot s_T(\tau) d\tau = \frac{A}{2}. \quad (5.8)$$

Note that A can be canceled from both sides, so the dead time is independent of the signal’s amplitude. That is, the dead time of the filter does not depend on the difference in RMS between states.

⁴Thanks to Matt Johnson for the outline of this derivation.

Since $s_\tau(\tau)$ is zero except between $-T_{\text{dead}}/2$ and $T_{\text{dead}}/2$, Eq. (5.8) becomes

$$\int_{-\frac{T_{\text{dead}}}{2}}^{+\frac{T_{\text{dead}}}{2}} g(\tau) d\tau = \frac{1}{2} \quad (5.9)$$

where we have already canceled A from both sides.

Because $g(\tau)$ is a Gaussian, we can rewrite the integral on the left-hand side of Eq. (5.9) in terms of the cumulative distribution function of a Gaussian, usually given the variable Φ , where

$$\Phi(x) = \int_{-\infty}^x g(t) dt, \quad (5.10)$$

and whose solution is given by

$$\int_{-\infty}^x g(t) dt = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x}{\sqrt{2}} \right) \right]. \quad (5.11)$$

Note that Φ is defined for a Gaussian whose standard deviation is 1; but we are considering a Gaussian with standard deviation σ_g . So when we write the integral in Eq. (5.9) in terms of $\Phi(x)$, we must write it as

$$\int_{-\infty}^{\frac{T_{\text{dead}}}{2}} g(\tau) d\tau - \int_{-\infty}^{-\frac{T_{\text{dead}}}{2}} g(\tau) d\tau = \Phi(T_{\text{dead}}/(2\sigma_g)) - \Phi(-T_{\text{dead}}/(2\sigma_g)), \quad (5.12)$$

expressing T_{dead} in terms of the σ_g of our filter. Given the solution to $\Phi(x)$ above, we have our final result for the condition on T_{dead} ,

$$\Phi \left(\frac{T_{\text{dead}}}{2\sigma_g} \right) - \Phi \left(\frac{-T_{\text{dead}}}{2\sigma_g} \right) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{T_{\text{dead}}/(2\sigma_g)}{\sqrt{2}} \right) \right] - \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{-T_{\text{dead}}/(2\sigma_g)}{\sqrt{2}} \right) \right] = \frac{1}{2}, \quad (5.13)$$

which simplifies to

$$\left[\operatorname{erf} \left(\frac{T_{\text{dead}}/(2\sigma_g)}{\sqrt{2}} \right) \right] - \left[\operatorname{erf} \left(\frac{-T_{\text{dead}}/(2\sigma_g)}{\sqrt{2}} \right) \right] = 1. \quad (5.14)$$

We can look up that the solution to this expression involving the error function ($\operatorname{erf}(x)$) happens

when $\frac{T_{\text{dead}}}{2\sigma_g} \approx 0.67$, or that

$$T_{\text{dead}} \approx 2 \cdot 0.67 \cdot \sigma_g. \quad (5.15)$$

Eq. (5.15) now gives us an expression for the dead time in terms of the standard deviation of the Gaussian filter. In Eq. (D.5) we relate this standard deviation to the cutoff frequencies we choose during our data analysis. Given the 0.0326 Hz cutoff frequency at which we analyze most of our data (and all the data discussed in this chapter), we can calculate that the dead time of our filter is 5.5 seconds.

5.A.3 A physical model for the observable distributions

In most hidden Markov analyses, including in the work by Bronson and coworkers on which our algorithm is based [163], it is assumed that the output observables (in our case, bead positions) for a given hidden state are Gaussian distributed and independent of observables at previous times. Therefore the emission parameters, the parameters that describe these observable distributions, are typically a set of means and variances for the Gaussians that characterize each hidden state. In our case, however, we know that the distribution of bead positions for a given hidden state, corresponding to a particular effective tether length, is not a Gaussian, because the restoring force from the DNA tether is not well characterized as that of a linear spring at short DNA lengths [105, 140, 162]. Because the emission parameter distributions we use are non-standard, and also because the K and B parameters that characterize these distributions are presented to the user in the GUI described in Section 5.2.2, we discuss the emission parameter model in some detail here, and develop some intuitions about the relationship of the K and B parameters to the observed motion of the bead.

What we want to obtain is a physical model of the bead's position for a given tether length, which will allow us to parameterize the observable distributions in a way that includes more details of the bead's motion than by simple time-independent Gaussians. Because in our TPM experiments we image the tethered beads from the top down, and so only track the beads in two dimensions, we will model only the bead's x and y positions; that is, the data to which we apply the HMM analysis will consist of a time series of two-component vectors, $\vec{x}_t = (x_t, y_t)$, that measure the bead's distance

from the anchoring point at every time point t .

We model the motion of the tethered bead with a term related to having a particle diffusing in a harmonic well, and a term that adds Gaussian (white) noise. The term related to a particle in a harmonic well accounts for the restoring force of the tether bringing the bead back to the anchoring point; and the noise term randomizes the bead’s position [140, 166]. Although an improvement over modeling the emission probability distribution as purely Gaussian, we note here that this model still does not capture the true motion of the bead—for example, there is not a positive probability of the bead’s position being more than some distance (determined by the length of the fully extended DNA tether) from the origin [162, 167]—but the approximation presented here was deemed sufficient.

If \vec{x}_t is the bead’s position at time t (consisting as above of an x-coordinate and a y-coordinate), then we will model the bead’s position as related to the position at time $t - 1$ as

$$\vec{x}_t = K_j \vec{x}_{t-1} + \frac{\vec{w}_t}{\sqrt{2B_j}}, \quad (5.16)$$

where K_j (unitless) and B_j (units of nm^{-2}) are the emission parameters that relate the bead’s position to the underlying hidden state j at time t (that is, $g_t = j$), and \vec{w}_t is a two-dimensional vector whose elements are drawn from a Gaussian distribution with mean 0 and variance 1. The term with K_j corresponds to a particle diffusing in a harmonic well: the values that K_j adopts are between 0 and 1, such that successive time points are forced closer to the origin (because after each time point, the bead’s position is multiplied by some fractional value, less than 1). The value of K depends on j , that is, the hidden state or effective tether length, because in the low-force regime of TPM, the DNA tether can be modeled as a spring with a linear restoring force in tether length. The term with B_j is the noise term, again dependent on the hidden state j , though not as intuitively as K_j (see discussion at the end of this section).

To gain insight into how K_j and B_j relate to aspects of the bead’s motion that are more familiar to us, we next derive the mean-squared motion of the bead based on this model (the square of the RMS motion that is plotted elsewhere in this work, e.g., in the sample trajectories in Appendix E

and Fig. 5.1), and the correlation time of the bead, for a given hidden state j . That is, for the purposes of these calculations, we will assume that the hidden state does not change. We will see that K_j is related to the correlation time of the bead, and a combination of K_j and B_j gives us the RMS motion of the bead.

By definition the mean-squared motion of the bead is $\langle \vec{x}_t^2 \rangle$, or

$$\langle \vec{x}_t^2 \rangle = \left\langle \left(K_j \vec{x}_{t-1} + \frac{\vec{w}_t}{\sqrt{2B_j}} \right)^2 \right\rangle, \quad (5.17)$$

which we can expand to

$$\langle \vec{x}_t^2 \rangle = K_j^2 \langle \vec{x}_{t-1}^2 \rangle + \frac{2K_j}{\sqrt{2B_j}} \langle \vec{w}_t \cdot \vec{x}_{t-1} \rangle + \frac{1}{2B_j} \langle \vec{w}_t^2 \rangle, \quad (5.18)$$

where we have moved constants out of the averages. Next we note that $\langle \vec{w}_t \cdot \vec{x}_{t-1} \rangle = 0$ because \vec{w}_t and \vec{x}_{t-1} are independent; that $\langle \vec{x}_{t-1} \rangle = \langle \vec{x}_t \rangle$ and $\langle \vec{x}_{t-1}^2 \rangle = \langle \vec{x}_t^2 \rangle$, because the hidden state does not change and so the average behavior of \vec{x}_t is independent of time (the process is “stationary”); and that $\langle \vec{w}_t^2 \rangle = 2$ because \vec{w}_t is a two-dimensional vector whose components are Gaussian variables with variance 1, and are independent both of each other and of components at previous times, so that $\vec{w}_t^2 = w_{x,t}^2 + w_{y,t}^2 = 1 + 1 = 2$. Therefore Eq. (5.18) simplifies to

$$\langle \vec{x}_t^2 \rangle = K_j^2 \langle \vec{x}_t^2 \rangle + \frac{1}{B_j}, \quad (5.19)$$

which we rearrange to solve for the mean-squared motion of the bead,

$$\langle \vec{x}_t^2 \rangle = \frac{1}{B_j(1 - K_j^2)}. \quad (5.20)$$

The correlation function of the time series is, again by definition, given by $\langle \vec{x}_t \cdot \vec{x}_{t-1} \rangle$, which we

can write in terms of Eq. (5.16) as

$$\langle \vec{x}_t \cdot \vec{x}_{t-1} \rangle = \left\langle \left(K_j \vec{x}_{t-1} + \frac{\vec{w}_t}{\sqrt{2B_j}} \right) \cdot \langle \vec{x}_{t-1} \rangle \right\rangle, \quad (5.21)$$

and which we then expand to

$$\langle \vec{x}_t \cdot \vec{x}_{t-1} \rangle = \left\langle K_j \vec{x}_{t-1}^2 + \frac{\vec{w}_t}{\sqrt{2B_j}} \vec{x}_{t-1} \right\rangle. \quad (5.22)$$

As before we recognize that $\langle \vec{w}_t \cdot \vec{x}_{t-1} \rangle = 0$ and $\langle \vec{x}_{t-1}^2 \rangle = \langle \vec{x}_t^2 \rangle$, and also that we have an expression for $\langle \vec{x}_t^2 \rangle$ in terms of K_j and B_j in Eq. (5.20), so the correlation function becomes

$$\langle \vec{x}_t \cdot \vec{x}_{t-1} \rangle = \frac{K_j}{B_j(1 - K_j^2)}. \quad (5.23)$$

By comparing Eq. (5.23) with Eq. (5.20), we can see that we can rewrite Eq. (5.23) as

$$\langle \vec{x}_t \cdot \vec{x}_{t-1} \rangle = \langle \vec{x}_t^2 \rangle K_j. \quad (5.24)$$

By another definition of the correlation function, we can write the right-hand side of this equation in terms of an exponential,

$$\langle \vec{x}_t \cdot \vec{x}_{t-1} \rangle = \langle \vec{x}_t^2 \rangle e^{-\delta t / \tau_{c,j}}, \quad (5.25)$$

where δt is the time between measurements (30 ms in our experiments) and $\tau_{c,j}$ is the correlation time. Since we know $\langle \vec{x}_t^2 \rangle$ in terms of K_j and B_j from Eq. (5.20), and $\langle \vec{x}_t \cdot \vec{x}_{t-1} \rangle$ in terms of K_j and B_j from Eq. (5.23), we can solve for the correlation time,

$$\tau_{c,j} = \frac{-\delta t}{\ln K_j}. \quad (5.26)$$

Intuitively we know that the correlation time should increase with longer tethers, because the correlation time is a measure of the time required to explore the configuration space of the tether (and

this space is bigger for longer tethers). So K_j increases with longer tethers. B_j has a more complicated relationship with the tether length, related to motional blur from the camera and the effect of hydrodynamic interactions with the wall on the drag on the bead. We note here simply that B_j tends to decrease with increasing tether length.

Chapter 6

The three operators of the wild-type *lac* system: A case study in combinatorial control

In the preceding chapters we have examined synthetic constructs that contain two binding sites for the Lac repressor, and have either unnatural or non-native sequences in the loop. While such constructs have provided important insights into the sequence dependence to short-length loop formation and the dynamics of the looping process, there in fact remain outstanding questions about the wild-type *lac* operon that these synthetic constructs cannot address (some of which are discussed below). In this chapter we will apply our combined TPM plus statistical mechanical model/kinetic analysis approach to the study of a set of constructs derived from the wild-type system (introduced in Chapter 1 in Section 1.5 and in Fig. 1.3), which include three operators rather than two, and contain the natural sequences in the loops.¹

Looping between two Lac operators as a function of interoperator spacing (i.e., loop length) as well as other parameters, both *in vivo* and *in vitro*, has been extensively studied (e.g., [32, 49, 50, 67, 68, 69, 70, 100, 101, 104, 108, 109, 113, 114, 115, 116]). However, few studies have been performed on systems containing all three operators, and none *in vitro*, where looping can be observed directly, instead of indirectly through its effects on gene expression. Two-operator studies have led to many hypotheses regarding the advantage conferred by two operators and therefore the ability for loops to form in transcriptional regulation (some of which are described in Chapter 1), but only one

¹This project was suggested by Jon Widom.

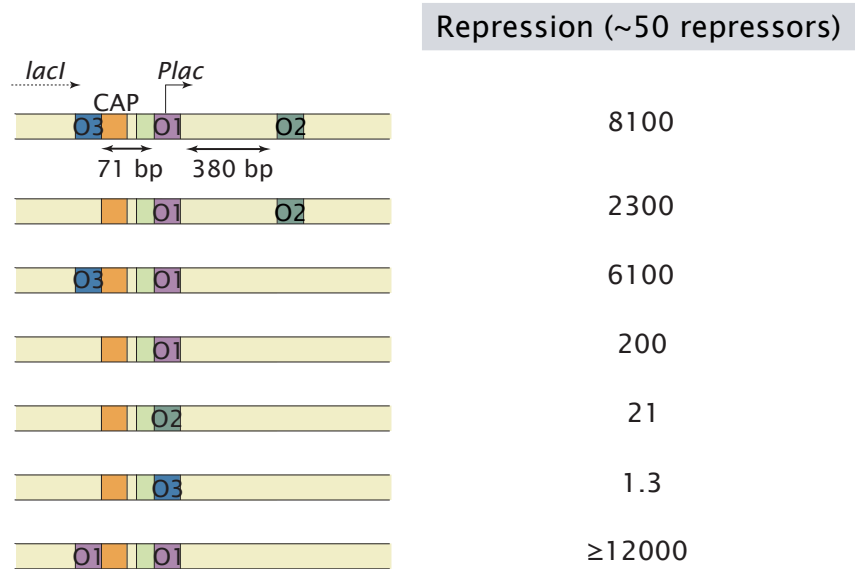


Figure 6.1: Wild-type levels of repression observed only with all three operators, as measured by *in vivo* β -galactosidase assays. In the DNA schematics to the left, the light green box that overlaps O1 is the *lac* promoter (related to the *lacUV5* promoter used in Chapter 4); the arrow labeled “*Plac*” indicates the direction of transcription of the operon genes and roughly the start site. The dashed, right-pointing arrow labeled “*lacI*” indicates that the gene that encodes the Lac repressor terminates near O3 (see Fig. 1.3). The orange box labeled “CAP” indicates the binding site for the catabolite activator protein which is the activator for this system, and is also known to bend the DNA of its binding site [168]. CAP was present in the experiments that are summarized by the data on the right. Repression of the expression of β -galactosidase (the product of the *lacZ* gene in Fig. 1.3) was measured in an *E. coli* strain harboring on average 50 Lac repressor molecules, as a function of the combinations of operators present (see Section 4.6 for a quantitative definition of repression). In a completely wild-type *E. coli* there are roughly 10 repressors per cell (see [93] for repression in the presence of wild-type levels of Lac repressor). Deletion of either auxiliary operator reduced the measured repression by one-quarter to three-quarters of its wild-type value. Moreover, as shown by the last row, wild-type levels of repression require the particular operator strengths of the wild-type system and (as shown in other data in [94] not shown here) their particular arrangement: for example, switching which operator overlaps the promoter in the third construct reduces repression by half. Note that repression is defined as the amount of gene expression in the absence of repressor divided by expression in the presence of repressor; therefore a repression of 1 (second to last row) is essentially no effect compared to the unregulated promoter. The error on these measurements is roughly 30%. Adapted from [94].

auxiliary operator, not two, is required for looping. Yet it has been shown by Oehler and coworkers, as demonstrated by the summary of their data in Figure 6.1, that both auxiliary operators as well as O1 are required for maximal repression by the *lac* operon *in vivo* [94]. In fact, not only the presence but the specific arrangement of all three operators has been shown to be necessary for wild-type repression: Oehler and coworkers demonstrated that, in the absence of O2, having two O_1 operators instead of O_1 and O_3 leads to unnaturally high levels of repression (see the last row of Fig. 6.1). Why this system should have evolved to contain not just one but two auxiliary operators, with their specific strengths and arrangement, remains a puzzle and outstanding question in the field regarding the wild-type *lac* system [39].

Several recent experimental and theoretical works have speculated on the role of the two auxiliary operators in the *lac* system. The most explicitly stated hypothesis about the presence of three operators is from Li and coworkers, who argued that auxiliary operators and DNA loop formation reduce the search time for transcription factors to find their specific binding sites on the *E. coli* genome [39], which can in turn contribute to a more efficient response in gene regulation to environmental cues. In their model, the presence of a single auxiliary operator can decrease target search time by a factor of 2 for low-copy transcription factors (which would include the Lac repressor, which is present in roughly 10 copies in the cell [90]); a second auxiliary operator decreases the search time by a factor of 3. This “antennae” effect mediated by DNA looping from auxiliary sites, as a means of increasing binding at the main operator, is preferential to simply increasing the number of Lac repressors in the cell and having only one binding site at the promoter, because of the crowding and road-blocking effects that result from having too many DNA binding proteins bound nonspecifically along the genome. Other recent theoretical and experimental studies have focused on the cooperativity between the Lac repressor and the CAP activator protein, known to bind in and bend the loop region between O_1 and O_3 (and thereby presumably to enhance the formation of the O_1 - O_3 loop) [169, 170], suggesting a special function of the O_1 - O_3 loop in allowing cross-talk between the activation and repression pathways of the operon.

The approach that we have developed here is particularly well suited to testing the predictions of these models for the role of two (rather than one) auxiliary operators in the *lac* operon. Not only do we employ an *in vitro* technique, which as noted above allows us to separate loop formation from its downstream effects on gene expression, but our technique is also a single-molecule one, by which we can resolve the loops that form between different combinations of operators, something that is difficult to do with other *in vitro* techniques such as gel shift assays. For example, the “antennae” model of Li and coworkers predicts a reduction in target-search time only if the rate of loop formation from both auxiliary sites is fast relative to other parameters of the system, something that we can explicitly test.

Beyond the particular question of the role of the three operators in the case of the *lac* operon,

we aim here also to begin to build up a model for how combinatorial control functions in other, more complicated systems. As discussed in Chapter 1, multi-loop regulatory regions are the rule rather than the exception in eukaryotic genes, and are probably common in prokaryotes as well (Fig. 1.1(A) and (C)). The *lac* operon is a particularly good initial case study for questions of combinatorial control. Not only does it have three binding sites for the Lac repressor, but it is also regulated by an activator, CAP, which as noted above is thought to cooperatively enhance looping by the Lac repressor, and thereby increase the sensitivity of the response of the Lac repressor to the inducer that removes it from the main operator so that transcription can occur [169]. We will show here that TPM, especially when combined with the hidden Markov model analysis of the previous chapter, should be able to dissect more complicated looping systems than the two-operator systems for which it is usually used.

As in previous chapters we begin by deriving a statistical mechanical model for a three-operator system, which in this case allows us not only to measure various parameters of the system through the concentration curves that are the main focus of Chapters 2, 3, and 4, but also to predict how looping in this system would change as we vary the K_d 's and J-factors away from their wild-type values. We then turn to experimental results with a three-operator DNA created directly from the regulatory region of the *lac* operon, as described in Appendix B.3. Finally, we use our model and some preliminary experiments with DNA-bending proteins to speculate on a surprisingly large difference between the *in vitro* results we present here and the classic *in vivo* work of Oehler and coworkers [94].

6.1 A statistical mechanical model of the wild-type *lac* system

A model that accounts for a three-operator construct can be derived in a manner analogous to that of [115] and Chapter 2. (See also Appendix A for a detailed derivation of another of this class of statistical mechanical models, and Ref. [170] for a similar three-operator statistical mechanical


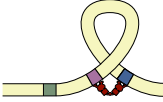



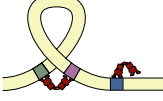
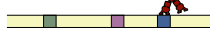
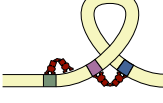





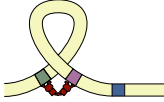
STATE	STATISTICAL WEIGHT	STATE	STATISTICAL WEIGHT
(i) 	1	(x) 	$\frac{1}{2} \frac{[R] J_{\text{loop},13}}{K_1 K_3}$
(ii) 	$\frac{[R]}{K_1}$	(xi) 	$\frac{1}{2} \frac{[R] J_{\text{loop},23}}{K_2 K_3}$
(iii) 	$\frac{[R]}{K_2}$	(xii) 	$\frac{1}{2} \frac{[R]^2 J_{\text{loop},12}}{K_1 K_2 K_3}$
(iv) 	$\frac{[R]}{K_3}$	(xiii) 	$\frac{1}{2} \frac{[R]^2 J_{\text{loop},13}}{K_1 K_2 K_3}$
(v) 	$\frac{[R]^2}{K_1 K_2}$	(xiv) 	$\frac{1}{2} \frac{[R]^2 J_{\text{loop},23}}{K_1 K_2 K_3}$
(vi) 	$\frac{[R]^2}{K_1 K_3}$		
(vii) 	$\frac{[R]^2}{K_2 K_3}$		
(viii) 	$\frac{[R]^3}{K_1 K_2 K_3}$		
(ix) 	$\frac{1}{2} \frac{[R] J_{\text{loop},12}}{K_1 K_2}$		

Figure 6.2: Schematized states and thermodynamic weights for a model that includes three operators. The nomenclature follows that of Fig. 2.2, although here, since there are three loops and therefore three J-factors to discriminate between, the J-factors are labeled with subscripts indicating the operators that a repressor must bind to form the corresponding loop (e.g., $J_{\text{loop},12}$ is the J-factor for the loop between O_1 and O_2). Note that these are total J-factors, but each loop has the potential to form two looped states (depending on the phasing of the operators at these lengths, as shown in Fig. 4.2), so there are actually six J-factors for this system and more than fourteen states. A model that separates out all of these looped states is a simple extension of the one presented here and is analogous to that of Section 2.2.

model applied to the *in vivo* data of Refs. [93] and [94] summarized in Fig. 6.1.) As with all of the models in this work, we start by enumerating the states that the system can be in, and deriving their corresponding weights, in terms of thermodynamic constants such as J-factors and K_d 's, as shown schematically in Fig. 6.2. With three operators in the system, there are now fourteen states, instead of the five states of the simple model. (As noted in the caption to Fig. 6.2, there are actually more than fourteen if the two looped states that can form from any given pair of operators are included as well, but we will not consider those here.)

We would next like to determine the three K_d 's and three J-factors for this system. To do so we note first that K_1 and K_2 are known already from the E8 and TA concentration curves of Chapter 4. To find K_3 and the J-factors for these loops, we constructed two-operator derivatives of the full three-operator construct (Appendix B.3), in which binding at either O_1 , O_2 , or O_3 was abolished. In principle, concentration titrations with these two-operator derivatives would have

allowed us to fit for K_3 and the three J-factors, using the method that we established in Chapters 3 and 4.

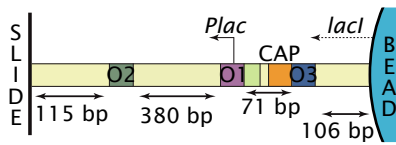
In practice, however, we found that we obtained no looping with either of the two-operator derivatives that contained O_3 as one of the two operators. This is perhaps not surprising: we can estimate the *in vitro* value of K_3 by noting that although absolute values of the *in vivo* K_d 's for these operators are quite different from their *in vitro* values (see Fig. 2 of [171], where *in vivo* $K_{id} \approx 130$ pM, $K_1 \approx 0.6$ nM, $K_2 \approx 2.7$ nM, and $K_3 \approx 200$ nM), the *ratios* of any two K_d 's are roughly the same *in vivo* and *in vitro*. Therefore by comparing Fig. 2 of [171] and Table 4.1 of this work, we can estimate the *in vitro* value of K_3 in our TPM experiments to be roughly 10–20 nM.² Even if we attempt to obtain K_3 by measuring the looping probability at $[R]_{\max}$ with a construct containing O_3 , O_{id} (the strongest known operator) and TA94 (which has one of the largest looping probability of the sequences we have examined), from Eq. (2.5) we predict a maximal looping probability of only 0.1, barely detectable in our assay.

This is already an important result. Note from Fig. 6.1 that all three operators, including O_3 , are necessary for obtaining wild-type levels of repression from the operon. Even with the O_1 - O_3 loop alone, which does not form in our *in vitro* assay, *in vivo* repression drops only by about a quarter compared to wild-type levels. The O_2 - O_3 loop apparently does not form *in vivo* [93]; but the O_1 - O_3 and O_1 - O_2 loops are roughly comparable in their contributions to looping (see Fig. 6.1, rows 2 and 3, and also [93]). We will return to this discrepancy between our *in vitro* data and the *in vivo* data of [93, 94] in the next section.

We can still obtain the J-factors for all three loops by replacing O_3 with O_1 in the various two-operator derivatives. We then know all of the parameters of these two-operator systems except the J-factors, since they contain only O_1 and/or O_2 . Although in principle data at only one concentration are sufficient for measuring J-factors, when both operators are known, in all but one case three concentrations were measured for each two-operator construct in order to reduce the error on the

² O_3 is difficult to measure *in vitro* because it is so weak, but its *in vitro* value has been estimated as 16 to 1000 times weaker than O_1 [99, 172, 173]. By that measure K_3 would be about 0.7–5 nM in our assay; 0.7 nM is too low of an estimate, because loops containing O_3 would be visible in our assay if K_3 were that low. It is possible that O_3 is closer to 5 nM than 10 nM, however, and so we use an estimate of 5 nM in some of the predictions in this chapter where a lower estimate of K_3 allows a better visualization of the predictions of the model.

(A) 3-operator TPM construct



(B) Concentration titrations

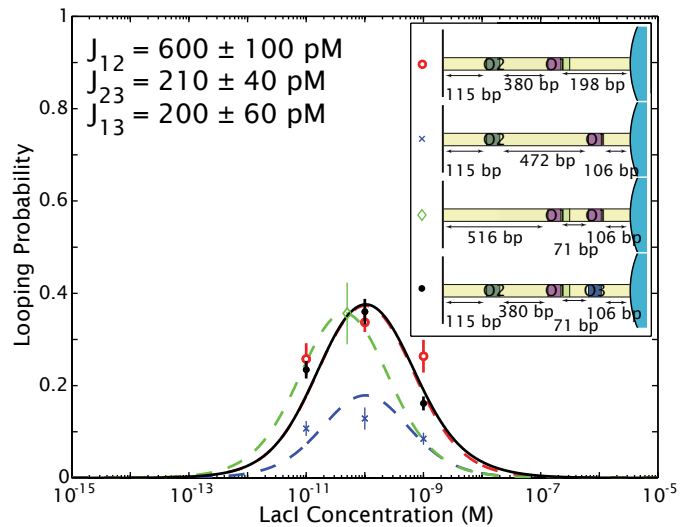


Figure 6.3: Looping with the three-operator, wild-type *lac* operon DNA, and with two-operator derivatives. **(A)** Schematic of the TPM construct. The linear DNA for these TPM constructs was extracted from *E. coli* directly by PCR (see Appendix B.3). As in Fig. 6.1, the light green box that overlaps O1 is the *lacZYA* promoter; the dashed, left-pointing arrow labeled “*lacI*” indicates that the gene that encodes the Lac repressor terminates near O3 (see also Fig. 1.3); and the orange box labeled “CAP” indicates the binding site for the CRP protein (not present in any of the work discussed here). **(B)** Concentration titrations with the three-operator construct in (A) (black), and two-operator derivatives, shown schematically in the legend (see Appendix B.3 for how these derivatives were obtained). The two-operator derivatives were used to obtain the J-factors for the various loops, and to examine the effect of having the O_3 operator present or absent. Three concentrations each were measured to obtain $J_{\text{loop},12}$ (red) and $J_{\text{loop},23}$ (blue). Only one concentration, however, was used to determine $J_{\text{loop},13}$ (green), at the maximum of looping as predicted by Eq. (2.3). This loop is so short (71 bp), relative to the total length of the tether (735 bp), as to be nearly undetectable, even at the maximum of looping; data at additional concentrations with even less looping would be difficult to obtain and probably not helpful in reducing the error on the J-factor for this loop. The J-factor obtained from this point should be considered an estimate only. Dashed lines indicate a global fit to the red, green and blue data simultaneously, enforcing the values of K_1 and K_2 from Table 4.1 obtained with the E894 and TA94 DNAs. The J-factors that result from these fits are given as total J-factors for simplicity, though we believe (see Fig 6.4) all three of these loops do show both looped states. With the results of these concentration curves here, all parameters of the three-operator model of Fig. 6.2 are known except K_3 . The solid black line in this figure is not a fit but a prediction of the three-operator model with all of the known K_d 's, and assuming $K_3 = 15 \text{ nM}$ (using $K_3 = 10 \text{ nM}$ or even 5 nM produces only a very small difference). Surprisingly, given the *in vivo* data of Fig. 6.1, the presence (black data) or absence (red data) of O_3 makes no detectable difference on looping *in vitro*. All looping probabilities were calculated by the thresholding method described in Appendix D.2.3, and nonloopers were subtracted according to Appendix D.2.5.

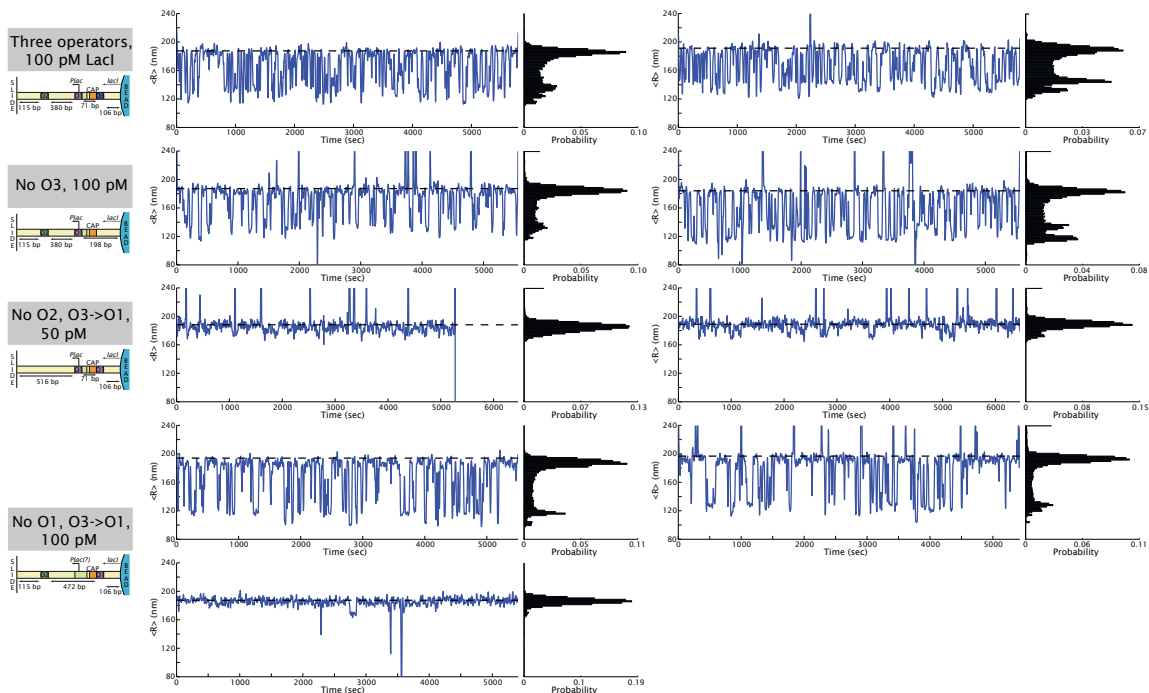


Figure 6.4: Representative examples of looping with the three-operator, wild-type construct and two operator derivatives, in the absence of HU. The plots are the same kind as in Figs. E.1 and E.2 in Appendix E; in particular, the black dashed line indicates the length of the particular tether in the absence of Lac repressor. Schematics to the left show which operators are present. Two looped states are clearly visible in all of the constructs except the one with the 71 bp loop; in the case of that very short loop, we expect, based on the phasing of E8 and TA near this length (see Fig. 4.3), that two looped states are present, but we probably cannot resolve the two given the long overall tether length, at least without employing the HMM approach discussed below and in Chapter 5. Sometimes, but not often, the two looped states in the other constructs are distinguishable in the histograms to the right of each trace. All of these DNAs were thresholded to obtain the looping probabilities in Fig. 6.3. The trace at the bottom demonstrates that binding at the O_1 operator has not been completely abolished in the “noO1” constructs (as has been claimed elsewhere for this deletion [170]); there are rare but noticeable trajectories in which the 71 bp loop—which, if the original O_1 operator is still present, is now flanked by two O_1 operators—is visible. However these data are easily discardable from our results.

J-factor measurements.

The results of these 3-point concentration titrations are shown in Fig. 6.3, which also shows the total (all possible loops) looping probability for the full-three operator construct at three concentrations. Figure 6.4 shows representative examples of TPM data with these constructs. Not surprisingly, given that any two-operator construct with O_3 shows no looping, the three-operator construct (black data in Fig. 6.3) exhibits only the O_1 - O_2 loop, and is indistinguishable in our assay from the two-operator derivative that is missing O_3 (red). The J-factors we obtain with these DNAs are all roughly the same as that of E894 (see Table 4.1), which could be an interesting result especially for the 71 bp O_1 - O_3 loop. A 71 bp loop is probably one with mostly out-of-phase

operators, consistent both with an extension of the data in Fig. 4.2(A) down to 71 bp, assuming a 10 bp periodicity, and with the observation that we believe we see two looped states at 71 bp with a preliminary HMM analysis of this construct (see also example traces in Fig. 6.4), whereas only one state predominates with fully in-phase or fully out-of-phase operators, if “in-phase” is defined by maximal looping probability (Fig. 4.2(B)). However with the E8 and TA no-promoter constructs discussed in Chapter 4, a 71 bp loop should have a J-factor almost an order of magnitude less than that of E894 (Fig. 4.3). Although this could indicate that the wild-type sequence of the O_1 - O_3 loop is a more flexible looping sequence than E8 or TA—or that the wild-type *lac* promoter induces a sequence dependence just like the synthetic derivative *lacUV5* does—data with a shorter overall tether length to better resolve this 71 bp loop will be necessary before any definite conclusions can be drawn. The value of J_{13} in Fig. 6.3 is currently an estimate only.

6.2 DNA-bending proteins may be essential elements of the *lac* regulatory system

If we are to comment on the models of Li and coworkers and others, it is necessary first to be able to reproduce, at least qualitatively, *in vivo* results such as those of Oehler and coworkers [93, 94]. So far, however, we have not done that: we have found *in vitro* that the O_3 operator has no effect on looping, and that the three-operator construct behaves the same as a two-operator derivative with O_1 and O_2 only. We speculated that either nonspecific bending proteins such as HU, or specific bending proteins such as the CAP activator, might enhance looping *in vivo* and cause the discrepancy between the *in vivo* and *in vitro* results. We therefore asked from both a theoretical and an experimental perspective what the effects of proteins like HU and/or CAP might have on this three-operator system. (Other cellular factors such as the supercoiled state of the DNA could also be contributing to this *in vivo-in vitro* discrepancy; in fact it has been shown *in vitro* that supercoiling greatly stabilizes the loop between O_1 and the auxiliary operators [175, 176]. These other potential contributing factors also need to be considered, though we will have not done so here

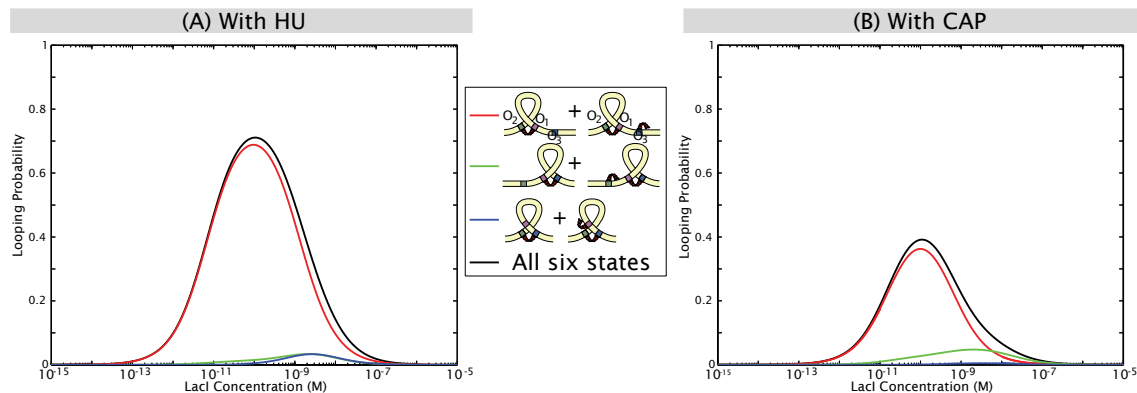


Figure 6.5: Predictions of effect of nonspecific or specific bending proteins such as HU or CAP on looping by the Lac repressor in the presence of three operators. The legend in the center shows schematics for which states of Fig. 6.2 are plotted; looped states versus states with a loop and the third operator bound by a separate repressor are probably not distinguishable in our assay and so are plotted together. They are not equally likely, however: for example, in (A), at repressor concentrations where the O_3 - O_2 loop forms, O_1 will also be bound (that is, the state with an O_3 - O_2 loop without O_1 also bound separately has zero looping probability at all concentrations). **(A)** Predictions of the model of Fig. 6.2 for the probabilities of the various looped states in the presence of some amount of HU, if we assume that HU does not change the repressor-operator dissociation constants but only increases all of the J-factors by some amount (here, we assume it increases all J-factors by a factor of 4, consistent with the results presented in the text here in the presence of 500 nM HU; see also [143], where HU increases looping for 60–100 bp loops by a factor of up to 6 *in vivo*). The amount by which HU increases a loop's J-factor may actually depend on length, such that, for example, longer loops would on average have more HU molecules bound and so would increase more in apparent flexibility, but we neglect such potential effects here. In these predictions we have used $K_3=5$ nM to make the trends more noticeable. **(B)** Predictions of the model of Fig. 6.2 for the probabilities of the various looped states in the presence of some amount of CAP, if we assume that CAP increases the J-factor for the O_1 - O_3 loop alone by a factor of 10, a value consistent with the $-1.4 k_B T$ to $-2.4 k_B T$ stabilization found by biochemical assays [174]. Unlike the addition of HU as shown in (A), the addition of CAP to the TPM assay could bring our *in vitro* results into better alignment with *in vivo* results in which O_3 is an essential component of the system, particularly if K_3 is closer to 5 nM than the 15 nM used here. It is also possible that CAP increases the J-factor of the O_2 - O_3 loop as well, but given the data in [93], in which repression in the absence of O_1 (but the presence of the other two operators) is negligible, it is reasonable to assume that CAP stabilizes only the O_1 - O_3 loop. Of course it is likely that *in vivo* both HU and CAP influence looping in the *lac* operon.

in these preliminary results.)

HU is known from both *in vitro* [177] and *in vivo* [19] studies to increase the flexibility of DNA, and to enhance DNA looping by the Lac repressor *in vivo* [19] and by the Gal repressor *in vitro* [122, 178]. Other nucleoid-associated proteins like IHF have also been shown to enhance looping by the Lac repressor *in vitro*, at least in some regimes [179]. As will be discussed in more detail in Chapter 7, more rigorous *in vitro* studies with HU and the Lac repressor must be done to precisely quantify the effects of HU on looping by the Lac repressor. However, Fig. 6.5(A) shows the prediction of our model for the simplest effect HU might have on looping, in which we assume that HU increases the J-factors for the three loops, leaving the dissociation constants unchanged. The result is that for reasonable values of the amount by which the J-factors might increase in the presence of HU (based

on the literature cited in the figure caption and on our results presented below), the effect of the O_3 operator are still negligible.

On the other hand, the CAP protein, which as noted above bends the DNA between the O_1 and O_3 operators [168] and stabilizes the loop between these two operators [174], might enhance looping between the O_1 and O_3 operators but not between the other operators, leading to a larger contribution of the O_1 - O_3 loop relative to the others *in vitro*. Fig. 6.5(B) shows the predictions of our model for the effect CAP might have on our TPM results, again assuming that CAP increases the J-factor for the O_1 - O_3 loop by an amount consistent with literature values, leaving the dissociation constants unchanged. The value of K_3 in that prediction is the relatively conservative value of 15 nM; particularly if K_3 is closer to 5 nM, CAP could make the stability of the O_1 - O_3 loop comparable to that of O_1 - O_2 at high repressor concentrations, potentially bringing the TPM results into better qualitative agreement with *in vivo* work where O_3 and O_2 are both important to the wild-type function of the system.

HU and CAP have both been purified and used in *in vitro* studies before (e.g., [122, 168, 177]), and so it should be feasible to add HU and/or CAP to our TPM assay and ask what their effects on looping with the wild-type three-operator *lac* system and its two-operator derivatives are. In Fig. 6.6 we show preliminary experimental results of the effect of adding HU to a TPM Lac repressor looping assay.³ HU alone compacts the DNA tethers, as has been observed previously using magnetic tweezers [177] (Fig. 6.6(A)); and, as we expect from *in vivo* assays [19], HU increases looping by the Lac repressor when both HU and Lac are present (compare Figure 6.4 and Fig. 6.6(B-E)). We can quantify the amount by which HU increases looping by the Lac repressor, again assuming that HU affects only the J-factor and not dissociation constants, by thresholding the traces from the data set represented by Fig. 6.6(B) (the two-operators-only construct that is missing O_3) to obtain a looping probability at 1 nM Lac repressor and 500 nM HU of 0.59 ± 0.3 . This looping probability corresponds to a J-factor for the O_1 - O_2 loop that is roughly 4 times higher than the J-factor determined without

³Purified HU was a kind gift from Remus Dame at Leiden University in the Netherlands, and was sent in a buffer of 25 mM Tris (pH 8.0), 200 mM NaCl, 1 mM EDTA, 5 mM β -mercaptoethanol, and 10% glycerol. The stock concentration is 94 μ M, so 500 nM HU, the concentration used in our assays, is only a 100- to 200-fold dilution into the Lac repressor buffer. Future work with HU and the Lac repressor should ensure that the small but significant amount of this HU buffer, particularly the glycerol, does not alter the activity of the Lac repressor.

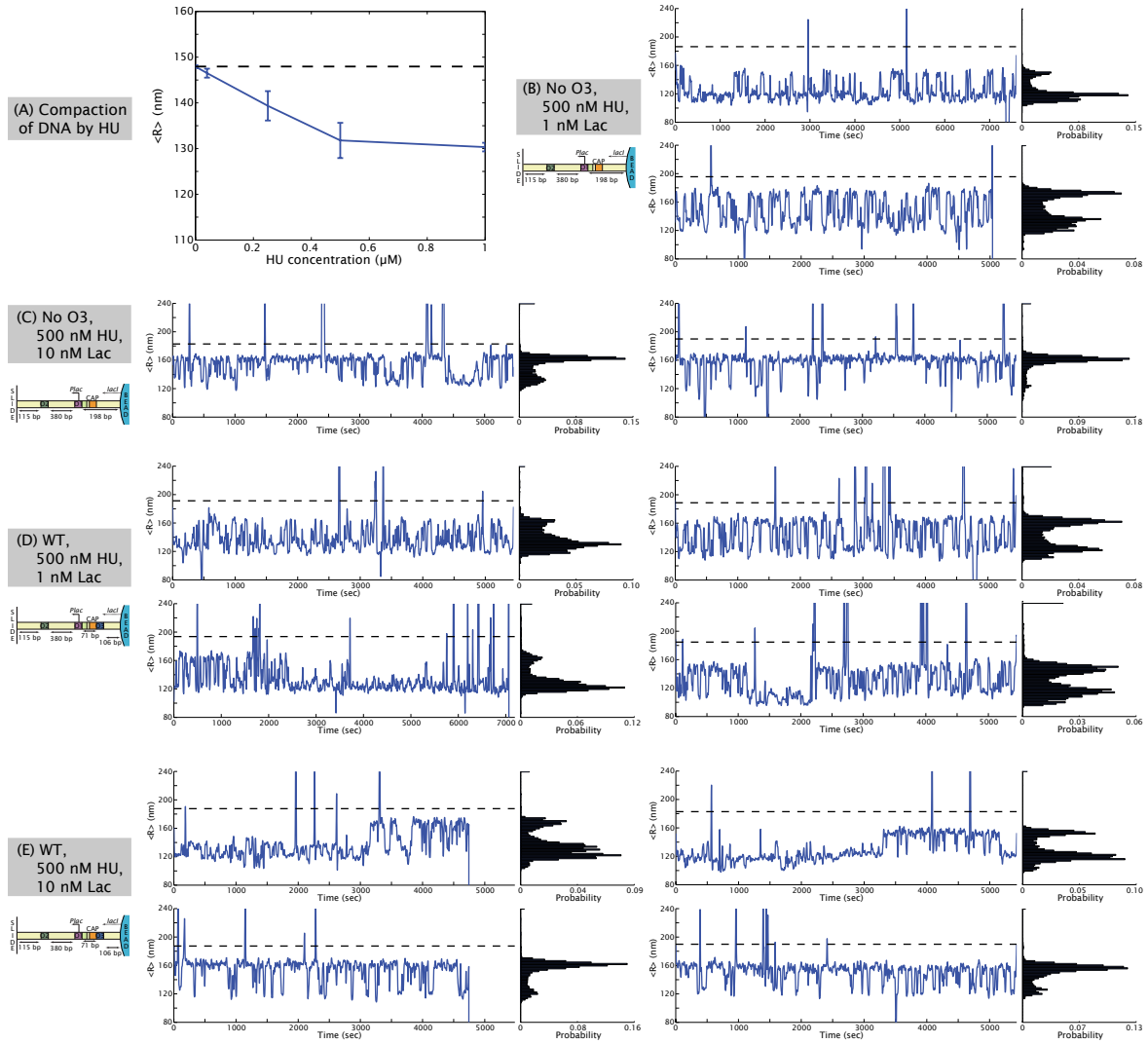


Figure 6.6: The DNA-bending protein HU increases looping by the Lac repressor and may lead to an observable effect from the O_3 operator *in vitro*. The format for the sample traces in (B–E) is the same as in Fig. 6.4, which shows representative traces for the constructs shown here but in the absence of HU. (A) The RMS motion of a tether in the absence of Lac repressor decreases with increasing HU. This result is comparable to that of [177], especially given the difference in salt concentrations between the two experiments (HU is sensitive to salt). Note that these data are for a shorter construct than in (B–E). (B) Sample trajectories with the construct that contains only the O_1 and O_2 operators (same construct as the red data of Fig. 6.3), in the presence of 500 nM HU and 1 nM Lac repressor. Stretches of long dwells as in the top trace are rare without the third operator, but do occur. The bottom trace is more representative of this data set, especially in that it looks like there might be more than two looped states. It is unclear if this is a real result or an artifact (two operators should yield only two looped states, according to the results of preceding chapters, though if these two looped states are superpositions of the four underlying loop topologies of Fig. 4.3, it is possible that HU changes how many different tether lengths the four states collapse into). Some traces without HU for this construct may also exhibit more than two looped states, though without HU the looping probability is so low as to make distinguishing looped states difficult. It is also possible that the deletion of O_3 was incomplete, as is probably the case for the deletion of O_1 (see caption to Fig. 6.4). The black dashed line in this and (C–E), represents the length of the particular tether in the absence of both HU and Lac repressor; note the compaction of the tether in the presence of HU, in that the unlooped state of the blue data is well below the black dashed line. (C) Sample trajectories with the construct that contains only the O_1 and O_2 operators, in the presence of 500 nM HU and 10 nM Lac repressor. (D) Sample trajectories with the full three-operator construct (black data in Fig. 6.3), in the presence of 500 nM HU and 1 nM Lac repressor. The top two traces are the most representative of this data set and are not obviously different than those in (B) that lack the third operator; however the bottom two traces show long dwells in one or more looped states that are more common than in the data set in (B), and may indicate the formation of the O_2 - O_3 loop. (E) Sample trajectories with the full three-operator construct in the presence of 500 nM HU and 10 nM Lac repressor. The top trajectories show the long dwells that are common at this repressor concentration, and are suggestive of both the O_1 - O_2 and O_2 - O_3 loops forming (and, interestingly, possibly directly interconverting). The bottom two traces look similar to those in (C) that lack the third operator.

HU (see Fig. 6.3(B)). As shown in Fig. 6.5(A), even if K_3 is as low as 5 nM, an increase in J-factors for all of the loops by a factor of 4 should still not allow us to reliably detect loops with O_3 , nor should we observe a difference between the full three operator construct versus the one that lacks O_3 .

As suggested by the examples in Fig. 6.6(B–E), there is a large bead-to-bead variation in looping behavior in the presence of HU, especially when all three operators are present, and so more data will be necessary to differentiate spurious behavior from real results before conclusions can be drawn about the effect of HU on this three-operator system. More importantly, a quantitative analysis and objective state identification is crucial, which we believe will be best accomplished by the hidden Markov model analysis discussed in the previous chapter. However, from the trajectories in Figs. 6.4 and 6.6(B–E) it does appear that the presence of the O_3 operator, with HU in the sample, alters the dynamics of looping: with HU and O_3 , long dwells in one or more looped states are observed, some of whose lengths are suggestive of the formation of the O_2 - O_3 loop, and possibly its direct interconversion with the O_1 - O_2 loop. Some traces also appear to have states at RMS values that would correspond to the O_1 - O_3 loop. It will be exciting to see if these trends hold with more data and a more rigorous analysis.

6.3 Conclusion

In this chapter we have presented preliminary results with a TPM construct derived from the wild-type *lac* operon, which unlike most previously studied looping constructs (by our lab and others), has three operators instead of two. Not only do we hope with this naturally derived construct to address some outstanding questions about the *lac* operon that are difficult to address by *in vivo*, rather than single-molecule *in vitro*, assays, but we also hope to set the stage for systematic dissections of other, more complicated cases of combinatorial control in transcriptional regulation.

However, one of our key preliminary findings is that unlike the *in vivo* repression data of Oehler and coworkers, as well as others, summarized by Fig. 6.1, *in vitro* we find that the O_3 operator has no detectable effect on looping. That is, the three-operator wild-type construct behaves identically

to a two-operator construct that lacks the weakest O_3 operator.

We have already shown in Chapter 4 (Section 4.6) that the non-specific DNA-bending protein HU masks a potential sequence dependence *in vivo* that we observe *in vitro*. We likewise suspect that the *in vivo* wild-type behavior of the *lac* operon depends strongly on the presence of HU or other DNA-bending proteins in the cell (and/or other cellular conditions such as the supercoiled state of the DNA), and that the absence of these proteins (or other conditions) in our experiments leads to the discrepancy we observe between our *in vitro* results and the *in vivo* results of [93, 94] and others. HU (or other DNA-bending proteins) is an essential component of the Gal looping system [180, 178]; although unlike the Gal repressor, the Lac repressor readily forms loops *in vitro*, its wild-type function may in fact depend more strongly than has been previously appreciated on interactions with other DNA binding proteins such as HU.

To begin to explore this hypothesis of the importance of proteins like HU to the wild-type function of the *lac* system, we have presented preliminary results from both theoretical (Fig. 6.5) and experimental (Fig. 6.6) approaches on the effect of HU on looping by the Lac repressor with the wild-type three-operator system. Although more data and more rigorous analysis are necessary before definite conclusions can be drawn, our preliminary results suggest that with the three-operator construct, the presence of HU leads to long dwells in one or more looped states, which do not occur without O_3 . Although according to our theoretical predictions with an estimated value of K_3 of 5 nM (Fig. 6.5(A)), neither of the loops containing O_3 should form to a detectable degree even in the presence of HU, nevertheless it appears (Fig. 6.6(E)) that we see more looped states with the third operator than without it.

This three-operator plus HU data is a clear example of the kind of data that the hidden Markov model analysis presented in the previous chapter are better suited to analyzing than the conventional thresholding method. The thresholding method requires the user to identify how many states are in a trajectory, while the HMM algorithm uses a more objective maximum likelihood approach, based on a physical model of the bead's motion (see Chapter 5), to determine how many states there are in a trajectory and what the most likely state at each time point is. An example of an HMM analysis

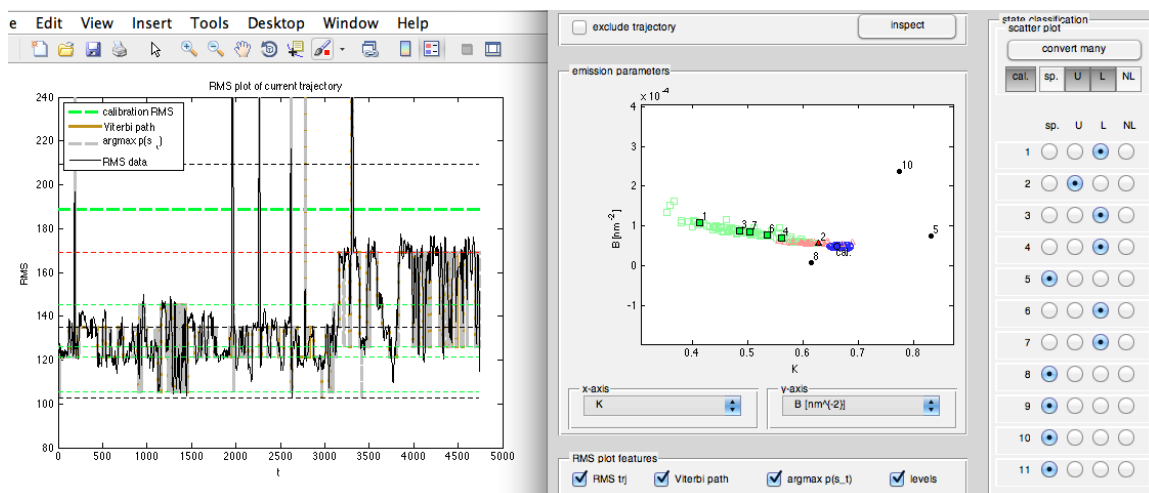


Figure 6.7: A screen shot from the graphical user interface (GUI) from our hidden Markov analysis code (see Section 5.2.2), showing a preliminary HMM analysis of a particular multi-state trajectory for the three-operator construct in the presence of 500 nM HU and 10 nM Lac repressor. This is the same trace as the top left trajectory of Fig. 6.6(E). In the left-hand plot, the thick dashed green line is the RMS motion of the tether in the absence of Lac and HU, the same as the black dashed line in Fig. 6.6(E). The other horizontal dashed lines indicate states found by the HMM algorithm: the unlooped state is in red, genuine looped states in green, and spurious states (corresponding to tracking errors or sticking events, or events with otherwise anomalous emission parameters) are in black. The Viterbi path represents the most likely sequence of states; the “ $\text{argmax } p(s_t)$ ” indicates the most likely state at any time point t . The right-hand plot shows the emission parameters K and B (related to the RMS and correlation time of the bead; see Chapter 5) for the current trajectory as well as for all genuine states from all trajectories: the looped states for this particular trajectory are shown as filled green squares, for all other trajectories as outlined green squares, the unlooped state for this trajectory as a filled red triangle, and for all other trajectories as outlined red triangles, and spurious states for this trajectory are black dots (not all are within the field of view). The blue “calibration” circles indicate the K and B parameters for all trajectories in this set in the absence of HU and Lac. Radio buttons to the left allow the user to alter the automatic state assignments from the algorithm (in this example, the user can assign states as “spurious”, “unlooped”, “looped”, or “never looped” (“NL”)); the GUI has the option of allowing additional state assignments as well, such as identities of particular loops). As discussed in more detail in Chapter 5, genuine states fall on a line in K - B space, whereas spurious states do not. Therefore even though by eye it looks like there is an additional looped state with an RMS around 135 nM, the HMM analysis identifies this state as anomalous. Similarly, though it is difficult to objectively identify how many looped states are in this trajectory by eye (as would be the case for the thresholding analysis used in the rest of this work), the HMM analysis identifies five looped states with different enough emission parameters that it classifies them as different states.

of the three-operator-plus-HU data is shown in Fig. 6.7, for one of the cases where it is difficult to threshold the trace by eye. We hope that further analysis with this HMM approach, especially of the kinetics of looping in the presence of two versus three operators, will shed more light on the role of the weakest operator on looping *in vitro*, and more generally on its role in gene expression *in vivo*.

Chapter 7

Conclusion

It is becoming increasingly apparent that the mechanical properties of the DNA polymer and its physical organization in cells are crucial to the regulation of the genetic information that the DNA encodes, but precisely what these mechanical properties are, especially at short length scales, and how they interact with the various proteins that decorate the DNA *in vivo*, are still open questions. Here we present a new way of measuring the J-factors that capture the mechanical properties of protein-mediated DNA loops, through a combination of careful single-molecule measurements and rigorous comparison to statistical mechanical theories, and have applied this technique primarily to the question of how sequence affects looping *in vitro* and *in vivo*. We have argued that “sequence flexibility” as a general term is misleading, and that both the shape of the deformation induced to measure sequence flexibility as well as other proteins that can interact with DNA *in vivo* play larger roles than previously anticipated on looping. Finally we also describe here a new hidden Markov model based on variational Bayesian inference for extracting kinetic information from our single-molecule data, which should be particularly important for dissecting more complicated multi-loop systems such as the wild-type *lac* operon that we describe in Chapter 6.

A key component of our approach has been a systematic dissection of single-molecule looping experiments, by varying the four key parameters shown in Fig. 1.4, both from theoretical (Chapter 2) and experimental (Chapters 3, 4, 5, and 6) standpoints. We have found such a systematic variation of biologically relevant parameters to be a powerful technique for measuring not only the J-factors that were a main focus of this work, but also other important features of the systems we discuss here,

such as repressor-operator dissociation constants, kinetic parameters, and even the effects of potential experimental artifacts. Moreover this systematic dissection, in conjunction with extensive dialogue between experiments and theory, have allowed us to make quantitative contact with parallel *in vivo* experiments. Indeed a quantitative comparison between our *in vitro* and *in vivo* experiments was made possible only by concrete theoretical frameworks for both systems in terms of experimentally tractable parameters. We have shown that such a quantitative comparison revealed an important discrepancy between the conclusions we would have drawn from either experimental system alone. Namely we found *in vitro* that two sequences with very different affinities for nucleosomes showed a significant sequence dependence to looping that followed the trends observed with nucleosomes, if a bacterial promoter sequence was present in the loop, as it is in our *in vivo* assay. We determined using our models for the *in vitro* and *in vivo* assays that the range of this sequence effect would be large enough to be detectable in *in vivo* repression assays. However, *in vivo* the nonspecific DNA-bending protein HU (and possibly others) masks any sequence dependence to repression that we might otherwise observe.

A key corollary experiment that must be done to fully understand how HU can mask this sequence dependence to looping, without abolishing the phasing (length dependence) that is a hallmark of loop formation, is to add purified HU to the TPM assay and quantify the effects of HU both on repressor-operator dissociation constants and on the J-factors of the E8, TA and poly(dA:dT) sequences that we examined here. In Chapter 6 we presented preliminary data with HU and the Lac repressor, demonstrating that HU affects both looping probabilities and dynamics. However a more systematic dissection of the role of HU in Lac repressor-mediated looping remains is warranted. For example, in Fig. 6.5(A) we assumed that HU alters only the J-factor and not the dissociation constants, but this is an assumption that has not yet been validated. Likewise, we assumed that HU increased the J-factors of each loop by the same factor, but it is plausible that the length of the loop, and therefore the number of (nonspecific) binding sites for HU, affects the amount by which HU increases the J-factor. HU is known to bind preferentially to distorted DNA structures [17]; it may therefore bind with higher probability to certain sequences, and thereby affect some J-factors more

than others, even at constant loop length; or HU may preferentially stabilize the structures of, for example, in-phase loops, or one looped state over the other, such that even single base-pair changes in loop length would be a factor in the effect of HU. Moreover, studies with the Lac repressor and a similar DNA-bending protein, IHF [17, 179], have shown at least two regimes to the effects of IHF on looping by the Lac repressor, one in which IHF inhibits looping and the other in which it enhances looping. HU may have similarly complex interactions with the Lac repressor.

Our combined statistical mechanical model plus concentration titration approach is ideally suited to answering these questions. By measuring repressor titration curves in the presence of varying amounts of HU, and by systematically tuning loop length, loop sequence, and operator strength as we have here, we can begin to fill in the gaps in our knowledge of how HU affects looping by the Lac repressor. Again, as we have shown in Sections 4.6 and 6.2, the picture emerging from our work is that DNA-bending proteins such as HU play crucial roles in loop formation *in vivo*, and we will not fully understand gene regulation *in vivo* until we understand the interactions between transcription factors and these architectural proteins both in cells and in isolation *in vitro*.

While the combined theory, *in vivo*, and *in vitro* approach presented here was able to resolve the apparent *in vivo/in vitro* sequence-dependence discrepancy by identifying the crucial role of HU in masking sequence dependence *in vivo*, we have not so far solved the mystery of the complex sequence dependence to looping observed *in vitro*. Namely we have shown here that two sequences with very different affinities for nucleosomes behave the same in the context of looping, unless a bacterial promoter sequence is added to the loop, in which case there is a significant sequence dependence to looping that follows the trends observed with nucleosomes. Moreover, a third sequence that contains poly(dA:dT) tracts does not follow the trends either of the E8 and TA loops or of the nucleosome formation assays that inspired our use of all three kinds of sequences. An important next step in understanding this complex sequence dependence will be to ask how general these trends are in terms of the protein that mediates the loop, or the promoter included in the loop. Do other looping proteins, such as the lambda repressor, which might impose different boundary conditions on the loop and so might change the resulting loop shape, show the same sequence-(in)dependent trends

as the Lac repressor? Are the promoter-dependent sequence effects that we observe peculiar to the *lacUV5* promoter only? We hope that future work with additional looping proteins and promoters will shed light on the property of the promoter sequence that causes the sequence effects that we observe, and how robust our loop-shape hypothesis is when confronted with different loop boundary conditions.

An equally intriguing question hinted at by some of the work presented here is whether or not there are in fact sequences that *do* alter the looping probability *in vitro* even in the absence of the *lacUV5* promoter (the poly(dA:dT) sequence of Chapter 4 being a potential candidate), or *in vivo* even in the presence of HU. The sequences we have studied here are all derived from nucleosome affinity assays; however if our shape-dependent hypothesis is correct, then “better” looping sequences will most likely not be found by attempting to adapt additional nucleosome-favoring sequences to looping, but rather by accessing entirely new regions of sequence space. Indeed, there are hints that especially favorable or unfavorable looping sequences do exist even *in vivo*: as discussed in Chapter 1, A-tracts are known to affect NtrC-mediated activation loops in bacteria [65, 66], and phased A-tracts create hyperstable Lac repressor loops *in vitro* [67, 68, 69, 70].

To more efficiently explore sequence space and address these questions of whether or not optimal and sub-optimal looping sequences exist *in vivo*, we are developing a high-throughput approach based on a newly described technique called “Sort-Seq” from Justin Kinney and coworkers [181]. Sort-Seq was originally designed to identify important regulatory regions of promoters such as transcription factor binding sites, as well as the sequence-dependent binding energies and interaction energies of transcription factors and polymerases, and relies on the simple premise that mutations to protein binding sites will result in larger effects on gene expression than mutations to other sites on the DNA. In a Sort-Seq experiment, random mutations are introduced into the region of a promoter that has been altered to drive a fluorescent reporter gene instead of its natural product, and cells containing these mutations are sorted by FACS (fluorescence-activated cell sorting, a kind of flow cytometry) according to the level of expression of the fluorescent reporter. The promoter regions of the sorted cells are sequenced and correlated to the expression levels measured in the FACS sorting.

More precisely, the *mutual information* between a particular mutation at a particular site and the level of gene expression resulting from that mutation is calculated. Important regions of a promoter, such as binding sites for transcription factors, are highly informative about gene expression levels, whereas other sites are not.

A key aspect of this Sort-Seq approach is its high-throughput nature. Many (on the order of ten thousand) data points on the relationship between sequence and gene expression are obtained in one experiment, through the use of both high-throughput sequencing and a large initial library of randomized constructs. Such a high-throughput approach is ideal for searching sequence space for sequences that are particularly good or particularly bad at loop formation. We are therefore developing a modified Sort-Seq for looping, with the eventual goal of determining the “rules” that govern the sequence dependence to loop formation just as studies of nucleosome preferences have resulted in the establishment of a set of sequence rules that predict nucleosome affinity [14].

In this modified Sort-Seq experiment, we will produce a library containing a large number of randomly chosen sequences for a loop region, and then measure gene expression with these randomly chosen loops in cells harboring the Lac repressor. By comparing the FACS profile generated by these new looping sequences to that generated by the E8 or TA sequences in one of the constructs whose repression level we have already measured, we will be able to identify any sequences that alter gene expression (presumably by altering looping), in a way that E8 versus TA does not. A population of cells containing one of the E8 (or, equivalently in the absence of HU, TA) loops that we discuss in Chapter 4, when sorted by FACS, will generate a relatively narrow distribution of fluorescence levels centered on the mean value for that construct shown in Fig. 4.5. If the pool of new, randomly chosen looping constructs contains any sequences that are better or worse for looping *in vivo* than E8/TA, they will lead to a broadening of the distribution of fluorescence compared to that of E8/TA. That is, sequences that are poorer loop formers than E8/TA will show a higher fluorescence in the FACS sorting than E8/TA (because they will lead to less repression), and sequences that are better loop formers will show lower fluorescence. By sequencing the strains that appear in the tails of this broadened distribution, we can identify those sequences that are particularly good or particularly

bad at loop formation, and thereby begin to develop the rules that govern ease of loop formation *in vivo*.

We suspect that any such interesting looping sequences will be rare in the pool of randomly selected sequences, and so we plan to collect cells that fall in the tails of the distribution, grow these cells overnight, and re-sort them, to enrich the FACS profile in any rare but very good or very poor looping sequences. We can also subject the sequences in the tails of the distribution to error-prone PCR, thereby exploring sequence space around these slightly better or slightly worse loopers, much as in the SELEX experiment that led to the determination of the strongest known nucleosome positioning sequence [86]. Even so, we do not expect randomly chosen sequences of lengths on the order of 100 bp to effectively explore very much of sequence space, and so we also anticipate needing to bias our search by starting from sequences that we have reason to believe might have high or low looping probabilities compared to E8/TA. For example, sequences enriched in dA-dT steps are thought to be especially flexible, so we can design our randomly selected starting sequences to have more dA and dT nucleotides than the other 3 bases. (The TA sequence, though already one such dA-dT enriched sequence, has these A-T steps precisely spaced in a way that may not be optimal for looping, though it is optimal for nucleosome formation.) Alternatively, because the removal of HU allows a sequence dependence to repression to appear for the E8 versus TA sequences, we also can construct a pool of sequences mutagenized around the TA sequence and sort them in an HU deletion strain.

In parallel we plan to use our single-molecule TPM assay to measure the *in vitro* looping free energies of any interesting sequences that arise from the Sort-Seq experiment, to further our efforts to understand how sequence governs DNA flexibility in the context of loop formation in the absence of complicating factors such as the DNA-bending proteins present *in vivo*. However, a high-throughput screen for interesting sequences will necessitate an equally high-throughput single-molecule assay by which to measure the *in vitro* J-factors of these sequences. To that end we have begun a collaboration with the group of Laurence Salomé at the Université de Toulouse in France, who have developed a high-throughput tethered-particle technique which allows the observation and tracking of up to 500

tethers at once [182].

Though we believe a deeper understanding of the complex relationship between DNA sequence and mechanical properties that such studies will provide is vital, we are ultimately interested in whether such mechanical properties are in fact a “knob” that bacteria tune to control gene expression *in vivo*. Therefore regardless of what our Sort-Seq search of sequence space reveals, whether or not we find a set of sequences that are particularly good or particularly bad at Lac repressor-mediated looping *in vitro* and/or *in vivo*, it will be even more informative to compare whatever results we obtain to the sequences that are found in naturally occurring bacterial loops—for example, those of the wild-type *lac* operon discussed in Chapter 6. A bioinformatics approach to comparing the sequences of the loops of the *lac* operon and those of other bacterial transcription factor-mediated loops may hint at common sequence motifs that we could exploit to enhance our search for particularly good or bad looping sequences, with the caveat that if different looping proteins impose different boundary conditions on the loops they form, there may not be any universal looping sequence rules to discover. But at least for the case of the *lac* operon we have already presented preliminary results on the J-factors of those natural loops, and though more careful studies will be necessary in order to draw conclusions about the flexibilities of these loops compared to the E8 and TA loops that we have focused on so far, the pieces are all in place for such a study. Indeed, the pieces are all in place for a thorough exploration of many of the ways in which the mechanical properties of DNA impact the regulation of the genetic information it encodes, and we anticipate a proliferation of quantitative and physics-minded approaches, such as those described here, to tackle this new way of thinking about the chief information molecule of the cell.

Appendices

Appendix A

Detailed derivation of the model that includes the dimer-to-tetramer transition

In this Appendix we describe in greater detail the derivation of the looping probability as a function of repressor concentration with a mixture of dimeric and tetrameric repressors present, summarized in Section 2.4. We first describe the assumptions about dimeric Lac repressor made in this derivation, and then derive a model that takes into consideration the dimer-to-tetramer transition at low repressor concentrations. In the last part we consider the case where there is a constant fraction of dimers due not to low concentration but to damaged protein. Figure A.1 summarizes the derivation and the variables which will be used throughout (the same notation is used as in [115]).

A.1 Assumptions

First, in order to write the statistical weights of states that include a dimeric form of the Lac repressor, we must define the energies associated with the dimeric form. The equilibrium constant associated with the dissociation of tetramers into dimers, K_{DT} , is thought to occur at such a low concentration of repressors that it has not been measured *in vitro* [131]. However, several mutant forms of the Lac repressor have been developed that are unable to form tetramers. Although some of these dimeric forms bind DNA less tightly than the tetrameric form, others have operator dissociation constants comparable to that of the wild-type tetramer [132, 133, 134]. Here we will assume that the

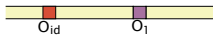




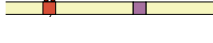
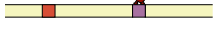
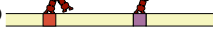
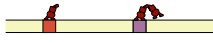
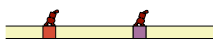
STATE	STATISTICAL WEIGHT	
	STATISTICAL MECHANICS	THERMODYNAMICS
(i) 	$\left(\frac{8\pi^2}{\delta\omega}\right)^{(T+D)} \frac{\Omega!}{T!D!(\Omega-T-D)!} e^{-\beta T\epsilon_{sol}} e^{-\beta D\frac{\epsilon_{sol}}{2}}$	1
(ii) 	$4\left(\frac{8\pi^2}{\delta\omega}\right)^{(T-1+D)} \frac{\Omega!}{(T-1)!D!(\Omega-T+1-D)!} e^{-\beta(T-1)\epsilon_{sol}} e^{-\beta D\frac{\epsilon_{sol}}{2}} e^{-\beta(\epsilon_1+\frac{\epsilon_{sol}}{2})}$	$\frac{[T]}{K_1}$
(iii) 	$4\left(\frac{8\pi^2}{\delta\omega}\right)^{(T-1+D)} \frac{\Omega!}{(T-1)!D!(\Omega-T+1-D)!} e^{-\beta(T-1)\epsilon_{sol}} e^{-\beta D\frac{\epsilon_{sol}}{2}} e^{-\beta(\epsilon_{id}+\frac{\epsilon_{sol}}{2})}$	$\frac{[T]}{K_{id}}$
(iv) 	$16\left(\frac{8\pi^2}{\delta\omega}\right)^{(T-2+D)} \frac{\Omega!}{(T-2)!D!(\Omega-T+2-D)!} e^{-\beta(T-2)\epsilon_{sol}} e^{-\beta D\frac{\epsilon_{sol}}{2}} e^{-\beta(\epsilon_{id}+\epsilon_1+2\frac{\epsilon_{sol}}{2})}$	$\frac{[T]^2}{K_{id}K_1}$
(v) 	$8\left(\frac{8\pi^2}{\delta\omega}\right)^{(T-1+D)} \frac{\Omega!}{(T-1)!D!(\Omega-T+1-D)!} e^{-\beta(T-1)\epsilon_{sol}} e^{-\beta D\frac{\epsilon_{sol}}{2}} e^{-\beta(\epsilon_{id}+\epsilon_1+\Delta F_{loop})}$	$\frac{1}{2} \frac{[T]J_{loop}}{K_{id}K_1}$
(vi) 	$2\left(\frac{8\pi^2}{\delta\omega}\right)^{(T+D-1)} \frac{\Omega!}{T!(D-1)!(\Omega-T-D+1)!} e^{-\beta T\epsilon_{sol}} e^{-\beta(D-1)\frac{\epsilon_{sol}}{2}} e^{-\beta\epsilon_{id}}$	$\frac{1}{2} \frac{[D]}{K_{id}}$
(vii) 	$2\left(\frac{8\pi^2}{\delta\omega}\right)^{(T+D-1)} \frac{\Omega!}{T!(D-1)!(\Omega-T-D+1)!} e^{-\beta T\epsilon_{sol}} e^{-\beta(D-1)\frac{\epsilon_{sol}}{2}} e^{-\beta\epsilon_1}$	$\frac{1}{2} \frac{[D]}{K_1}$
(viii) 	$8\left(\frac{8\pi^2}{\delta\omega}\right)^{(T+D-2)} \frac{\Omega!}{(T-1)!(D-1)!(\Omega-T-D+2)!} e^{-\beta(T-1)\epsilon_{sol}} e^{-\beta(D-1)\frac{\epsilon_{sol}}{2}} e^{-\beta(\epsilon_{id}+\epsilon_1+\frac{\epsilon_{sol}}{2})}$	$\frac{[T][D]}{K_{id}K_1}$
(ix) 	$8\left(\frac{8\pi^2}{\delta\omega}\right)^{(T+D-2)} \frac{\Omega!}{(T-1)!(D-1)!(\Omega-T-D+2)!} e^{-\beta(T-1)\epsilon_{sol}} e^{-\beta(D-1)\frac{\epsilon_{sol}}{2}} e^{-\beta(\epsilon_{id}+\epsilon_1+\frac{\epsilon_{sol}}{2})}$	$\frac{[T][D]}{K_{id}K_1}$
(x) 	$4\left(\frac{8\pi^2}{\delta\omega}\right)^{(T+D-2)} \frac{\Omega!}{T!(D-2)!(\Omega-T-D+2)!} e^{-\beta T\epsilon_{sol}} e^{-\beta(D-2)\frac{\epsilon_{sol}}{2}} e^{-\beta(\epsilon_{id}+\epsilon_1)}$	$\frac{1}{4} \frac{[D]^2}{K_{id}K_1}$

Figure A.1: States and weights for a model which includes the presence of dimers. The left-hand column diagrams the ten different states of the system, the first five of which are the same as the simple model with just tetramers and the last five of which involve dimers binding to one or both operators. The middle column shows the statistical weights for each of the ten states in the language of statistical mechanics: here T and D are the number of tetramers and dimers in solution, Ω is the number of lattice sites assigned to the solution, β is the reciprocal of the Boltzmann constant times the temperature, $\epsilon_{sol}/2$ is the energy of a repressor head in solution, ϵ_1 and ϵ_{id} the energies of a repressor head bound to the O_1 operator or the O_{id} operator, ΔF_{loop} is the energy cost of deforming the DNA into a loop, and $\delta\omega$ is an infinitesimal angle, such that $8\pi^2/\delta\omega$ are the rotational degrees of freedom of a tetramer or dimer in solution (4π for the directions a dimer or tetramer can point on the unit sphere, and 2π for the rotation around a dimer's or tetramer's axis). The left-hand column shows the statistical weights of the ten states in the language of thermodynamics: here $[T]$ and $[D]$ are concentrations of tetrameric and dimeric repressor, K_{id} and K_1 are the dissociation constants for a dimer or tetramer bound to one of the operators, and J_{loop} is the looping J-factor (defined in terms of ΔF_{loop} in the text).

binding constants for both nonspecific DNA and operator DNA have values which are independent of whether the repressor is in dimeric or tetrameric form. Note that this assumption results in a larger perturbation to the simple model in which dimers are not considered, compared to the case where dimers bind more weakly than tetramers (and therefore have less of an effect on looping).

The derivation in [115] assumes a tetramer in solution has total energy ϵ_{sol} , and a tetramer with one head bound has total energy $\epsilon_b + \frac{\epsilon_{sol}}{2}$, where ϵ_b is the energy of one head bound to an operator. It is implied, then, that the remaining free head in solution has energy $\frac{\epsilon_{sol}}{2}$. Therefore we assume a dimer free in solution has energy $\frac{\epsilon_{sol}}{2}$ (and ϵ_b when bound to an operator), since it only has one binding head. Second, we assume that tetramers and dimers have the same number of rotational configurations in solution, or $\frac{8\pi^2}{\delta\omega}$ per dimer.

The previous two points contain an implicit assumption that the unbound head of a tetramer does not bind nonspecifically to non-operator DNA when the other head is bound to an operator; otherwise dimers and tetramers would have different dissociation constants, either through a change to the energy of the one-head-bound state for tetramers, or through the restriction of the entropy of tetramers with one head bound compared to dimers bound to an operator. This assumption about nonspecific binding of free heads is a reasonable first approximation, particularly given the equivalence of the dissociation constants for some dimeric mutants and wild-type tetramers described above. However, these kinds of simplifications can all be relaxed if desired without changing the essence of the calculations.

Finally, as discussed in more detail in Section 2.4, we assume that the binding of repressors to DNA does not affect the equilibrium between dimers and tetramers in solution.

A.2 Derivation of $p_{\text{loop}}([R])$, taking into account $T \Leftrightarrow 2D$

We start by enumerating the possible states of the system and their Boltzmann weights, and summing the weights to obtain the partition function. We will do this first in terms of numbers of dimers and tetramers (D and T , respectively), and then use the equilibrium constant for the dissociation reaction of tetramers into dimers, K_{DT} , to write the partition function in terms of K_{DT} and the

total concentration of repressor $[R]$.

When dimers are present in solution, there are 5 new states that the system can be in, in addition to the 5 states of the simple model that does not include dimers. All 10 states are shown schematically in Fig. A.1, along with the statistical weights which we will now derive.

To construct the weight of each state, we first note that the energy of each state is $\epsilon_{\text{sol}}/2$ times the number of tetramer heads in solution, plus $\epsilon_{\text{sol}}/2$ times the number of dimers free in solution, plus ϵ_{id} per dimer or tetramer bound to the O_{id} operator, plus ϵ_1 per dimer or tetramer bound to the O_1 operator, plus ΔF_{loop} if a loop is formed. So, for example, if a tetramer is bound at O_{id} and a dimer at O_1 , as in state (viii), the argument of the exponential is $-\beta[(T-1)\epsilon_{\text{sol}}+(D-1)\epsilon_{\text{sol}}/2+\epsilon_1+\epsilon_{id}+\epsilon_{\text{sol}}/2]$, where β is the reciprocal of the Boltzmann constant times the temperature.

The multiplicity of each state consists of three parts. First, as in the simple model of [115] (and, implicitly, of Chapter 2), there are $\frac{8\pi^2}{\delta\omega}$ rotational configurations per repressor in solution; so, for example, in the state where there are neither dimers nor tetramers bound to the DNA (state (i)), there are T tetramers and D dimers in solution, or $\left(\frac{8\pi^2}{\delta\omega}\right)^{T+D}$ total rotational configurations. Second, since each tetramer head has 2 orientations in which it can bind to the DNA and each tetramer has 2 heads, a tetramer bound at an operator contributes a factor of 4 to the multiplicity (2 heads times 2 configurations per head). However a dimer bound at either operator contributes a multiplicative factor of 2, since each dimer has only one binding domain. Finally we account for the ways of arranging the tetramers and dimers in solution by using a lattice model to describe the solution, as in [115]. The lattice has Ω lattice sites, so for example in state (viii), with a tetramer bound at O_{id} and a dimer bound at O_1 , there are $\frac{\Omega!}{(T-1)!(D-1)!(\Omega-T-D-2)!}$ ways of arranging the remaining $T-1$ tetramers and $D-1$ dimers in solution.

We next apply some simplifications. If we assume a dilute solution, such that $\Omega \gg T + D$, then

$$\frac{\Omega!}{(\Omega - (T + D))!} \approx \Omega^{T+D} \quad (\text{A.1})$$

and likewise for similar terms. The parts of the multiplicities of each state that correspond to the

ways of arranging the dimers and tetramers in solution then become:

$$(i) \quad \frac{\Omega!}{T!D!(\Omega - T - D)!} \rightarrow \frac{\Omega^{T+D}}{T!D!} \quad (\text{A.2})$$

$$(ii), (iii) \quad \frac{\Omega!}{(T-1)!D!(\Omega - T - D + 1)!} \rightarrow \frac{\Omega^{T+D-1}}{(T-1)!D!} \quad (\text{A.3})$$

$$(iv) \quad \frac{\Omega!}{(T-2)!D!(\Omega - T - D + 2)!} \rightarrow \frac{\Omega^{T+D-2}}{(T-2)!D!} \quad (\text{A.4})$$

$$(v) \quad \frac{\Omega!}{(T-1)!D!(\Omega - T - D + 1)!} \rightarrow \frac{\Omega^{T+D-1}}{(T-1)!D!} \quad (\text{A.5})$$

$$(vi), (vii) \quad \frac{\Omega!}{T!(D-1)(\Omega - T - D + 1)!} \rightarrow \frac{\Omega^{T+D-1}}{T!(D-1)!} \quad (\text{A.6})$$

$$(viii), (ix) \quad \frac{\Omega!}{(T-1)!(D-1)(\Omega - T - D + 2)!} \rightarrow \frac{\Omega^{T+D-2}}{(T-1)!(D-1)!} \quad (\text{A.7})$$

$$(x) \quad \frac{\Omega!}{T!(D-2)(\Omega - T - D + 2)!} \rightarrow \frac{\Omega^{T+D-2}}{T!(D-2)!} \quad (\text{A.8})$$

where the Roman numerals correspond to the numbering of the states in Fig. A.1.

Next we divide each weight by the weight of state (i) so that the weights of the 10 states become:

$$(i) \rightarrow 1 \quad (\text{A.9})$$

$$(ii) \rightarrow 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \left(\frac{\Omega^{T+D-1}}{(T-1)!D!} \frac{T!D!}{\Omega^{T+D}} \right) e^{-\beta[(T-1)\epsilon_{sol} + D\frac{\epsilon_{sol}}{2} + \epsilon_1 + \frac{\epsilon_{sol}}{2} - T\epsilon_{sol} - D\frac{\epsilon_{sol}}{2}]} \quad (\text{A.10})$$

$$= 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta(\epsilon_1 - \frac{\epsilon_{sol}}{2})} \quad (\text{A.11})$$

$$(iii) \rightarrow 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \left(\frac{\Omega^{T+D-1}}{(T-1)!D!} \frac{T!D!}{\Omega^{T+D}} \right) e^{-\beta[(T-1)\epsilon_{sol} + D\frac{\epsilon_{sol}}{2} + \epsilon_{id} + \frac{\epsilon_{sol}}{2} - T\epsilon_{sol} - D\frac{\epsilon_{sol}}{2}]} \quad (\text{A.12})$$

$$= 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta(\epsilon_{id} - \frac{\epsilon_{sol}}{2})} \quad (\text{A.13})$$

$$(iv) \rightarrow 16 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \frac{T^2}{\Omega^2} e^{-\beta[T\epsilon_{sol} - 2\epsilon_{sol} + D\frac{\epsilon_{sol}}{2} + 2\frac{\epsilon_{sol}}{2} + \epsilon_1 + \epsilon_{id} - T\epsilon_{sol} - D\frac{\epsilon_{sol}}{2}]} \quad (\text{A.14})$$

$$= 16 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \frac{T^2}{\Omega^2} e^{-\beta(\epsilon_1 + \epsilon_{id} - 2\frac{\epsilon_{sol}}{2})} \quad (\text{A.15})$$

$$(v) \rightarrow 8 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta[T\epsilon_{sol} - \epsilon_{sol} + D\frac{\epsilon_{sol}}{2} + \epsilon_1 + \epsilon_{id} + \Delta F_{loop} - T\epsilon_{sol} - D\frac{\epsilon_{sol}}{2}]} \quad (\text{A.16})$$

$$= 8 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta(\epsilon_1 + \epsilon_{id} + \Delta F_{loop} - 2\frac{\epsilon_{sol}}{2})} \quad (\text{A.17})$$

$$(vi) \rightarrow 2 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \left(\frac{\Omega^{T+D-1}}{T!(D-1)!} \frac{T!D!}{\Omega^{T+D}} \right) e^{-\beta[T\epsilon_{sol} + D\frac{\epsilon_{sol}}{2} - \frac{\epsilon_{sol}}{2} + \epsilon_1 - T\epsilon_{sol} - D\frac{\epsilon_{sol}}{2}]} \quad (\text{A.18})$$

$$= 2 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{D}{\Omega} e^{-\beta(\epsilon_1 - \frac{\epsilon_{sol}}{2})} \quad (\text{A.19})$$

$$(vii) \rightarrow 2 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \left(\frac{\Omega^{T+D-1}}{T!(D-1)!} \frac{T!D!}{\Omega^{T+D}} \right) e^{-\beta[T\epsilon_{sol} + D\frac{\epsilon_{sol}}{2} - \frac{\epsilon_{sol}}{2} + \epsilon_{id} - T\epsilon_{sol} - D\frac{\epsilon_{sol}}{2}]} \quad (\text{A.20})$$

$$= 2 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{D}{\Omega} e^{-\beta(\epsilon_{id} - \frac{\epsilon_{sol}}{2})} \quad (\text{A.21})$$

$$(viii), (ix) \rightarrow 8 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \left(\frac{\Omega^{T+D-2}}{(T-1)!(D-1)!} \frac{T!D!}{\Omega^{T+D}} \right) e^{-\beta[T\epsilon_{sol} - \epsilon_{sol} + D\frac{\epsilon_{sol}}{2} - \frac{\epsilon_{sol}}{2} + \epsilon_1 + \epsilon_{id} + \frac{\epsilon_{sol}}{2} - T\epsilon_{sol} - D\frac{\epsilon_{sol}}{2}]} \quad (\text{A.22})$$

$$= 8 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \frac{TD}{\Omega^2} e^{-\beta(\epsilon_1 + \epsilon_{id} - 2\frac{\epsilon_{sol}}{2})} \quad (\text{A.23})$$

$$(x) \rightarrow 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \left(\frac{\Omega^{T+D-2}}{T!(D-2)!} \frac{T!D!}{\Omega^{T+D}} \right) e^{-\beta[T\epsilon_{sol} + D\frac{\epsilon_{sol}}{2} - 2\frac{\epsilon_{sol}}{2} + \epsilon_1 + \epsilon_{id} - T\epsilon_{sol} - D\frac{\epsilon_{sol}}{2}]} \quad (\text{A.24})$$

$$= 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \frac{D^2}{\Omega^2} e^{-\beta(\epsilon_1 + \epsilon_{id} - 2\frac{\epsilon_{sol}}{2})}. \quad (\text{A.25})$$

Finally we define $\Delta\epsilon_1 \equiv \epsilon_1 - \frac{\epsilon_{sol}}{2}$ and $\Delta\epsilon_{id} \equiv \epsilon_{id} - \frac{\epsilon_{sol}}{2}$. (Note that this is the same convention as in [115]: there $\Delta\epsilon \equiv \epsilon_b + \epsilon_t - \epsilon_{sol}$, where ϵ_t is the energy of the unbound head in solution, and then it is assumed that there is no cooperativity to the binding of the second head, such that $\epsilon_t = \epsilon_{sol}/2$, as we have assumed here.) Note that in the arguments of the exponentials there is always an $\frac{\epsilon_{sol}}{2}$ to go with each ϵ_{id} and ϵ_1 , so all the ϵ_{sol} 's disappear and all $\epsilon_1, \epsilon_{id}$ become $\Delta\epsilon_1, \Delta\epsilon_{id}$.

We can now write the total partition function as

$$\begin{aligned} Z = & 1 + 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta\Delta\epsilon_1} + 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta\Delta\epsilon_{id}} + 16 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \frac{T^2}{\Omega^2} e^{-\beta(\Delta\epsilon_1 + \Delta\epsilon_{id})} \quad (\text{A.26}) \\ & + 8 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta(\Delta\epsilon_1 + \Delta\epsilon_{id} + \Delta F_{loop})} + 2 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{D}{\Omega} e^{-\beta\Delta\epsilon_1} + 2 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{D}{\Omega} e^{-\beta\Delta\epsilon_{id}} \\ & + 16 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \frac{TD}{\Omega^2} e^{-\beta(\Delta\epsilon_1 + \Delta\epsilon_{id})} + 4 \left(\frac{8\pi^2}{\delta\omega} \right)^{-2} \frac{D^2}{\Omega^2} e^{-\beta(\Delta\epsilon_1 + \Delta\epsilon_{id})} \end{aligned}$$

where we have combined states (viii) and (ix) into one term since they are mathematically identical.

To convert from numbers of dimers and tetramers to concentrations, we start by defining the

total amount of repressor in the TPM sample, $[R]$, as

$$[R] = \frac{[D]}{2} + [T], \quad (\text{A.27})$$

as in Chapter 2. As in [115], we will define $[T]$ and $[D]$ in terms of the number of lattice sites as

$$[T] \equiv \frac{T}{\Omega v} \quad \text{and} \quad [D] \equiv \frac{D}{\Omega v}, \quad (\text{A.28})$$

where v is the size of each lattice site, chosen such that

$$\frac{1}{v} \frac{8\pi^2}{\delta\omega} = 1\text{M}. \quad (\text{A.29})$$

Also in keeping with [115] we define the dissociation constants as

$$K_d \equiv \frac{1}{4v} \frac{8\pi^2}{\delta\omega} e^{\beta\Delta\epsilon}, \quad (\text{A.30})$$

where $\Delta\epsilon$ is either $\Delta\epsilon_{id}$ or $\Delta\epsilon_1$. In units of concentration this becomes

$$K_d = \frac{1}{4} \text{M} e^{\beta\Delta\epsilon}. \quad (\text{A.31})$$

Finally we define the J-factor as

$$J_{\text{loop}} \equiv \frac{1}{v} \frac{8\pi^2}{\delta\omega} e^{-\beta\Delta F_{\text{loop}}}, \quad (\text{A.32})$$

or in units of concentration,

$$J_{\text{loop}} = 1 \text{ M} e^{-\beta\Delta F_{\text{loop}}}. \quad (\text{A.33})$$

We can now write the partition function in terms of $[T]$, $[D]$, K_{id} , K_1 , and J_{loop} . Consider first the looped state (state (v)),

$$8 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta(\Delta\epsilon_1 + \Delta\epsilon_{id} + \Delta F_{\text{loop}})}. \quad (\text{A.34})$$

If we replace T/Ω with $[T]v$ and group terms that can be combined into J-factors and dissociation constants we can rewrite Eq. (A.34) as

$$8 \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} \frac{T}{\Omega} e^{-\beta(\Delta\epsilon_1 + \Delta\epsilon_{id} + \Delta F_{loop})} = [T] \left(8v \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} e^{-\beta\Delta\epsilon_1} \right) \left(e^{-\beta\Delta\epsilon_{id}} \frac{4v}{\frac{8\pi^2}{\delta\omega}} \right) \left(\frac{\frac{8\pi^2}{\delta\omega}}{4v} e^{-\beta\Delta F_{loop}} \right). \quad (\text{A.35})$$

Then replacing grouped terms with J_{loop} or the appropriate K_d yields

$$[T] \left(8v \left(\frac{8\pi^2}{\delta\omega} \right)^{-1} e^{-\beta\Delta\epsilon_1} \right) \left(e^{-\beta\Delta\epsilon_{id}} \frac{4v}{\frac{8\pi^2}{\delta\omega}} \right) \left(\frac{\frac{8\pi^2}{\delta\omega}}{4v} e^{-\beta\Delta F_{loop}} \right) = [T] \frac{2}{K_1} \frac{1}{K_{id}} \frac{J_{loop}}{4}, \quad (\text{A.36})$$

which is the same as in the original model, with $[R]$ replaced by $[T]$ (which makes sense since only tetramers can form loops). Similar manipulations can be applied to the rest of the states, yielding a new partition function of

$$Z = 1 + \frac{[T]}{K_1} + \frac{[T]}{K_{id}} + \frac{[T]^2}{K_1 K_{id}} + \frac{[T]J_{loop}}{2K_1 K_{id}} + \frac{[D]}{2K_1} + \frac{[D]}{2K_{id}} + \frac{[T][D]}{K_1 K_{id}} + \frac{[D]^2}{4K_1 K_{id}} \quad (\text{A.37})$$

$$= 1 + \frac{2[T] + [D]}{K_{id}} + \frac{2[T] + [D]}{K_1} + \frac{[T]^2 + [T][D]}{K_1 K_{id}} + \frac{[T]J_{loop}}{2K_1 K_{id}} + \frac{[D]^2}{4K_1 K_{id}}. \quad (\text{A.38})$$

Since $[D]/2 + [T] = [R]$ (Eq. (A.27)),

$$Z = 1 + \frac{[R]}{K_{id}} + \frac{[R]}{K_1} + \frac{[R]^2}{K_1 K_{id}} + \frac{[T]J_{loop}}{2K_1 K_{id}}, \quad (\text{A.39})$$

which is the same as Eq. (2.16) in Chapter 2.

Finally we will use K_{DT} , the equilibrium constant for $T \Leftrightarrow 2D$, to express $[T]$ in terms of $[R]$ and K_{DT} . We start with the definition of K_{DT} :

$$K_{DT} = \frac{[D]^2}{[T]}, \quad (\text{A.40})$$

or

$$[T] = \frac{[D]^2}{K_{DT}}. \quad (\text{A.41})$$

Substituting this into Eq. (A.27) yields

$$\frac{[D]}{2} + \frac{[D]^2}{K_{DT}} = [R], \quad (\text{A.42})$$

which we can solve for $[D]$:

$$[D] = \frac{1}{4} \left(-K_{DT} + \sqrt{K_{DT}^2 + 16K_{DT}[R]} \right) \quad (\text{A.43})$$

where the positive root is chosen because $[D]$ must be positive and the discriminant is always positive.

However we want $[T]$, not $[D]$; so we again use Eq. (A.27) to say that

$$[T] = [R] - \frac{[D]}{2} = [R] - \frac{1}{8} \left(-K_{DT} + \sqrt{K_{DT}^2 + 16K_{DT}[R]} \right). \quad (\text{A.44})$$

By substituting this into Eq. (A.39) we obtain our final partition function of

$$Z = 1 + \frac{[R]}{K_{id}} + \frac{[R]}{K_1} + \frac{[R]^2}{K_1 K_{id}} + \frac{[R]J_{\text{loop}}}{2K_1 K_{id}} \left(1 - \frac{1}{8[R]} \left(-K_{DT} + \sqrt{K_{DT}^2 + 16K_{DT}[R]} \right) \right). \quad (\text{A.45})$$

Since the looping probability is the weight of the looped state divided by the partition function, we have our final result that

$$p_{\text{loop, dimers}} = \frac{\frac{[R]J_{\text{loop}}}{2K_1 K_{id}} \left(1 + \frac{K_{DT}}{8[R]} - \frac{1}{8[R]} \sqrt{K_{DT}^2 + 16K_{DT}[R]} \right)}{1 + \frac{[R]}{K_{id}} + \frac{[R]}{K_1} + \frac{[R]^2}{K_1 K_{id}} + \frac{[R]J_{\text{loop}}}{2K_1 K_{id}} \left(1 + \frac{K_{DT}}{8[R]} - \frac{1}{8[R]} \sqrt{K_{DT}^2 + 16K_{DT}[R]} \right)}. \quad (\text{A.46})$$

A.3 Dimers due to damaged protein

A mixture of dimeric and tetrameric repressors could be present in a TPM experiment for two different reasons. The first, already discussed in the previous section, stems from the use of such low concentrations of repressor that the tetramer-to-dimer dissociation reaction needs to be taken into consideration. A second case is one in which a fraction of the repressors is damaged in some way due to the purification, storage, or thawing process, leading to an inability of some repressors

to form tetramers. This will lead to a fraction of dimers, ν , that is constant with the total repressor concentration $[R]$. (As noted in Chapter 2, we could consider a third case, in which a fraction of monomers that are damaged such that when incorporated into a tetramer they result in a head that is unable to bind DNA. We will not consider this case here but it is well within the scope of scenarios that can be captured by the class of models presented here.)

Note that since we assumed in the previous section that the binding of dimers and tetramers to the DNA did not affect the equilibrium reaction between tetramers and dimers in solution, we can start with Eq. (A.39), since the derivations for the two cases are the same up to this point.

We will define the concentration of repressors in tetrameric form as

$$[T] = (1 - \nu)[R], \quad (\text{A.47})$$

where ν is the fraction that are dimers. Again we are considering here the case where the dimeric fraction is constant with the total concentration. Then because $\frac{[D]}{2} + [T] = [R]$, we must define ν as

$$\nu = \frac{[D]}{2[R]}, \quad (\text{A.48})$$

so that $[D]/2 + [T] = 2\nu[R]/2 + (1 - \nu)[R] = [R]$. We can now use this expression for $[T]$ in Eq. (A.39), so that when we form the looping probability we obtain

$$p_{\text{loop, dimers}} = \frac{\frac{(1-\nu)[R]J}{2K_1K_{id}}}{1 + \frac{[R]}{K_{id}} + \frac{[R]}{K_1} + \frac{[R]^2}{K_1K_{id}} + \frac{(1-\nu)[R]J}{2K_1K_{id}}}. \quad (\text{A.49})$$

Note that this is the same result we obtained in the previous section, if we use the expression for $[D]$ in Eq. (A.43) in the definition of ν as $\frac{[D]}{2[R]}$, except that here ν is a scalar, not a function of $[R]$.

Appendix B

DNA_s

B.1 Constructs containing E8 and 601TA

The E8- and TA-containing constructs discussed in Chapters 3 and 4 are PCR products of plasmids pZS25' Oid-E/T(89-116)-O1₋₄₅-YFP, where “E/T(89-116)” indicates that the sequence of the loop is either from the random E8 sequence or the 601TA sequence from [85] and has a length of 89 to 116 bp. The original constructs used in [115] and [120] (lengths 89, 94, and 100 bp) were constructed by site-directed mutagenesis as described in [115]. Jonathan Widom kindly provided the E8 and TA sequences used in [85], which are a subset of those used here and from which the other E8 and TA lengths were derived. QuikChange site-directed mutagenesis (Agilent Technologies) was used to make the operator changes O_{id} to O_1 and O_{id} to O_2 , additional loop lengths, and the promoter-containing constructs. Linear labeled DNAs used in tethering assays were created by PCR with primers labeled at the 5' ends with digoxigenin (forward primers) or biotin (reverse primers) (Eurofins MWG Operon); a PCR of the pZS25' plasmids resulted in approximately 450 bp tethers

Name	Sequence
O_1	AATTGTGAGCGGATAACAATT
O_2	GGTTGTTACTCGCTCACATTT
O_3	GGCAGTGAGCGCAACGCAATT
O_{id}	AATTGTGAGCGCTCACAATT

Table B.1: Sequences of the three naturally occurring Lac repressor operators O_1 , O_2 , and O_3 , and of the synthetic O_{id} (“Oideal”) operator. All sequences are 5' to 3' and are from [94]. Note that O_{id} is perfectly symmetric about its midpoint, whereas the naturally occurring, weaker operators are only pseudo-symmetric, with O_1 being the strongest, O_2 weaker, and O_3 the weakest. In this work, the loop is to the 3' end of the appropriate ($O_{id}/O_1/O_2$) operator sequence shown here; the O_1 that is constant in all constructs (nearest to the bead) has the loop 5' to the sequence given here.

E89: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E90: GGCCG-----TGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E91: GGCCG-----CTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E92: GGCCG-----CTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E93: GGCCG-----CTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E94: GGCCG-----TGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E95: GGCCG-----GCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E96: GGCCG-----AGGCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E97: GGCCG-----AGGCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E98: GGCCG-----AGGCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E99: GGCCG-----GAGGCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E100: GGCCG-----GAGGCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E101: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E102: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E103: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E104: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E105: GGCCG-----TGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E106: GGCCG-----CTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E107: GGCCG-----GCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E108: GGCCG-----GCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E109: GGCCG-----AGGCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E110: GGCCG-----AGGCTGCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E111: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E112: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E113: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E114: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E115: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C
E116: GGCCG-----GCTGCGTAGAACTACTTTTATTATCGCCTCCACGGTGTGATCCCTGTGCTTTGGCCGTGTATCTCGAGT**AGT**ACGAC-----C

T89: -----GGCCG-----GGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T90: -----GGCCG-----TGCTGCTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T91: -----GGCCG-----TTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T92: -----GGCCG-----TTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T93: -----GGCCG-----GGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T94: -----GGCCG-----GGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T95: -----GGCCG-----ATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T96: -----GGCCG-----AATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T97: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T98: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T99: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T100: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T101: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T102: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T103: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T104: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T105: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T106: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T107: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T108: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T109: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T110: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T111: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T112: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T113: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T114: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T115: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C
T116: -----GGCCG-----TAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGT-----C

601TA: ctggagatacCCGGTCTAAGCCGCTTAATTGGTCGTAGCAAGCTCTAGCACCCTTAAACGCACGTACGGCGTGTCTACCGCGTTTAAACGCCAATAGGATTACTTACTAGTCTCTAGG
CACGTG**T**-agatatatacatctgtgatgta

Figure B.1: Sequences of the no-promoter E8 and TA constructs used in this work (Fig. 4.2(A–C)). All sequences are 5′ to 3′ and are listed such that the O_{id} operator (or O_1 or O_2 operator in the case of the E894 sequence) is immediately 5′ of these sequences, and O_1 is immediately 3′. Bolded sequence labels indicate constructs examined by cyclization in [85], which were incorporated into the pZS25′ plasmid by Hernan Garcia; the rest were designed and created by Stephanie Johnson. In the top section containing the E8 sequences, dashes indicate bases missing relative to the 116 bp E8 sequence listed at the bottom of that section. In the bottom section containing the TA sequences, dashes indicate bases missing relative to the full 154 bp 601TA sequence (provided to us by Jon Widom; see also [86]) listed below the TA sequences; in that 601TA sequence, the dash indicates where a C has been inserted at the end of all of the TA sequences used in both cyclization [58, 85] and in the looping work presented here. Upper-case letters in the full 601TA sequence indicate the region from which all TA sequences in this work were derived. The 601TA sequence is so named because of the TA dinucleotide steps which occur every 10 bp and which are thought to confer its affinity for nucleosome formation [14]; these TA steps have been highlighted in red. Note that the E8 sequence also has several TA steps spaced 10 bp apart; however this pattern does not repeat across the entire sequence as it does in the 601TA sequence, nor does the E8 sequence have other characteristics of the 601TA sequence such as GC pairs between the TA pairs which are also supposed to be important for its particular properties [14, 52].

```

E92: -----CTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E93: -----CCTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E94: -----GCCTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E95: -----CGCCTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E96: -----TCGCCTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E97: -----ATCGCCTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E101: -----ATTATCGCCTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E103: -----TTATTTATCGCCTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E105: -----TTTTATTTATCGCCTCCACGGTGTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E114: -----TAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E115: -----GTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E116: -----CGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E117: -----GCGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E118: -----TGGGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E119: -----CTGGGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E120: -----GCTGGGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E121: -----GGGTGGGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E122: -----CGGTGGGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E123: -----CCGGTGGGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC
E124: G-CCGGTGGGTAGAACTACTTTTATTTATCGCCTCCACGGTGTGCTGATCCCCTGTGCTGTTGGCCGTGTTATCTCGAGTTAGTACGACC

T92: -----ACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T93: -----AACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T94: -----AAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T95: -----TAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T96: -----TAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T97: -----CTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T101: -----ACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T103: -----GCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T105: -----TAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T106: -----CTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T107: -----TCTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T108: -----CTCTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T114: -----AGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T115: -----TAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T116: -----GTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T117: -----CGTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T118: -----TCGTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T119: -----GTCTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T120: -----GGTCTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T121: -----GGGTCTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T122: -----CGGGTCTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T123: -----CCGGTCTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC
T124: G-CCGGTCTAGCAAAGCTTAGCACCGCTTAAACGCAGTACGGCGTGTCTACCGGTTTTAACCGCAATAGGATTACTACTAGTC

```

Figure B.2: Sequences of the with-promoter E8 and TA constructs used in this work (Fig. 4.2(D–F)). All sequences are 5′ to 3′ and are listed such that the O_{id} operator is immediately 5′ of these sequences, and O_2 is immediately 3′ (but O_2 is the reverse complement of the sequence given in Table B.1). The *lacUV5* promoter is to the 3′ end, before the O_2 operator; its sequence is TTTACAATTAATGCTTCCGGCTCGTATAATGTGTGG. As in Fig. B.1, TA steps have been highlighted in red. Dashes indicated bases missing from the 89 bp no-promoter equivalents shown in the previous figure.

```

PolyA105:  ACCTTGATTGTATTTCCTTTGCGTGATGAAAAAAAAAACTGAAAAAGAGAAAAATAAGAAAATCTTCTAGAACGTTCCGAAACAGGAC-gtgctgatccccctgtgc--
PolyA106:  ACCTTGATTGTATTTCCTTTGCGTGATGAAAAAAAAAACTGAAAAAGAGAAAAATAAGAAAATCTTCTAGAACGTTCCGAAACAGGAC-gtgctgatccccctgtgc-
PolyA107:  ACCTTGATTGTATTTCCTTTGCGTGATGAAAAAAAAAACTGAAAAAGAGAAAAATAAGAAAATCTTCTAGAACGTTCCGAAACAGGACggtgctgatccccctgtgc-
PolyA108:  ACCTTGATTGTATTTCCTTTGCGTGATGAAAAAAAAAACTGAAAAAGAGAAAAATAAGAAAATCTTCTAGAACGTTCCGAAACAGGACggtgctgatccccctgtgc-

PolyA105(prom): -----TTGCGTGATGAAAAAAAAAACTGAAAAAGAGAAAAATAAGAAAATCTTCTAGAACGTTCCGAAACAGGAC
PolyA106(prom): -----TTGCGTGATGAAAAAAAAAACTGAAAAAGAGAAAAATAAGAAAATCTTCTAGAACGTTCCGAAACAGGAC
PolyA107(prom): -----CTTTGCGTGATGAAAAAAAAAACTGAAAAAGAGAAAAATAAGAAAATCTTCTAGAACGTTCCGAAACAGGAC
PolyA108(prom): -----CCTTTGCGTGATGAAAAAAAAAACTGAAAAAGAGAAAAATAAGAAAATCTTCTAGAACGTTCCGAAACAGGAC

```

Figure B.3: Sequences of the poly(dA:dT)-rich sequences used in Chapter 4, which are derived from the nucleosome-free region of the *S. cerevisiae* promoter given in Fig. 4E of [149]. This 88 bp poly(dA:dT)-rich sequence is shown in capital letters, with stretches of more than four consecutive A's highlighted in green, where we have chosen to define an A-tract as 4 or more A's in a row because this is the shortest length that shows special structural properties under a variety of methods [148]. The top section lists no-promoter sequences; the bottom section, sequences to which the 36 bp *lacUV5* promoter was added to the loop as in Fig. B.2. As in that figure and Fig. B.1, the O_{id} operator is to the left of all of these sequences, and the O_1 operator (for the no-promoter sequences) or the promoter and then O_2 (for the with-promoter sequences) to the right. The DNA flanking the loop region are the same as those for the E8 and TA constructs (lengths given in Fig. 1.4). As the poly(dA:dT)-rich region of [149] is only 88 bp, the no-promoter constructs were padded with a portion of the E8 sequence (which should be a random sequence); these bases are shown in lower-case letters. Dashes in the no-promoter construct indicate where bases were removed relative to the 108 bp construct. In the with-promoter construct, dashes indicate bases removed relative to the 88 bp sequence from [149].

(see Fig. 1.4 for flanking DNA lengths). Primer sequences can be found in Table 3 of [115]. The PCR product was gel purified using a QIAquick Gel Extraction Kit (Qiagen), and the concentration determined by quantitative gel electrophoresis.

Table B.1 gives the sequences of the three naturally occurring operators of the *lac* operon and the strong synthetic operator O_{id} used in some of the work discussed here. Figures B.1 and B.2 shows the E8 and TA sequences that form the loops in the constructs discussed in Chapters 3 and 4. All constructs were verified by sequencing (Laragen).

B.2 Constructs containing poly(dA:dT)

The poly(dA:dT)-rich sequence used in Section 4.5 was taken from the top row (the *S. cerevisiae* sequence) of Fig. 4E of [149], and ligated into the AatII and EcoRI restriction sites that fall just outside the operators of the pZS25' plasmids described in the previous section. The 106 bp no-promoter and with-promoter sequences given in Fig. B.3, plus the sequences of the relevant operators (and promoter, if applicable), and the restriction sites were ordered as single-stranded oligonucleotides from IDT. The oligonucleotides were annealed and then ligated into one of the pZS25' plasmids which had been doubly digested with AatII and EcoR1 (NEB) and gel purified. Successful ligation was confirmed by sequencing (Laragen), and the approximately 450 bp dig- and

5' - GACTGTCTGGCCGTAACCGACCCAGCGCCCGTTGCACCACAGATGAAACGCCGAGTTAACGCCATCAAAAATAATTCGGCTCTGGCCTTCTGTAGCCAGCTTTCATCAACATTAATGTTGAGCGAGTAACAACCCGTCGGATTCTCCGTGGGAACAAACGGCGGATTGACCGTAATGGGATAGGTCACGTTGGTGTAGATGGGGCCATCGTAACCGTGATCTGCCAGTTTGAGGGGACGACGACAGTATCGGCCTCAGGAAGATCGCACTCCAGCCAGCTTTCGGCACCGCTTCTGGTGCCTGAAACCGCAAGGCCAATTCGCCATTTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGG GCCTTTCGCTATTACGCCAGCTGGCGAAAGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTCCAGTCACGACGTTGTAACGACGCGCCAGTGAATCCGTAATCATGGT CATAGCTGTTTCTGTGTGAATGTTATCCGCTCAATT↓CCACAC aacata CGAGCCGGAAGCATAAAG tgaata GCCTGGGTGCCTAATGAGTGAGCTAACTCACATTAATTCGGCTTTCGCTACTGCGCTTTCAGTGGGAAACCTGTCTGCGCAGCTGCATTAATGAATCGGCCAACGCGGGGAGAGGGCGTTTTCGCTATTGGGCGCCAGGGTGGTTTTTCTTTTCAC - 3'

Figure B.4: Sequence of the region of the *lac* operon amplified by colony PCR for the three-operator (and derivative two-operator) constructs in Chapter 6. Note that this sequence is shown such that, as with all the other sequences in this section, the left-most end of the sequence would be attached to the surface in the TPM assays, and the right-most end to the bead. This is reversed, however, from the way promoter regions are usually presented, with the downstream (coding regions) of the gene to the right and the upstream (regulatory) regions to the left (see Fig. 6.1). Here, transcription would occur *right to left*, beginning at the arrow next to O_1 . Colors are the same as in the schematics of Chapter 6: O_1 is shown in purple, O_2 in green, and O_3 in blue. The -10 and -35 regions where RNA Polymerase binds are shown in lower-case dark green letters. The CAP binding site is in orange.

biotin-labeled TPM construct created by PCR as described for the E8- and TA-containing constructs in the previous section. As PCRs often introduce mutations into repetitive sequences (such as these AT-rich DNAs), the TPM constructs were again confirmed by sequencing (Laragen). Additional lengths were created via site-directed mutagenesis (Qiagen).

B.3 Constructs derived from the naturally occurring *lac* operon

The three-operator, wild-type construct used in Chapter 6 was derived from a colony PCR of MG1655 *E. coli* followed by a second PCR with digoxigenin- and biotin-labeled primers. The colony PCR consisted of the dilution of a single colony from an overnight growth on an LB plate into 20 μ L water; 1 μ L of this mixture was then used as the template in a 25 μ L PCR reaction performed using an AccuPrime Pfx SuperMix kit (Invitrogen), with 200 nM of each of the primers “wtLac_extract_fwd” and “wtLac_extract_rev” (Table B.2). The cycling parameters were: 5 minutes initial denaturation at 95° C, and 35 cycles of: 15 seconds denaturation at 95° C, 30 seconds annealing at 60° C, and 1 minute extension at 68° C. The ~1 kb product of this colony PCR was purified with a QIAquick PCR Purification Kit (Qiagen), and 80 ng of this product were used in a second, 50 μ L PCR to add the digoxigenin and biotin primers, with 2.5 U Taq DNA Polymerase (Roche), 1x PCR Reaction Buffer with 15 mM MgCl₂ (Roche), 20 μ M of each of the “TPM_wtLac_fwddig” and “TPM_wtLac_revbio primers” (Table B.2), and 10 mM each dNTP. The cycling parameters were: 5 minutes initial denaturation at 94° C, 10 cycles of: 30 seconds denaturation at 94° C, 45 seconds annealing at 58° C, and 1 minute extension at 72° C, and a final 10 minute extension at 72° C. The

Primer Name	Sequence
wtLac_extract_fwd	CACGGAAAATGCCGCTCATC
wtLac_extract_rev	GGGATACGACGATACCGAAGACAG
TPM_wtLac_fwddig	Dig-GACTGTCTTGGCCGTAACCGACC
TPM_wtLac_revbio	Bio-GTGAAAAGAAAAACCACCCTGGCG
wtLac_noO1_rev	cgtatggtgtgtgg GATTGTTAGCGGAGAAGAATT tcacacaggaacagc
wtLac_noO2_rev	cccacggagaatccgacg GGGTGCTATTCATTAACATT Caatggtgatgaaagctggc
wtLac_noO3_rev	ggtttcccgactggaaagcg AACCTCGAGCTCAACGCAAT Taatgtgagttagctcac
wtLac_O3toO1_rev	ggtttcccgactggaaagcg AATTGTGAGCGGATAACAAT Taatgtgagttagctcac

Table B.2: Sequences of the PCR primers used to create the TPM constructs based off of the wild-type, three operator *lac* promoter. Mutagenesis primers were obtained from IDT and were based on [94, 93]; bases that form the operators are shown in capital letters, with bold indicating mutated bases (compare Table B.1). Dig- and bio-labeled primers were obtained from MWG Biotech. All sequences are 5' to 3'. Only reverse primers are needed for the megaprimer mutagenesis that was used to eliminate operators and change O_3 to O_1 .

735 bp resulting PCR product was gel purified as described above for the E8- and TA-containing constructs. Its sequence is shown in Fig. B.4.

The region of the *E. coli* genome containing the three Lac operators is difficult to clone (K. Matthews, personal communication), so the additional two-operator constructs discussed in Chapter 6 derived from the wild-type operon were created through a process called megaprimer mutagenesis [183, 184, 185, 186, 187], which, unlike site-directed mutagenesis, does not require the insertion of the DNA to be modified into a plasmid. The standard megaprimer mutagenesis method was modified such that the final product was dig- and bio-end-labeled as necessary for TPM. The megaprimer method consists of two PCR reactions.¹ The first reaction uses a forward primer (digoxigenin labeled in our case) that anneals to one end of the template to be mutated, and another primer that anneals to the middle of the template and that carries the mutation to be introduced. This first PCR creates a double-stranded “megaprimer” identical to roughly half of the template, except where the mutation has been introduced. The second PCR then involves this double-stranded “megaprimer”, purified by gel extraction, the original template, and a reverse primer that anneals to the opposite end of the original template (biotin labeled). Only one strand of the mutation-carrying megaprimer can be used for extension, thus the dominant product of this second PCR is a double-stranded DNA molecule of the same length as the original product but containing the desired mutation in

¹Ling and Robinson [183] as well as others have proposed “one-tube” or “one-step” megaprimer procedures that greatly expedite the process, mostly by eliminating the megaprimer purification step. However since the single-molecule TPM assay is very sensitive to even small amounts of contaminants, the original method was used to ensure the minimal amount of original, unmutated template would be present in the TPM sample.

the middle. Both PCRs were performed in 50 μL reactions in the presence of 2.5 U of PfuUltra Hotstart DNA Polymerase (Stratagene). The first reaction amplified 5 ng DNA and contained 0.2 mM each dNTP (Stratagene), 125 ng each “TPM_wtLac_fwddig” and the mutation-carrying primer (Table B.2), and 1x PfuUltra HF buffer (Stratagene). Cycling parameters were: 95°C 30 seconds, and 30 cycles of: 95°C 30 seconds, 55°C 1 minute, 68°C 1 minute. The second reaction again amplified 5 ng of the original template and contained the same amounts of dNTPs and buffer; however, the primers for this reaction were 156 ng “TPM_wtLac_revbio” and 20 μL (\approx 240 ng) of the “megaprimer” generated by the first PCR. Cycling parameters were: 94°C 3 minutes, 94°C 2 minutes, 60°C 2 minutes, 66°C 2 minutes, 72°C 6 minutes, and 30 cycles of: 94°C 30 seconds, 58°C 30 seconds, 72°C 50 seconds. The final product was gel extracted and sequenced (Laragen), and the concentration determined by quantitative gel electrophoresis.

Appendix C

Lac repressor purification

As discussed in Section 3.1.1, we obtained reproducible TPM results only with Lac repressor purified in-house. Our purification protocol was modified from one received from the Kathy Matthews lab in May 2009, similar to that described in [188]. Plasmid pJC1 containing the gene for wild-type, tetrameric LacI was transformed into *E. coli lacI⁻* BLIM cells (both cells and plasmid were kind gifts from the Matthews lab). Cultures were propagated in successively larger LB cultures supplemented with 0.05 $\mu\text{g}/\text{mL}$ ampicillin for either one or two days, after which they were grown in 3 L 2x YT medium (16 g/L tryptone, 10 g/L yeast extract, 5 g/L NaCl) with ampicillin for 20–24 hours at 37°C with shaking. The cells were collected by centrifugation, resuspended in ~ 45 mL cold Breaking Buffer (0.2 M Tris-HCl, pH 7.6, 0.2 M KCl, 0.01 M magnesium acetate, 5% (w/v) glucose, 0.3 mM DTT, and 50 $\mu\text{g}/\text{L}$ PMSF), supplemented with 0.5 mg/mL lysozyme (Sigma), and frozen at -20°C for at least 12 hours.

The cells were slowly thawed on ice, then cold Breaking Buffer with fresh DTT but without PMSF was added until the total volume of the cells was ~ 75 mL. 120 μL DNaseI (Sigma), at 2000 Kunitz units/mL in 0.15 M NaCl, and 3 mL 1 M MgCl_2 were added to the thawed cells, which were allowed to sit on ice for ~ 1 hour. Cell debris was pelleted by centrifugation at 14,784 rcf for 45 minutes at 4°C, then ammonium sulfate was slowly added to the supernatant at 4° to a final level of 37% saturation to precipitate the protein. After 1 hour, the precipitate was collected by centrifugation at 7700 rcf for 40 minutes at 4°C, and the pellet resuspended in 20 mL cold 0.09 M KP buffer (0.09 M potassium phosphate, pH 7.5–7.6, obtained from 0.015 M monobasic potassium

phosphate and 0.075 M dibasic potassium phosphate, 5 % (w/v) glucose, 0.3 mM DTT). The protein was dialyzed in a Spectra/Por RC membrane with MWCO 12-14,000 (Spectrum Labs) against 2 L 0.09 M KP buffer for ~4 hours at 4°C, then against a fresh 2 L of 0.09 M KP buffer overnight, then against 1 L fresh 0.09 M KP buffer for several hours the following morning.

The dialysate was spun at 7700 rcf for 30 minutes at 4°C and the supernatant purified over a phosphocellulose (Whatman P-11 Phosphocellulose) gravity-flow column equilibrated with 0.09 M KP buffer. The phosphocellulose had been charged by first suspending 12.5 g phosphocellulose in 750 mL water, letting the resin settle, pouring off the supernatant, and repeating for a total of 6 washes. The resin was then resuspended in 750 mL 0.5 M NaOH, incubated at room temperature for 5 minutes, then washed with ddH₂O in a Buchner funnel with Whatman #541 filter paper until the pH reached neutral. The resin was then suspended in 750 mL 0.5 M HCl, incubated for 5 minutes, and washed with water until the pH reached that of the water (pH ~5). Finally the resin was suspended in ~300 mL 0.09 M KP buffer without DTT, allowed to settle for 10–20 minutes, washed in the Buchner funnel with 0.09 M KP buffer until the pH of the resin was that of the 0.09 M KP (pH 7.5), resuspended in 125 mL 0.09 M KP buffer without DTT, and stored at 4°C.

After the dialysate supernatant had been loaded onto the phosphocellulose column and washed with 0.09 M KP buffer to re-establish the baseline, loosely bound proteins were washed off the column with 0.12 M KP buffer (0.12 M potassium phosphate, pH 7.5–7.6, obtained from 0.02 M monobasic potassium phosphate plus 0.1 M dibasic potassium phosphate, 5% (w/v) glucose, 0.3 mM DTT) until the baseline was re-established, and then LacI was eluted from the column with a 140 mL linear gradient formed from equal amounts of 0.12 M KP buffer and 0.3 M KP buffer (0.3 M potassium phosphate, pH 7.5–7.6, obtained from 0.05 M monobasic potassium phosphate plus 0.25 M dibasic potassium phosphate, 5% (w/v) glucose, 0.3 mM DTT). 5 mL fractions were collected; LacI eluted over ~10–15 fractions around 0.18 M KP, with the peak concentration between 1 and 2 mg/mL, using a monomer extinction coefficient of $0.6 \text{ (mg/mL)}^{-1} \text{ cm}^{-1}$ [189]. The protein was $\geq 99\%$ pure by SDS-PAGE. In one case some repressor was also purified over a Superdex 200 10/300 GL size-exclusion column (GE Healthcare) using an AKTA system and eluted as a single peak at a

molecular weight corresponding to the expected weight of a LacI tetramer. 5.5–6 μL aliquots were made from the peak fraction(s) and stored immediately at -80°C . Once removed from -80°C LacI aliquots were stored at -20°C for not more than two weeks and were thawed not more than 3 times in total.

Appendix D

Tethered particle motion: Methods

D.1 TPM sample preparation

D.1.1 Method summary

The DNA tethering protocol used for this work was essentially that of [115], with the following modifications (a full detailed protocol follows below): (1) 0.2% Tween-20 (Sigma) was added to the PTC buffer (called “TBP” in [115]) that some batches of beads were washed in, to reduce aggregation and nonspecific binding. (2) Unless otherwise indicated, the beads used in this work were 0.49- μm -diameter, streptavidin-coated polystyrene beads (Bangs) at 1.5×10^{11} beads per mL and with a binding capacity of 1.14 or 1.8 μg biotin-FITC/mg microspheres. For some controls in Chapter 3, 0.27- μm -diameter beads from Indicia Biotechnology, at 9.24×10^{11} per mL, were used instead.

D.1.2 Detailed protocol

A 1.55 mm, plated-diamond flat-tip drill bit (CRLaurence) was used to drill either two or four holes in a glass microscope slide. Slides and 24x60 mm No. 1.5 coverslips were plasma cleaned on high for 2 minutes, after which 0.02 in. ID/0.06 in. OD tygon microbore tubing (Cole-Parmer) was threaded through the holes in the slides and epoxied to the slides. One or two rounded-edge rectangles were cut out of 0.12-mm-thick double-sided tape (Grace Bio-Labs) and secured to the slide. Two-chamber slides formed this way had the chambers parallel to each other along the long axis. The chambers

were sealed with a coverslip and heated for about 30 seconds at 130°C to firmly adhere the tape to the glass. Flow chambers were prepared not more than 1 day in advance for optimal tether density.

To construct tethers, 50 μL of 20 $\mu\text{g}/\text{mL}$ polyclonal anti-digoxigenin from sheep (Roche) in MgCl_2 - and CaCl_2 -free Dulbecco's PBS (Sigma) were flowed into a chamber and incubated at room temperature for 25 minutes. The chamber was then washed with 750 μL of PTC buffer (20 mM Tris-acetate, pH 8.0, 130 mM KCl, 4 mM MgCl_2 , 0.1 mM DTT, 0.1 mM EDTA, 20 $\mu\text{g}/\text{mL}$ acetylated BSA [Sigma], and 80 $\mu\text{g}/\text{mL}$ heparin sodium salt [Sigma]), and then with 750 μL of PTC supplemented with 3 mg/mL biotin-free casein (RDI-Fitzgerald). DTT was added fresh each day to all buffers used that day, from a 0.1 μM stock in Tris-EDTA, pH 7.4 made that day. 250 μL of approximately 1 pM DNA in PTC with 3 mg/mL casein were then flowed into the chamber and incubated for 1 hour. DNA concentration was optimized empirically for each construct to maximize tether density while not creating substantial amounts of multiple tethers.

The beads to be added to the slides were first washed to exchange the storage buffer and to remove any free streptavidin. Unless otherwise indicated, the beads used in this work were 0.49- μm -diameter, streptavidin-coated polystyrene beads (Bangs) at 1.5×10^{11} beads per mL and with a binding capacity of 1.14 or 1.8 μg biotin-FITC/mg microspheres. Where indicated, 0.27- μm -diameter beads from Indicia Biotechnology, at 9.24×10^{11} per mL, were used. To wash the beads, 6 μL of the 0.49- μm -diameter beads or 12–24 μL of the 0.27- μm -diameter beads were first diluted to 30 μL in PTC with 3 mg/mL casein. Some lots of the 0.49 μm beads had improved performance when 0.2 % (v/v) Tween-20 (Sigma) was added to the wash buffer. The beads were centrifuged for 3 (0.49 μm) or 5 (0.27 μm) minutes to pellet the beads, resuspended in PTC with 3 mg/mL casein (and Tween-20 when needed), and centrifuged again for a total of 3 spins. The final resuspension was in 50 μL PTC with 3 mg/mL casein (and Tween-20 when needed) for the 0.49 μm beads, and 30 μL for the 0.27 μm beads. After the DNA incubation, excess DNA was removed from the chamber by washing with 750 μL PTC with 3 mg/mL casein, and then all 50 or 30 μL of beads were introduced into the chamber and incubated for 20 minutes. Excess beads were removed by washing with 500 μL PTC with 3 mg/mL casein.

Immediately prior to the start of the data acquisition, chambers were washed with 500 μL of LRB buffer (10 mM Tris-HCl, pH 7.4, 200 mM KCl, 0.1 mM EDTA, 0.2 mM DTT, 5% [v/v] DMSO, and 1 mg/mL casein). All dilutions of the protein stock were into LRB. After an initial 500 seconds of data were obtained in the absence of protein, as described below, LacI at the desired concentration was then flowed into the chamber, and beads were tracked for about 1.5 hours. All data were obtained at 22–24 $^{\circ}\text{C}$.

D.2 TPM data acquisition and analysis

D.2.1 Acquiring data

Data acquisition essentially followed that of [115], with the following modifications: (1) Tethers were imaged using brightfield microscopy, instead of differential interference contrast (the results are equivalent), on inverted Olympus IX71 microscopes with either a 100x oil objective (as in [115]), or a 60x oil objective with a 1.6x magnifier (again the results are equivalent). (2) A Basler A602f camera was used to acquire images at a native frame rate of 60 frames per second; however for consistency with previous results [115], every other frame was dropped for a final frame rate of 30 fps but an exposure time of 10 ms per frame. (3) Improvements to the speed of the acquisition code that allowed up to 45 beads to be tracked at once, which corresponds to the maximal tether density obtainable in the field of view of the camera without a significant number of multiply tethered particles. (4) In addition to the symmetry-of-motion and length-of-motion checks that were used as initial screens for acceptable tethers in [115], data were first acquired for 500 seconds in the LRB buffer but in the absence of protein, in order to characterize each tether in the unlooped state. Not only does this allow a more rigorous screening of tethers for anomalous behavior (e.g., unphysically short or long lengths, non-uniformity of tether length over time) but it also records the unlooped length of each individual bead, which allows easier identification of looped states, especially in DNAs with short loops that have high looping probabilities. This must be done on a tether-by-tether basis due to the significant variability of tether lengths that we see, and allows

us to observe small differences in tether length in the presence versus absence of looping, which we attribute to operator bending (see Section 4.A.1).

D.2.2 Particle tracking and calculation of the root-mean-squared motion of the bead

As in [115], beads were tracked with custom Matlab code by cross-correlating each frame with the initial frame in a time series on a bead-by-bead basis. This results in “raw” x and y positions of each bead, relative to the tethering point, as a function of time. Drift was removed from these raw data as in [115], by subtracting the results of a low-pass first-order Butterworth filter with a cutoff frequency $f_{cB} = 0.05$ Hz for the $0.49 \mu\text{m}$ beads or $f_{cB} = 0.07$ Hz for the $0.27 \mu\text{m}$ beads. The root-mean-square motion was obtained by applying a Gaussian filter with a -3 dB frequency of $f_{cG} = 0.0326$ Hz for the $0.49 \mu\text{m}$ beads or $f_{cG} = 0.461$ Hz for the $0.27 \mu\text{m}$ beads, corresponding to a 4 second or 2.8 second standard deviation of the filter, to the mean-squared displacement of the data (that is, to the quantity $(\bar{x}^2 + \bar{y}^2)$). The root-mean-squared (RMS) motion of the bead is then the square root of the result of the convolution of $(\bar{x}^2 + \bar{y}^2)$ and the Gaussian filter.

More precisely, filters were applied in Fourier space (so that convolutions become simple multiplication). This means for both the drift-correction and the Gaussian filter smoothing, we first Fourier transformed the data (which in the case of the Butterworth, applied separately to the x and y position coordinates, means we Fourier transformed the raw \bar{x} and \bar{y} position data; whereas the Gaussian filter is applied to the transform of $(\bar{x}^2 + \bar{y}^2)$), multiplied by the appropriate filter, and then inverse Fourier transformed the result. In frequency space the Butterworth filter takes the form

$$B(f) = \frac{1}{(1 + (f/cf_B)^{2n})}, \quad (\text{D.1})$$

where f is frequency, $n = 1$ so that this is a first-order filter (which determines the sharpness of the transition at the cutoff frequency), and cf_B is a rescaled cutoff frequency of the filter based on how we define our frequency axis. We choose to establish our frequency axis from $-\text{num. frames}/2$ to

+num. frames/2, where “num. frames” is the number of image frames in a trajectory. If we want a cutoff frequency for the filter at $f_{cB} = 0.05$ Hz, then we must define

$$cf_B = \frac{\text{num. frames}}{fps \times f_c^{-1}}, \quad (\text{D.2})$$

where fps is the frame rate of the camera (30 Hz in our case). This is essentially a unit conversion, since our frequency axis is unitless but f_c is in Hz. Note that now the units work out, because the Matlab command that Fourier transforms the data (the `fft` function) returns a vector the same length as the input vector, in frequencies from 0 to $fps/2$.

In the case of the Gaussian filter, the filter has the form in Fourier space of

$$G(f) = e^{-0.3466(f/cf_G)^2}. \quad (\text{D.3})$$

The factor of 0.3466 in the exponent is chosen to give 3 dB of attenuation at the cutoff frequency [158]. That is, when $f = cf_G$, the attenuation is half (3 dB corresponds to a change in power ratio of a factor of 2). For this to be the case, we must have a pre factor in the exponent of $\ln 2/2 = 0.3466$. As with the Butterworth filter, here we also have the problem of needing a unit conversion between f_{cG} , which is in Hz, and cf_G . This conversion is the same as in Eq. (D.2) because we define the f axis for the Gaussian filter in the same manner as for the Butterworth.

Finally we note that the Fourier transform of a Gaussian is a Gaussian, so in time space the Gaussian filter defined in Eq. (D.3) becomes

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma_g} e^{-\frac{t^2}{2\sigma_g^2}}, \quad (\text{D.4})$$

where σ_g defines the width of the filter and is related to f_{cG} by [158, 120]

$$\sigma_g = \frac{\sqrt{\ln 2}}{2\pi f_{cG}}. \quad (\text{D.5})$$

Therefore the 0.0326 Hz Gaussian filter we apply to most of our data corresponds to a Gaussian-shaped smoothing profile with a 4 second standard deviation in time space, and the 0.461 Hz filter applied to the smaller beads results in a 2.8 second smoothing window. The choice of σ_g is directly related to the temporal resolution of our analysis, as discussed in Section 5.A.2.

D.2.3 Determining the looping probability for each trajectory

For the constructs of Chapter 3 and Fig. 4.1(D–E) in Chapter 4 (the constructs with 94 bp loops, which have primarily only the “middle” looped state; or in Chapter 3, the PUC306 construct), data for each tether were histogrammed separately and fit to one (all looped or unlooped), two (unlooped and one looped state), or three (two looped states and an unlooped state) Gaussians. The looping probability was determined as the area under Gaussian(s) corresponding to the looped state(s) divided by the sum of the areas under all the Gaussians. This was done on a tether-by-tether basis and the mean looping probabilities and standard errors on these means for a population of tethers were reported.

However, for the predominantly three-state DNAs in Fig. 4.1(F), Fig. 4.2, Fig. 4.4, and Chapter 6 (excluding the three-operator constructs), the two looped states were often not well described by Gaussians (see Figs. E.1 and E.2 for examples). We therefore investigated a thresholding approach to calculating each bead’s looping probability [158, 113, 109, 104, 111]. Thresholding was performed subsequent to the Gaussian-fitting method described above. The intersection points of the fitted Gaussians were used to identify initial threshold values, which were adjusted manually as needed, such that the thresholds split the trajectories into the unlooped state and any looped state(s). A threshold above which data were excluded was set to the mean RMS of the tether in the absence of repressor plus three times the standard deviation of the no-repressor RMS; data above this point were usually due to tracking errors or free beads in solution temporarily entering the field of view. An empirically determined lower bound was set at 80 nm to exclude sticking events. The looping probability was then determined as the number of data points between the thresholds that delineated looped state(s), divided by the total number of points in the trajectory below the topmost threshold

and above the sticking-state threshold. For traces with well-separated and well-populated states, the looping probabilities calculated by this thresholding method were comparable to those calculated by the Gaussian fitting method; where they differed, we believe the thresholding method to better represent the behavior of the trajectory. Therefore all looping probabilities for the constructs in Fig. 4.1(F), Fig. 4.2, Fig. 4.4, and Chapter 6 were obtained by this thresholding method. As with the Gaussian fitting approach, where this thresholding method was used, it was done on a tether-by-tether basis and the mean looping probabilities and standard errors on these means for a population of tethers were reported.

D.2.4 Minimum number of trajectories and minimum observation time

In order to make our measurements of dissociation constants and J-factors as precise as possible, we considered how many trajectories needed to be included in each population mean looping probability, and how long each trajectory needed to be, in order to obtain reproducible mean looping probabilities. We will briefly summarize our methods and conclusions here.

The minimum number of trajectories needed to measure the mean looping probability of a population of tethers under a given set of conditions depends both on intrinsic tether-to-tether variation in looping probability, and on the minimum observation time discussed below. With regards to the latter: consider, for example, a case where it takes 2000 seconds to obtain an accurate measure of a tether's looping probability. Only 20 trajectories may be needed to sample the intrinsic spread in looping probabilities in a population of tethers; but if trajectories that only last 1000 seconds are also included, those trajectories will increase the spread of the data and more than 20 tethers will be needed to accurately measure the mean.

Our approach to choosing a minimum number of trajectories is depicted in Figure D.1. After histogramming the set of looping probabilities for a population of tethers under certain conditions of repressor concentration, etc. (Fig. D.1(A)), we chose, with replacement, progressively larger subsets of these looping probabilities and recalculated the mean looping probability and the standard error of this mean, repeating this procedure 10^4 times per subset size (Fig. D.1(B,C)). If very few tethers

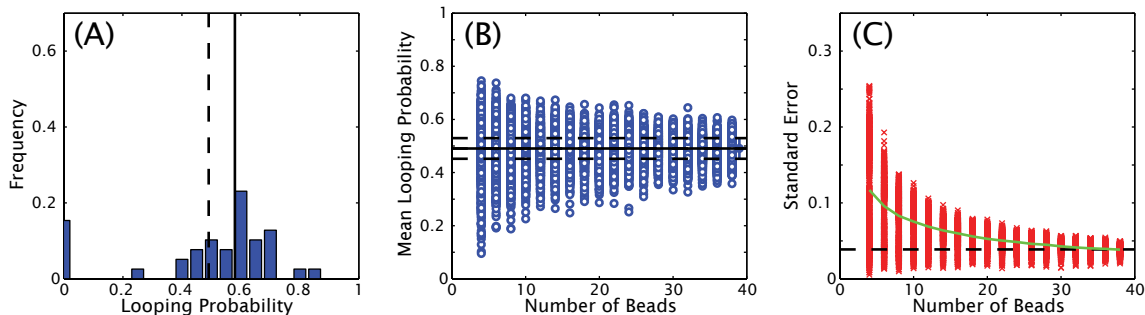


Figure D.1: Distribution of looping probabilities obtained with Oid-E894-O1 at 10 pM repressor, as representative of the issues discussed in Sections D.2.4 and D.2.5. (A) Histogram of looping probabilities of all trajectories that lasted at least 3000 seconds, normalized by the total number of tethers included in the distribution. Black vertical dashed line is the mean of the distribution; solid black vertical line is the mean not including tethers that have zero looping probability (the “nonloopers” discussed in Section D.2.5). (B) Mean looping probability as a function of the number of tethers included in the mean, resampled with replacement 10^4 times per number included (that is, there are 10^4 blue points per x-value). Solid horizontal black line is the mean of the distribution that includes all 39 tethers in the distribution in (A) (so it is the same as the vertical dashed line in (A)); dashed lines here indicate the mean plus or minus the standard error of the distribution with all tethers. (C) Standard errors of the resampled distributions whose means are plotted in (B); that is, there is a red “x” for every blue circle. Horizontal dashed line indicates the standard error with all tethers included; green curve is the standard deviation of the 10^4 blue points at each x-value tested (so it is a measure of the vertical spread of the blue points in (B)).

are used to calculate the mean, the mean fluctuates wildly between resamplings (indicated by the vertical spread in the blue points); but as the number of tethers included in the mean surpasses 20, the spread in the blue points remains constant, suggesting we need about 25 beads to accurately calculate the true mean. Similarly, as the number of trajectories included exceeds 20–25, the standard error of the distribution decreases until it reaches a constant value, which is the error associated with intrinsic tether-to-tether variation in looping probability and will not decrease with more data. We found this number of 20–25 tethers to be consistent across repressor concentrations and DNA constructs, and so all reported means and standard errors in this work are obtained from sample sizes of at least 20 tethers.

As mentioned above, we asked not only how many trajectories were needed, but also how long each trajectory needed to be in order to be included in the analysis. As discussed in the next section, we considered several schemes for calculating the mean looping probability for a population of tethers. Most of these methods, and the analysis in [115], involved weighting each tether’s looping probability equally in the calculation of the population mean (the other strategy, discussed below, is to weight each tether’s looping probability by the observation time). For the schemes in which each tether was weighted equally, it is important to include in the population mean only those that

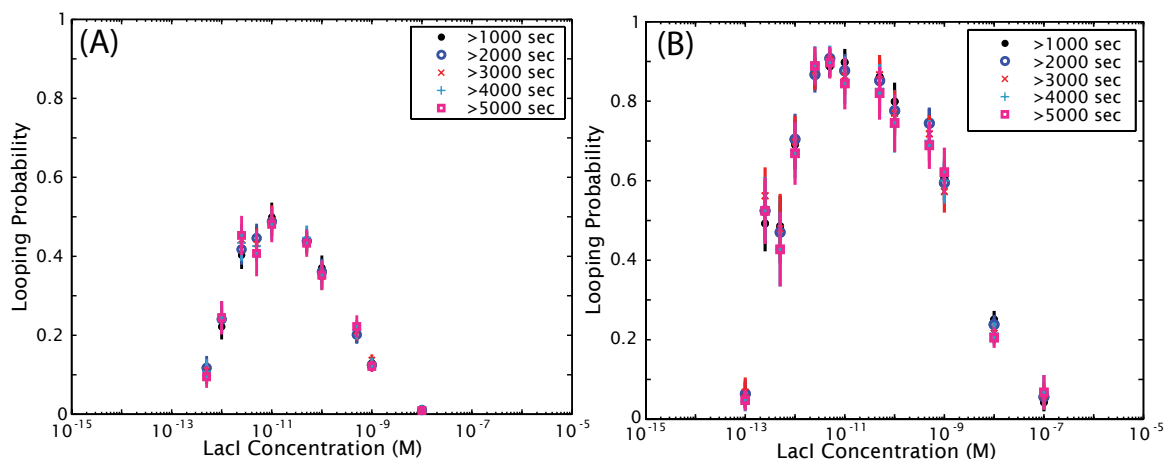


Figure D.2: Looping probability and observation time, method one. Means and standard errors of looping probability distributions as a function of minimum observation time for (A) Oid-E894-O1 and (B) Oid-TA94-O1. The legends indicate the minimum observation time for a bead to be included in the corresponding data point; for example, “>1000 seconds” means all trajectories at least 1000 seconds long. For the majority of these points it does not matter how strict the minimum observation time is—all of the means are within error of each other at all concentrations. Note that these data have not had “nonloopers” subtracted (Section D.2.5).

have been observed long enough to obtain an accurate measure of their looping probability.

We took two approaches to analyzing the minimum observation time. The first is analogous to the approach used to determine the minimum number of trajectories: the mean and standard error for sets of looping probabilities with successively stricter minimum observation requirements were calculated, to determine at what cutoff the mean and standard error of the distributions converges to a constant (Fig. D.2). We found these calculated means and standard errors to be surprisingly insensitive to minimum observation time, except at very low repressor concentrations where looping events are rare. We therefore considered a second approach, in which we considered each tether’s trajectory individually, and asked how its individual mean looping probability varied as data were removed from the end of the trajectory (Fig. D.3). Here again we found the looping probabilities for individual tethers to be surprisingly insensitive to observation time; where they did vary, an ideal minimal observation time (above which a tether’s looping probability generally remained constant) was about 3000 seconds. Therefore all tethers included in the analyses in this work lasted at least 3000 seconds.

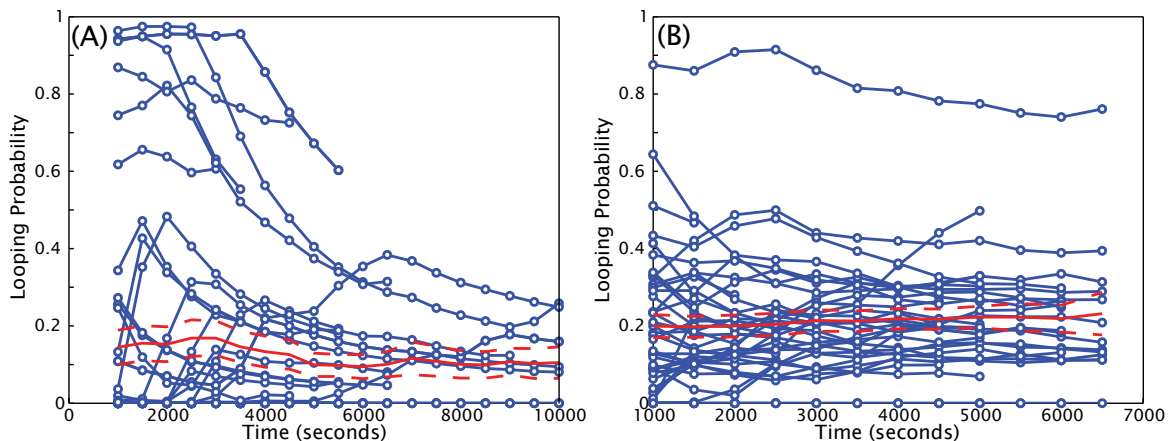


Figure D.3: Looping probability and observation time, method two. The looping probabilities of individual beads are shown as a function of how much of the trace is included in calculating the looping probability, for Oid-E894-O1 at (A) 500 fM LacI and (B) 500 pM LacI. Each blue line corresponds to a different bead. Each bead’s looping probability was calculated by Gaussian fitting (Section D.2.3) including successively more 500 second intervals of its total trajectory, starting at 1000 seconds, then 1500, etc., until the bead breaks (blue points). Solid red line is the mean of all the beads’ looping probabilities at each time interval; dashed lines are the mean plus or minus the standard error. Note that in neither (A) nor (B) does the mean include more than 20 beads past about 5500 seconds, and is therefore unreliable past this point as an accurate estimate of the true looping probability (see previous section). Note also that to eliminate user oversight, the Gaussian fit parameters for each bead’s total trajectory were used as the starting point to this analysis, and at each other time point, only the amplitude of the Gaussian fit to each (looped or unlooped) state, not the mean value or the width, was refit; therefore in a small number of cases the re-fitting was suboptimal and the looping probability should be considered an estimate in all cases. There is no clear minimum observation time after which each bead’s looping probability remains constant, though it is clear that 1000–2000 seconds is too short a minimum, and it is also clear that lower repressor concentrations (e.g., (A)) require longer observation times before the majority of beads’ looping probabilities cease to change dramatically. Note that the red line, the mean looping probability over all beads, remains roughly constant, even at 1000–2000 seconds.

D.2.5 Calculating the average looping probability for a set of trajectories

The histogram of looping probabilities in Fig. D.1(A) illustrates an aspect of our data that is not captured by the simple mean and standard error of a distribution, namely, the clustering of the looping probabilities of most trajectories around a single peak but a substantial fraction of trajectories that have zero looping probability. We see this bi-modal behavior, with a fraction of trajectories with no looping activity, at all concentrations where the rest of the distribution is sufficiently nonzero, for all DNAs derived from the low-copy pZS25 plasmid used in this work (that is, E8 and TA of all lengths and with various operators). However we do not see this nonlooping population with DNAs derived from the high-copy pUC19 plasmid used in [115] and discussed in Section 3.1.2. Moreover we observe the same fraction of these “nonloopers” with the different repressors batches discussed in Section 3.1.1 in Chapter 3. We therefore suspect that these nonloopers

are caused by multiple DNAs tethering one bead, defects in a DNA tether, or some other DNA-specific factor.

As a result, we investigated several methods of describing the average behavior of the distribution of looping probabilities that involved different ways of handling the nonlooper population. If we assume the nonloopers to be DNA-specific but concentration independent, then for the concentration titrations of Fig. 4.1(D–F), we can calculate the average number of nonloopers at concentrations where most trajectories have significantly nonzero looping probabilities, and subtract this average fraction from the distributions at all concentrations. The problem with this method is the relatively small number of concentrations at which the nonloopers are sufficiently well separated from the rest of the distribution from which to calculate the average. We found a more robust method to be to remove *all* trajectories with zero looping probability for those samples where the nonloopers were well separated from the rest of the distribution; and then to use the average number of these nonloopers (about 10% for most constructs) for distributions with significant weight near zero. We find that subtracting nonloopers using this second approach results in a mean parameter that best represents the distribution of looping probabilities; and, moreover, it results in fitted operator dissociation constants that agree well with values found in the literature—while including nonloopers does not (see Table D.1). Therefore this second method was used for all data reported in this work. (We note here two cases in which the approach had to be modified: first, in the case of O2-E894-O1, no concentration resulted in a mean looping probability significantly above zero to completely separate the nonloopers. We therefore calculated the average fraction of nonloopers between 50 pM and 500 pM, where the zero bin was clearly in the tail of the distribution, and subtracted this average from all concentrations. Second, for the length series shown in Fig. 4.2 in Chapter 4, where we do not have multiple concentrations for each construct, we applied the same approach but as a function of loop length: all nonloopers were subtracted from those lengths with clearly separated nonloopers, and an average number derived from these clear cases were subtracted from the rest.)

We note here an alternate approach to calculating the population mean (applicable when nonloopers have been kept or excluded): weighting each tether’s looping probability by the amount of

time it was observed, so that longer trajectories contribute more to the calculation of the mean, and to calculate the errors on each mean by bootstrapping the distribution (similar to the process described in Section D.2.7 below). If all tethers are drawn from the same population, that is, if all tethers behave the same way, then observing many tethers for shorter amounts of time should be equivalent to observing only a few tethers for longer times. This method could be particularly relevant at low repressor concentrations, which may require very long observation times to measure the equilibrium looping probability due to the limiting amount of repressor. However, we found that weighting by observation time, versus weighting all tethers equally, has no statistically significant effect on the calculation of the mean.

D.2.6 Fitting concentration curves

One of the assumptions behind Eq. (2.1) is that binding and looping are independent, i.e., that the binding constants do not depend on the DNA outside the operator sites, and that the looping J-factor does not depend on the operator sites. Hence, we should be able to model all our data for the TA and E8 sequences with three binding constants, K_1 , K_2 , and K_{id} , and two J-factors, $J_{loop,TA}$ and $J_{loop,E8}$.

To see that this is indeed possible, we fit Eq. (2.1) to looping data in two different ways. “Individual fits” involved independent parameters for each data set (we only used one binding constant if both operator sequences were the same), while “global fits” involved fitting several data sets simultaneously with the five parameters mentioned in the previous paragraph. That is, an individual fit to Oid-E894-O1 had three free parameters, K_{id} , K_1 , and $J_{loop, E8}$; but a global fit to the three E8-containing data sets has only four, because it requires that all three data sets be fit with the same J-factors and the same K_1 , reflecting the reality of the DNA constructs.

Fitting was performed in Matlab using custom routines with a weighted nonlinear least-squares

method,¹ which means that we minimized

$$\chi^2 = \min_{\theta} \sum_{k=1}^n \left(\frac{\bar{p}_{\text{loop},k} - p_k(\theta)}{\sigma_k} \right)^2 \quad (\text{D.6})$$

with respect to the free parameters θ (which would include dissociation constants and J-factors). In Eq. (D.6), n is the number of data points to fit (i.e., concentrations and, in the case of the global fits, DNAs as well), $\bar{p}_{\text{loop},k}$ is the mean looping probability for one DNA at one concentration, σ_k is the standard error of $\bar{p}_{\text{loop},k}$, and $p_k(\theta)$ is the theoretical prediction (using Eq. (2.1)) for data point k .

Fig. D.4 shows the results of individual versus global fits for all of the constructs in Fig. 4.1(D–E) in Chapter 4. Fit parameters, plus or minus the standard errors described in the next paragraph, are given in Table D.1. The difference in looping probability titration curves from the individual and global fits are within the experimental uncertainty almost everywhere, which indicates that the different data sets are indeed consistent with a single set of parameters.

We estimated the standard errors of the fit parameters using a bootstrap method [190], in which we constructed 10^4 resampled data sets (with replacement) at each concentration, computed the mean and standard error of the looping probability for each of these resampled sets, and redid both the individual and the global fits for each set. We then estimated the standard error of the fit parameters by the standard deviation of the bootstrap parameters.

As noted in Chapter 4, we privilege the global fits over the individual fits, for several reasons: they better reflect the physical reality of the DNA constructs, they better constrain the parameter values in many cases, and they match more closely with values in the literature obtained through traditional bulk biochemical assays (see Table D.1). For similar reasons, we consider the subtraction of some fraction of nonloopers to be justified; as noted in the previous section, we suspect these nonloopers to be experimental artifacts derived from the DNAs.

The Oid-E107-O1 concentration curve shown in Fig. 4.1(F) in Chapter 4 was fit separately from the other concentration curves, assuming the values of K_{id} and K_1 obtained from the global fit to

¹Thanks to Martin Lindén for help developing these fitting routines and the bootstrapping error method.

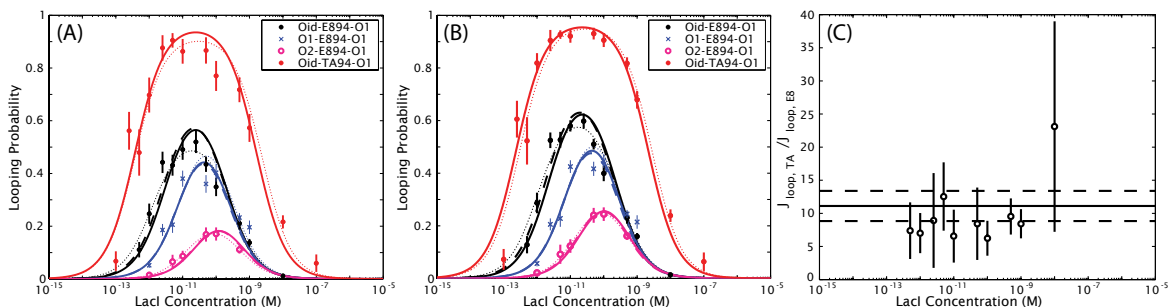


Figure D.4: Comparison of individual and global fits to E894 and TA94 data, and of the effect of excluding some fraction of “nonloopers” (see Section D.2.5). **(A)** Looping probabilities and individual and global fits for all data (that is, nonloopers included) for the constructs shown in Fig. 4.1(D–E) in Chapter 4. **(B)** Same as (A), but with some nonloopers excluded, according to the scheme (see Section D.2.5 for details) where all nonloopers were discarded for concentrations where they are clearly separated from the rest of the distribution, and a constant fraction of nonloopers were discarded for the other concentrations. The data in this panel are the same as in Fig. 4.1(D) and (E) in Chapter 4. In (A) and (B) here, the dotted lines represent individual fits to the data set of the corresponding color; the dashed lines correspond to a global fit with the three E8 data sets only; and the solid lines to a global fit to all 4 data sets simultaneously. Fit parameters are given in Table D.1. For all data sets aside from Oid-E894-O1, the global fits with and without the TA data are essentially indistinguishable from the individual fits. **(C)** Ratio of J-factors as a function of concentration including nonloopers, shown for completeness with Fig. D.5(A) below. See the caption of that figure for details.

the three E894 data sets and the TA94 data set (minus nonloopers). The looping probabilities for the two looped states were fit simultaneously, like the global fits above, but to Eqs. (2.9) and (2.10) in Section 2.2, assuming the two looped states have the same K_d 's and differ only in J-factor. The errors were computed according to a similar bootstrapping method as that used above: data at each concentration were resampled with replacement 10^4 times, and then the 10^4 K_d 's obtained from the global fits to the other concentration curves were used to refit the 10^4 new Oid-E107-O1 data sets. As above, we estimated the standard error of the fitted J-factors for the two states by the standard deviation of the bootstrap parameters.

D.2.7 Calculating J-factors without concentration curves for each construct

This section describes the use of the theoretical results presented in Section 2.7 above to calculate absolute J-factors for the DNAs in the length series presented in Fig. 4.2(C) in Chapter 4. First, however, the validity of Eq. (2.29) was examined by using it to calculate a ratio of J-factors as a function of concentration for the Oid-E894-O1 and Oid-TA94-O1 data in Fig. 4.1(E) in Chapter 4. If

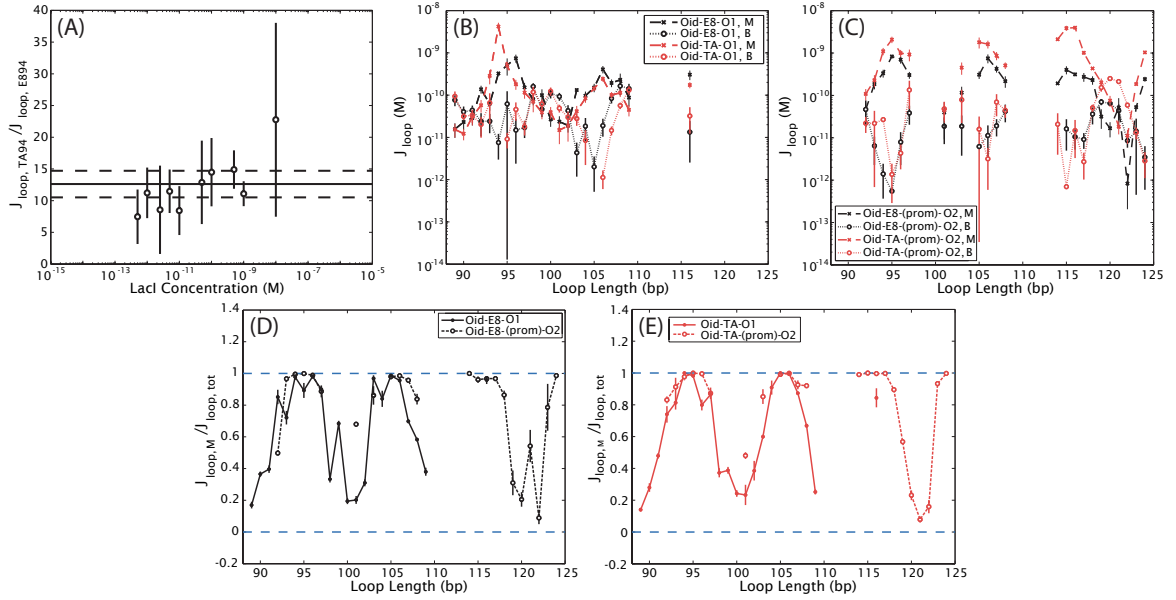


Figure D.5: Relative J-factors and the J-factors for the two looped states, with and without promoter. **(A)** Ratio of J-factors for the Oid-TA94-O1 and Oid-E894-O1 data shown in Fig. 4.1(E) in Chapter 4, as a function of concentration, calculated from Eq. (2.29); the errors are bootstrapped as described in the text here. The solid horizontal line corresponds to the ratio of J-factors obtained from the global fit to the concentration titrations in Fig 4.1(E) (that is, the ratio of the $J_{\text{loop,TA}}$ to the $J_{\text{loop,ES}}$ given in row six of Table 4.1 in Chapter 4); the dashed lines are this ratio plus or minus a bootstrapped error, described in the text here. Note that it is necessary if Eq. (2.29) is to hold that the ratio of J 's is constant across all concentrations: that is, it does not matter at what concentration we measure the looping probabilities for the two sequences, we will still obtain the same ratio of J-factors. It is clear, however, that some concentrations are better choices than others—for example, where both constructs show looping probabilities not too close to 0 or 1. **(B)** J-factors for the two looped states separately for the no-promoter constructs shown in Fig. 4.2(A–C) in Chapter 4. Because of the results in (A) here, we can use Eq. (2.29) to compute absolute J-factors from J-factor ratios, when one of the two components of the ratio is known. The result of using Eq. (2.29) to calculate J-factors for the E8 and TA constructs of varying lengths (without a concentration curve for each construct) is shown in Fig. 4.2(C) in Chapter 4; here we show J-factors for the two looped states of these constructs separately. The computation of these J-factors and associated errors are described in the text here. The E8 J-factors shown here are the same as are plotted in Fig. 4.3 in Chapter 4 in comparison with theoretical results for the looping J-factor. **(C)** J-factors for the two looped states separately for the with-promoter constructs shown in Fig. 4.2(D–F) in Chapter 4. Here we cannot use Eq. (2.29) because these constructs have a different combination of operators than any of the constructs used in the concentration titrations of Fig. 4.1. However, we can use the fitted parameters for K_2 and K_{id} obtained from those data to calculate total J-factors for the with-promoter constructs, as shown in Fig. 4.2(F) in Chapter 4, or J-factors for the two states separately (described in the text here). **(D)** The fraction of the total J-factor that is contributed by the J-factor of the middle state for E8, with and without the promoter. If only the middle state is occupied at a certain loop length, this fraction is 1; if only the bottom state is occupied, this fraction is 0. Where this ratio is not already close to 1, the addition of the promoter shifts it closer to 1 at most lengths, meaning that the promoter increases the probability that the middle looped state, rather than the bottom looped state, will form. **(E)** Same as (D) but for the TA constructs.

Data	K_{id}	K_1	K_2	$J_{loop, E8}$	$J_{loop, TA}$
Oid-E894-O1	3 (\pm 1)	130 (\pm 30)	-	320 (\pm 50)	-
O1-E894-O1	-	50 (\pm 4)	-	350 (\pm 30)	-
O2-E894-O1*	-	20 (8, 130)	300 (80, 1500)	200 (100, 600)	-
Oid-TA94-O1*	5 (2, 50)	200 (\pm 100)	-	-	4400 (\pm 700)
Global Fit, E8	12 (\pm 2)	42 (\pm 3)	310 (\pm 70)	260 (\pm 20)	-
Global Fit, E8 & TA	14 (\pm 4)	44 (\pm 4)	340 (\pm 70)	280 (\pm 20)	3000 (\pm 600)
Oid-E894-O1	3 (\pm 1)	90 (\pm 20)	-	350 (\pm 40)	-
O1-E894-O1	-	47 (\pm 4)	-	380 (\pm 30)	-
O2-E894-O1	-	26 (11, 125)	300 (\pm 200)	320 (\pm 90)	-
Oid-TA94-O1	10 (5, 46)	80 (\pm 40)	-	-	5500 (\pm 600)
Global Fit, E8	9 (\pm 1)	42 (\pm 3)	210 (\pm 40)	300 (\pm 20)	-
Global Fit, E8 & TA	12 (\pm 3)	44 (\pm 3)	240 (\pm 50)	330 (\pm 30)	4200 (\pm 600)
Literature values	8.3 ± 1.7 [125]	37 ± 5 [137, 138, 139]	350 ± 130 [137]	-	-

Table D.1: Fit parameters for individual and global fits, in pM, with and without nonloopers subtracted. “Global fit, E8” includes only Oid-E894-O1, O1-E894-O1, and O2-E894-O1; “Global fit, E8 & TA” includes these three data sets and Oid-TA94-O1 as well. The top section is with all data (Fig. D.4(A)); the bottom is with nonloopers subtracted (Fig. D.4(B)) and is the same as Table 4.1 in Chapter 4 (shown here for comparison). The last row here is the same as the last row of Table 4.1 as well. An asterisk (*) indicates that the distributions of fit parameters obtained from bootstrapped data were multimodal. In most cases, the best fit parameter plus or minus an error that is the standard deviation of the fit parameters to bootstrapped data is reported, as described in the text; however, for those cases in which the standard deviation includes negative parameter values, a 95% confidence interval is reported in parentheses instead.

Eq. (2.29), and our theoretical framework in general, are to hold, then the ratio of J-factors calculated from Eq. (2.29) should be independent of concentration, and, moreover, should be consistent with the ratio of J-factors for E894 and TA94 obtained from the global fits to the concentration titrations. This is indeed what we find. The J-factor ratios computed from Eq. (2.29) are plotted in Fig. D.5(A), along with a solid horizontal line that indicates the ratio of J-factors obtained from a global fit to the E8 and TA data. The errors on the ratios were calculated using a bootstrap method similar to that described for fitting concentration curves in Section D.2.6 above: the distribution of looping probabilities for each DNA at each concentration was resampled with replacement 10^4 times. After each resampling, a new mean looping probability was computed, and then a new ratio of J-factors was computed from these new means, for each E8-TA concentration pair. This resulted in a distribution of 10^4 new J-factor ratios. The error was taken to be the standard deviation of this distribution. A similar procedure was used to compute the horizontal dashed lines in Fig. D.5(A), which represent the error on the ratio of J-factors obtained from the global fit. As described in Section D.2.6 above, these global fits were performed on bootstrapped data (resampled with replacement 10^4 times), which yielded distributions of 10^4 J-factor values for Oid-E894-O1 and Oid-TA94-O1. From each of the 10^4 rounds of fitting, a new ratio of the fitted J-factors for TA and E8 were computed. The

error plotted as horizontal dashed lines in Fig. D.5(A) is the standard deviation of these 10^4 ratios.

We now turn to the use of Eq. (2.29) to compute J-factors for the constructs plotted in Fig. 4.2(C) in Chapter 4, for which we do not have concentration titrations. The absolute J-factors in that figure were obtained from ratios with the 94 bp E8 construct whose J-factor is known from the concentration titrations of Fig. 4.1(D–E); that is, Eq. (2.29) was used with DNA i being one of the E8 or TA constructs of variable loop length whose looping probability is given in Fig. 4.2(A), and DNA ii being the 94 bp E8 construct used in the concentration titrations of Fig. 4.1(D) and (E). However, instead of using the measured looping probability for Oid-E804-O1, we used the looping probability predicted for 100 pM based on the global fit to the three E8 data sets plus the TA data. We thereby obtained the ratio of the J-factor for the sequences in Fig. 4.2(A) to the J-factor for Oid-E894-O1 given by the global fit; and since this latter J-factor is known, we could then calculate J-factors for the other E8 and TA lengths.

To estimate the errors on these length-series J-factors, we bootstrapped the 4 sets of data used in the global fit for the Oid-E894-O1 J-factor, as well as the looping probabilities for each length-series construct in Fig. 4.2(A). This resulted in 10^4 new sets of predicted looping probabilities for Oid-E894-O1 at 100 pM, and 10^4 new looping probabilities for each loop lengths in Fig. 4.2. For each of the 10^4 new values, we computed absolute J-factors for the E8 and TA length series, as described in the previous paragraph, and then took the standard deviation of these new absolute J-factors. This standard deviation became the errors plotted in Fig. 4.2(C). We note that it is equally possible to use the looping probability for Oid-TA94-O1 as the reference instead of Oid-E894-O1; however, doing so results in much larger errors on the calculated J-factors, probably because the looping probability for the TA-containing sequence is very close to 1.

The J-factors shown in Fig. 4.2(C) in Chapter 4 are for both looped states combined. It is also possible to calculate the J-factors separately for each of the two looped states (see Section 2.2 above). To do so we note that the sum of J-factors of the two looped states is the total J-factor calculated in the previous paragraph; and that the ratio of the J-factors of the two looped states is simply the ratio of their looping probabilities (which can be seen if Eq. (2.10) is divided by Eq. (2.9)). Fig. D.5(B)

shows the result of this calculation. The errors in Fig. D.5(B) follow the usual formulation: for each DNA construct, the pair of looping probabilities corresponding to the two states was resampled with replacement 10^4 times (that is, the resampling was done in such a way that the looping probability for the bottom looped state for one bead was not severed from the looping probability for the middle state for that particular bead), and 10^4 new J-factors were calculated in the same way as each state's mean J-factor. The error was then the standard deviation of these 10^4 J-factors.

In the case of the promoter-containing constructs of Fig. 4.2(D–F) in Chapter 4, the change of operator precluded the use of the above procedure for calculating either total or separate looping probabilities. However the dissociation constants for both O_{id} and O_2 are known from the concentration curves of Fig. 4.1(D), and so the total J-factors for each of the constructs in Fig. 4.2(D) could be calculated directly from their looping probabilities by solving Eq. (2.1) for J_{loop} . Then the errors were computed using a bootstrapping procedure similar to that described above, utilizing the 10^4 bootstrapped fit values for K_2 and K_{id} from the concentration curve fits. Finally the J-factors for the two looped states separately were computed by taking advantage, as in the no-promoter case, of the fact that the ratio of the J-factors for the two states is the ratio of their looping probabilities; and the errors were again bootstrapped. Fig. D.5(C) shows these with-promoter J-factors.

Appendix E

Representative traces

Figures E.1 and E.2 give representative examples of RMS motion versus time for the constructs used in Chapters 2–5 of this work. (See Fig. 6.4 in Chapter 6 for examples of the three-operator constructs.) Fig. E.1, which shows examples from different repressor concentrations, illustrates several points mentioned elsewhere in this work: the presence of two experimentally distinguishable looped states at some loop lengths and concentrations (see also Fig. E.2(A)); the difficulty of using a Gaussian-fitting method to obtain looping probabilities for some of these three-state trajectories (see Section D.2.3 and Fig. E.2(B)); and the shortening of the unlooped state observed at high repressor concentrations, discussed in Section 4.A.1. We note here also the difference in looped and unlooped state lifetimes at low and high repressor concentrations, even when a low concentration and a high concentration result in almost equal looping probabilities (compare Oid-E894-O1 at 1 pM and at 500 pM), discussed in more detail in Chapter 5: at low concentrations, long dwells in the unlooped state are interspersed with bursts of looping transitions (or long dwells in the looped state, as for the Oid-TA94-O1 data at 500 fM), whereas there are more transitions and shorter dwell times at higher concentrations. We attribute the larger error bars on low-concentration data for some constructs to finite observation time combined with few transitions between states, which leads to larger tether-to-tether variability in looping probability (see also Section 3.2.2). Finally we note that it appears that the two looped states can directly interconvert, which suggests that the two states differ in repressor conformation instead of operator binding orientation; however these trajectories have been downsampled to reduce file size (in addition to being Gaussian filtered), and so it is possible that

very short transitions to the unlooped state between looped state interconversions are not visible. (See the discussion in Section 5.1.3.)

Figure E.2 shows representative examples of DNA constructs with only the bottom state (to complement the middle-state-only and both-states examples in Fig. E.1), trajectories without clearly separated states, and distributions of looping probabilities. In Fig. E.2(B), the top two trajectories illustrate a behavior observed in a minority of Oid-E894-O1 and O1-E894-O1 tethers: instead of the clear two states seen for most tethers with these constructs (see representative examples in Fig. E.1), the looped and unlooped states have such similar RMS motions that they overlap in the histograms to the right of each trace. It is unclear what causes this behavior. The middle two trajectories in Fig. E.2(B) illustrate a similar behavior that is more prevalent in the O_2 -containing construct, which we attribute to the shorter looped-state lifetimes with this weaker operator compared to constructs with O_1 or O_{id} . The bottom two trajectories in (B) illustrate the kinds of Gaussian fits to poorly separated looped states that motivated a thresholding method, instead of a Gaussian-fitting method, for calculating the looping probabilities of the three-state DNAs in Figs. 4.1(F) and 4.2 in Chapter 4 (see Section D.2.3). We were also interested in assessing whether thresholding the length series data of Fig. 4.2 would reduce the spread in the distributions of looping probabilities obtained from populations of otherwise identical tethers, shown in Fig. E.2(C). All of the distributions of looping probabilities for the middle-state-only constructs used in the concentration titrations of Fig. 4.1(D–E) in Chapter 4, for most of the with-promoter constructs of Fig. 4.2(D–F), and for some of the bottom-state only or both-states no-promoter constructs of Figs. 4.1(F) and 4.2(A–C), showed a clustered set of looping probabilities with a clear peak, as with the Oid-E8103-O1 example here. However, many of the no-promoter constructs in Fig. 4.2(A–C) in Chapter 4 showed broad distributions of looping probabilities, as with the Oid-TA106-O1 example here, and for a minority of constructs the distribution was so broad as to include almost all probabilities from 0 to 1, as with Oid-E109-O1. In some cases, the spread was reduced when the two states were histogrammed separately, as with Oid-E898-O1. The thresholding method, as compared to the Gaussian fitting method, reduced the spread in some but not all looping probability distributions. We note that

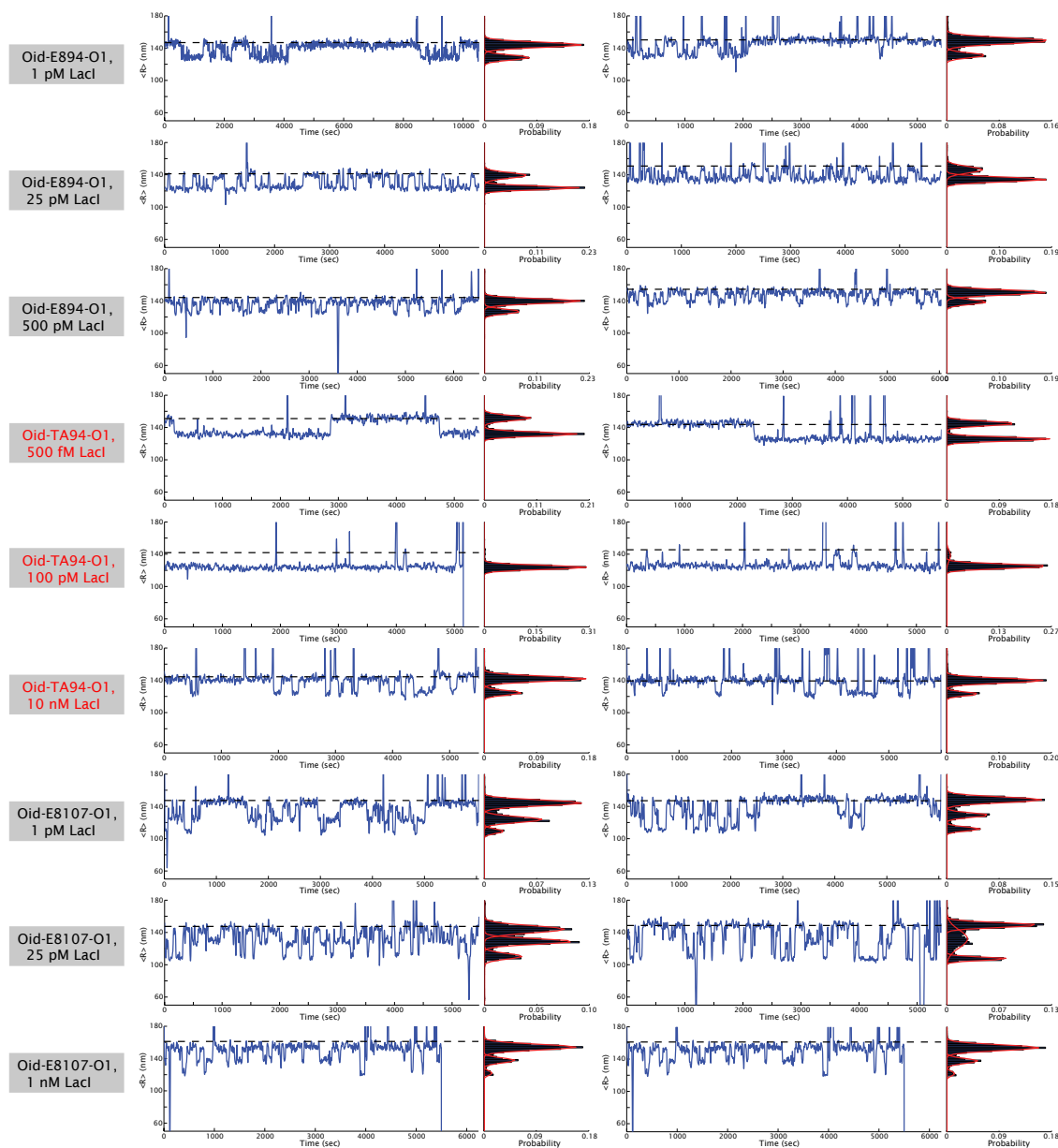


Figure E.1: Representative examples of the Gaussian-filtered root-mean-squared motion (“ $\langle R \rangle$ ”) of selected tethers as a function of time, for several of the DNA constructs shown in Fig. 4.1(D–F) in Chapter 4, at several Lac repressor concentrations. To the right of each $\langle R \rangle$ -vs.-time trajectory is a histogram showing the probability of finding a given $\langle R \rangle$ over the whole trajectory. Red lines on this histogram show the results of a Gaussian fit, one way of determining the looping probability (see Section D.2.3). The horizontal black dashed line in the plots of $\langle R \rangle$ -vs.-time indicate the average length of that particular tether in the absence of repressor. Excursions to RMS values less than about 80 nm are attributed to non-specific, transient adsorption of the bead to the surface, the DNA to the surface, or the DNA to the bead (“sticking” events); excursions to RMS values well above the horizontal black dashed lines are due to tracking errors (e.g., due to free beads in solution transiently entering the field of view). No trajectories for the O1-E894-O1 or O2-E894-O1 constructs are shown because they are essentially the same as those for Oid-E894-O1 (but see Fig. E.2 for difficulties particular to the O2-E894-O1 construct).

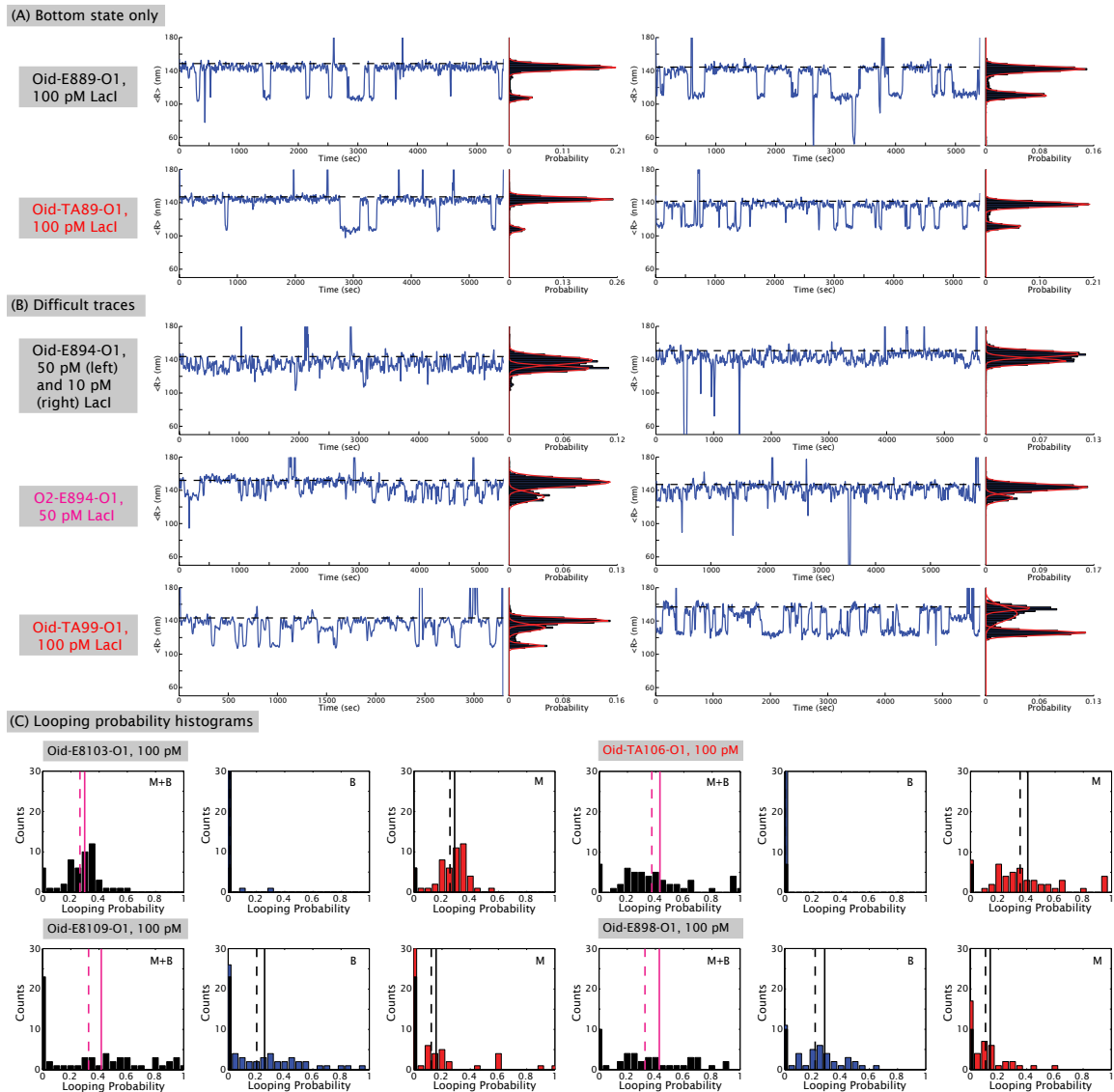


Figure E.2: Representative examples of **(A)** DNA constructs with only the bottom state (to complement the middle-state-only and both-states examples in Fig. E.1), **(B)** trajectories without the clear states shown in (A) here and in Fig. E.1, and **(C)** distributions of looping probabilities. The plots in (A) and (B) are of the same kind as in Fig. E.1 (see the caption of that figure for description). In (C), *unnormalized* looping probability histograms show the distributions of total (“M+B”) looping probabilities, or the probabilities of the bottom (“B”) or middle (“M”) looped states, for four different DNA constructs at 100 pM repressor. The dashed line shows the mean looping probability with nonloopers included (see Section D.2.5); the solid line shows the mean looping probability with nonloopers subtracted, which is the mean looping probability reported in the figures in Chapter 4. In each “B” and “M” panel, the number of tethers that never loop at all are shown as a black bar in the zero bin; a blue or red bar above the black bar in the zero bin indicates the number of tethers that show no “B” state in the “B” panel, or no “M” state in the “M” panel.

even in cases where the looping probability distributions for the no-promoter constructs were very broad, the E8 and TA histograms were still so similar that we could conclude there was no sequence dependence to looping at those lengths.

Figure 6.4 in Chapter 6 shows representative examples of trajectories with the DNA constructs derived from the wild-type three-operator *lac* operon regulatory region, and whose looping probabilities are shown in Fig. 6.3.

Bibliography

- [1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.
- [2] R. D. Fleischmann *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512, 1995.
- [3] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.
- [4] W. C. Warren *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175–256, 2008.
- [5] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer. A vision for the future of genomics research. *Nature*, 422(6934):835–847, 2003.
- [6] C. Gonzaga-Jauregui, J. R. Lupski, and R. A. Gibbs. Human genome sequencing in health and disease. *Annu Rev Med*, 63:35–61, 2012.
- [7] S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Trevino-Quintanilla, and J. Collado-Vides. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res*, 36:D120–D124, 2008.

- [8] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muniz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp. EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res*, 39:D583–D590, 2011.
- [9] T. Misteli. Beyond the sequence: cellular organization of genome function. *Cell*, 128(4):787–800, 2007.
- [10] T. Takizawa, K. J. Meaburn, and T. Misteli. The meaning of gene positioning. *Cell*, 135(1):9–13, 2008.
- [11] T. Misteli. Higher-order genome organization in human disease. *Cold Spring Harb Perspect Biol*, 2(8):a000794, 2010.
- [12] R. Kanwal and S. Gupta. Epigenetic modifications in cancer. *Clin Genet*, 81(4):303–311, 2012.
- [13] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997.
- [14] J. Widom. Role of DNA sequence in nucleosome stability and dynamics. *Quart Rev Biophys*, 34:1–56, 2001.
- [15] M. Radman-Livaja and O. J. Rando. Nucleosome positioning: How is it established, and why does it matter? *Dev Biol*, 339(2):258–266, 2010.
- [16] U. Moran, R. Phillips, and R. Milo. SnapShot: Key numbers in biology. *Cell*, 141(7):1262–1262.e1, 2010.
- [17] M. S. Luijsterburg, M. C. Noom, G. J. L. Wuite, and R. T. Dame. The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: a molecular perspective. *J Struct Biol*, 156(2):262–272, 2006.

- [18] L. Czaplá, J. P. Peters, E. M. Rueter, W. K. Olson, and L. J. Maher 3rd. Understanding apparent DNA flexibility enhancement by HU and HMGB architectural proteins. *J Mol Biol*, 409(2):278–289, 2011.
- [19] N. A. Becker, J. D. Kahn, and L. J. Maher 3rd. Bacterial repression loops require enhanced DNA flexibility. *J Mol Biol*, 349:716–730, 2005.
- [20] N. A. Becker, J. D. Kahn, and L. J. Maher 3rd. Effects of nucleoid proteins on DNA repression loop formation in *Escherichia coli*. *Nucleic Acids Res*, 35(12):3988–4000, 2007.
- [21] X. J. Chen and R. A. Butow. The organization and inheritance of the mitochondrial genome. *Nat Rev Genet*, 6(11):815–825, 2005.
- [22] M. F. White and S. D. Bell. Holding it together: chromatin in the Archea. *Trends Genetl*, 18(12):621–6, 2002.
- [23] K. Sandman and J. N. Reeve. Archaeal chromatin proteins: different structures but common function? *Curr Opin Microbiol*, 8(6):656–661, 2005.
- [24] H. Echols. Nucleoprotein structures initiating DNA replication, transcription, and site-specific recombination. *J Biol Chem*, 265(25):14697–14700, 1990.
- [25] R. Schleif. DNA looping. *Annu Rev Biochem*, 61:199–223, 1992.
- [26] K. S. Matthews. DNA looping. *Microbiol Rev*, 56(1):123–136, 1992.
- [27] H. G. Garcia, P. Grayson, L. Han, M. Inamdar, J. Kondev, P. C. Nelson, R. Phillips, J. Widom, and P. A. Wiggins. Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity. *Biopolymers*, 85(2):115–130, 2007.
- [28] N. Ptashne. Gene regulation by proteins acting nearby and at a distance. *Nature*, 322(6081):697–701, 1986.
- [29] K. Rippe, P. H. von Hippel, and J. Langowski. Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem Sci*, 20:500–506, 1995.

- [30] B. Tolhuis, R.-J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. Looping and interaction between hypersensitive sites in the active β -globin locus. *Molecular Cell*, 10:1453–1465, 2002.
- [31] P. J. Hagerman. Flexibility of DNA. *Annual Review of Biophysics and Biophysical Chemistry*, 17:265–286, 1988. PMID: 3293588.
- [32] J. Müller, S. Oehler, and B. Müller-Hill. Repression of *lac* promoter as a function of distance, phase, and quality of an auxiliary *lac* operator. *J Mol Biol*, 257:21–29, 1996.
- [33] R. Amit, H. G. Garcia, R. Phillips, and S. E. Fraser. Building enhancers from the ground up: a synthetic biology approach. *Cell*, 146:105–118, 2011.
- [34] N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci USA*, 100(9):5136–5141, 2003.
- [35] T. M. Dunn, S. Hahn, S. Ogden, and R. F. Schleif. An operator at -280 base pairs that is required for repression of *araBAD* operon promoter: Addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proc Natl Acad Sci USA*, 81(16):5017–20, 1984.
- [36] R. B. Lobell and R. F. Schleif. DNA looping and unlooping by AraC protein. *Science*, 250(4980):528–532, 1990.
- [37] A. Hochschild and M. Ptashne. Cooperative binding of λ repressors to sites separated by integral turns of the DNA helix. *Cell*, 44(5):681–687, 1986.
- [38] P. J. Choi, L. Cai, K. Frieda, and S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–446, 2008.
- [39] G.-W. Li, O. G. Berg, and J. Elf. Effects of macromolecular crowding and DNA looping on gene regulation kinetics. *Nature Phys*, 5:294–297, 2009.
- [40] H. G. Garcia, A. Sanchez, T. Kuhlman, J. Kondev, and R. Phillips. Transcription by the numbers redux: experiments and calculations that surprise. *Trends Cell Biol*, 20(12):723–733, 2010.

- [41] S. Small, A. Blair, and M. Levine. Regulation of even-skipped stripe 2 in the *Drosophila* embryo. *EMBO J*, 11(11):4047–4057, 1992.
- [42] C. V. Kirchhamer, C.-H. Yuh, and E. H. Davidson. Modular *cis*-regulatory organization of developmentally expressed genes: Two genes transcribed territorially in the sea urchin embryo, and additional examples. *Proc Natl Acad Sci USA*, 93:9322–9328, 1996.
- [43] M. Rappas, D. Bose, and X. Zhang. Bacterial enhancer-binding proteins: unlocking σ^{54} -dependent gene transcription. *Curr Opin Struct Biol*, 17:110–116, 2009.
- [44] M. Bulger and M. Groudine. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev Biol*, 339(2):250–257, 2010.
- [45] M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. Mediator and cohesion connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, 2010.
- [46] R. Yasmin, K. T. Yeung, R. H. Chung, M. E. Gaczynska, P. A. Osmulski, and N. Noy. DNA looping by RXT tetramers permits transcriptional regulation “at a distance”. *J Mol Biol*, 343(2):327–338, 2004.
- [47] J. E. Stenger, P. Tegtmeyer, G. A. Mayr, M. Reed, Y. Wang, P. Wang, P. V. Hough, and I. A. Mastrangelo. p53 oligomerization and DNA looping are linked with transcriptional activation. *EMBO J*, 13(24):6011–6020, 1994.
- [48] D. H. Lee and R. F. Schleif. *In vivo* DNA loops in *araCBAD*: size limits and helical repeat. *Proc Natl Acad Sci USA*, 86(2):476–480, 1989.
- [49] H. Krämer, M. Niemöller, M. Amouyal, B. Revet, B. von Wilcken-Bergmann, and B. Müller-Hill. *lac* repressor forms loops with linear DNA carrying two suitably spaced *lac* operators. *J Mol Biol*, 267(5):1305–1314, 1997.

- [50] G. R. Bellomy, M. C. Mossing, and M. T. Record Jr. Physical properties of DNA *in vivo* as probed by the length dependence of the *lac* operator looping process. *Biochemistry*, 27:3900–3906, 1988.
- [51] J. P. Peters and L. J. Maher 3rd. DNA curvature and flexibility *in vitro* and *in vivo*. *Quart Rev Biophys*, 43(1):22–63, 2010.
- [52] W. K. Olson and V. B. Zhurkin. Working the kinks out of nucleosomal DNA. *Curr Opin Struct Biol*, 21:348–357, 2011.
- [53] O. Kratky and G. Porod. Röntgenuntersuchung gelöster fadenmoleküle. *Recl Trav Chim Pays-Bas*, 68(12):1106–1122, 1949.
- [54] D. Shore, J. Langowski, and R. L. Baldwin. DNA flexibility studied by covalent closure of short fragments into circles. *Proc Natl Acad Sci USA*, 78(8):4833–4837, 1981.
- [55] J. Kahn and D M. Crothers. Protein-induced bending and DNA cyclization. *Proc Natl Acad Sci USA*, 89(14):6343–6347, 1992.
- [56] H. Jacobson and W. H. Stockmayer. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J Chem Phys*, 18(12):1600–1606, 1950.
- [57] J. Shimada and H. Yamakawa. Ring-closure probabilities for twisted wormlike chains. Application to DNA. *Macromolecules*, 17(4):689–698, 1984.
- [58] T. E. Cloutier and J. Widom. Spontaneous sharp bending of double-stranded DNA. *Mol Cell*, 14(3):355–362, 2004.
- [59] Q. Du, C. Smith, N. Shiffeldrim, M. Vologodskaia, and A. Vologodskii. Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proc Natl Acad Sci USA*, 102(15):5397–5402, 2005.
- [60] R. A. Forties, R. Bundschuh, and M. G. Poirier. The flexibility of locally melted DNA. *Nuc Acids Res*, 37(14):4580–4586, 2009.

- [61] M. Hogan, J. LeGrange, and B. Austin. Dependence of DNA helix flexibility on base composition. *Nature*, 304(5928):752–754, 1983.
- [62] C. M. Collis, P. L. Molley, G. W. Both, and H. R. Drew. Influence of the sequence-dependent flexure of DNA on transcription in *E. coli*. *Nucleic Acids Res*, 17(22):9447–9468, 1989.
- [63] A. Schulz, J. Langowski, and K. Rippe. The effect of the DNA conformation on the rate of NtrC activated transcription of *Escherichia coli* RNA Polymerase- σ^{54} holoenzyme. *J Mol Biol*, 300:709–725, 2004.
- [64] M. Serrano, I. Barthelemy, and M. Salas. Transcription activation at a distance by phage Φ 29 protein p4. Effect of bent and non-bent intervening sequences. *J Mol Biol*, 219(3):403–414, 1991.
- [65] A. K. Cheema, N. R. Choudhury, and H. K. Das. A- and T-tract-mediated intrinsic curvature in native DNA between the binding site of upstream activator NtrC and the *nifLA* promoter of *Klebsiella pneumoniae* facilitates transcription. *J Bacteriol*, 181(17):5296–5302, 1999.
- [66] A. E. Lilja, J. R. Jenssen, and J. D. Kahn. Geometric and dynamic requirements for DNA looping, wrapping and unwrapping in the activation of *E. coli glnAp2* transcription by NtrC. *J Mol Biol*, 342:467–478, 2004.
- [67] R. A. Mehta and J. D. Kahn. Designed hyperstable Lac repressor-DNA loop topologies suggest alternative loop geometries. *J Mol Biol*, 294:67–77, 1999.
- [68] L. M. Edelman, R. Cheong, and J. D. Kahn. Fluorescence resonance energy transfer over 130 basepairs in hyperstable Lac repressor-DNA loops. *Biophys J*, 84(2):1131–1145, 2003.
- [69] M. A. Morgan, K. Okamoto, J. D. Kahn, and D. S. English. Single-molecule spectroscopic determination of Lac repressor-DNA loop conformation. *Biophys J*, 89(4):2588–2596, 2005.
- [70] A. R. Haeusler, K. R. Goodson, T. D. Lillian, X. Wang, S. Goyal, N. C. Perkins, and J. D. Kahn. FRET studies of a landscape of Lac repressor-mediated DNA loops. *Nucleic Acids Res*, 2012.

- [71] H. R. Widlund, P. N. Kuduvalli, M. Bengtsson, H. Cao, T. D. Tullius, and M. Kubista. Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *J Biol Chem*, 274(45):31847–31852, 1999.
- [72] M. Roychoudhury, A. Sitlani, J. Lapham, and D. M. Crothers. Global structure and mechanical properties of a 10-bp nucleosome positioning motif. *Proc Natl Acad Sci USA*, 97(25):13608–13613, 2000.
- [73] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.
- [74] Y. Zhang, Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, and K. Struhl. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*. *Nat Struct Mol Biol*, 16(8):847–852, 2009.
- [75] Y. Field, N. Kaplan, Y. Fondufe-Mittendorf, I. K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol*, 4(11):e1000216, 2008.
- [76] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, 2009.
- [77] M.-L. Visnapuu and E. C. Greene. Single-molecule imaging of DNA curtains reveals intrinsic energy landscapes for nucleosome deposition. *Nat Struct Mol Biol*, 16(10):1056–1062, 2009.
- [78] P. Partensky and G. J. Narlikar. Chromatin remodelers act globally, sequence positions nucleosomes locally. *J Mol Biol*, 391:12–25, 2009.
- [79] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal. Nucleosome sequence preferences influence *in vivo* nucleosome organization. *Nat Struct Mol Biol*, 17(8):918–920, 2010.

- [80] Y. Zhang, Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, and K. Struhl. Evidence against a genomic code for nucleosome positioning. *Nat Struct Mol Biol*, 17(8):920–923, 2010.
- [81] W. K. Olson, D. Swigon, and B. D. Coleman. Implications of the dependence of the elastic properties of DNA on nucleotide sequence. *Phil Trans R Soc Lond A*, 362:1403–1422, 2004.
- [82] A. V. Morozov, K. Fortney, D. A. Gaykalov, V. M. Studitsky, J. Widom, and E. D. Siggia. Using DNA mechanics to predict *in vitro* nucleosome positions and formation energies. *Nuc Acids Res*, 37(14):4704–4722, 2009.
- [83] S. Balasubramanian, F. Xu, and W. K. Olson. DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences. *Biophys J*, 96(6):2245–2260, 2009.
- [84] S. Geggier and A. Vologodskii. Sequence dependence of DNA bending rigidity. *Proc Natl Acad Sci USA*, 107(35):15421–15426, 2010.
- [85] T. E. Cloutier and J. Widom. DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc Natl Acad Sci USA*, 102(10):3645–3650, 2005.
- [86] P. T. Lowary and J. Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol*, 276:19–42, 1998.
- [87] T. R. Blosser, J. G. Yang, M. D. Stone, G. J. Narlikar, and X. Zhuang. Dynamics of nucleosome remodeling by individual ACF complexes. *Nature*, 462(7276):1022–1027, 2009.
- [88] G. R. Bellomy and M. T. Record Jr. Stable DNA loops *in vivo* and *in vitro*: roles in gene regulation at a distance and in biophysical characterization of DNA. *Prog Nucleic Acid Res Mol Biol*, 39:81–128, 1990.
- [89] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, 1961.

- [90] B. Müller-Hill. *The lac operon: A short history of a genetic paradigm*. De Gruyter, 1996.
- [91] C. J. Wilson, H. Zhan, L. Swint-Kruse, and K. S. Matthews. The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding. *Cell Mol Life Sci*, 64(1):3–16, 2007.
- [92] S. L. Roderick. The *lac* operon galactoside acetyltransferase. *Comptes Rendus Biologies*, 328(6):568–575, 2005.
- [93] S. Oehler, E. R. Eismann, H. Krämer, and B. Müller-Hill. The three operators of the *lac* operon cooperate in repression. *EMBO J*, 9(4):973–979, 1990.
- [94] S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill. Quality and position of the three *lac* operators of *E. coli* define efficiency of repression. *EMBO J*, 13(14):3348–3355, 1994.
- [95] M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, 271(5253):1247–1254, 1996.
- [96] K. C. Cone, M.A. Sellitti, and D. A. Steege. Lac repressor mRNA transcription terminates *in vivo* in the *lac* control region. *J Biol Chem*, 258(18):11296–11304, 1983.
- [97] W. S. Reznikoff, R. B. Winter, and C. K. Hurley. The location of the repressor binding sites in the *lac* operon. *Proc Natl Acad Sci USA*, 71(6):2314–2318, 1974.
- [98] W. Gilbert. In H. Sund and G. Blauer, editors, *Protein-ligand interactions*. de Gruyter, 1975.
- [99] M. Pfahl, V. Guide, and S. Bourgeois. “Second” and “third operator” of the *lac* operon: an investigation of their role in the regulatory mechanism. *J Mol Biol*, 127(3):339–344, 1979.
- [100] M. C. Mossing and M. T. Record Jr. Upstream operators enhance repression of the *lac* promoter. *Science*, 233(4766):889–892, 1986.

- [101] L. M. Bond, J. P. Peters, N. A. Becker, J. D. Kahn, and L. J. Maher 3rd. Gene repression by minimal *lac* loops *in vivo*. *Nucleic Acids Res*, 38(22):8072–8082, 2010.
- [102] D. A. Schafer, J. Gelles, M. P. Sheetz, and R. Landick. Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature*, 352:444–448, 1991.
- [103] H. Yin, R. Landick, and J. Gelles. Tethered particle motion method for studying transcript elongation by a single RNA polymerase molecule. *Biophys J*, 67(6):2468–2478, 1994.
- [104] L. Finzi and J. Gelles. Measurement of Lactose repressor-mediated loop formation and breakdown in single DNA molecules. *Science*, 267:378–380, 1995.
- [105] P. C. Nelson, C. Zurla, D. Brogioli, J. F. Beausang, L. Finzi, and D. Dunlap. Tethered particle motion as a diagnostic of DNA tether length. *J Phys Chem B*, 110(34):17260–17267, 2006.
- [106] S. Blumberg, A. Gajraj, M. W. Pennington, and J.-C. Meiners. Three-dimensional characterization of tethered microspheres by total internal reflection fluorescence microscopy. *Biophys J*, 89:1272–1281, 2005.
- [107] C. Zurla, A. Franzini, G. Galli, D. D. Dunlap, D. E. A. Lewis, S. Adhya, and L. Finzi. Novel tethered particle motion analysis of CI protein-mediated DNA looping in the regulation of bacteriophage lambda. *J. Phys: Condens Matter*, 18:S225–S234, 2006.
- [108] D. Rutkauskas, H. Zhan, K. S. Matthews, F. S. Pavone, and F. Vanzi. Tetramer opening in LacI-mediated DNA looping. *Proc Natl Acad Sci USA*, 106(39):16627–16632, 2009.
- [109] O. K. Wong, M. Guthold, D. A. Erie, and J. Gelles. Interconvertible Lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biol*, 6(9):e232, 2008.
- [110] J. P. Mumm, A. Landy, and J. Gelles. Viewing single lambda site-specific recombination events from start to finish. *EMBO J*, 25(19):4586–4595, October 2006.
- [111] N. Laurens, S. R. W. Bellamy, A. F. Harms, Y. S. Kovacheva, S. E. Halford, and G. J. L. Wuite. Dissecting protein-induced DNA looping dynamics in real time. *Nucleic Acids Res*, 37(16):5454–5464, 2009.

- [112] J.-F. Chu, T.-C. Chang, and H.-W. Li. Single-molecule TPM studies on the conversion of human telomeric DNA. *Biophys J*, 98(8):1608–1616, April 2010.
- [113] F. Vanzi, C. Broggio, L. Sacconi, and F. S. Pavone. Lac repressor hinge flexibility and DNA looping: single molecule kinetics by tethered particle motion. *Nucleic Acids Res*, 34(12):3409–3420, 2006.
- [114] D. Normanno, F. Vanzi, and F. S. Pavone. Single-molecule manipulation reveals supercoiling-dependent modulation of *lac* repressor-mediated DNA looping. *Nucleic Acids Res*, 36(8):2505–2513, 2008.
- [115] L. Han, H. G. Garcia, Blumberg S., K. B. Towles, J. F. Beausang, P. C. Nelson, and R. Phillips. Concentration and length dependence of DNA looping in transcriptional regulation. *PLoS ONE*, 4(5):e5621, 2009.
- [116] J. N. Milstein, Y. F. Chen, and J.-C. Meiners. Bead size effects on protein-mediated DNA looping in tethered-particle motion experiments. *Biopolymers*, 95(2):144–150, 2011.
- [117] S. Johnson, M. Lindén, and R. Phillips. Sequence dependence of transcription factor-mediated DNA looping. *accepted to Nucleic Acids Res*, 2012.
- [118] J. Q. Boedicker, H. G. Garcia, S. Johnson, and R. Phillips. DNA-bending protein HU masks the sequence-dependence of repressor-mediated loop formation *in vivo*. *under review at PNAS*, 2012.
- [119] M. Linden, S. Johnson, C. Wiggins, and R. Phillips. HMM analysis of tethered particle motion. *in preparation*, 2012.
- [120] L. Han. *In vitro DNA mechanics in gene regulation: one molecule at a time*. PhD thesis, California Institute of Technology, 2008.
- [121] Y.-F. Chen, J. N. Milstein, and J.-C. Meiners. Femtonewton entropic forces can control the formation of protein-mediated DNA loops. *Phys Rev Lett*, 104(4):048301, 2010.

- [122] G. Lia, D. Bensimon, V. Croquette, J.-F. Allemand, D. Dunlap, D. E. A. Lewis, S. Adhya, and F. Finzi. Supercoiling and denaturation in Gal repressor/heat unstable nucleoid protein (HU)-mediated DNA looping. *Proc Natl Acad Sci USA*, 100(20):11373–11377, 2003.
- [123] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15(2):116–124, 2005.
- [124] J. M. G. Vilar and S. Leibler. DNA looping and physical constraints on transcriptional regulation. *J Mol Biol*, 331(5):981–989, 2003.
- [125] D. E. Frank, R. M. Saecker, J. P. Bond, M. W. Capp, O. V. Tsodikov, S. E. Melcher, M. M. Levandoski, and M. T. Record Jr. Thermodynamics of the interactions of *lac* repressor with variants of the symmetric *lac* operator: Effects of converting a consensus site to a non-specific site. *J Mol Biol*, 267(5):1305–1314, 1997.
- [126] H. Zhan, Z. Sun, and K. S. Matthews. Functional impact of polar and acidic substitutions in the Lactose repressor hydrophobic monomer-monomer interface with a buried lysine. *Biochemistry*, 48(6):1305–1314, 2009.
- [127] K. B. Towles, J. F. Beausang, H. G. Garcia, R. Phillips, and P. C. Nelson. First-principles calculation of DNA looping in tethered particle experiments. *Phys Biol*, 6(2):025001, 2009.
- [128] D. Swigon, B. D. Coleman, and W. K. Olson. Modeling the Lac repressor-operator assembly: the influence of DNA looping on Lac repressor conformation. *Proc Natl Acad Sci USA*, 103(26):9879–9884, 2006.
- [129] Y. Zhang, A. E. McEwen, D. M. Crothers, S. D. Levene, and P. Fraser. Analysis of *in vivo* LacR-mediated gene repression based on the mechanics of DNA looping. *PLoS ONE*, 1:e136, 2006.
- [130] S. Goyal, T. Lillian, S. Blumberg, J.-C. Meiners, E. Meyhöfer, and N. C. Perkins. Intrinsic curvature of DNA influences LacR-mediated looping. *Biophys J*, 93(12):4342–4359, 2007.

- [131] J. K. Barry and K. S. Matthews. Thermodynamic analysis of unfolding and dissociation in Lactose repressor protein. *Biochemistry*, 38:6520–6528, 1999.
- [132] J. Chen and K. S. Matthews. Subunit dissociation affects DNA binding in a dimeric *Lac* repressor produced by C-terminal deletion. *Biochemistry*, 33:8728–8735, 1994.
- [133] J. Chen and K. S. Matthews. Deletion of Lactose repressor carboxyl-terminal domain affects tetramer dissociation. *J Biol Chem*, 267(20):13843–13850, 1992.
- [134] M. Brenowitz, N. Mandal, A. Pickar, E. Jamison, and S. Adhya. DNA-binding properties of a *Lac* repressor mutant incapable of forming tetramers. *J Biol Chem*, 266(2):1281–1288, 1991.
- [135] M. M. Levandoski, O. V. Tsodikov, D. E. Frank, S. E. Melcher, R. M. Saecker, and M.T. Record Jr. Cooperative and anticooperative effects in binding of the first and second plasmid *O_{sym}* operators to a *LacI* tetramer: Evidence for contributions of non-operator DNA binding by wrapping and looping. *J Mol Biol*, 260:697–717, 1996.
- [136] M. Plischke and B. Birger. *Equilibrium statistical physics*. World Scientific, Hackensack, NJ, 3rd edition, 2006.
- [137] W.-T. Hsieh, P. A. Whitson, K. S. Matthews, and R. D. Wells. Influence of sequence and distance between two operators on interaction with the *lac* repressor. *J Biol Chem*, 262(30):14583–14591, 1987.
- [138] P. A. Whitson, J. S. Olson, and K. S. Matthews. Thermodynamic analysis of the Lactose repressor-operator DNA interaction. *Biochemistry*, 25(13):3852–3858, 1986.
- [139] P. A. Whitson and K. S. Matthews. Dissociation of the Lactose repressor-operator DNA complex: Effects of size and sequence context of operator-containing DNA. *Biochemistry*, 25(13):3845–3852, 1986.
- [140] D. E. Segall, P. C. Nelson, and R. Phillips. Volume-exclusion effects in Tethered-Particle experiments: bead size matters. *Phys Rev Lett*, 96(8):088306, 2006.

- [141] A. D. Hirsh, T. D. Lillian, T. A. Lionberger, and N. C. Perkins. DNA modeling reveals an extended Lac repressor conformation in classic *in vitro* binding assays. *Biophys J*, 101:718–726, 2011.
- [142] M. A. Watson, D. M. Gowers, and S. E. Halford. Alternative geometries of DNA looping: an analysis using the *Sfi*I endonuclease. *J Mol Biol*, 298:461–475, 2000.
- [143] L. Saiz and J. M. G. Vilar. Multilevel deconstruction of the *in vivo* behavior of looped DNA-protein complexes. *PLoS ONE*, 2:e355, 2007.
- [144] E. Villa, A. Balaeff, and K. Schulten. Structural dynamics of the lac repressor–DNA complex revealed by a multiscale simulation. *Proc Natl Acad Sci USA*, 102(19):6783–6788, May 2005.
- [145] H. R. Drew and A. A. Travers. DNA bending and its relation to nucleosome positioning. *J Mol Biol*, 186(4):773–790, 1985.
- [146] K. J. Dechering, K. Cuelenaere, R. N. H. Konings, and J. A. M. Leunissen. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res*, 26(17):4056–4062, 1998.
- [147] V. Iyer and K. Struhl. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J*, 14(11):2570–2579, 1995.
- [148] E. Segal and J. Widom. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol*, 19(1):65–71, 2009.
- [149] G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, 309:627–630, 2005.
- [150] D. Rhodes. Nucleosome cores reconstituted from poly(dA-dT) and the octamer of histones. *Nucleic Acids Res*, 6(5):1805–1816, 1979.
- [151] T. E. Shrader and D. M. Crothers. Effects of DNA sequence and histone-histone interactions on nucleosome placement. *J Mol Biol*, 216(1):69–84, 1990.

- [152] J. D. Anderson and J. Widom. Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol Cell Biol*, 21(11):3830–3839, 2001.
- [153] K. Struhl. Naturally occurring poly(dA:dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proc Natl Acad Sci USA*, 82(24):8419–8423, 1985.
- [154] H. C. M. Nelson, J. T. Finch, B. F. Luisi, and A. Klug. The structure of an oligo(dA)-oligo(dT) tract and its biological implications. *Nature*, 330(6145):221–226, 1987.
- [155] T. E. Haran, J. D. Kahn, and D. M. Crothers. Sequence elements responsible for DNA curvature. *J Mol Biol*, 244(2):135–143, 1994.
- [156] E. D. Ross, A. M. Keating, and L. J. Maher 3rd. DNA constraints on transcriptional activation *in vitro*. *J Mol Biol*, 297(2):321–334, 2000.
- [157] E. D. Ross, P. R. Hardwidge, and L. J. Maher 3rd. HMG proteins and DNA flexibility in transcriptional activation. *Mol Cell Biol*, 21(19):6598–6605, 2001.
- [158] D. Colquhoun and F. J. Sigworth. Fitting and statistical analysis of single-channel records. In B. Sakmann and E. Neher, editors, *Single Channel Recording*, pages 191–263. Plenum Press, 1983.
- [159] C. Manzo and L. Finzi. Quantitative analysis of DNA-looping kinetics from Tethered Particle Motion experiments. *Methods Enzymol*, 475:199–220, 2010.
- [160] S. A. McKinney, A.-C. Déclais, D. M. J. Lilley, and T. Ha. Structural dynamics of individual Holliday junctions. *Nat Struct Biol*, 10(2):93–97, 1981.
- [161] J. F. Beausang, C. Zurla, C. Manzo, D. Dunlap, L. Finzi, and P. C. Nelson. DNA looping kinetics analyzed using diffusive hidden Markov model. *Biophys J*, 92(8):L64–L66, 2007.
- [162] J. F. Beausang and P. C. Nelson. Diffusive hidden Markov model characterization of DNA looping dynamics in tethered particle experiments. *Phys Biol*, 4:205–219, 2007.

- [163] J. E. Bronson, J. Fei, J. M. Hofman, R. L. Gonzalez Jr., and C. H. Wiggins. Learning rates and states from biophysical time series: A Bayesian approach to model selection and Single-Molecule FRET data. *Biophys. J.*, 97(12):3196–3205, 2009.
- [164] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*, 77:257–286, 1989.
- [165] D. Colquhoun and B. Sakmann. Fluctuations in the microsecond time range of the current through single acetylcholine receptor ion channels. *Nature*, 294:464–466, 1981.
- [166] J. F. Beausang, C. Zurla, L. Finzi, L. Sullivan, and P. C. Nelson. Elementary simulation of tethered Brownian motion. *American Journal of Physics*, 75(6):520–523, June 2007.
- [167] M. Manghi, C. Tardin, J. Baglio, P. Rousseau, L. Salomé, and N. Destainville. Probing DNA conformational changes with high temporal resolution by tethered particle motion. *Phys Biol*, 7(4):046003, 2010.
- [168] S. C. Schultz, G. C. Shields, and T. A. Steitz. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, 253(5023):1001–1007, 1991.
- [169] T. Kuhlman, Z. Zhang, M. H. Saier Jr., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc Natl Acad Sci USA*, 104(14):6043–6048, 2007.
- [170] L. Saiz and J. M. G. Vilar. *Ab initio* thermodynamic modeling of distal multisite transcriptional regulation. *Nucleic Acids Res*, 36(3):726–731, 2008.
- [171] H. G. Garcia and R. Phillips. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci USA*, 108(29):12173–12178, 2011.
- [172] M. Fried and D. M. Crothers. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 9(23):6505–6525, 1981.
- [173] R. B. Winter and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The *Escherichia coli* repressor-operator interaction: equilibrium measurements. *Biochemistry*, 20(24):6948–6960, 1981.

- [174] J. M. Hudson and M. G. Fried. Co-operative interactions between the catabolite gene activator protein and the *lac* repressor at the lactose promoter. *J Mol Biol*, 214(2):381–396, 1990.
- [175] P. A. Whitson, W.-T. Hsieh, R.D. Wells, and K. S. Matthews. Supercoiling facilitates *lac* operator-repressor-pseudooperator interactions. *J Biol Chem*, 262(11):4943–4946, 1987.
- [176] P. A. Whitson, W.-T. Hsieh, R.D. Wells, and K. S. Matthews. Influence of supercoiling and sequence context on operator DNA binding with *lac* repressor. *J Biol Chem*, 262(30):14592–14599, 1987.
- [177] J. van Noort, S. Verbrugge, N. Goosen, C. Dekker, and R. T. Dame. Dual architectural roles of HU: formation of flexible hinges and rigid filaments. *Proc Natl Acad Sci USA*, 101(18):6969–6974, 2004.
- [178] T. Aki, H. E. Choy, and S. Adhya. Histone-like protein HU as a specific transcriptional regulator: co-factor role in repression of *gal* transcription by GAL repressor. *Genes Cells*, 1(2):179–188, 1996.
- [179] C. Zurla, T. Samuely, G. Bertoni, F. Valle, G. Dietler, L. Finzi, and D. D. Dunlap. Integration host factor alters LacI-induced DNA looping. *Biophys Chem*, 128:245–252, 2007.
- [180] S. Kar and S. Adhya. Recruitment of HU by piggyback: a special role of GalR in repressosome assembly. *Genes Dev*, 15(17):2273–2281, 2001.
- [181] J. B. Kinney, A. Murugan, C. G. Callan Jr., and E. C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A*, 107(20):9158–9163, 2010.
- [182] T. Plénat, C. Tardin, P. Rousseau, and L. Salomé. High-throughput single-molecule analysis of DNA-protein interactions by tethered particle motion. *Nucleic Acids Res*, 2012.
- [183] M. Ling and B. Robinson. A one-step polymerase chain reaction site-directed mutagenesis method for large gene-cassettes with high efficiency, yield, and fidelity. *Anal Biochem*, 230(1):167–172, 1995.

- [184] M. Ling and B. Robinson. Approaches to DNA mutagenesis: an overview. *Anal Biochem*, 254(2):157–178, 1997.
- [185] M. Kammann, J. Laufs, J. Schell, and B. Gronenborn. Rapid insertional mutagenesis of DNA by polymerase chain reaction (PCR). *Nucleic Acids Res*, 17(13):5404, 1989.
- [186] G. Sarkar and S. S. Sommer. The “megaprimer” method of site-directed mutagenesis. *Biotechniques*, 8(4):404–407, 1990.
- [187] O. Landt, H.P. Grunert, and U. Hahn. A general method for rapid site-directed mutagenesis using the polymerase chain reaction. *Gene*, 96(1):125–128, 1990.
- [188] J. Xu and K. S. Matthews. Flexibility in the Inducer Binding Region is Crucial for Allostery in the *Escherichia coli* Lactose repressor. *Biochemistry*, 48(2):4988–4998, 2009.
- [189] A. P. Butler, A. Revzin, and P. H. von Hippel. Molecular parameters characterizing the interaction of *Escherichia coli lac* repressor with non-operator DNA and inducer. *Biochemistry*, 16(22):4757–4768, 1977.
- [190] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999.