Studying Looping-Based Transcriptional Regulation Using Synthetic Biology Tools

Michal Brunwasser-Meirom

Studying Looping-Based Transcriptional Regulation Using Synthetic Biology Tools

Research Thesis

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

Michal Brunwasser-Meirom

Submitted to the Senate of the Technion- Israel Institute of Technology

Av, 5777 Haifa July 2017

The Research Thesis Was Done Under the Supervision of Associated Prof. Roee Amit in the Department of Biotechnology and Food Engineering.

The generous financial help of the Technion is gratefully acknowledged.

I would like to express my deep gratitude to Prof. Roee Amit that escorted me throughout this journey. This PhD would have not been possible without his invaluable guidance and continuous support. I truly appreciate his encouragement, patience and advice whenever I needed it.

I would also like to thank my husband Eli, for escorting me during all my academic path and for never stopping believing in me.

List of Publications

1. Using synthetic bacterial enhancers to reveal a looping-based mechanism for quenching-like repression. **Brunwasser-Meirom, M.**, Pollak Y., Goldberg S., Levy L., Atar O., Amit R. *Nature Communications*. 7:10407, 2016.

Conferences

Oral Presentations

1. **Brunwasser-Meirom, M.**, Pollak Y., Goldberg S., Levy L., Atar O., Amit R. ‘‘Using Synthetic Biology to Reveal a Mechanism for Quenching Repression”. Ilanit 2017 Eilat, Israel.

Poster Presentations

1. **Brunwasser-Meirom, M.**, Pollak Y., Goldberg S., Levy L., Atar O., Amit R. ‘‘Using Synthetic Biology to Reveal a Mechanism for Quenching Repression”. Pearl Seiden conference 2016. Haifa, Israel.
2. **Brunwasser-Meirom, M.**, Pollak Y., Goldberg S., Levy L., Atar O., Amit R. ‘‘Deciphering Enhancers Regulatory Code using Synthetic Biology”. Total Transcription 2015 Hinxton, UK.
3. **Brunwasser-Meirom, M.**, Pollak Y., Goldberg S., Levy L., Atar O., Amit R. ‘‘Deciphering Enhancers Regulatory Code using Synthetic Biology”. Ilanit 2014 Eilat, Israel.
4. **Brunwasser-Meirom, M.**, Pollak Y., Goldberg S., Atar O., Amit R. “Studying Transcription Regulation using Synthetic Biology”. DNA60 conference 2013 Haifa, Israel.
5. **Brunwasser-Meirom, M.**, Pollak Y., Goldberg S., Atar O., Amit R. ‘‘Studying Transcription Regulation using Synthetic Biology”. RBNI Winter Symposium “Studying Transcription Regulation using Synthetic Biology”. In Nanotechnology, 2012 Hagoshrim, Israel.

Table of Contents

List of Figures

List of Tables

Abbreviations

bp - base pair

Indels - insertion deletions

RNAP - RNA polymerase

PolII - RNA polymerase II

TF - transcription factors

EBP - enhancer binding protein

TSS - transcription start site

UAS - upstream activating sequence

GST - gluthatione-S- transferase

WLC - worm-like chain

pspG - phage shock protein

σ^{54} - sigma 54

σ^{70} - sigma 70

DNA - deoxyribonucleic acid

dsDNA - double-stranded DNA

RNA - ribonucleic acid

IHF - integration host factor

OD - optical density

LB - lysogeny broth

BA - bioassay media

DDW - double-distilled water

PBS - phosphate-buffered saline

IPTG - isopropyl b-D-1-thiogalactopyranoside

aTc - anhydroTetRacycline

Kb - kbp 10^3 basepairs

Mbp - 10^6 basepairs

kDa - 10^3 daltons

qrr -quorum regulatory RNA

g - gram

ml - milliliter

µl - microliter

µM, mM, M - micromolar, millimolar, molar

YPD - yeast extract peptone-dextrose

SD - synthetic defined

FACS - flow cytometry cell sorter

SDS-PAGE - sodium dodecyl sulfate polyacrylamide gel electrophoresis

RT - real time

WT - wild type

CRISPR - Clustered regularly interspaced short palindromic repeats

gRNA - guide RNA

HDR - homologous recombination

Ori - origin of replication

Kan - kanamycin

Amp - ampicillin

3OC8 - N-(3-Oxooctanoyl)-L-homoserine lactone

C_4HSL - N-butanoyl-L-homoserine lactone

EMSA - electrophoretic mobility shift assay

O/N - overnight

dNTPs - deoxyribonucleic acid

Δ - Gene deletion

MBW - molecular biology water

NtrC - Nitrogen regulatory protein C

PCR - polymerase chain reaction

TetR - Tetracycline repressor

RPM - rounds per minute

TBE - Tris/Borate/EDTA

V - volt

EDTA - Ethylenediaminetetraacetic acid

DTT - Dithiothreitol

CDS - coding sequence

GFP - green fluorescent protein

Spec- spectinomycin

MPRA- massive parallel reporter assay

Abstract

Distal regulation by transcription factors is a regulatory phenomenon ubiquitous in all organisms. To develop a deeper understanding of distal regulatory regions, not only is there a need for additional gene expression data sets but models describing the underlying mechanisms must be formulated as well. To address this problem, we explore a model for 'quenching-like' repression by studying synthetic bacterial enhancers, each characterized by a different binding-site architecture. To do so, we take a three-pronged approach: first, we compute the probability that a protein-bound dsDNA molecule will loop. Second, we use hundreds of synthetic enhancers to test the model's predictions in bacteria. Finally, we verify the mechanism bioinformatically in native genomes. Our combined findings suggest that excluded volume of bound proteins can account for both up-regulating and down-regulating (quenching) effects in bacterial enhancers. We found that the nature and magnitude of the regulatory effect are influenced by the size of the TF, the number of bound TFs and their relative arrangement within the enhancer. The nature and magnitude of the effect are highly sensitive to the location of the TF binding site and exhibit an oscillating pattern whose period matches the dsDNA helical repeat. Additionally, bound TFs can augment or diminish the effect, depending on their relative orientation to the other TFs. The implications of these results are that enhancers should be insensitive to 10–11 bp insertion deletions (INDELs) and sensitive to 5–6 bp INDELs. We test this prediction on 61 σ^{54}-regulated *qrr* genes from the *Vibrio* genus and confirm the tolerance of these enhancers' sequences to the DNA's helical repeat. Furthermore, genomic editing of σ^{54} -regulated gene *pspG* looping region, gave additional support to our predictions. Additionally, we attempt to examine whether the mechanism that we characterized also plays a role in yeast. In our system, looping is not observed in yeast, pointing out that perhaps other mechanisms play a role in the action-at-a-distance regulation present in this organism.

Introduction

Enhancers are non-translated DNA sequences which play a fundamental role in gene regulation. They were first identified in viral DNA as regions of non-coding DNA capable of increasing transcription [1]. Soon after, they were recognized as important players in transcription regulation of virtually all life systems [2,3]. An enhancer is typically comprised of multiple binding sites for transcription factors (TF) for activators and repressors, and functions as a type of molecular integrator that determines when, where, and how much of a certain gene is expressed. Enhancers can be located very close to their regulated promoter or hundreds of bases away, and the regulatory output is often independent of their location or orientation relative to the basal promoter. That is, enhancers can be found upstream or downstream of genes, or within introns[4,5]. Additionally, enhancers do not necessarily regulate the closest promoter but can act on genes located more distantly. Although enhancers have been extensively studied for a few decades, they are still poorly understood. Most structural features, such as the importance of having several binding sites for a given transcription factor, the genomic distance of these binding sites from the basal promoter and the functional significance of particular arrangements of the binding sites remain poorly understood.

Enhancers' Repression Mechanisms

Enhancers can execute complex regulatory functions. However, not much is known about regulatory mechanisms that take place within enhancers, and the rules that connect the internal structure or arrangement of the binding sites to the regulatory output. Several different modes of repression have been proposed, which include "short-range" and "long-range" repression[6]. In

the long-range mechanism, TFs can generate repression from hundreds to thousands of base-pairs away from the nearest activator or relevant promoter by an unknown mechanism. Short-range repression can be achieved by competition, which occurs when an activator and a repressor have overlapping binding sites or their binding sites are very close to one another (Figure 1a). This was shown in bacteria in the well-studied LacI repressor, which has a binding site close to the transcription initiation site to block RNA polymerase access to the promoter [7], and was also shown for phage lambda cl and LexA repressors [8,9] . A second short-range mechanism is called quenching-repression and is discussed below.

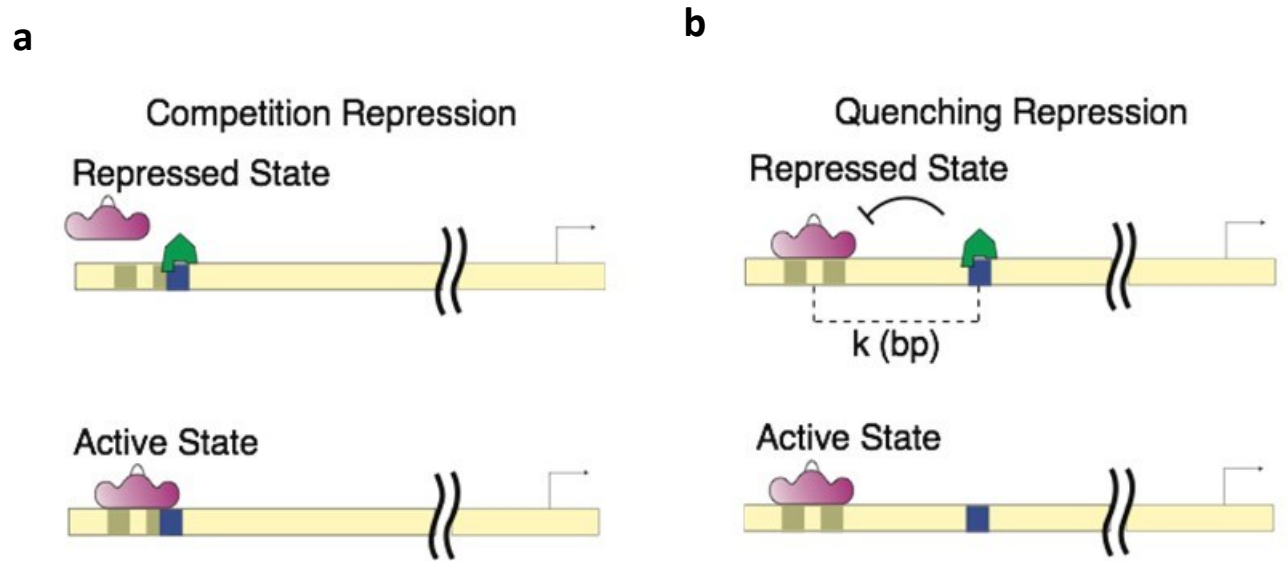


Figure 1. Mechanisms of short-range repression.

In competition repression, the repressor competes with the activator for the same binding sites (a) whereas in quenching repression the binding sites do not overlap, the repressor can bind up to 100 bp away from the activator and there is no evidence for physical interaction between the two (b).

Quenching Repression

'Quenching' is a form of repression originally observed in fly enhancers, where repressors such as Snail [10], Kruppel [11], Knirps [12] or Giant [13] downregulate expression not via a competition with an activator for binding, but rather through having their binding sites positioned several 10 to ~100 bp away from the nearest activator (Figure 1b and Figure 2). 'Quenching-like' repression effects have also been reported for eukaryotic promoter-proximal regulatory regions and in bacterial enhancers. Here, repressors are bound in-between the activators and the core promoter in repressed complexes. Well-documented examples include the YY1 repressor in the c-fos and other promoters in mammalian cells [14,15]; the a2 repressor that was found to be co-bound with the Gal4 activator in a tightly repressed complex in *S. cerevisiae* [16], the glnAp2 σ^{54} promoter in *Escherichia coli* [17] and the Nac σ^{54} promoter in *Klebsiella aerogenes* [18]. Quenching-like effects have been attributed to an inadequate positioning of an IHF-binding site with respect to the activator or σ^{54} promoter sequences in other promoters as well [19]. However, despite the many observations of closely bound ensembles of proteins on distal regulatory elements, which interact in a repressive manner to regulate gene expression, establishing a broadly applicable mechanistic explanation has remained elusive.

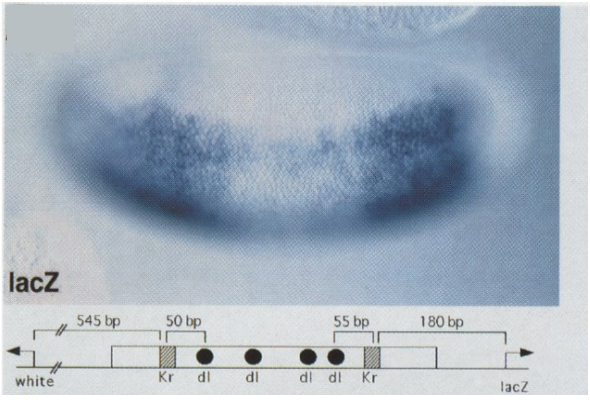


Figure 2. Quenching repression is the Drosophila embryo.

LacZ staining pattern of Rho is controlled by multiple binding sites of dorsal activator which are flanked by Krupple repressor located ~50bp upstream/ downstream. LacZ is not observed in the center part of the embryo where Krupple is specifically expressed. Picture from [11].

Enhancer Subtypes

Enhancers can be separated into a few subclasses based on binding-site architecture and proximity to the promoter. While in bacteria there is only one enhancer sub-type, in eukaryotes we can differentiate between a few types: near/distal-promoter enhancers, enhanceosomes, and shadow enhancers.

Bacterial Enhancers

Bacterial enhancers are highly modular objects, whose binding-site architecture can be grossly divided into three distinct modules: the driver (activator), the basal promoter, and the region in-between the two aforementioned modules which typically contains a multitude of binding sites for several (1–5) transcription factors. The driver module is typically associated with either two or three specialized binding sites that are located between 50 and 500 bp upstream of the basal promoter.

Eukaryotic Enhancers

Eukaryotic enhancers range in size from several hundred base pairs to 1 kb, and can be located very close to the promoter or tens of kilobases away. They contain multiple clustered binding sites for transcription factors. Eukaryotic enhancers can be classified into the following subtypes: near/distal-promoter enhancers, enhanceosomes, and shadow enhancers.

Near/distal Promoter Enhancers

The “near-promoter” sub-class of eukaryotic enhancers share many features with both bacteria and eukaryotic examples. These enhancers are composed of regulatory sequences made of a handful of binding sites that are located <500 bp from a basal promoter and are often transcribed via the promoter proximal pausing mechanism [20].

Enhanceosomes

Enhanceosomes refer to tightly-clustered binding sites for transcription factors which bind the enhancer in a highly cooperative manner, leading to gene activation by recruitment of RNA Pol II [21,22]. There are a few important differences between enhanceosomes and developmental enhancers. First, they are typically located near the promoter, usually <300 bp from the TATA box. Second, they do not operate in an additive fashion, as disruption or displacement of a single binding site or the absence of even one regulatory TF result in an inactive complex. And lastly, while developmental enhancers are associated with poised/paused polymerase and DNA looping, enhanceosomes operate by recruiting the Pol II machinery to allow transcription initiation.

Shadow Enhancers

Shadow enhancers represent another subclass of eukaryotic enhancers, and are found to control genes involved in critical developmental processes [23]. These enhancers are considered secondary enhancers, since they are located many kilobases from the primary enhancer that is often closer to the gene being regulated. Nevertheless, they produce patterns of gene expression that are the same as or similar to those produced by more proximal primary enhancers.

Super Enhancers

A recently emerging sub class of enhancers, super enhancer, has been used to describe groups of putative enhancers in close genomic proximity with unusually high levels of Mediator binding[24]. These regions are thought to play a part in controlling cell identity and disease genes, probably through shaping miRNA expression [25] .

Action-at-a-Distance in Bacterial and Eukaryotic Enhancers

Enhancers use various mechanisms for regulating their target genes. One important aspect is the activation of the promoter from a distance by an “action-at-a-distance” mechanism. While some enhancers are located close enough to the polymerase to make direct contact, others may be

located hundreds and thousands of base pairs upstream to the promoter and are thought to make contact by looping of the DNA (Figure 3). DNA looping has been implicated in distal regulation in eukaryotes [26,27] and has been shown directly to be involved in σ^{54} promoter expression in bacteria [28].

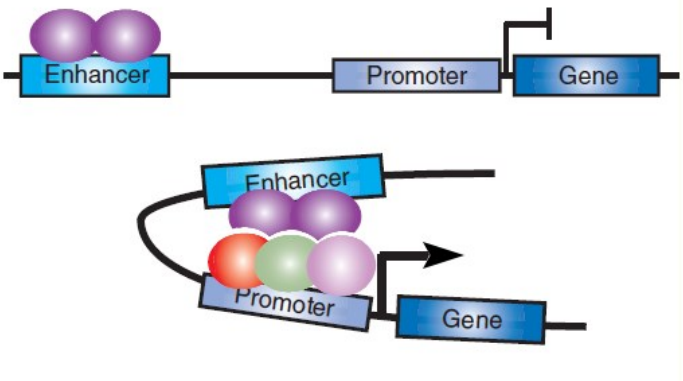


Figure 3. Action at a distance via DNA looping.

DNA looping brings transcription factors bound at the enhancer site to the promoter. Picture adapted from (*48*).

Yeast Upstream Activating Sequences

The metazoan enhancers' counterparts in yeast are upstream activating sequences (UASs), which can activate transcription by recruiting gene-specific activators. They are usually positioned within a few hundred base pairs from the core promoter [29]. While some evidence indicates that UASs cannot act if positioned at a great distance from their target gene, this area has not been fully investigated. To date, it has not been shown directly that activation at a distance in yeast is mediated by a looping mechanism that is controlled by transcription factors. It has been shown that, in principle, looping can occur in yeast in other contexts involving telomeres, that are known to form back-folding, looped-structures[30]. Others have shown DNA looping in a reporter system, where GAGA, a transcription factor from *Drosophila* known to facilitate DNA loop formation, enables enhancer action in yeast over a distance of 3000 bp [29,31]. Hence, there is no direct evidence that DNA looping is the mechanism for transcription activation at-a-distance in yeast[32,33]. Since most of this work aims to study bacterial enhancers, the next sections will focus on

bacterial transcription regulation, bacterial enhancers and common features of bacterial and eukaryotic enhancer.

Regulation of Gene Expression in Bacteria

Transcription is an important regulatory step in bacterial gene expression. Initiation of transcription is mediated by RNA polymerase (RNAP) along with a modular subunit called sigma factor required for both directing the polymerase to a specific promoter and DNA melting [34], which together form a holoenzyme. Sigma factors can be separated into main two classes based on regulation properties and promoter specificity. The first class, sigma factor σ^{70} family, is represented by a group of diverse sigma factors which are involved in the expression of most genes in exponential growth [35,36]. σ^{70} members bind to conserved -10 and -35 bp promoter elements, and direct the polymerase binding to this consensus sequence**.** Once the holoenzyme is bound, a closed complex is no longer energetically favored and is converted to an open complex to initiate transcription. The second class is the sigma factor σ^{54} class, which is composed of only one member. σ^{54} binds to different consensus sequences at -24 and -12 bp. As opposed to σ^{70}-mediated initiation, σ^{54} is unable to initiate transcription by itself. It requires the presence of a bacterial enhancer binding protein (EBP) that we term “driver” that couples the energy generated from ATP hydrolysis to the isomerization of the holoenzyme closed complex (Figure 4). This effectively causes the polymerase to be poised at the gene of interest awaiting the arrival of the driver. The driver typically binds to an enhancer, which is a DNA sequence typically located a large distance upstream (100–1000 bp) to the promoter, precluding it from forming direct contact with the poised polymerase [5]. The interaction of the poised polymerase with the driver has been shown to be mediated via DNA looping [28] which is an action-at-a-distance mechanism. One notable example in bacteria is the glnAp2 promoter which requires the σ^{54} factor. This promoter is regulated by NtrC that binds near position -110 relative to TSS, and cannot interact with the polymerase without looping [37]. Both ATP hydrolysis and DNA looping are required to induce open complex formation and transcription initiation [35]. On its own, a poised

promoter has the capability to execute little or no transcriptional regulation, but together with enhancers, it can achieve its full regulatory potential [38,39].

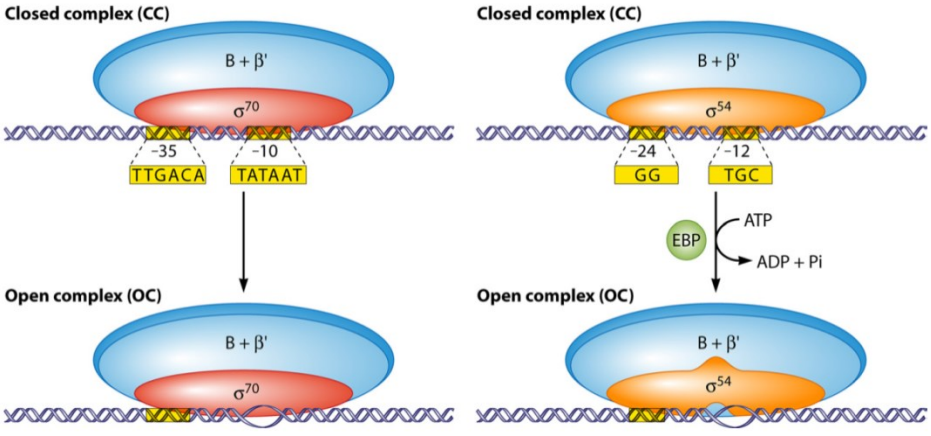


Figure 4. Transcription initiation by RNAP-σ^{70} and RNAP-σ^{54}holoenzymes.

σ^{54} mediated initiation (right) requires both hydrolysis of ATP and an addition activator (EBP) which drives initiation, whereas σ^{70} (left) binding itself serves as a driving force for initiation. Picture adapted from [17].

σ^{54}-Dependent Gene Expression in Bacteria

σ^{54}-mediated transcription is associated with a wide range of cellular processes including survival under different stress conditions. The analysis of [40] has shown that σ^{54} promoters predominantly regulate genes that control the transport and biosynthesis of the molecules that constitute the bacterial exterior, thus affecting cell structure, developmental phase, and interaction potential with the environment Some well-known examples are the nitrogen regulatory protein C (NtrC) and the nitrogen fixation protein A (NifA), both required for nitrogen

metabolism, the phage shock protein F (PspF) which stabilizes cells under stress, and xylene catabolism regulatory protein (XylR) and 3,4-dimethylphenol catabolism regulatory protein (DmpR) required for xylem and phenol catabolism, respectively [3,40]. In addition, σ^{54} regulated genes also play an important role in developmental-like processes, as seen in *M. Xanthus*. When nutrition is limited, *M. Xanthus* undergoes a developmental stage which results in its assembly into aggregates that grow into complex structures called fruiting bodies and spores [41].

Common Features of Bacterial and Eukaryotic Enhancers

Poised Transcription

One of the more striking features of enhancer-regulated transcription in bacteria is the "poised-polymerase" phenomenon, where the holoenzyme complex is stalled at the TSS unable to initiate transcription on its own without the driver, which is typically bound a large distance upstream. In eukaryotes, for the most part the generation of mRNA from protein-coding genes is mediated by RNA polymerase II (Pol II), along with many auxiliary factors. However, recent genome-wide studies of Pol II distribution indicated that another "stalled" or "poised" mechanism may also be prevalent in many enhancer-regulated gene expression processes [42]. The findings suggest that in 20-30% of the genes, Pol II is concentrated in the promoter regions at times and in places where the genes are known to be off. In these promoters, the Pol II holoenzyme complex has initiated transcription but has paused indefinitely in the promoter proximal region 20-60 nucleotides from the TSS [43]. Interestingly, of the genes identified to harbor a "paused" polymerase by the genomic studies, many are associated with developmental control [44].

Similarly, in bacteria, poised σ^{54} transcription [35] has been associated with developmental-like processes such as nitrogen regulation [39] and fruiting body development in *M. Xanthus*. Thus, a form of paused or poised transcription is ubiquitous to many biological kingdoms, and seems to be over-represented in genes that are known to play an important role in executing "developmental" programs.

Common Architecture

One example, which conveys the similarities between a sample bacterial enhancer and a near-promoter eukaryotic enhancer, is seen in Figure 5. Another example has shown that the eukaryotic YY1 DNA-bending protein orientation and relative positioning determines whether this TF either represses or activates expression [15]. Similarly, IHF, a prokaryotic protein known to bend DNA, acts in an orientation-dependent manner [4]. Since IHF and YY1 bend DNA, their effect is an elastic one and therefore plays a role in enhancers located relatively close to the promoter. In contrast, distal promoter enhancers are located thousands of bp from the promoter and presumably rely on entropic regulation of looping, which implies that elastic phenomenon should not be observed in their regulatory regulation.

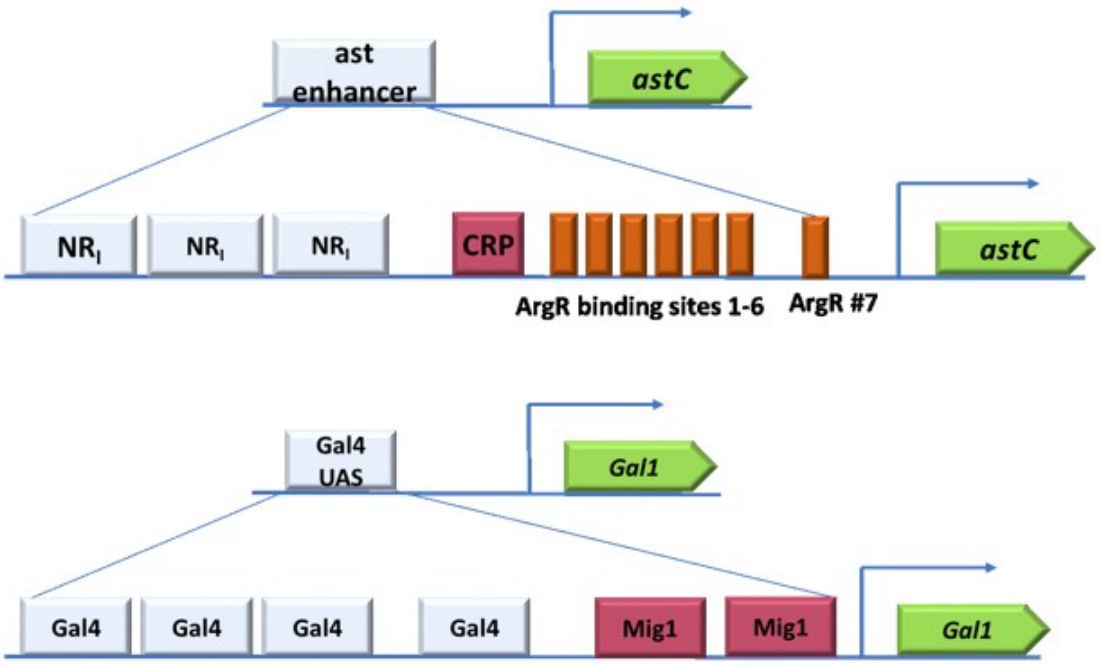


Figure 5. Bacteria and yeast enhancers.
(a) The astCp2 enhancer in E. coli, exhibiting a ~200 bp looping region, at least ten binding sites for three different kinds of TFs, and NRI~P driver binding sites [24]*. (b) The Gal1 UAS in yeast*[23] *showing a similar architecture to bacterial enhancers in terms of binding sites, proximity to promoter, and binding-site separation.*

Synthetic Biology as a Basic Research Tool for Studying Regulation

Synthetic biology is an emerging field of research where scientists construct new biological systems and redesign existing biological systems. It relies on the fact that biological systems, despite their complexity, have some basic rules and modularity that can be manipulated through engineering processes. This field is inspired by electrical engineering, computer science and information theory, using fundamental elements from these fields to help guide the designs.

At the core of Synthetic biology, is the idea of utilizing “biological parts” (e.g. promoters, RBSs repressors etc.) in order to construct genetic networks to exert control of cellar behavior. The use of the knowledge and elements from different doctrines has enabled researchers to create various logic gates (e.g. AND, OR, XOR etc.) [45] demonstrating the potential to harness the molecular biology parts that evolution produced to form the back-bone of a new hard-wired programming language.

A few examples of these new functions are, (i) modular counter circuit that can count inducible events according to programmed input [46], and are used in cells which needs accurate count accuracy of tightly controlled processes. (ii) Toggle switch circuit which can switch between two states in a rapid fashion [47], and (iii) an artificial clock which shows oscillation behavior that may lead to engineering new biological functions in cells [48].

The synthetic biology approach allows us not only to create computer-based circuitry, but also test fundamental basic questions and study regulatory mechanisms in a modular orderly fashion. While the classical biochemical approaches to study enhancers are low-throughput and laborious, the synthetic approach enables us to test many synthetic enhancers simultaneously in a high- throughput manner. Moreover, the synthetic biology approach combines modeling which is crucial for the rational design of synthetic systems. The model-based approach enables us to focus our design based on a set of predictions that we test experimentally in a high-throughput synthetic enhancer experiment. Since enhancers occupy a large genomic space, testing each and every possible combination of certain synthetic enhancer designs, will generate a very large number of possibilities. The model enables us to focus on a particular sequence space.

In recent years, new DNA synthesis technologies have enabled a new form of enhancer studies, where large libraries of synthetic enhancers are constructed and studied. Many recent works mainly in eukaryotic systems, have utilized hundreds to tens of thousands of designed enhancers, leveraging the synthetic approach, to ask basic questions on their features. Examples included mapping out regulatory functional variation by perturbing the sequence space around a particular binding architecture, or designing cis-regulatory regions from the ground up starting from a minimal promoter [49–51]. These studies aimed to define "grammatical rules" such as the number, location, orientation and order of TFs binding sites.

One aspect of their findings revealed that expression levels increase with the addition of more binding sites and seem to saturate at a specific number of sites. Both the number of binding sites at which saturation was observed and the expression value at saturation differed among TFs and sequence contexts[51,52] . Another important insight is concerned with the binding site affinity. The expression could be affected by the presence of weak binding sites, which can serve as a sensor for TF's concentration [51,53,54]. Along with multiplicity and affinity, the effect of the TF identity has proven to play a major role as well. Heterotypic clustering- multiple binding sites for different TFs was shown to be important in several organisms and contexts, specifically in developmental processes that require precise regulation of morphogen gradients to form proper spatiotemporal gene expression patterns[55,56]. In addition, heterotypic clustering can result in higher expression levels than their homotypic clusters, a cluster of multiple transcription factor binding sites for the same transcription factor. The results from these works have shed some light on parts of the regulatory region that are truly important or irreducible, while also illuminated the parts which may be redundant and play a secondary role. However, despite the promise the synthetic-

approaches have yet to realize their potential in developing data sets, which can massively illuminate the underlying regulatory rules. In this work we devised the model-guided synthetic enhancer approach. We realized that even though current technology allows us to rapidly study tens of thousands of synthetic enhancer variant, that number is still infinitesimally smaller than the potential allowed by the vastness of sequence space. As result, we wanted to develop an approach that would, on the one hand, constrain the size of the possible sequence, while taking advantage of low-cost synthesis technology. To do so, we design our synthetic enhancers based on the predictions produced by numerical simulation, which utilize the Self-Avoiding Wormlike chain model to predict the probability of looping based on particular binding sites. After several rounds of synthetic biology experiments and model improvement, we then use the model to make genomic predictions, and use both bioinformatic analysis and genome-editing techniques to test our predictions on real genomes. With the model-based synthetic enhancer approach, we were able to establish with ~300 synthetic enhancer variants that looping-based regulation is a key determinant in σ^{54} promoter-based gene expression levels.

Major Objectives of the Proposed Research

My research focused on investigating enhancer's regulatory behavior by utilizing synthetic biology tools. Efforts in the last few decades aimed to characterize enhancers by classical biochemical assays using knock-down and rescue methods done, along with recent top-down dissection using modern NGS-based approaches, resulted in few findings[49,50]. The synthetic approach allows us to build up the complexity one step at a time, from a hypothetical minimal enhancer to more complex structures that mimic naturally-occurring examples. This approach allowed us to uncover detailed design rules that are common to enhancers in general. Previous works have used synthetic enhancers in order to address basic questions in regulation [4,28,57]. Accordingly, we used this approach for addressing important questions regarding enhancer features.

Research Goals:

1. Study looping-based transcription, specifically the rules of enhancer organization of σ^{54}-mediated genes using a synthetic approach. This approach is based on first modelling and giving predictions for a given architecture, then testing the predictions experimentally, and lastly, looking at genomes to find relevancy for our results.
2. Elucidate quenching repression mechanisms in bacteria. Use a looping-based model to suggest a possible mechanism for quenching repression.
3. Build a system in higher eukaryotes to test whether our suggested looping mechanism plays a similar regulatory role in different organisms.

Research Importance

Enhancers play a fundamental role not only in gene regulation but also in developmental processes. Therefore, understanding the fundamental aspects of enhancer regulation and organization is of great importance. In addition, a better understanding of enhancers will enable to construct synthetic gene circuits for various applications.

Materials and Methods

Bacterial Strains

- *Escherichia coli* Top10 (Invitrogen) cells (Genotype: F– *mcrA* Δ(*mrr-hsdRMS-mcrBC*) *Φ80lacZΔM15* ΔlacX74 *recA1 araD139* Δ(araleu) 7697 *galU galK rpsL* (StrR) *endA1 nupG*), was used for cloning purposes.
- *Escherichia coli* Δ*rpoN* Top10 cells (Genotype: F– *mcrA* Δ(*mrr-hsdRMS-mcrBC*) *Φ80lacZΔM15 ΔlacX74 recA1 araD139* Δ(araleu) 7697 *galU galK rpsL* (*StrR*) *endA1 nupG*, Δ*rpoN*), was used for the Δ*rpoN* silencing real-time experiments.
- *Escherichia coli* 3.300LG cells [58] (Genotype: Δ*glnL*:Δ*glnG*) used in all synthetic enhancers experiments.

Synthetic Enhancer Cassette Design

Synthetic enhancer cassettes were ordered as double-stranded DNA minigenes from Gen9 Inc. Each minigene ordered was ~500 bp long and contained the following parts: BamHI restriction site, tandem NRI-binding sites from glnAp2 promoter (also containing the σ^{70} glnAp1 promoter), the σ^{54} glnAp2 promoter and a HindIII restriction site. In addition, each minigene contained a looping segment in between the NRI tandem binding sites and the σ^{54} promoter. The looping segment was of variable length (N) and contained either one or two binding sites for TraR, TetR or LacI. The binding sites' sequences are depicted in Table 1. The binding sites were positioned in varying inter-site spacing (s) from one another, and distance from the NRI binding sites (k) within the looping region. For insertion into synthetic enhancer plasmids, minigene cassettes were first

double-digested with BamHI/HindIII before being used as an insert in the cloning step. Cloning was then carried out into a basic template synthetic enhancer plasmid as previously described [59,60]. Briefly, synthetic enhancer sequences were computationally designed to have a minimal probability to bind DNA-binding proteins. This was done by using an algorithm that randomly generated a set of sequences, which were compared with the roughly 2,000 known specific DNA-binding sites for *E. coli* transcription factors obtained from RegulonDB (http://regulondb.ccg.unam.mx). Sequences were designed with 40–50% AT/GC content.

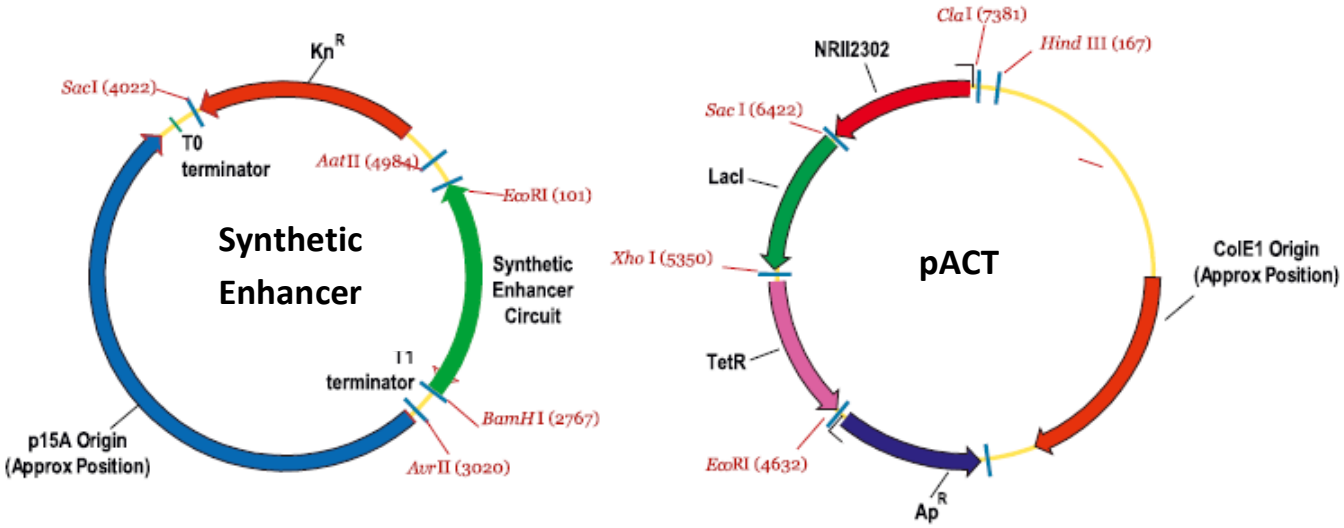


Figure 6. Plasmid schematic.

The left drawing corresponds to a schematic of the synthetic enhancer plasmid containing kanamycin resistance and low copy number p15A Origin of replication. The right drawing corresponds to a schematic of the pACT plasmid, which contains the NRII2302 mutant, LacI, and TetR or TraR genes. This plasmid has the high copy number ColE1 Origin of replication that provides high concentrations of the proteins encoded by those genes.

Transcription Factor	Binding site sequence
TraR 1	ACGTGCAGATCTGCACGT
TraR 2	ATGTGCAGATCTGCACAT
LacI _O1	AATTGTTATCCGCTCACAATT
LacI _Oid	AATTGTGAGCGCTCACAATT
Tet 1	TCCCTATCAGTGATAGAGA
Tet 2	ACTCTATCATTGATAGAGAT

Table 1. Sequences of transcription factors' binding sites used in the synthetic libraries.

Strain Construction

The synthetic enhancer strains were constructed as described by Amit et al. [59,60]. Briefly, *E. coli* strain 3.300LG with deletions for *glnL* and *glnG* genes was transformed with sequence-verified pACT and synthetic enhancer plasmids. The pACT family of plasmids was constructed by modifying p3Y15[58]. We inserted a *LacI* gene and either a *TetR* or *TraR* genes into the parent plasmid, under the control of the same *glnL* promoter controlling the NRII2302 mutant. The TetR sequence that we used is that of TetR-B[61], which we refer to as TetR. Selection was carried out via double Kan/Amp resistance (20 and 100 mg/ml, respectively). Candidate synthetic enhancer strains were tested for fluorescence in the presence and absence of the suitable inducer (see below) on a plate reader (Tecan, Infinite F200), to ensure that a proper strain was constructed.

Expression Level Ratio Measurement Assay

Expression level measurements for all synthetic enhancers without LacI-binding sites were carried out as follows: first, synthetic enhancer strains were grown in fresh Luria-Bertani with

appropriate antibiotics (Kan/Amp) to midlog range (OD600 of ~0.6) as measured by a spectrophotometer (Novaspec III, Amersham Biosciences) and were resuspended in low growth/low-autofluorescence bioassay buffer (for 1 liter: 0.5 g Tryptone (Bacto), 0.3 ml glycerol, 5.8 g NaCl, 50 ml of 1M $MgSO_4$, 1 ml 10xPBS buffer at pH7.4 and 950 ml double distilled water). Isopropyl b-D-1-thiogalactopyranoside (IPTG; 1mM) was added at this point, to deactivate the LacI protein that represses the glnAp2 promoter in the pACT plasmid. Two milliliters (ml) of resuspended culture with IPTG were dispensed into each well of a 48-well plate. Appropriate concentrations of anhydrotetracycline (aTc, Cayman Chemical 10009542) or N-(3-Oxooctanoyl)-L-homoserine lactone (3OC8, Sigma-Aldrich O1764) were dispensed into each well, spanning four to six orders of magnitude. Up to 24 levels of aTc or 3OC8 concentration were used for each strain. The plates were then incubated in a 37^{o}C shaker until cultures reached steady-state growth. Measurements of fluorescence levels were taken by dispensing 200 ml of culture into each well of a 96-well plate and were carried out on a plate reader (Tecan F200). Two wells were used as IPTG controls. We carried out each measurement in duplicate. All fluorescence measurements were divided by optical density, to measure the normalized expression, and autofluorescence levels (cells with no plasmid) were subtracted from the normalized values. Some TetR and TraR synthetic enhancers were also tested in strains lacking the LacI protein and no distinguishable difference in the regulatory response was observed.

To compute the expression level ratio $R_1(N,k)$ for a synthetic enhancer, we take the ratio in fluorescence expression levels between the protein-unbound regime to the protein bound case for each measurement of a synthetic enhancer's regulatory response curve. Typical regulatory-

response curves for individual synthetic enhancers are shown in Figure 11. All expression level ratio measurements in our experiments were obtained using such a procedure with the protein-unbound case regimes set at low 3OC8, high aTc and high IPTG for TraR, TetR and LacI, respectively.

LacI Experiment

Synthetic enhancer cassettes containing LacI-binding sites were cloned into a similar template plasmid, as other cassettes, containing glnAp2 promoter and two NRI-binding sites, except the LacI-binding site was removed from the glnAp2 promoter. The removal of the exogenous LacI sites were done to ensure that no tetramerization could take place between the site on the enhancer and these sites, thus affecting the regulatory outcome. The cassettes were cloned into the 3.300LG strain along with the pACT plasmid (expressing LacI). The experiment was carried out as described above using IPTG as an inducer for LacI removal from the synthetic-enhancer binding site.

Fusion Proteins Construction

Fusion proteins were designed by connecting the GST coding sequence to either the C or N terminus of TraR, N terminus of TetR and C terminus of LacI coding sequences. GST was PCR amplified from PGEX-4T1 and cloned into pACT, pACT-TetR or pAct-TraR plasmid using Gibson assembly [62]. His-GST-TetR fusion was created by cloning GST-TetR into PGEX-4T. Fusions were sequence verified and their expression was verified by Coomassie or western blot analysis.

Testing Functionality of Fusion Proteins

To test additional protein size-related predictions of our model, we constructed a small set of fusion proteins, where a GST epitope was added at either the N-terminus or C-terminus of each of the TFs used in the experiment. All proteins were tested for functionality before a larger-scale synthetic enhancer experiment was performed.

TraR Fusions

Fusion protein activity was tested by transforming cells with each of the pAct-Tra fusions as well as a synthetic enhancer bearing a TraR binding site. Both C-terminus and N-terminus fusions showed no effect upon 3OC8 induction. We concluded that the TraR-GST fusions lost their DNA binding ability and thus were not useful for our purposes.

LacI Fusions

We chose to construct a C-terminus GST fusion, since it was reported that N-terminus fusions disrupt DNA binding of LacI and are inactive[63]. The C-terminus LacI-GST fusion was sequence-verified, and its activity was tested in by transforming cells each of the pAct-LacI-GST fusions as well as a synthetic enhancer bearing a LacI binding site. In Figure 10G, we compare the dose response functions for LacI-GST and LacI, for the LacI synthetic enhancer with the binding site set at k = 95 bp. Both dose response functions show a specific interaction with the LacI binding site, with the transition to unbound state occurring at nearly the same concentration of IPTG. The

figure also shows that the expression level ratio is significantly smaller for LacI-GST as compared with LacI.

TetR Fusions

His-TetR and His-GST-TetR activity were tested by transforming cells with each of the pAct-TetR fusions as well as a synthetic enhancer bearing a TetR binding site. DNA binding was further demonstrated in a gel shift assay.

Electrophoresis Mobility Shift Assay (EMSA)

Electrophoretic mobility shift assay (EMSA) reactions were performed in binding buffer (10 mM Tris–HCl pH 7.4, 50 mM NaCl, 0.5 mM EDTA pH 8.0, 1 mM $MgCl_2$, 2 mM DTT and 4% glycerol). Binding reactions with DNA cassettes with a single TetR binding site containing either His-TetR or His-GST-TetR proteins were performed at room temperature over the course of 20min. As a control, we used a DNA cassette with a single TraR binding site, and aTc was added to the sample to a final concentration of 0.1 mM. Samples were analyzed using an Agilent 2200 TapeStation and High Sensitivity D1K ScreenTape or by loading samples onto 8% non-denaturing polyacrylamide gel at 80 V in TBE buffer containing 0.09 M Tris, 0.09 M boric acid and 5 mM $MgCl_2$. DNA run on PAGE was detected by ethidium bromide staining.

His-Tag Protein Purification

His-TetR and His-GST-TetR were sub-cloned into the pet28a vector for over-expression and transfected into *E. coli* BL21-Gold (DE3) cells. Following IPTG induction of expression, lysis of cells

by homogenization and clearing of the supernatant by centrifugation, the lysate was loaded on a Ni++-Sepharose 6 Fast Flow resin (GE Healthcare). The matrix was washed, eluted with A300 buffer [20 mM Tris (pH 7.5), 10% glycerol, 300mM KCl, 0.1mM EDTA] containing 0.2M imidazole. The eluted fractions were analyzed by SDS–PAGE and those containing His-TetR or His-GST-TetR were pooled and dialyzed against A100 buffer [20mM Tris (pH 7.5), 10% glycerol, 100mM KCl, 0.1 mM EDTA and 2mM DTT]. Purified proteins were verified by Coomassie staining. Fractions containing purified proteins were dialyzed against buffer containing 20 mM Tris‑HCl, pH 7.5, 150 mM NaCl, 1 mM DTT and stored in −80 °C in the presence of 50% glycerol.

Gibson Assembly Cloning

Gibson assembly was done by following the manufacturer's manual [62] using a home-made mix of the three enzyme components. Following Gibson assembly, products were directly transformed into competent cells.

Plasmid Purification

Plasmids were produced and purified using NucleoSpin Plasmid Easy Pure Kit (Macherey-Nagel) for plasmid DNA extraction and purification according to the manufacturer's protocol. In brief, bacteria transformed with the proper plasmids were grown over night (maximum 16 hr) in 5 ml of LB with 100µg/ml Ampicillin or 25µg/ml Kanamycin. Following centrifugation (Thermo Scientific, Heraeus Megafuge 16R, 5000 rpm for 5 min), purification continued according to the manufacturer's protocol. Following purification, DNA amounts were quantified using nano-drop (Thermo Scientific NanoDrop 2000 Spectrophotometer).

PCR Reaction and Product Purification:

All PCR reactions were done using Q5 enzyme (NEB), primers from IDT and dNTP mix (Takara). Following PCR, salts and proteins were removed using Wizard SV Gel and PCR Clean-Up system (Promega) for subsequent enzymatic reactions. The procedure was done according to the manufacturer's protocol.

DNA Extraction and Purification from Gel

DNA extraction and purification from gel was done according to the manufacturer's protocol of Wizard SV Gel and PCR Clean-Up system (Promega).

Crude Cell Extraction

Prior to PCR, crude cell extract containing bacterial genomic DNA was prepared by suspending a bacteria colony with 15ul of 0.1% Triton x-100. Then, the suspension was incubated in the PCR machine with the following program:

Step 1 – 99°C, 5 minutes

Step 2 – 30°C, 1 minute

Step 3 – 4°C

RNA Extraction

For total RNA isolation, an overnight culture was diluted 1:100 and grown to an OD600 of 0.6. 1.5 ml of cells were centrifuged, resuspended in Max Bacterial enhancer (Thermo) and incubated in 95^{o}C for 4 min to facilitate cell lysis. Following lysis, RNA was isolated using phenol-chloroform extraction protocol. After recovering aqueous phase, RNA was cleaned by means of ethanol precipitation. RNA integrity was determined by running 500 ng in 1% agarose gel. RNA samples were stored at -80^{0}C or immediately subjected to subsequent DNaseI reactions.

DNaseI Treatment

To avoid DNA contamination RNA (1200ng) was subjected to DNaseI treatment using the TURBO DNA-free kit for 30 minutes in 37°C (catalog number: Cat# AM1907, Ambion life technologies).

Reverse Transcription

To generate cDNA from total RNA, we used the High capacity cDNA reverse transcription kit (Cat# 4368814, Applied Biosystems by life technologies). Duplicates of 400ng DNA-free RNA were taken from each sample. The reaction was done accordingly to the manufacturer manual to a final volume of 20µl.

The PCR instrument was programmed as follows:

Step 1 – 25°C, 10 minutes

Step 2 – 37°C, 120 minutes

Step 3 – 85°C, 5 minutes

Step 4 – 4°C

Real-Time PCR

Primers pairs for *rpoN*, *pspG*, *pspF* and *pspA* and the normalizing gene *idnT* were chosen using the Primer Express software and BLASTed (NCBI) with respect to the *E. coli* K-12 sub-strain DH10B (taxid: 316385) genome (which is similar to TOP10) to avoid off-target amplicon. Real time (RT) PCR was performed using SYBR-mix (Applied Biosystems). 5-fold serial dilutions were measured for each inspected gene for the primer calibration curve, and DNase control samples were measured to exclude genomic contamination is samples. Data including the standard and amplification curves value were acquired by QuantStudio 12k flex Real-Time PCR system (Applied Biosystems). Three technical replicates were measured for each of the three biological replicates. A C_t threshold of 0.2 was chosen for all genes. Table 2 lists the analyzed genes in RT-PCR and their designed primer pairs.

Table 2 Primers used in RT-PCR reaction with their sequences

Gene	Primer sequence	Remarks
idnT	F - CTGTTTAGCGAAGAGGAGATGC R- ACAAACGGCGGCGATAGC	Normalizing gene
rpoN	F - CGTGAGTCGCTGTATCGTTGA R - CGGCCAGTACCATCGGTTT	Test deletion of σ^{54} in Δ*rpoN* strain
pspG	F-GATGGTCACCGGCGTTTC R - GCATACCGCCGAGGAACATA	Measure *pspG* levels in control and edited strains
pspA	F - CGAACGTCGTATTGACCAGATG R - CAAACTGATCGTCCAGCGATT	Control of another σ^{54} regulated gene
pspF	F - GAGCAGGTCAGCGCGAAA R - TGAATGTGCGGTTGGTATGC	Measure *pspF* levels in control and edited strains

CRISPR-Cas9 Genome Editing

The editing was based on the components of the lambda-RED system, for improving the frequency of recombination, using phage-derived recombinases[64,65]. For easier screening of positive colonies, we first created a "base" edit where we inserted a resistance gene (Amp^R) in the target genomic site. Next, we performed an additional edit of our desired genomic insertion, removing the initial Amp^R insertion, enabling us to screen for positive colonies of the second edit. The pTarget plasmid containing the donor DNA was constructed as follows. Genomic segments upstream and downstream to the edited region were amplified using PCR with a genomic template and with primers containing either EcoRI/BsaI (for upstream segments) or BsaI /HindIII restriction sites (for downstream segments). Edited segments were ordered as two complementary ssDNA oligos, with BsaI sites on both ends (IDT). Upstream, downstream, and edited segments were cut with EcoRI/BsaI, HindIII/BsaI, or BsaI only, respectively (NEB), and cleaned on a column (Promega) to remove ends. Upstream, downstream, and edited segments were then ligated into the EcoRI/HindIII-opened pTarget plasmid encoding the relevant guide RNA. Top10 cells containing the pCas9 plasmid (Figure 7) were made electro-competent, as follows. Cells were grown overnight (O/N) in LB at 30°C. The starter was diluted 1:100 into fresh LB containing 10 mM L-arabinose and grown to OD600 of 0.8, to induce recombinase expression. After transformation of pTarget via electroporation (200 ng DNA into 50 µl competent cells), cells were recovered for 2 hours in LB only at 30^0C, after which 10 mM L-arabinose was added and growth was continued O/N (14-16 hrs total). In the morning, 50 µl were seeded on Kan/Spec plates, and grown for 48 hours in 30^oC. pTarget-MT encoding an empty RNA guide was used as a

control for Cas9 activity (this transformation has visibly higher OD after O/N growth). Approximately 100 Kan^R+/$Spec^R$+ colonies were selected and grown O/N in 2 ml LB+Kan + 2 µl of 1M IPTG (1:1000), without Spec, in 48-well format, in 30°C. All 100 colonies were screened on Kan+Amp plates, Kan+Spec plates, and Kan plates, to confirm that editing occurred (change in Amp^R) and that the pTarget plasmid was cured (no $Spec^R$). For curing of pCas9 plasmid, Kan^R+/$Spec^R$-/correct Amp^R colonies were grown O/N in 37^0C with Amp (for the base edit) or without antibiotics (for the final edits), and isolated colonies were collected and tested to be Kan^R-/$Spec^R$-/correct Amp^R. To confirm edited clones, positive-screened colonies were subjected to colony-PCR for validation by length estimation. All PCR products were sequenced for further verification of the genomic edits.

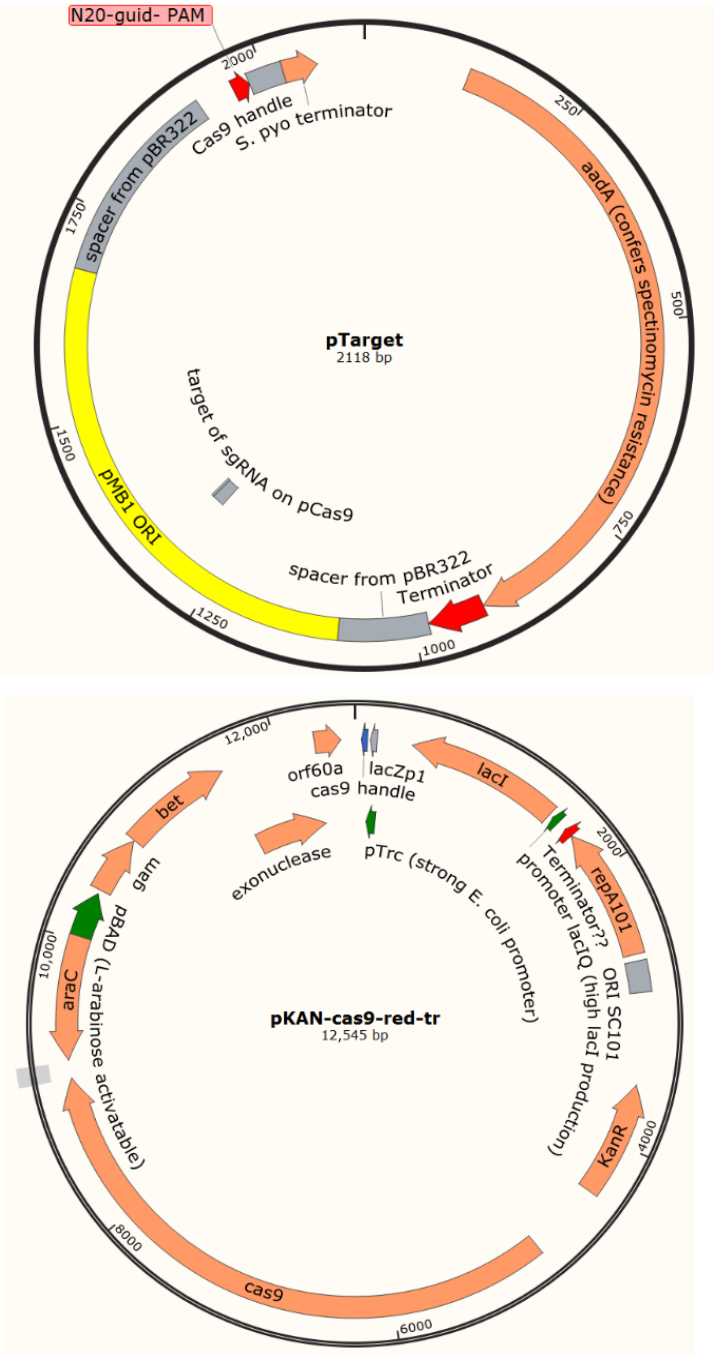


Figure 7. CRISPR-cas9 plasmids used for genomic editing.

pTarget plasmid encodes the 20 nt PAM-adjacent sequence upstream to the Cas9 handle, the donor DNA for the genomic edit, a target for Cas9 for plasmid curing, and Spec resistance. pKan-cas9 plasmid has three lambda red components exo, beta, and gam, under the regulation of an inducible arabinose promoter, and the CDS for Cas9 protein and Kan resistance. This plasmid also has an IPTG-inducible guide RNA targeting the ORI of pTarget for curing pTarget, and a heat-sensitive Ori for self-curing following the editing step.

Testing of Genomic Edits

To induce *pspG* and NRI/NRII expression, *their* coding sequence (CDS) were cloned under the expression of RhlR inducible promoter using restriction cloning. Competent WT Top10,

genomically-edited Top10 and Δ*rpoN* Top10 cells were transformed with the A133-RhlR-pspF or A133-RhlR-NRII/NRI plasmid. Cells were grown O/N in LB+AMP, diluted 1:100 -/+ 116 mM C4HSL (1:1000) to induce *pspF* or NRI/NRII expression, and grown to OD600 of 0.6. Next, cells were lysed in Max Bacterial Enhancer (Thermo) and subjected to RNA purification.

Bioinformatic Tools

Annotating the *qrr* σ^{54} Promoters

Given that it was shown that the *qrr2-4* in *V. cholerae* are activated by the NtrC-like phosphorylated version of LuxO (LuxO-P) [66,67], we hypothesized that other genes in other *Vibrio* species are also activated by LuxO from σ^{54} promoters. To test this, we first conducted a BLASTN[68] search of nucleotide databases using a conserved sequence of 32 nucleotides common to all known *qrr* genes (query: GGGTCACCTAKCCAACTGACGTTGTTAGTGAA) [67,69], and discarded all hits above an e-value threshold of 10^{-5}. We downloaded the full sequences of all remaining hits (169 hits from 86 unique sequences) and extracted the *qrr* gene and 500 bp upstream of the gene for each hit. We then discarded all sequences shorter than 105 bp and all identical 500 bp sequences (leaving one of each). The remaining set contained 114 hit sequences with unique 500 bp sequences upstream of the hit start position.

After identifying the relevant genes, we extracted the sequence that was located immediately upstream of the *qrr* genes in order to annotate the putative σ^{54} promoters. σ^{54} promoters are located within 30 bp of the transcriptional start site (TSS), and exhibit strongly conserved sequences, especially around the -24/-12 positions. We based our search on a consensus

sequence derived from[70], a large-scale analysis of 186 -24/-12 promoter elements from 47 different bacterial species. Interestingly, unlike σ^{70} promoters, the consensus sequence for the σ^{54} promoter is highly conserved across many bacterial species, which makes annotating these promoters a straightforward exercise. Consequently, to narrow our search, we examined the first 50 bp upstream of the BLASTN hit start position. Each position in the 50 bp window was scored with the likelihood of the promoter to start at its site based on 16-bp σ^{54} consensus matrix taken from[70], where we divided the non-consensus weight equally among the non-consensus nucleotides at each position. Scores were then normalized to a 0-1 range by subtracting the minimum possible score and dividing by the maximum possible score. Only promoters with normalized scores greater than 0.744 were accepted, based on the distribution of scores for 4000 randomly-sampled sequences. This allowed the identification of a single putative σ^{54} promoter for all 114 hits. We continued analysis for the 112 of these sequences that had unique 500 bp windows upstream of the promoter.

Extracting the LuxO Binding Sites

After identifying the promoters, we next proceeded to identify the LuxO-driver binding sites. Since LuxO is considered to be homologous to members of the NtrC family of response regulators[71], we hypothesized that as for the *V. cholerae* case, LuxO should drive expression from a tandem of binding sites that are spaced approximately two helical repeats (22 bp) apart. Finally, as with other NtrC-family members, we expected the tandems to be located several tens to a couple of hundred bp upstream of the putative σ^{54} promoter. To find the LuxO binding site tandems, we searched the 500 bp upstream of the respective putative σ^{54} promoters. These

upstream sequences were first scanned for 13 bp sequences that were candidate LuxO binding sites based on the only annotated LuxO sites in V. *cholerae*: TTGCATTTTGCAA and TTGCAATTTGCAA [66]. We chose a threshold of 9 of the 13 bases that had to match one of the two V. *cholerae* LuxO sites. We then computed the separation between the centers of any two candidate sites in the same upstream region. The center-to-center separation distribution we observe three distinct peaks: at 7 bp, 13 bp, and 21 bp. The 7 bp peak corresponds overlapping binding sites, which are highly probable since each binding site is almost an exact double repeat of TTGCAA, and thus this peak can be discounted. The peak at 13 bp is likely also an artifact due to the repetitive nature of the binding sites, with the first of the two sites overlapping the first TTGCAA of the real binding site, and the second binding site overlapping the second TTGCAA. This peak may additionally correspond to adjacent binding sites without separation, but these are not of interest due to the low likelihood of physical binding of proteins next to one another. Finally, the peak at 19-25 bp corresponds to the desired tandem separation, and these pairs of sites were chosen to be the putative LuxO binding site tandems. Some of the upstream sequences had more than one potential tandem. In these cases, we chose the tandem that maximized the following criterion: score(binding site 1) + score(binding site 2) - 1.01 x (spacing - 22), where score of a binding site is the number of base pairs matching the LuxO binding site (maximized over both LuxO binding sites), spacing is the distance between the binding sites in bp, and where the 1.01 factor gives a slight advantage to optimal spacing when everything is equal. Using this analysis, we found 61 of 112 sequences with putative LuxO binding tandems and unique putative loop sequences, where we define loop sequences to be all bases between the last base of the putative tandem binding site and the first base of the putative promoter. This also allowed us to determine

a putative consensus LuxO binding sequence to find bases at particular positions (depicted pictorially with the binding site logo graph in the inset of Figure 17A: T T G C A A/T T/A T T G C A A.

Yeast Growth Media

Yeast Extract Peptone-Dextrose (YPD): 1% Bacto Yeast Extract (Becton Dickinson), 2% Bacto Peptone (Becton Dickinson), 2% D-(+)-Glucose (Sigma-Aldrich).

Synthetic Defined (SD): 0.17% Difco Yeast Nitrogen Base w/o Amino Acids and Ammonium Sulfate (Becton Dickinson), 5% Ammonium Sulfate (Merck), 0.14% Yeast Synthetic Drop-out medium Supplements without histidine, leucine, tryptophan and uracil (Sigma-Aldrich). Carbon source-varies, depending on the experiment's purpose: 2% D-(+)-Glucose (Sigma-Aldrich), 2% or 1% D-(+)-Raffinose pentahydrate (Alfa Aesar). For inductive medium: 0.5% or 2% D-(+)-Galactose (Acros Organics).

For agar plates: 1.5% Bacto Agar (Becton Dickinson).

Amino acids: Added as supplements to the yeast growth media to a final concentration of: 20mg/L L-Histidine, 80mg/L L-Leucine and 20mg/L Uracil. All were purchased from Sigma-Aldrich.

Yeast Strain

The strain used in the experiments was the engineered S288C laboratory strain BY4741 (Genotype: MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0).

Growth Conditions

BY4741 strain was grown in appropriate medium plate in 30°C overnight. A 10 ml starter from one grown colony was made, the starter grew overnight before diluting the cells in a fresh liquid

medium, for inductive conditions a medium contained galactose was added to the cells. Cells were grown to OD600 of ~0.4-0.8 for yeast transformation, or OD600 ~0.6 for induction experiments.

Yeast transformation: cells were made competent and transformed using the Yeast frozen-EZ Yeast Transformation II Kit (Zymo) following the manufacturer's protocol. Briefly, cells were grown to OD600 ~0.6-0.8, washed by the supplied buffers and transformed via electroporation with 1ug of DNA. Cells were plated on appropriate plates and incubated in 30°C for at least two days.

Yeast Library

Synthetic enhancer cassettes were ordered as double-stranded DNA minigenes from Gen9 Inc. Each minigene ordered was ~500 bp long and contained the Gal1 UAS. The parts were clones into pUG34 vector, which harbors an AMP selection marker, *His3* CDS for selection and yeGFP reporter gene.

Chromosomal Integration into Yeast Genome

HIS3 chromosomal locus, located on chromosome XV, was chosen as the integration locus in the genome. PCR reactions were performed to amplify the cassette and HIS3 overlapping regions at the cassette's tails. Following PCR, Gibson assembly reaction was carried out into PstI digested pUC19 vector. All fragments were confirmed by sequencing. The fragments were amplified using PCR, resulting in linear fragments, which were individually transformed to WT component BY4714 yeast cells. Clones were screened by growth on plates without histidine. The colonies

were verified by colony PCR with appropriate primers specifically for the insert. Clones were confirmed by measurement of GFP *via* FACS.

Flow Cytometry

Cloned yeast cells were inoculated overnight (30°C, 250rpm) in SD media with 2% raffinose and appropriate amino acids supplementations. The day after, cells were diluted to OD600 0.6 and incubated in 30°C, 250rpm. Cells were centrifuged and resuspended with low growth medium and induced with 0.5% galactose for 4 hours. GFP expressing cells were analyzed using a BDTM LSR II flow cytometer with excitation at 488nm and 530/30 emission filter. Results were analyzed and graphs were generated using the FlowJo software and MATLAB.

Materials

Enzymes:

All enzymes (restriction enzymes, ligases and polymerases) were purchased from New England Biolabs (NEB).

Bacterial growth media:

Luria-Bertani (LB): 1%Bacto Tryptone (Becton Dickinson), 0.5% Bacto Yeast Extract (Becton Dickinson), 1%NaCl (Merck).

For agar plates: 1.5% Bacto Agar (Becton Dickinson).

Super Optimal Broth (SOB): 2% Bacto Tryptone (Becton Dickinson), 0.058% NaCl (Merck), 0.5% Bacto Yeast Extract (Becton Dickinson) and 0.019% Potassium Chloride (Merck).

For recovery after bacterial transformation, the following materials were added to SOB: 1% 1M $MgSO_4$ (Merck), 1% 1M $MgCl_2$ (Merck) and 2% 1M D-(+)-Glucose (Sigma-Aldrich).

Antibiotics: All antibiotics were purchased from Sigma-Aldrich.

Kits:

- NucleoSpin Plasmid Easy Pure Kit (Macherey-Nagel) for plasmid DNA extraction and purification.
- Wizard SV Gel and PCR Clean-Up system (Promega) for DNA purification from gels and in-vitro enzymatic reactions.
- Wizard Genomic DNA Purification Kit (Promega) for both bacterial and yeast genomic DNA isolation.
- Hylab Taq Ready Mix (2X) for bacterial colony PCR.
- Thermo-Fisher DreamTaq PCR Master Mix (2X) for yeast colony PCR.
- TURBO DNA-free Kit (Ambion by life technologies) for removal of DNA contamination after total RNA isolation from yeast cells.
- High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) for cDNA synthesis.
- Fast SYBR Green Master Mix Kit (Applied Biosystems) to perform real-time PCR.
- Frozen-EZ Yeast Transformation II Kit from Zymo Research.

Results

Part I: Studying Quenching Repression in Bacteria

Simulating Quenching Repression

As quenching is a phenomenon that is associated with proteins that seem to bind DNA several tens of base pairs away from an activator, we hypothesized that the underlying mechanism for repression might be an excluded volume effect, where a bound protein alters the propensity of DNA to form a loop by its mere presence. A collaboration was established with a physicist colleague in our lab, who opted to explore this hypothesis by devising a numerical simulation using the worm-like chain (WLC) model as a basis [72]. To do so, we modified the wormlike chain model to generate chains made of finite volume links. Such 'thick' chains can be used to probe excluded volume effects, as only configurations where parts of the chain do not cross each other are considered. In addition, thick chains can be 'deformed' locally by additional volumes or protrusions and using numerical simulations the effects of these local protrusions on various chain properties can be estimated. Since these protrusions can be likened to proteins bound to DNA, the model's results can be used to estimate the likelihood that a protrusion-bound looped configuration will occur. We term this approach the self-avoiding WLC model [73]. To obtain an initial set of predictions, we generated ensembles containing 10^7-10^9 configurations of thick chains with protrusions, up to a chain length of 300 links, with one link corresponding to 1 bp. To model quenching effects, we generated a thick chain architecture containing three protrusions: one at each end of the chain, simulating the activator and holoenzyme complex, and an additional protrusion simulating a generic transcription factor (TF) positioned somewhere along the chains

(Figure 8).For the predictions shown in this study, we chose a looping boundary condition that mimics the actual geometry of the bacterial s54 interaction with its upstream activator [74]. In our simulations, the TF binding site or protrusion is located some k links away from the 'activator' and N-k links away from the promoter. The looping probability ratio is defined by R_1 (N,k) and is calculated by the equation: $\hat{R}_1(N,k) = \frac{P_{\text{looped},1}(N,k)}{P_{\text{looped},0}(N)}$

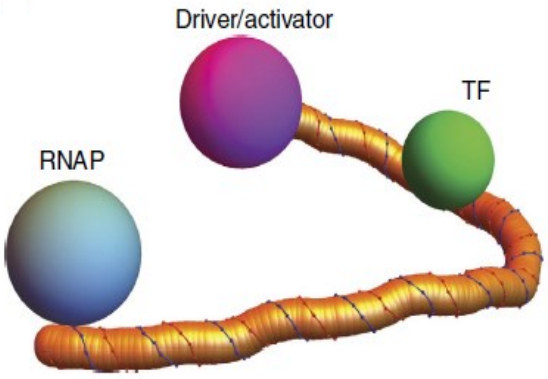


Figure 8. Modelling quenching effects.

Sample conformation of a thick chain with three protrusions generated using the self-avoiding WLC (SAWLC) algorithm.

Testing the Excluded-Model for Quenching Repression Experimentally

We explore the transcriptional regulatory behavior associated with enhancers using a synthetic approach which allows us to design synthetic enhancers *de novo*. Our experiments utilize a biological circuit [28] based on bacterial NRI/NRII (NtrC/NtrB) two-component system (see Figure 9). We hypothesized that it is possible to describe an enhancer via a set of "footprints" or control parameters that can be read from the enhancer's sequence. By varying these parameters, we believe that we can affect transcriptional activity in a systematic and tractable fashion.

Among these parameters are:

- The number of TF binding sites.
- The spacing between the TF binding sites
- The length of the sequence that bridges the σ^{54} promoter and the enhancer (looping length).
- The location of TF binding sites relative to the enhancer.

The synthetic enhancer we use has been previously used to construct synthetic enhancers and genetic oscillators [28,75] and is based on the bacterial NRI/NRII (NtrC/NtrB) two-component system. This system controls nitrogen assimilation in many prokaryotes. Since my first set of experiments is based on this system, I will briefly describe it. NRI and NRII are the gene products of *glnG* and *glnL* respectively. *glnG* is activated first by a σ^{70} promoter glnAp1, which overlaps the NRI #1, 2 sites. This promoter keeps a low basal level of the protein product NRI available to the cell. Phosphorylation of NRI by NRII is crucial for its DNA binding. Since NRII can function both as a phosphatase and kinase, the expression levels of endogenous NRI are therefore tightly coupled to a complex signaling pathway. To decouple the circuit from the signaling pathway, we used a deficient mutant NRII2302 in a 3.300 *E. coli* strain with deletions of the endogenous *glnL* and *glnG* genes (3.300LG [58]) to ensure that NRI remains phosphorylated at all times [17]. To activate the σ^{54} promoter, the cell must accumulate a sufficient amount of phosphorylated NRI proteins in order to assemble a hexamer on the DNA, which serves as the driver for the reaction. Once a hexamer is formed, it interacts with the σ^{54} promoter glnAp2 by looping, generating more NRI and

providing positive feedback within the circuit. Downstream to NRI an mCherry reporter gene is transcribed, allowing to measure the transcriptional activity of the glnAp2 promoter.

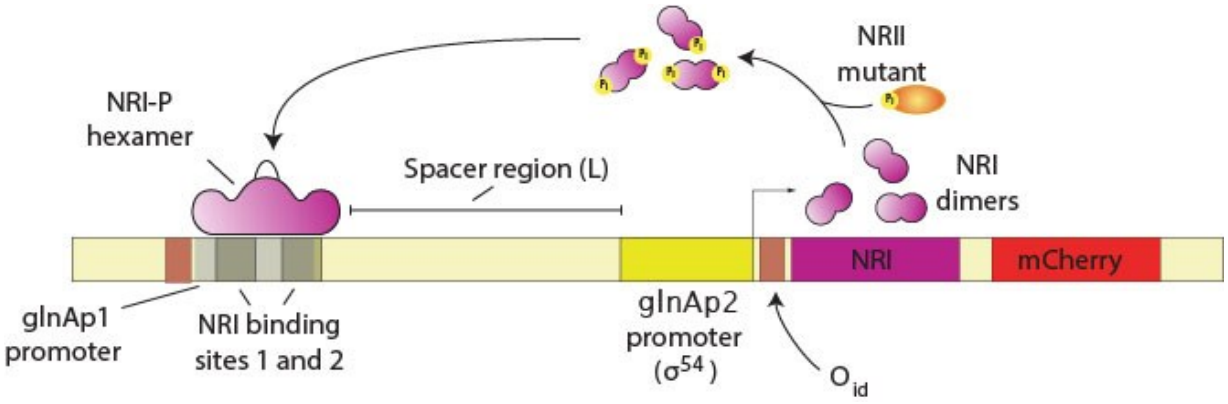


Figure 9. Schematic for the basic enhancer circuit.

The circuit expresses via a σ^{54} promoter the glnG (ntrC) gene, whose protein product (NRI) remains phosphorylated at all times via the action of the phosphatase-deficient mutant NRII2302, which also serves to decouple the NRI/NRII system from the nitrogen assimilation pathway. The synthetic enhancer circuit was transformed into a ΔglnL: ΔglnG:3.300 E. coli strain (3.300LG) on a low-copy plasmid (~10/cell). Picture from [50] .

We start with a minimal enhancer made of driver binding-sites and a poised promoter region (Figure 10), and progressively increase the synthetic enhancer's complexity. This is done by addition of discrete sets of defined binding sites for transcription factors within the enhancer that can alter the looping probability. These transcription factors are not thought to interact directly with either the driver protein or the poised RNA polymerase.

To compute the expression level ratio $R_1(N,k)$ for a synthetic enhancer as a function of N (looping length and k (distance from the activator), we take the ratio in fluorescence expression levels between the protein-unbound regime to the protein-bound case for each measurement of a synthetic enhancers' regulatory response curve, as can be seen in the following equation:

$$\text{Expression ratio} = \frac{\text{Fluorescence with bound TF}}{\text{Fluorescence with NO TF bound}} = \frac{\text{Fluorescence with max 3OC8}}{\text{Fluorescence with 0 3OC8}}$$

Typical regulatory-response curves for individual synthetic enhancers are shown in Figure 11. All expression level ratio measurements in our experiments were obtained using the equation above, with the protein-unbound regimes correspond to low 3OC8, high aTc and high IPTG for TraR, TetR and LacI, respectively.

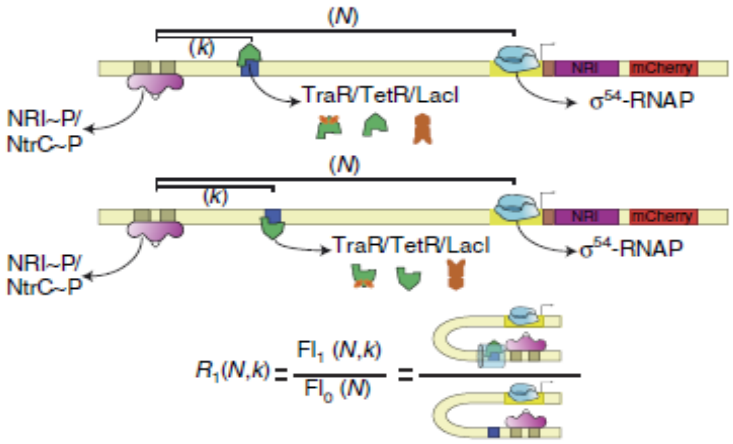


Figure 10. Schematic for the minimal bacterial enhancer.

The system used in our experiments, showing the poised holoenzyme complex at the σ^{54} promoter, NtrC activator and the additional binding site for either TraR, TetR or LacI. The schemas represent the two 'extreme' configurations. Top: binding sites are positioned 'out-of-phase' relative to the activator. Bottom: binding sites are positioned 'in-phase' relative to the activator. The schematic for TraR is drawn with two ovals corresponding to the 3OC8 ligand.

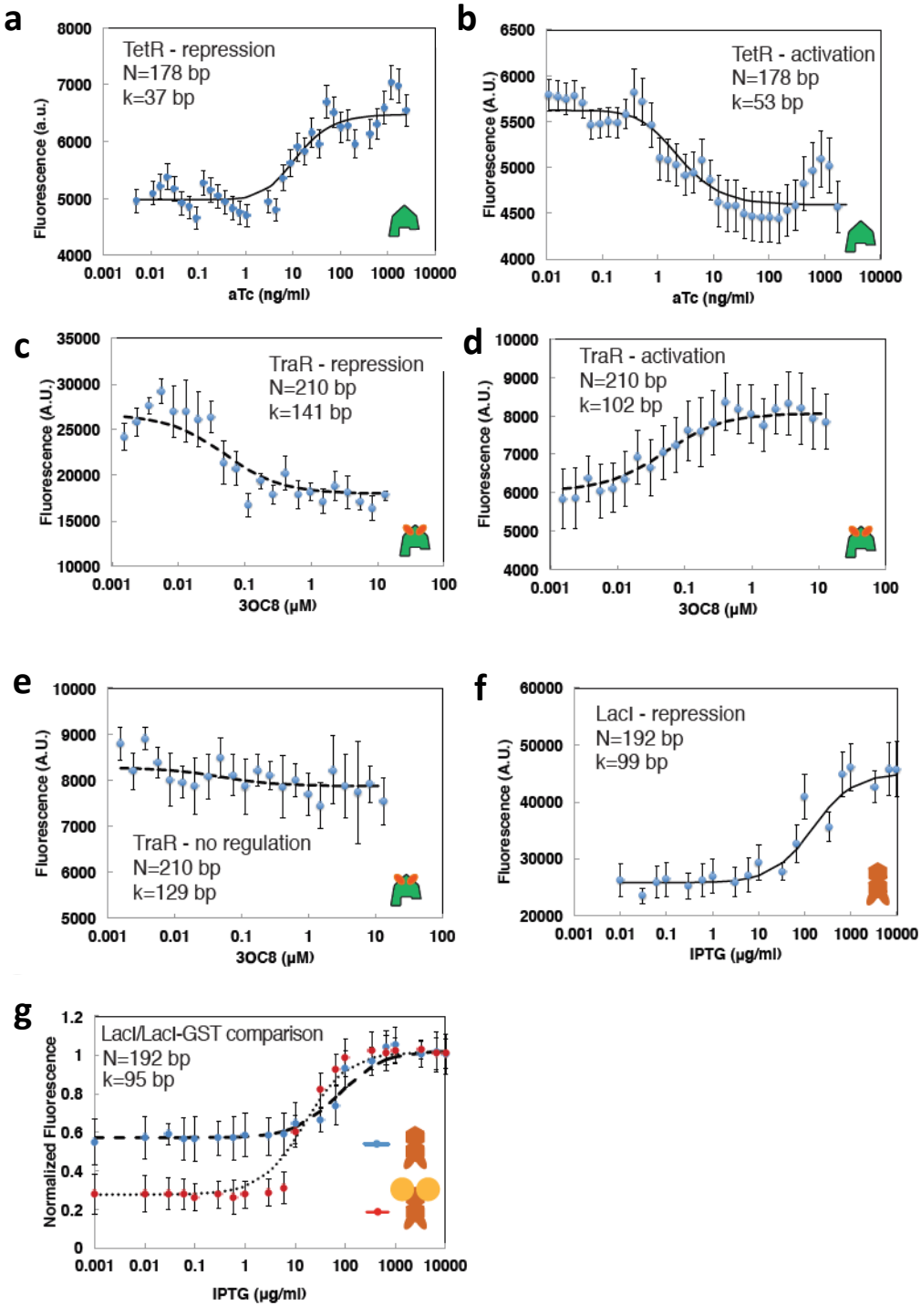

a
TetR - repression
N=178 bp
k=37 bp
Fluorescence (a.u.)
aTc (ng/ml)
b
TetR - activation
N=178 bp
k=53 bp
Fluorescence (A.U.)
aTc (ng/ml)
c
TraR - repression
N=210 bp
k=141 bp
Fluorescence (A.U.)
3OC8 (µM)
d
TraR - activation
N=210 bp
k=102 bp
Fluorescence (A.U.)
3OC8 (µM)
e
TraR - no regulation
N=210 bp
k=129 bp
Fluorescence (A.U.)
3OC8 (µM)
f
LacI - repression
N=192 bp
k=99 bp
Fluorescence (A.U.)
IPTG (µg/ml)
g
LacI/LacI-GST comparison
N=192 bp
k=95 bp
Normalized Fluorescence
IPTG (µg/ml)

Figure 11. Expression level ratio measurements.

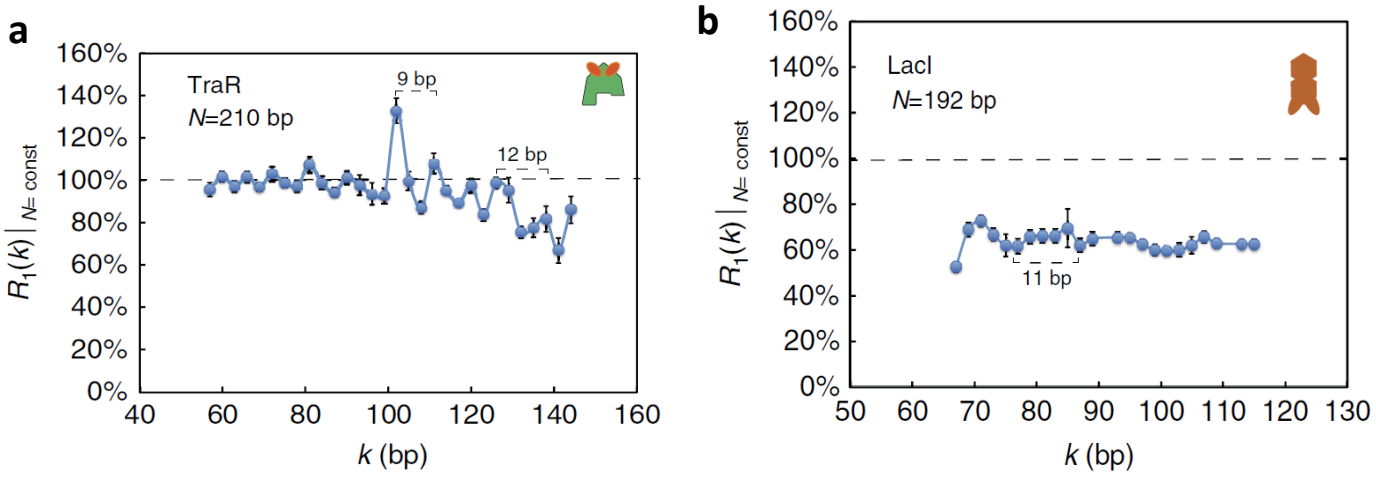
Six sample data sets showing fluorescence measurements that are subsequently used for the experimental determination of the expression level ratio. The panels plot inducer concentration on the x-axis and raw fluorescence measurements on the y-axis for the synthetic enhancers with looping length N with a single TF binding site displaced k bp away from the driver. All datasets were fit by Hill-1 functions to determine expression level ratio values with error-bars. (a) TetR repression. (b) TetR-activation. (c) Tra-repression. (d) TraR-activation. (e) TraR - no regulation. (f) LacI-repression. The expression level ratio values determined for these data sets were: (a) 77% ± 5%, (b) 120% ± 5%, (c) 67% ± 6%, (d) 133% ± 3%, (e) 95% ± 5% and (f) 57% ± 3%. Note that increasing inducer levels has different effects for the different proteins: IPTG and aTc remove LacI and TetR, respectively, from DNA, while 3OC8 enables binding of TraR to DNA. (g) Comparison of dose response functions for LacI and LacI-GST for the single LacI binding site synthetic enhancer set at k=95 bp. LacI-bound synthetic enhancers (blue) and LacI-GST-bound synthetic enhancers (red).

Synthetic Enhancers with a Single Binding Site

Our first library design consisted of 81 regulatory sequences containing a single binding site for one of three different TFs: LacI [76], TetR [77] and TraR [78]. The choice of LacI, TetR and TraR was predicated on their ability to bind DNA dependent on the absence (LacI and TetR) or presence (TraR) of a ligand, whose concentration we controlled externally.

In addition, we varied two control parameters: the looping length N, defined as the distance from the center of the NtrC/NRI activator-binding site tandem to the center of the σ^{54} promoter, and the distance k separating the center of the transcription factor-binding site from the middle of the NtrC/NRI-binding site tandem. We limited our enhancers to looping lengths N>150 bp but did or at some intermediate position.

In Figure 12, we plot expression-level ratio $R_1(N,k)$ for constant looping length (N) as a function of binding-site position (k) , for synthetic enhancers containing a single binding site for the three transcription factors: TraR, LacI and TetR (Figure 12 a-c). In addition, for all three data sets the position of the binding site (k) was varied across the looping region at 1-3 bp intervals. For the TetR and TraR cases, we observe a long-range oscillatory function in the expression-level ratio between quenching and upregulation with a period ~10-11 bp, which is consistent with the accepted value for the DNA helical repeat (~10.5-10.9 bp [79–82]). Here, we define quenching as expression-level ratio values that are <100%, as the protein-bound case yields a lower total mCherry reporter level than the unbound case. Conversely, we define upregulation as the case in which the expression-level ratio is >100%. Interestingly, maximal quenching seems to occur at k-values that are roughly integer multiples of the DNA helical repeat, whereas binding-site positions that are displaced 5-6 bp away from the minimas resulted in either weaker quenching repression or slight upregulation (>100%) of the expression level of the bound enhancer with respect to the unbound case.



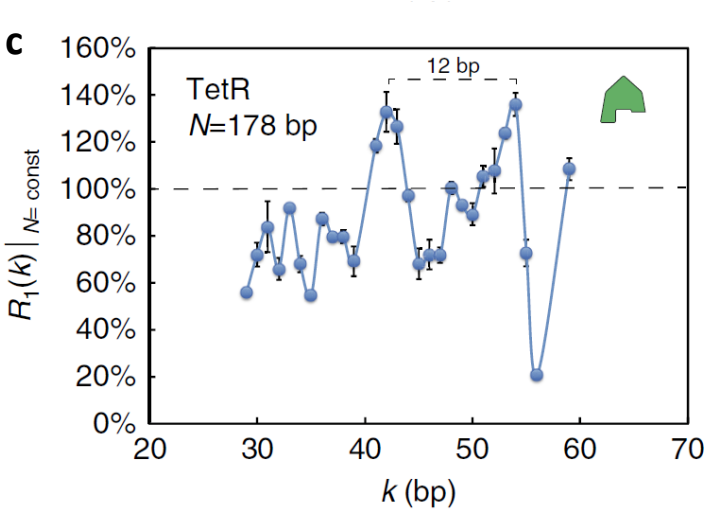


Figure 12 . Expression level ratio results for synthetic enhancers with a single binding site for TraR (a), LacI (b) and TetR (c) at constant N.

k was varied in 3, 2 and 1 bp steps for TraR, LacI and TetR, respectively. The expression-level ratio observable as defined here is approximately equal to the probability-of-looping ratio given known rates for the NtrC- σ^{54} system, thus allowing us to quantitatively compare experimental results to theoretical predictions. Error bars correspond to the s.d. from multiple measurements.

A closer examination of the regulatory response curves shows distinct differences that are strongly dependent on the TF type. For TraR synthetic enhancers (Figure 12a), the effect of the transcription factor on the probability of looping is small and the total regulatory effect observed varies between weak quenching to slight upregulation (~75‑120%). For LacI synthetic enhancers (Figure 12b), a barely detectible oscillatory behavior with ~11 bp periodicity is observed. Here, the small-amplitude oscillations vary between intermediate (~50%) to weak quenching (~70%). Moreover, the amplitude of the oscillations seems to diminish as k increases, settling on an intermediate quenching level of ~60%. Finally, a third distinct regulatory response is observed for TetR in Figure 12c. Here, the regulatory effects persist for the entire segment of the loop tested and both significant quenching and upregulation effects are observed (20‑140%). Thus, although the oscillatory quenching/upregulation phenomenon observed clearly for two of the three

proteins (TraR and TetR) supports the excluded volume looping-based regulatory mechanism, the differences in the expression-level ratio responses suggest that additional protein-specific aspects need to be added to the model, to better explain the data.

Additivity in Synthetic Enhancers

We reasoned that the simplest protein-specific parameter that can affect the regulatory response is the total volume in the loop. Our single-binding-site data suggested that this parameter is indeed relevant, as both TetR (25 kDa) and TraR (26 kDa) are significantly smaller than LacI (38 kDa). To try to compensate for the mass and volume difference, we hypothesized that an increase in the number of binding sites for a smaller protein such as TraR should lead to a larger cumulative quenching effect, which would be comparable to the maximal affect achieved by LacI. To do this, we constructed a second synthetic enhancer library with two TraR-binding sites. To determine the optimal binding site arrangement for quenching, we first scanned the expression-level ratio values for a set of tandem TraR synthetic enhancers characterized by simultaneously varying values of the inter-site spacing s and the looping length N at 1 bp increments, while keeping k constant. We plot the results for the expression-level ratio as a function of the spacing s in Figure 13a. The figure shows an oscillating function with a significantly stronger maximal quenching response (~30%) observed for s=23 than the one observed for the single TraR-binding site (~70%) and slightly larger maximal quenching than the response obtained for the single LacI-binding site. The figure also shows expression-level ratio minima at inter-site spacing values that are integer multiples of the helical repeat (that is, s=23‑24, 34‑35 and 44‑45), whereas no quenching is

observed for odd half-integer multiples of the spacing, as predicted by the model. Interestingly, the values of the expression-level ratio minima and maxima shift to higher values (from 30 to 60% and 80 to 100%, respectively) as the number of helical repeats between binding sites increases from two to three, to four. In Figure 13b, we plot the model's predictions for the effects of inter-protrusion spacing (s) on the probability of looping. As in the experiment, the cross-section was taken for values of s and N that varied together by one-link increments for each successive point, whereas the protrusion position (k) was kept constant at four helical repeats (42 links). Remarkably, the simulation exhibits not only oscillations in the probability ratio levels, as expected, but also an overall upward shift in the values of the probability ratio minima and maxima, in close agreement with the experimental data.

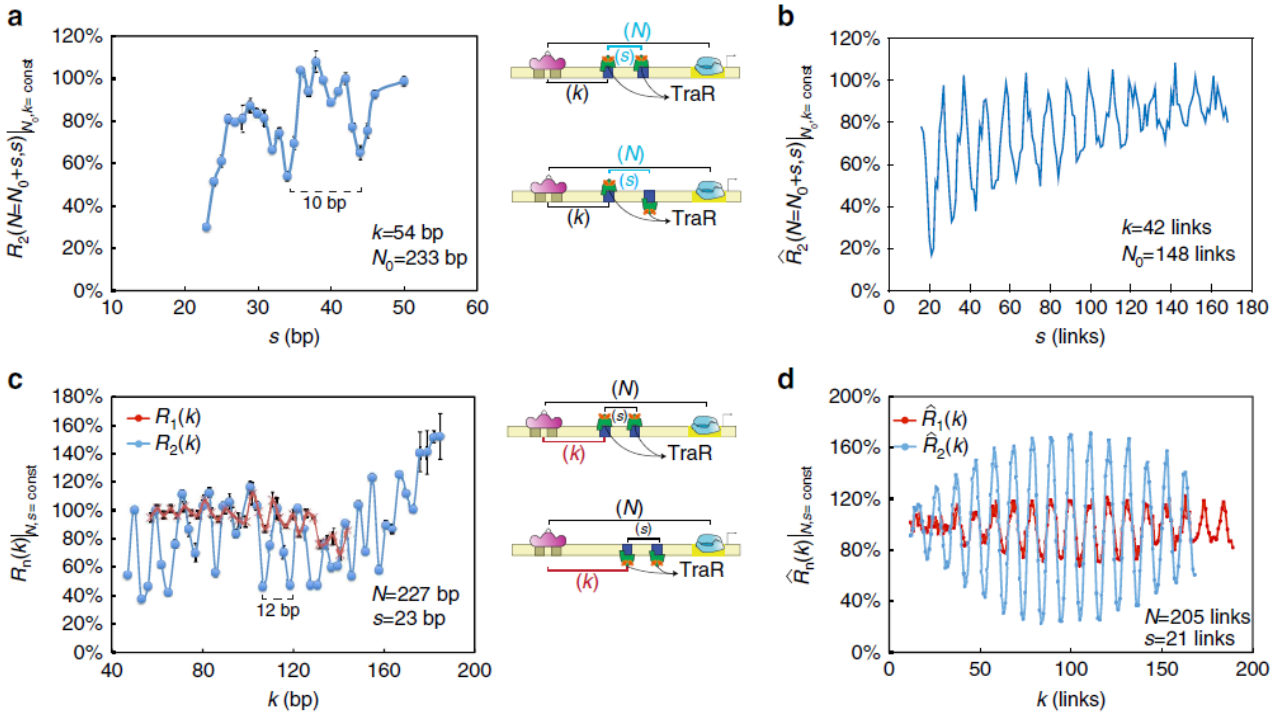


***Figure 13. Excluded volume is additive**.*
(a) Expression level ratio for synthetic enhancers with a tandem of TraR-binding sites at various inter-TF spacing (s) and looping lengths (N); see schematic on the right. (b) Model predictions for looping level probability ratio as a function of intra-binding site spacing s and looping length N. (c) Expression ratio measurements for synthetic enhancers with a tandem of TraR-binding sites (blue) that are in-phase with fixed inter-TF spacing (s=23 bp); see schematic on the right. The single TraR-binding site synthetic enhancer data are overlayed as reference in red. (d) Model predictions for looping probability ratio as a function of distance to the chain origin (k) for the following configurations: blue, in phase tandem protrusions (s=21 bp); red, single protrusion. Error bars correspond to the s.d. from multiple measurements.

To further validate the volume additivity prediction, our second library also included synthetic enhancers with a tandem of TraR binding characterized by a fixed inter-site spacing (s=23 bp, in phase), while the placement of the proximal TraR-binding site from the (k) was varied. In Figure 13c, we plot the expression-level ratio results (blue circles). The data stably oscillate from a strong quenching value of ~40% to no-quenching or slight upregulation values of 100‑110%. Although this behavior persists for binding site positions that are spread over 100 bps, as k increases further so that the location of the tandem binding sites approaches the promoter, the amplitude of the oscillations diminishes and a clear bias towards upregulation emerges, with a maximal upregulation value of 160% observed for k=189 bp. The oscillatory pattern is highly repetitive with a periodicity of 10.5±0.3 bp, an expression-level ratio amplitude that is approximately twice as large as for the synthetic enhancer with a single TraR-binding site (Figure 13c, red Xs) and persists for nearly the entire looping length (~227 bp) with little dependence on the position of the first binding site. Comparing the experimental data with one-dimensional cross-sections of the modelling results (Figure 13d) for thick chains with a single protrusion (red line) and a tandem of in-phase protrusions (s=21 links, blue line), the model captures the experimental trends nicely.

Here, the in-phase tandems exhibit a probability ratio response, which is characterized by a significantly larger amplitude of oscillations, as compared with the chain containing a single protrusion. In addition, the in-phase tandems exhibit oscillations whose amplitude first increases as k varies from small values, reaches a maximum at k~N/2, decreases for k>N/2 and increases again at k~N, which agrees with similar trends observed experimentally.

We carried out additional measurements for tandem TraR synthetic enhancers, keeping the binding site spacing (s) and distance to the driver (k) constant while varying the looping length (N), as follows: in-phase - s=23 bp spacing and k=54 bp, and out-of-phase - s = 28 bp and k = 54 bp. The expression level ratio data and probability ratio model predictions are plotted in Figure 14a and 14c, respectively, with a schematic of the two synthetic enhancer designs shown in Figure 14b. Figure 14a shows that both experimental data sets exhibit similar trends. The in-phase synthetic enhancers (s = 23 bp, blue dots) exhibit strong repression at small looping lengths (~30%), and for larger looping lengths a fluctuating regulatory response that converges on a repression value of 80% independent of the looping length. Despite the strong fluctuations, we also seem to be observing a faint periodic signature characterized by 4 cycles with a periodicity of ~11 bp centered on peaks in expression level ratio for looping lengths N = 148, 160, 171, 181, 192. For the out-of-phase data set (s = 28 bp, red squares), the expression level ratio shows a similar trend, which is characterized by a weaker over-all regulatory response. The repression at short looping lengths is significantly weaker (70%) as compared with the in-phase case (30%), the convergence to a constant repression value seems to occur at a shorter looping length (N = 120 bp vs N = 140 bp), and the overall long looping length regulatory response seems to hover around

the barely detectible expression level ratio value of 95%. Thus, the out-of-phase arrangement of the TraR binding site produces an expression level ratio behavior which is weaker than the in-phase arrangement, and overall is distinguishable from the in-phase expression level ratio response.

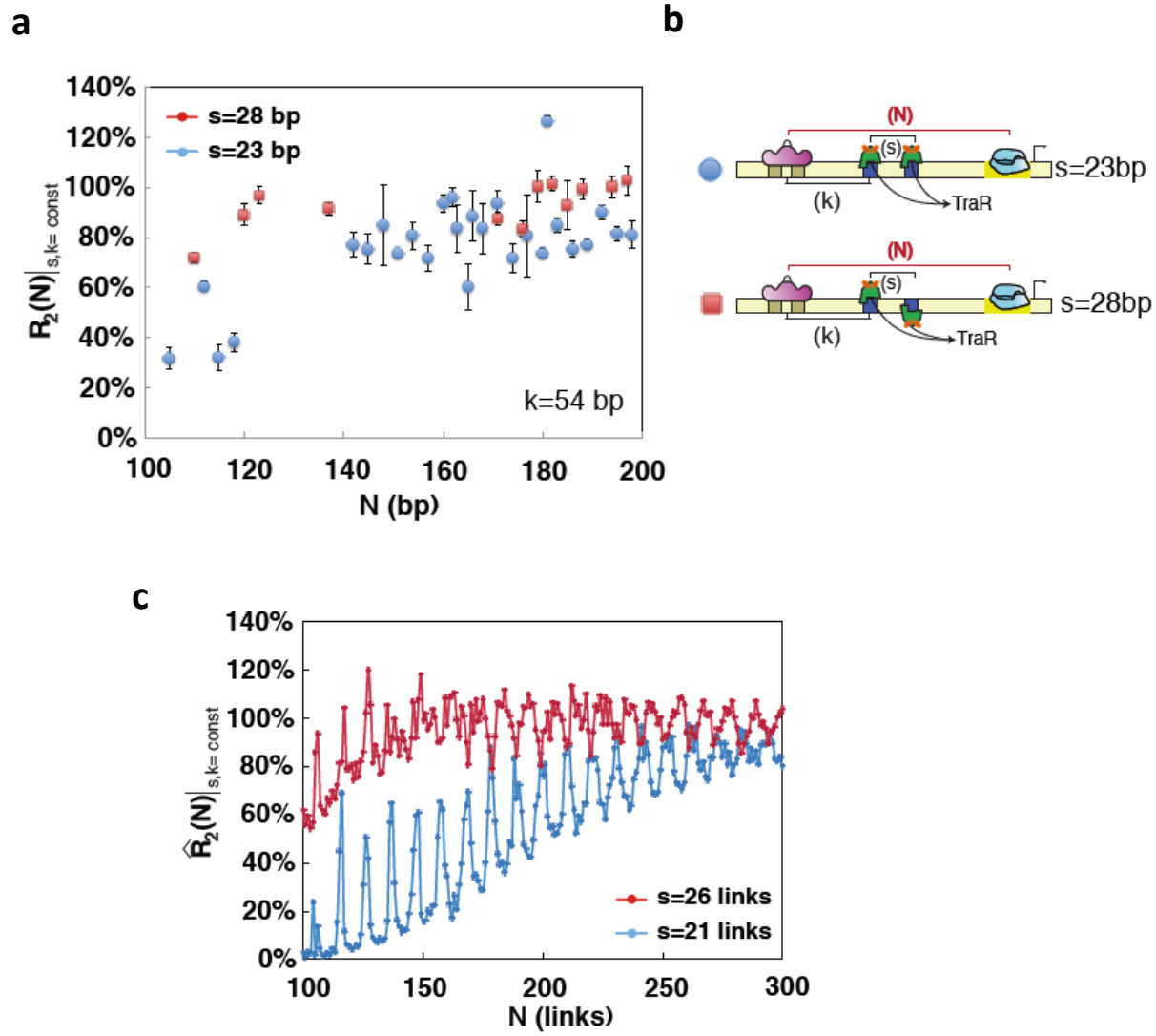


Figure 14. Looping length variation: measurement and model.
(a) Data showing the expression level ratio for a tandem of TraR binding sites with spacing s = 23 bp (blue) and s =28 bp (red) with constant location k of the first binding site. The looping length N is varied in 2-3 bp jumps. (b) Schematic of the two constructs showing the s = 23 in-phase orientation (top), and s = 28 bp out-of-phase orientation (bottom). (c) 1D contour maps of (N, s, k) probability ratio level space taken at k = 53 links, with s = 21 links (blue) and s = 26 links (red).

Stiffening and Bending Effects

An anomaly observed in the experimental data for the tandem-TraR and LacI synthetic enhancers as compared with model predictions is the lack of significant upregulation in the former (except near the promoter for large values of k) and a complete absence thereof in the latter. We hypothesized that a bias towards quenching can emerge if the transcription factors also 'stiffens' the DNA, making it slightly harder to bend locally. Based on our model, we expect such a bias to be dependent on the loop length (diminishing quickly for large N), the extent of the stiffened region (that is, the number of stiffening binding sites) and the binding sites' proximity to the center of the loop (as predicted by the model). To experimentally test the validity of the stiffening-excluded volume-looping regulatory model and provide further support for the additivity finding, we fused the 25-kDa glutathione S-transferase (GST) domain to the carboxy terminus of LacI to make a new 63 kDa TF: LacI-GST. This allowed us to generate a significantly larger LacI (see Figure 15a top schema), while not affecting its capacity to bind DNA. As both LacI (38 kDa, Figure 12b) and a tandem of TraR proteins bound to two binding sites (2x25 kDa, Figure 13c) seem to stiffen DNA as compared with a single bound TraR (25 kDa), we reasoned that a larger transcription factor may also add to the stiffening effect. Thus, according to the excluded volume portion of the model, the larger protein should generate a larger amplitude of oscillations between the regulatory minima and maxima, while the stiffening effect should shift the mean regulatory levels of these oscillations towards quenching. To quantify the regulatory effect induced by LacI-GST when bound to the synthetic enhancers and compare with the effect generated by the native LacI, we measured the expression-level ratio for LacI-GST on the same

synthetic enhancer library as the one used for LacI. The data are plotted in Figure 14a. The figure shows that the expression-level ratio for LacI-GST (blue) exhibits an oscillatory function that varies from very strong quenching values (25–30%) to intermediate quenching (40–50%). The oscillations exhibit the 10.5-bp periodicity observed for the TraR tandems and the overall extent of the expression-level ratio indicates that LacI-GST generates significantly stronger mean quenching response (38%±7%) than the native LacI (64%±4%, red). Moreover, the amplitude of the oscillations that are observed for LacI-GST-bound synthetic enhancers (19%±4%) is approximately twice as large as the amplitude exhibited by the LacI synthetic enhancers (9%±5%). Finally, the LacI-GST expression level ratio oscillatory function is phase flipped. Namely, the peaks of the LacI-GST data set appear at the minima of the LacI data set, and vice versa. In Figure 15b we show that by increasing the stiffness parameter (blue versus red line) the probability ratio gains an additional bias towards quenching. In Figure 15c we show that for the same stiffness value, increasing the size of the protrusion by a small amount (blue versus red line) leads to an increase in the amplitude of the oscillations, as expected. However, to account for the phase flipping, another mechanism is needed. One possibility is bending, in which the transcription factor also 'bends' the DNA locally. We find using the model that when the bending protrusion is inside the loop, the probability of looping is upregulated, whereas when the protrusion is outside of the loop the probability of looping is reduced. Thus, to account for the phase flipping observed for LacI-GST as compared with LacI, we plot (Figure 15d) the looping probability ratio for two scenarios. In the first, we simulate a thick chain with a small protrusion inside the loop that also bends the chain by 10^0 (red line). In the second, we simulate a thick chain with a 3x larger protrusion positioned inside the loop, which bends the DNA by the same amount (blue line). The

data show that for the thick chain with a smaller bending protrusion, the oscillations are consistent with a dominant bending effect generating an upregulation prediction for in-phase locations k. However, for the thick chain with the larger protrusion, the oscillations are consistent with a dominant excluded volume and stiffening effect generating a phase-flipped signal, which is similar to the one observed when comparing the LacI and LacI-GST synthetic enhancers.

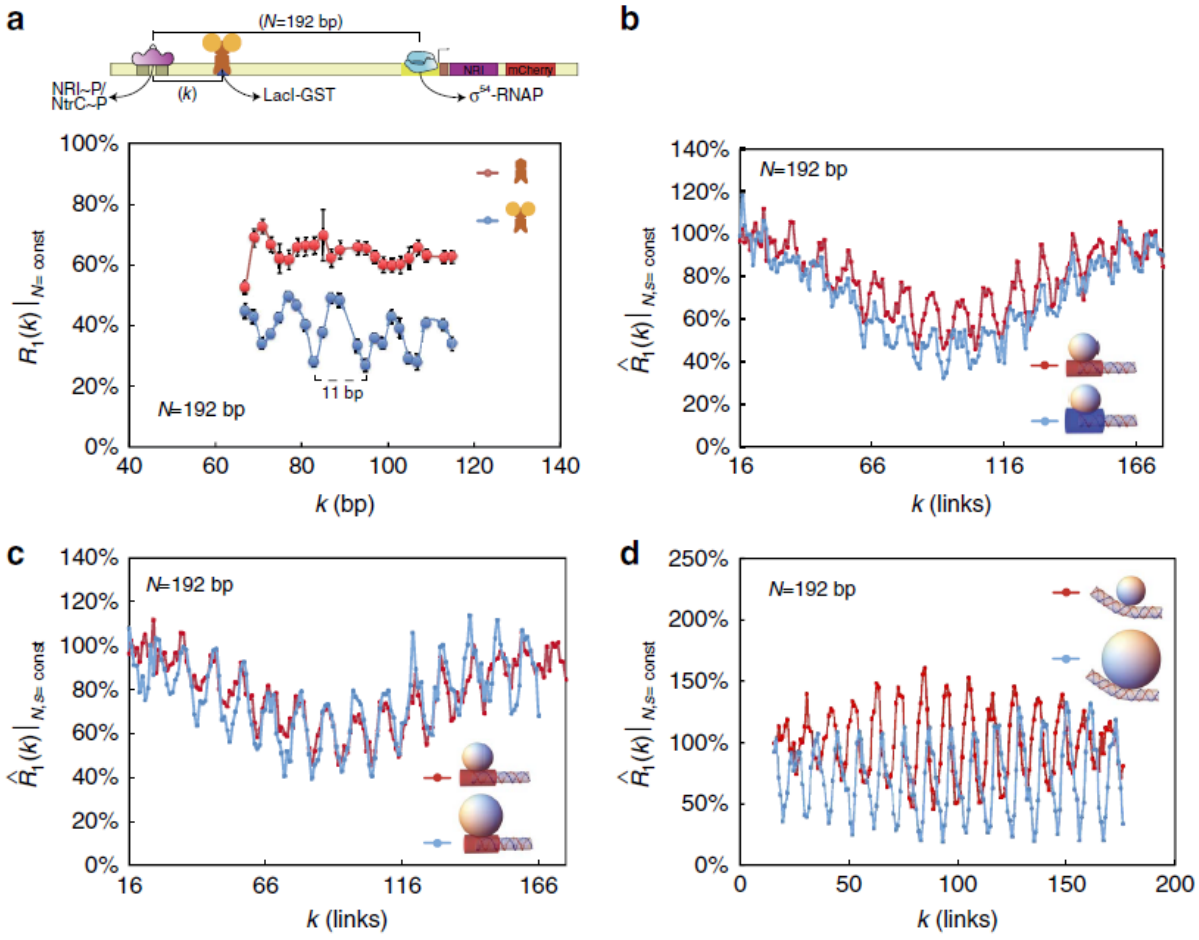


Figure 15. Combined elastic and entropic effects on looping.

(a) Expression-level ratio measurements for the synthetic enhancers with a single LacI binding site. Blue and red circles correspond LacI-GST and LacI expression-level ratios, respectively. (b–d) One-dimensional cross- sections for N=192 links comparing: (b) two values of stiffening with constant protrusion volume (blue) 2b and (red) 1.5b, where b is the DNA's

Kuhn length (106 nm). (c) Two protrusion volumes (blue, 8.16 nm; red, 5.44 nm) at constant stiffening (1.5b). (d) Two protrusion volumes that can also bend the thick chain by 10^0 (blue, 8.16 nm; red, 5.44 nm). Error bars correspond to the s.d. from multiple measurements.

Deciphering Higher-Order TF-Binding Configurations

We next hypothesized that a pair of binding sites arranged in an out-of-phase configuration should generate a regulatory response with a periodicity that differs from the in-phase-tandems or single-binding-site cases. In Figure 16a we plot model predictions for the regulatory output for an out-of-phase 'Z-shape' binding configuration showing that 5‑6 bp oscillations are generated with a pattern of alternating strong and weak maxima/minima. This periodicity is a result of the fact that a 180^0 rotation of the Z-shaped tandem around the thick chain axis yields a similar configuration. The deviation between these two configurations is responsible for the alternating extrema. Although the overall regulatory effect predicted by the model is relatively small for such a configuration, for larger TFs a detectable signature may be observed. We constructed a final synthetic enhancer library to test the 5‑6 bp periodicity and alternating weak/strong extrema predictions for synthetic enhancers bound by Z-shaped TF structures. First, we characterized 18 synthetic enhancers with tandems of LacI-binding sites whose center-to-center spacing was set at 38 bp. Such an arrangement not only places the LacI dimers in opposite orientation, but also strongly restricts their ability to tetramerize. The binding sites' positions inside the loop were shifted together in 2 bp increments, thus covering a range of 36 bp of intra-loop positions. In Figure 16b (red) we plot the results. The data show that the tandem-LacI synthetic enhancer strains exhibit a fluctuating regulatory response with a distinct 4‑6 bp periodicity for the majority

of intra-loop positions of the tandems and a slight increase in the overall magnitude of quenching as the binding sites are moved towards the center of the loop.

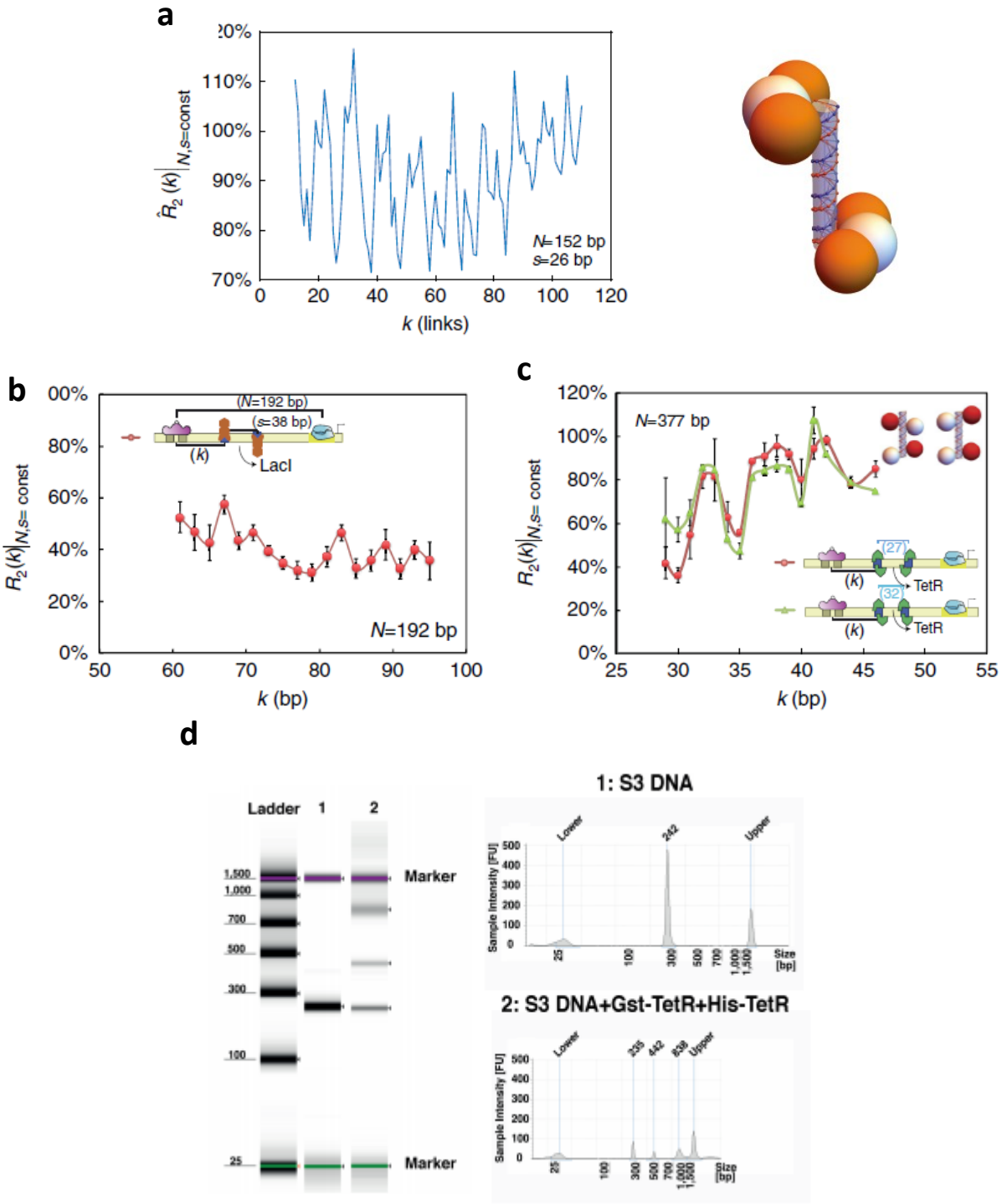


Figure 16. Periodicity of a half-helical repeat.

(a) probability-ratio level for a thick chain with two 'out-of-phase' protrusions taken at a fixed inter-site distance (s), showing oscillations with a 5- to 6-bp periodicity (lef) and 3D models (bottom) showing two potential Z-shaped binding architectures (right). (b) Expression level tandem binding sites spaced at 38 bp. (c) Expression level ratio data for synthetic enhancers with tandems of TetR-binding sites that are spaced by half integer (s=27 bp, red line) and integer multiples of the helical repeat (s=32 bp, green line), showing identical regulatory function. Inset: the structural binding model showing a thick chain with two dumbbell-shaped protrusions positioned in-phase (right) and out-of-phase (left), respectively. (d) Gel shift data for TetR and TetR-GST as read by an Agilent tape-station. Left panel: gel-like depiction with left, middle, and right lanes corresponding to the ladder, pure DNA, and TetR/TetR-GST bound to DNA. TetR-GST bound to DNA is the thick band below the 1500 bp marker band, and TetR bound to DNA is the band above the DNA band. Alternative depiction of data via absorption plots. Right panel, top: pure DNA absorption. Right panel, bottom: DNA, His-TetR, and GST-TetR absorptions.

Given these results and Z-shape model predictions, we wondered whether there was something amiss with our interpretation of the single binding-site expression-level ratio results for the TetR synthetic enhancer shown in Figure 12c. This data set shows a strong regulatory response with sharp fluctuations between quenching and upregulation even though TetR is a small protein (25 kDa [83]), which is nearly the size of TraR (26 kDa [84]). In addition, a closer look at the expression-level ratio scan (1 bp increments) reveals that the oscillations do not exhibit the ~11-bp periodicity expected from a single binding-site synthetic enhancer. Rather, a complex pattern of strong peak/weak trough‑weak peak/strong trough seems to emerge with a 5- to 6-bp periodicity between adjacent peaks and an 11-bp periodicity between 'strong' peaks or troughs is also apparent. These results are reminiscent of the model predictions shown in Figure 16 a for the out-of-phase tandems. Consequently, to account for the periodicity and size effect

anomalies, we hypothesized that our form of TetR (TetR-B [61]) might bind its binding site not as a dimer but rather as a dimer-of-dimers oriented in dumbbell-like configuration. This binding architecture is known for a member of the TetR family QacR [77] and in this interpretation an additional cryptic binding site overlaps the major site, allowing a dumbbell-like bound TetR structure to form. To test our hypothesis, the second part of our final library was designed with synthetic enhancers containing tandems of TetR-binding sites. We designed two sets: (i) first, with the TetR-binding sites in-phase (s=32 bp) and (ii) second, with the binding sites out-of-phase (s=27 bp). If the dimer-of dimer structural interpretation was correct, then the expression level ratio for both binding site configurations should be nearly identical. This can be seen from a schematic of a thick chain with protrusions (Figure 16c inset). In both configurations, two dimer-of dimer protrusion structures are shown on the thick chain with an overall 'out-of-phase' arrangement of two dimers inside the loop and two outside for the chosen inter-site spacings s. In Figure 16c we plot the expression level ratio measured as a function of k for these synthetic enhancers (N~377 bp), with green triangles and red circles for the in-phase and out-of-phase inter-site spacing configurations, respectively. The figure show that the expression level ratio regulatory response generated by both tandems is nearly identical as predicted by the model, with a distinct 5‑6 bp periodicity over a range of values for (k) that spans 20 bp at a single base-pair resolution. The regulatory pattern for both cases exhibits three distinct peaks and four troughs with a slightly increasing overall expression level ratio trend. In addition, the data sets lack the strong peak‑weak peak pattern of the single binding-site synthetic enhancer. As a result, the dimer-of-dimer dumbbell binding structure for TetR may indeed be a possibility in vivo. In order to test the dumbbell binding structure of TetR(B) to dsDNA, we carried out gel-shift

experiments with purified His-tagged TetR and TetR-GST. We reasoned that purified TetR that binds to DNA as a dimer of dimers should do so not only as a homodimer of dimers (TetR-TetR-TetR-TetR), but as a heterodimer of dimers (TetR-TetR - TetR-GST-TetR-GST) as well. We assumed that both TetR and TetR-GST would already be in dimer form before mixing. Thus, a 200 bp piece of DNA containing the TetR binding site and the suspected secondary cryptic site should exhibit three discrete shifted gel bands for TetR-TetR, TetR-TetR-GST, and TetR-GST-TetR-GST complexes. In Supp. Figure 16d we show the results for the gel shift experiment carried out with a mix of His-TetR-GST and His-TetR after properly calibrating for a 1:1 binding ratio for the two TetR protein species. The data show only two shifted bands with respect to the non-bound DNA (242 bp): the shifted band that appears for purified TetR (440 bp – middle band, presumably TetR dimer), and the band that appears for purified TetR-GST (850 bp – top band, presumably TetR GST dimer). Thus, while we have strong evidence for the formation of a dimer-of-dimer structure *in vivo*, the gel shift experiment at the very least seems to rule out the formation of a heterodimer-of-dimer structure *in vitro*, and does not provide additional support for the structural interpretation of our *in vivo* data.

INDEL Mutations in Natural Bacterial Enhancers

Finally, we wondered whether we could find evidence for the excluded-volume regulatory model in bacterial genomes. As this type of regulation depends on a fixed relative arrangement of the self-avoiding volumes, we speculated that naturally-occurring enhancers should exhibit a conserved evolutionary signature for this mechanism if it does indeed play a biological role. Specifically, we speculated that bacterial enhancers with similar regulatory function should be

insensitive to ~11 bp INDEL mutations that conserve both the regulatory TF and activator orientations relative to the promoter, and sensitive to function-altering INDELs that are 5 or 6 bp long. As a result, we expected that looping sequences should reflect this tolerance to the DNA helical repeat. To test this hypothesis, we analyzed the *qrr* (quorum regulatory RNA) genes in the *Vibrio* genus. The *Vibrio* genus of bacteria contains many sequenced genomes of marine pathogens. These bacteria have been used as model systems extensively, particularly for the characterization of quorum sensing systems. Quorum regulating RNA (Qrr) are part of the complex regulatory pathway of quorum sensing in *Vibrio*. Some of the *qrr* genes in this genus are known to be regulated by LuxO, an NtrC-like activator, which drives σ^{54} promoters. This system was implicated in the quorum-sensing pathway and was characterized for *Vibrio cholerae* [69,85]. The *qrr* genes are located on both chromosomes: *qrr1* is on the large chromosome, while *qrr2*-5 are located on the smaller chromosome. Quite a few studies have been carried out on these genes [66,67]. All sequenced *Vibrio* species carry the *qrr1* gene on the larger chromosome immediately adjacent to the sequence encoding LuxO. However, only a subset of the species carry either an additional three or four *qrr* genes on the smaller chromosome[69]. Not much is known about the promoters regulating the transcription of these genes, except for one notable study carried out on the *V. cholerae qrr4* promoter[66]. Given this partial understanding of the promoters, this system can be used as a case study to test the INDEL prediction.

Using standard bioinformatic tools (see materials and methods), we annotated 61 *qrr* enhancers, which included a putative LuxO-binding site tandem, a looping region and a single putative σ^{54} promoter. In Figure 17a we plot the distribution of the looping lengths for all 61 putative *qrr* enhancers. The figure shows a set of clustered looping lengths (N) ranging from an average length

of ~80 to ~120 bp, which are displaced from one another by ~ 11bp, thus providing initial support for our hypothesis. To further test the sensitivity of the *qrr* looping regions to integer multiples of the helical repeat, we checked the average identities of each loop sequence to itself and to all other loop sequences. To do this, we calculated the relative identity of each 9 bp window within a given loop sequence to all other 9 bp windows either on the same sequence, or on all other sequences, noting the distance between the positions of the first bases of compared windows (the relative identity is defined as the number of positions [from 1 to 9] for which both windows contain the same base, divided by the window length). We then computed the mean relative identity for each window separation by averaging over all relative identities in a particular window separation value. In Figure 17b we plot the results. The figure shows that the mean relative identity for the annotated *qrr* enhancers exhibits an oscillatory behavior, which persists for all possible values of the distance between the windows. Interestingly, the oscillatory pattern is detected not only for cross-correlated enhancers, but also for each enhancer to itself (self, red; other, black), with the first maxima appearing at ~0 bp displacement and with periodicity of 10.45 bp. Next, we checked whether there was some underlying signature for a conserved sequence within the looping region. To that end, we computed the average AT/GC content of each position within the loop and plotted the results in Figure 17c (top). The figure shows that the AT content is enriched at positions that are integer multiples of 10.6 bp with at least six distinct peaks visible in the data. In addition, the minima between the positions of AT enrichment converge on a content value of ~0.5, which is the value expected for a random allocation of AT or GC at those particular positions. Thus, loop sequences are similar either at the same relative position or alternatively at positions displaced by an integer multiple of the helical repeat from the position

of the reference sequence, with a preference in this case for AT segments. Finally, in Figure 17c (bottom) we plot the average AT/GC content and in Figure 16d the average relative identities of the *qrr* enhancer ‘upstream sequences' that are immediately adjacent to the annotated LuxO-binding site tandem (Figure 16d schematic). Unlike the looping region, the analysis on the non-looping region results in no particular repetitive pattern of AT/GC content within the upstream sequence. In addition, a monotonic or slightly varying signal across all possible values of the window displacement is observed in the upstream region without any detectable characteristic oscillations. Thus, the striking difference between Figure 17b, d and in Figure 17c (red versus blue line) provides further support to the special sensitivity of *qrr* enhancer sequences to the helical periodicity as compared with non-looping sequences.

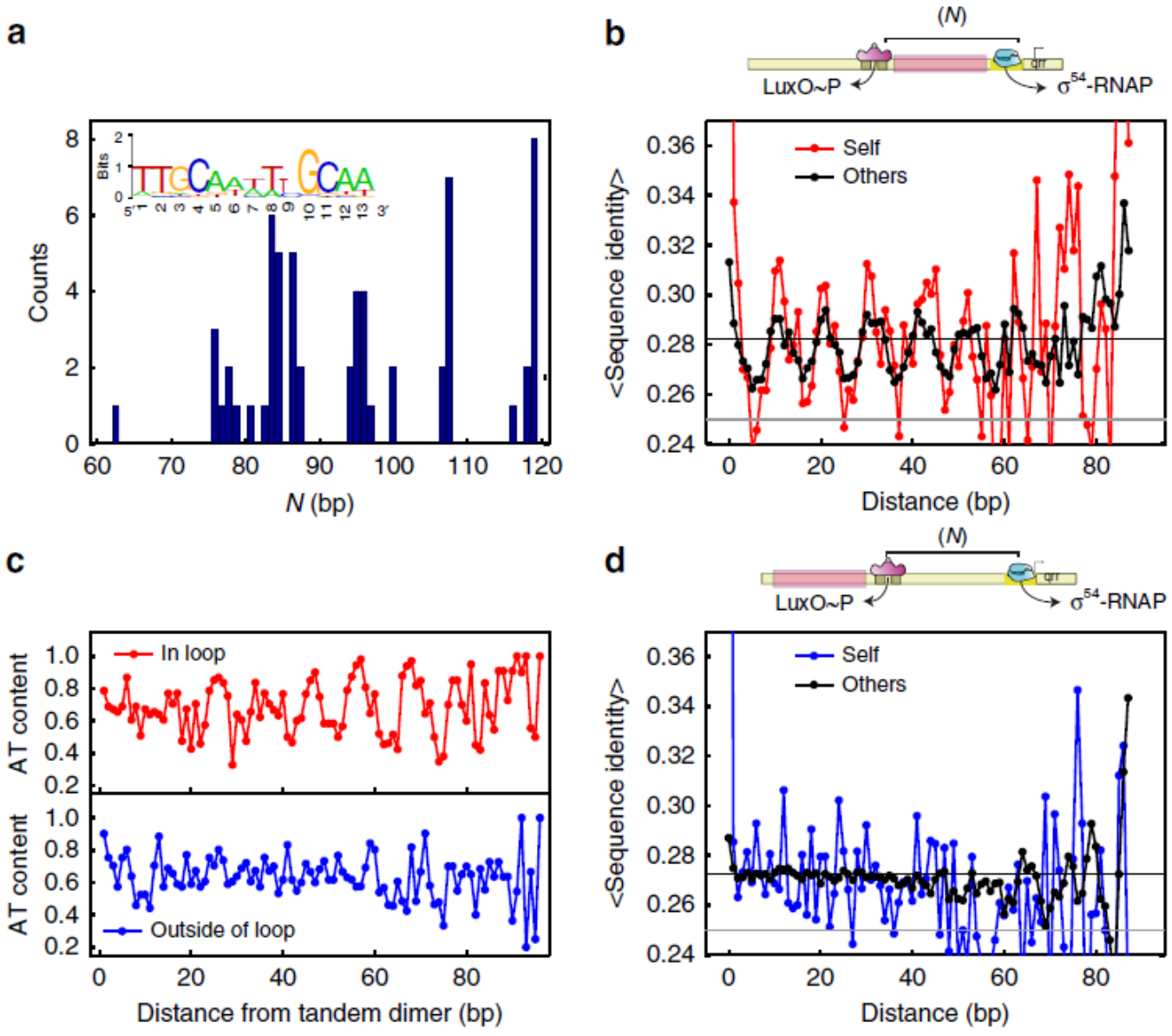


Figure 17. ***Sensitivity to the DNA's helical repeat in the Vibrio qrr enhancers***.
(a) Histogram of loop lengths for all putative loop sequences. Inset, LuxO consensus sequence. (b) Relative identity of all putative loop sequences to themselves (red) and to all other putative loop sequences (black). Horizontal lines indicate the expected identity level for sequences with equal probability for all four nucleotides (grey) and for the putative loop sequences (black). (c) AT nucleotide content as function of position for the loop sequences (top, red) and for the non-looping upstream sequences (bottom, blue). (d) Relative identity for all non-looping upstream sequences (from the tandem LuxO binding sites, see top schematic) to themselves (blue) and to all other upstream sequences (black).

AT/GC Variation Within the *qrr* Enhancer Sequences

The average relative identity values for both enhancer and upstream-enhancer sequences, averaged over all displacements (see horizontal lines in Figure 17c and 17d), are on the order of 0.25, which is the value expected for samples with equal probabilities for the four nucleotides at all loop positions, P(A) = P(C) = P(G) = P(T) = 0.25. For the case of the *qrr* enhancer sequences, the distribution of nucleotides are P(A,C,G,T) = 0.3505, 0.1431, 0.1798, 0.3266, indicating AT enrichment. In Figure 17c, we plot the AT and GC content as function of position within the loop sequences. The AT content is enriched at positions that are integer multiples of 10.6 bp, consistent with the helical periodicity of 10.45 bp found for the loop sequences using Fourier transform (FFT). While it is not clear what the significance of this enrichment is (e.g., sequences that are amenable for bending, binding site of some sort, etc.), it is interesting to note that we found no significant AT enrichment either at odd half-integer multiples of the helical periodicity, or for the non-looping enhancer upstream sequences.

To provide further support to the observations described in Figure 17, we carried out a comparative study on the annotated pAstC/AruC promoter in E. *coli*, S. *typhimurium*, and P. *aeruginosa*. In all three cases, a σ^{54} promoter is driven by NtrC or its P. aeruginosa homologue CbrB to generate expression of the *ast*/*aru* genes. In Table 3 we depict the structure of these bacterial enhancers as a function of their location on the genomes, and values of control parameters. By examining this table, it is clear that once again structure is remarkably conserved. The enhancers for E. *coli* and S. *typhimurium* are almost precisely conserved, in terms of looping length and positioning of the ArgR sites. The only difference seems to be the shifting of the site

by roughly a helical periodicity upstream. It is important to note that ArgR binds a tandem of sites to form a dimer-of-trimers, which is capable of bending DNA by roughly 70°. This implies that the four sites really should be viewed as a tandem of binding sites. For S. *typhimurium* and P. *aeruginosa*, this architecture is particularly conserved, with each tandem separated by approximately three helical periodicities. In E. *coli*, this symmetry is broken by having three half-sites closely positioned, and an additional half-site separated from the arg432 cassette. The significance of this is unclear. However, the conservation of ArgR arrangement between P. aeruginosa and S. *typhimurium*, and the overall conservation of enhancer size and binding site locations between S. *typhimurium* and E. *coli* (except for the shift in the binding site Arg2) are highly consistent with the previous observations for the *qrr* genes.

Table 3. Spacings of ArgR and NtrC binding sites in three different species

bacteria	# of NtrC sites	# of ArgR sites	NtrC-Arg4 spacing	Arg43 spacing	Arg32 spacing	Arg21 spacing	Arg1-promoter spacing
E. coli	2	4	110	22	21	31	24
S. typhimurium	2	4	110	21	32	21	22
P. aeruginosa	2	4	45	21	30	21	45

After concluding this set of experiments in bacteria, which we published in [86], we decided to proceed in two directions which are still ongoing work. The preliminary data are shown in parts II and III, below.

Part II: Verification of the Bioinformatics Analysis by Editing *E. coli's* Genome

Genomic Editing of σ^{54}-Regulated Genes

The bioinformatic analysis points to initial support for our findings within genomes, where the regulatory effect is governed by TF binding site position. Since all the previous experiments were done using plasmid and not genomic DNA, we decided to edit *E. coli*'s genome in σ^{54} looping regions to further validate the relevancy of our results. We selected σ^{54}-regulated genes, to test how introducing insertions in their looping regions affects transcription of the downstream gene. Genomic edits were done using the CRISPR/Cas9 system[64,87] (for elaborated protocol see Materials and Methods). The CRISPR/cas9 (clustered regularly interspaced short palindromic repeats) system enables to edit the genome in a specific location by the help of a RNA guide (gRNA). The gRNA directs the Cas9 protein, which can introduce double-stranded breaks in the DNA. Donor DNA is also added, harboring the desired insertion or modification and flanked by segments of DNA homologous to the segments immediately upstream and downstream of the cleaved DNA. Following the RNA-guided cleavage of a specific site of DNA, the donor DNA can be integrated via homologous recombination (HDR)[88]. The genes that we chose were *pspG* and *nac.*

pspG Gene

The phage shock protein (PspG) expression is driven by σ^{54}-RNA polymerase. It is activated by PspF and IHF and negatively regulated by PspA [89] (Figure 18). PspG is strongly induced by protein IV (pIV), a secretin from filamentous phage, which forms pores in the bacterial outer membrane, required for the export of the phage [90]. To induce the expression of PspG, we decided to overexpress its activator PspF, bypassing pIV induction. Hence, we cloned *pspF* coding sequence into an overexpressing plasmid harboring an inducible promoter (see materials and methods).

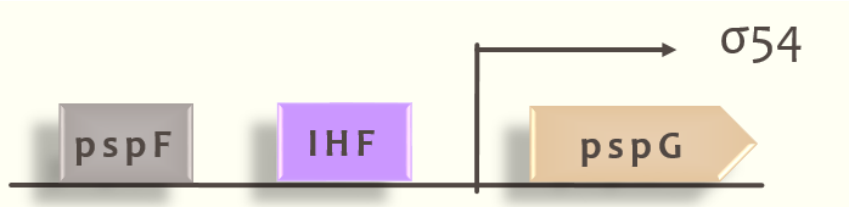


Figure 18. pspG enhancer.

PspG is positively regulated by PspF, which binds upstream to its CDS. An addition binding site for IHF is in vicinity to PspF.

We started with a control experiment, where we validated that *pspG* expression is indeed a result of σ^{54}-mediated expression. Since we wish to measure the effect of the genomic edits on looping, we wanted to be sure that σ^{70} (which does not form loops) is not involved in *pspG*'s expression. To do so, we used a σ^{54} deletion (Δ*rpoN*) *E. coli* strain, which has a deletion in the *rpoN* gene that codes for the σ^{54} protein. This strain was created in our lab using CRISPR/Cas9 system as well. RT-PCR results can be seen in Figure 19. *pspF* is indeed overexpressed due to induction.

Surprisingly, *pspG*'s levels were induced as well, although the strain does not express the *rpoN* gene. *pspA*'s expression, which is also regulated by σ^{54}, remains unchanged, as we expected. We could not find evidence in the literature for σ^{70} mediated expression of *pspG*, though running the sequence in a TSS locating tool [91] we found two possible σ^{70} sites upstream to the σ^{54} site, that can perhaps explain why we see an induction despite the lack of σ^{54}. Since unexpectedly, we found that *pspG* is regulated by both σ^{70} and σ^{54}, it might not be the ideal system to test the genomic edits.

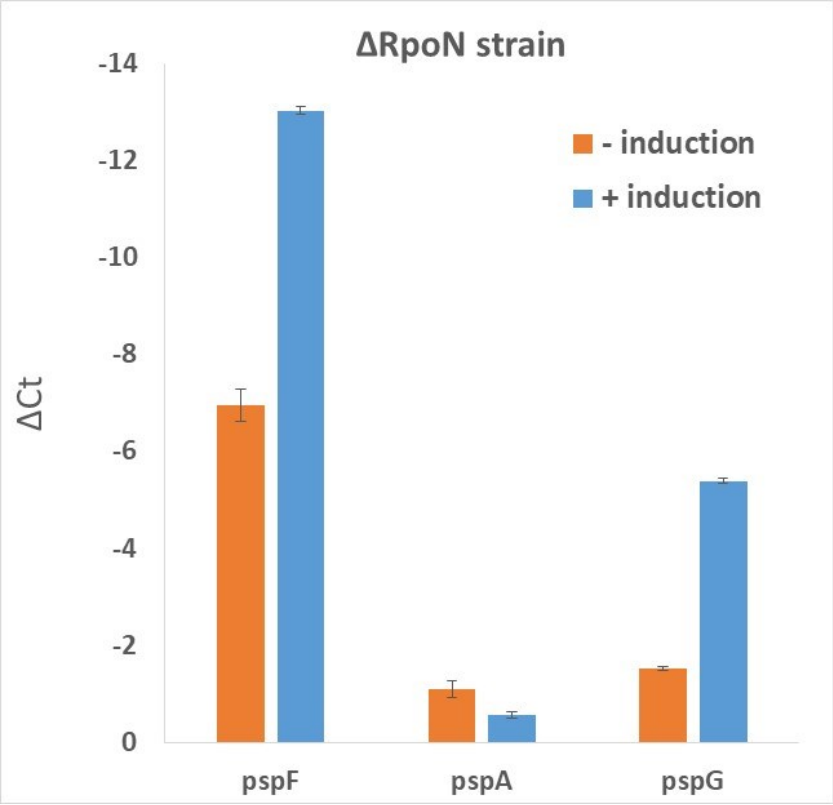
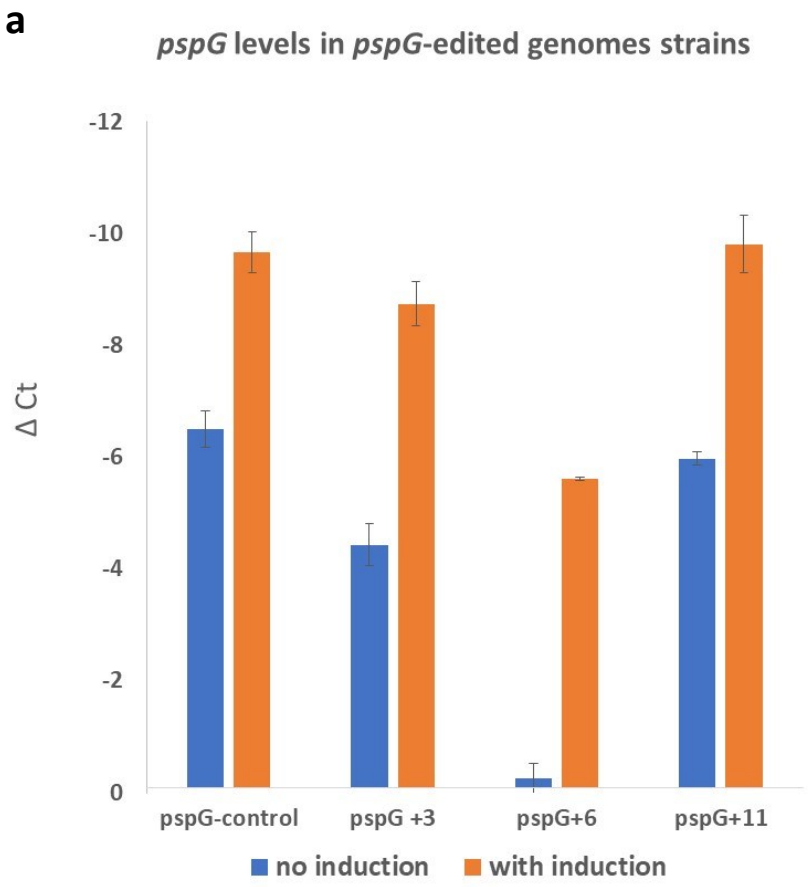


Figure 19. RT-PCR for pspG, pspF and pspA in ΔrpoN E. coli strain. *Ct values of the three genes with and without induction of pspF expression.*

Nevertheless, since we've already had the genomic edits ready, we decided to proceed with the experiment, and test whether we see differences in expression for different loop sizes. The genomic edits we created using the CRISPR/Cas9 system included insertions of 3, 6 and 11bp, downstream to the pspF binding site. In addition, we created a control edit- a strain where a change in a nucleotide (SNP) was introduced but the looping length was as the WT strain. Following induction of *pspF*, cells were subjected to RNA isolation and *pspG* levels were determined using RT-PCR. The results can be seen in Figure 20a.



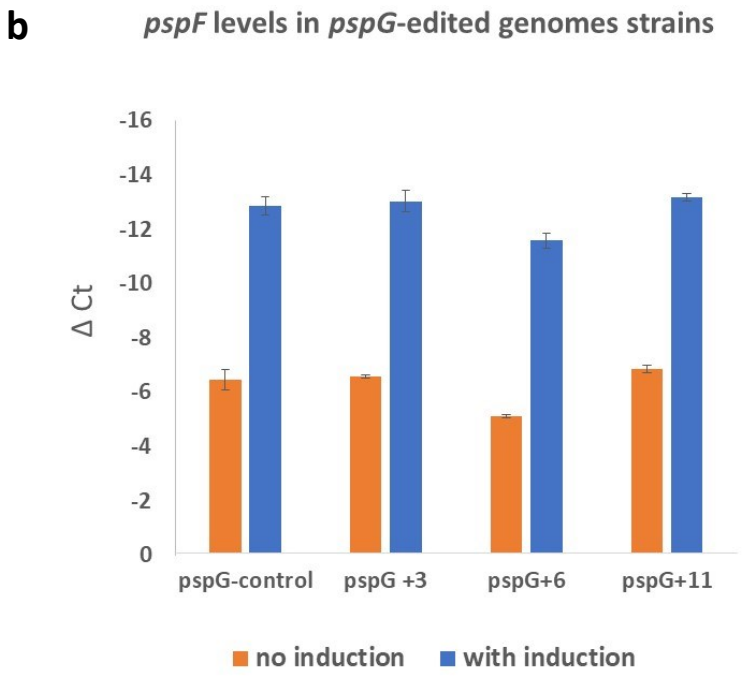


Figure 20. Real-time PCR of pspG in pspg-edited loops of E.coli genomes. *(a) Ct values of pspG's levels in different loop sizes with and without induction of pspF induction (b) Ct values pspF's level in different loop sizes with and without induction.*

The levels of *pspG* in the +11bp edit are as the control, with significant increase (8 fold) following pspF induction (Figure 20a.) This is expected, since 11 bp insertion retains the relative position of the *pspG* and the promoter site. While an addition of 3bp to the looping length has led to a slight decrease of pspG level, the +6bp edit had a dramatic decrease in *pspG's* levels, giving further support to our hypothesis. To verify the levels of induction, we also performed in the same experiment a measurement of *pspF's* levels to confirm the induction levels were even in all samples. As seen in Figure 20b, *pspF* was highly induced in all strains (~80-fold induction).

Nac Gene

The nitrogen assimilation control (Nac) gene expression is driven by σ^{54}-RNA polymerase as well. It is activated by nitrogen assimilation regulatory protein (NtrC), under nitrogen-limiting conditions, and is negatively autoregulated by *Nac* binding site located near the NtrC site (see Figure 21). Similar to *pspF*, the goal was to explore how changing the orientation of the regulatory elements affects transcription. To induce the expression of *Nac*, we decided to overexpress its activator NtrC along with NtrB, a regulatory protein crucial for its activation, bypassing nitrogen starvation induction. Hence, we cloned both NtrC and NtrB coding sequence into an overexpressing plasmid harboring an inducible promoter similarly to *pspF* (see materials and methods).

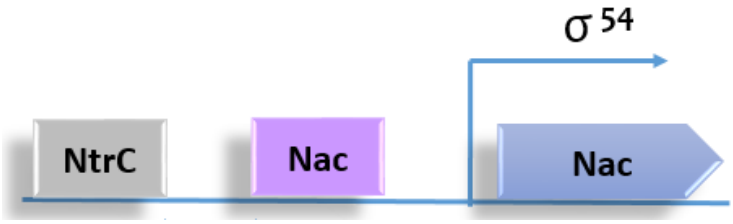
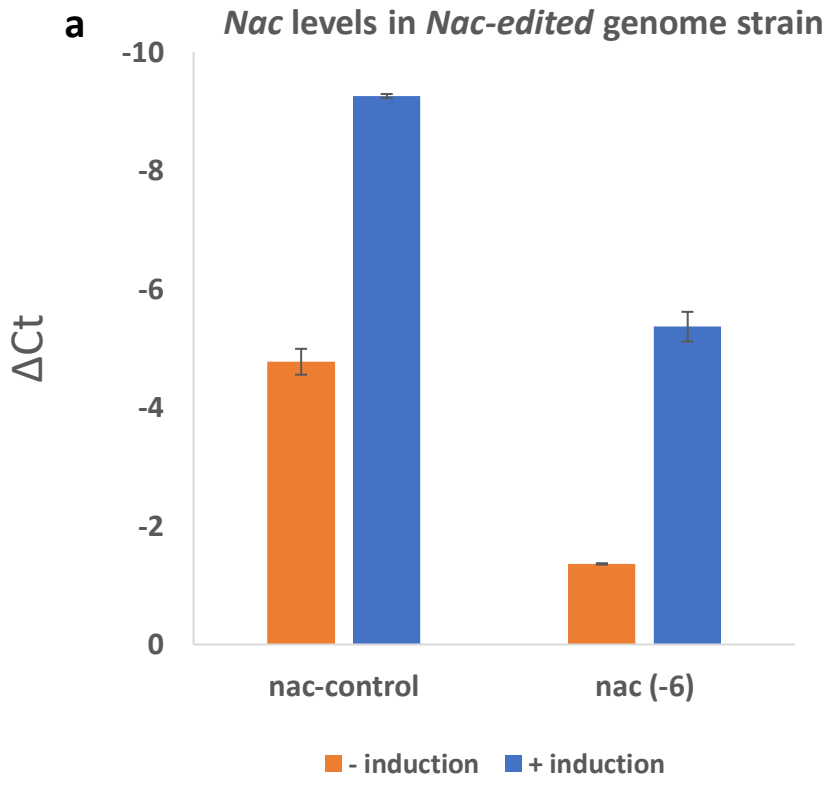


Figure 21. Nac enhancer. *Nac is positively regulated by NtrC, which binds upstream to its CDS. An addition binding site for Nac is in vicinity to Ntrc.*

The genomic edits we created for *Nac*, using the CRISPR/Cas9 system included deletions of (-3), (-6) and (-11) bp, downstream to the NtrC binding site. In addition, we created a control edit, a strain where a change in a nucleotide (SNP) was introduced but the looping length was as the WT

strain. To date, we were able to produce only (-6) edit, and currently we are screening for the remaining edits. The inducible plasmid was transformed into the genomic edit strain. Following induction of *Nac*, cells were subjected to RNA isolation and *Nac* levels were determined using RT-PCR. Figure 22a shows the results of the Nac (-6) edit versus the control. It is evident that there's a significant decrease in the levels of *Nac* edited strain as compared to the control in both induced and non-induced samples. As control we verified that the induction of *Nac* was even in both control and edited strains, by measuring the amounts of Ntrc/Ntrb (Figure 22b). Those levels were evenly expressed in both samples Thus, in this case deletion of half helical repeat had dramatic change on expression. Along with the *pspG's* results these findings provide very nice support to our INDEL's prediction



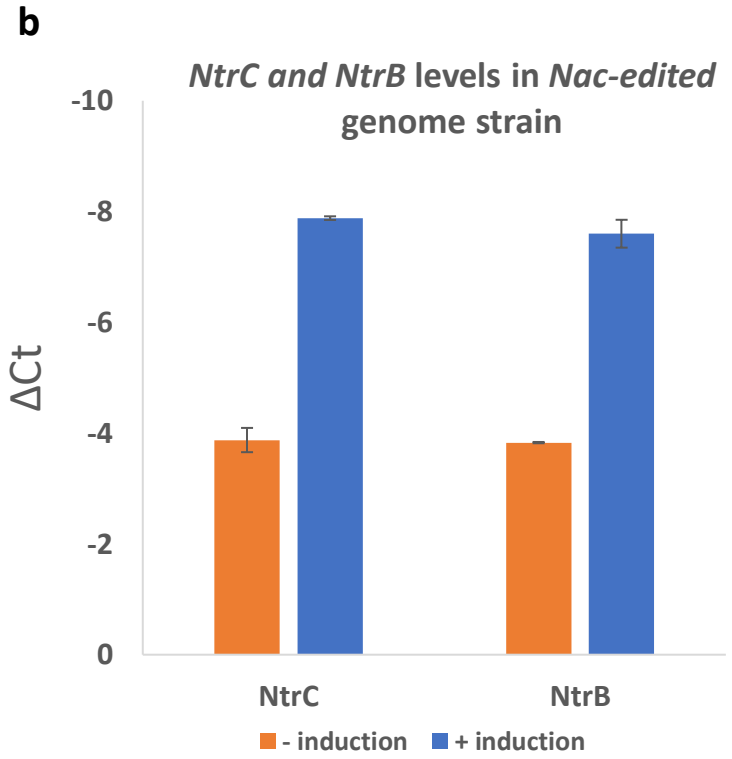


Figure 22. Real-time PCR of Nac in Nac-edited loops of E.coli genomes. *(a) Ct values of Nac's levels with and without NtrC/NtrB induction (b) Ct values NtrC/NtrB level with and without induction.*

Part III: Testing for DNA looping in Yeast

The structural findings in bacteria led us to examine whether we can find evidence for our excluded volume model in higher organisms. The motivation to proceed to a yeast system emanated from both the fact that yeast UASs are located hundreds of base-pairs from the core promoter, similar to driver-promoter distances in bacteria, and the overall resemblance of UASs to bacterial enhancers' architecture (Figure 3) . Although some papers point that DNA looping can occur in yeast in some circumstances, it was not shown to play a role in transcription in a direct or systematic fashion. Since our excluded volume model is based on DNA looping, we decided to first test *in vivo* whether DNA looping indeed plays a role in action at a distance in yeast. There are various methods to show looping of two regions, from 3C-based methods (chromosome conformation capture)[92], through microscopy methods[93] and single molecule experiments [94]. Nevertheless, these methods are indirect and prove contact that not necessarily results in actual transcription. Another approach is based on the requirement for the binding sites of the activator and polymerase. For the formation of DNA loops, it is necessary that the activator and RNA polymerase be in phase on the double-helix to allow the interaction between them. DNA has natural torsional rigidity and develops a resistive torque when it is twisted[95]. If the two sites are on opposite sides of the double helix, torsional energy is required to twist the DNA, which is energetically costly and thus disfavored[96]. This concept was examined in *E. coli*, in a high-throughput experiment done by Amit at el. [59]. Systematically varying the length of the DNA sequence between the driver-binding sites (activator) and σ^{54} promoter yielded an expression pattern that depends on the length of the looped DNA and thus on the phasing of the complex

(the orientation of the driver with respect to the polymerase bound to the promoter, that depends on the DNA helical periodicity). In order to check if model plays a role in yeast, we created a library of different looping lengths based on the well-characterized Gal1 UAS[97]. The UAS of Gal1 is composed of four different binding sites for the transcription factor Gal4 and two binding sites for the Mig1 repressor (Figure 23).

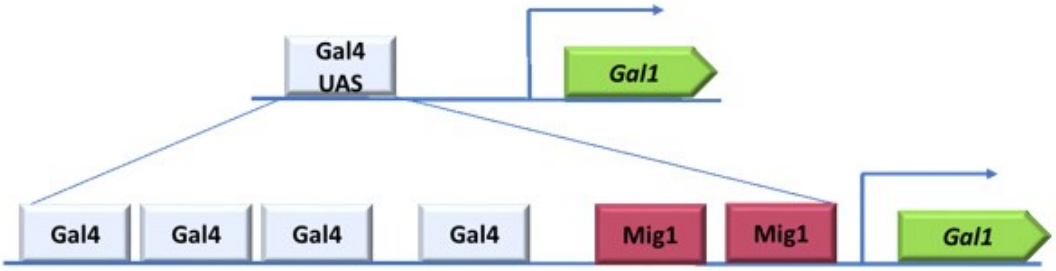


Figure 23. The Gal1 UAS.

Expression of the GAL1 gene is strongly repressed by growth on glucose. Upon galactose induction Gal4 becomes active and activates gal1 expression. MIG1 binds to the GAL4 promoter and represses its transcription in the presence of glucose.

We conducted the experiment in a very similar way to the bacterial enhancers' library described in Part I. We replaced the Gal1 CDS with a fluorescent protein, GFP (green fluorescent protein) and designed the library in such way that we change the distance of the UAS relative to the promoter region, in 3 bp intervals. To induce the circuit, we induce cells with galactose and measure the reporter levels using flow cytometry (for more details see Materials and Methods).

Figure 24 shows the results of the designed library of two different experiments. The data obtained was very noisy, and the two experiments showed very different results. Moreover, no

apparent oscillatory behavior is observed in either of the two experiments. Due to the high variability of our results, we decided to integrate the library into the genome, to minimize the noise that might emanate from variability that may arise from different number of plasmids per cell or different expression levels. Indeed, previous reports showed that plasmid-based systems are noisier than chromosomal integration, both in yeast and mammalian cells [98,99]. Therefore, we started integrating the library into the His3 locus (see Materials and Methods), and hoped to get more reproducible results for the integrated library.

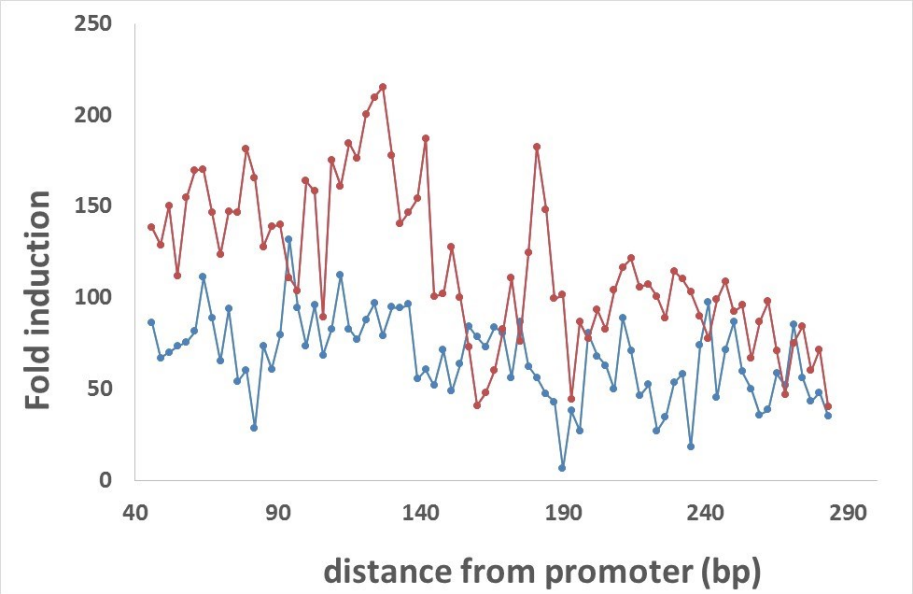


Figure 24. Relative fluorescence levels versus looping length.

For each looping length, fold induction is defined as the ratio between the measured fluorescence levels of the synthetic enhancer strain to the fluorescence level of the WT strain.

Figure 25 shows the expression of a subset of points from the library, compared to their corresponding point from Figure 24. It can be seen by the error bars, that the data that we received is still very noisy in the points that were integrated to the genome (red line). Here again, no apparent oscillations are observed. There is still large difference in the expression levels between the library that was expressed on the plasmid level to the genomic one. We are still waiting to get the full data set to draw final conclusions from this library. It could be that technical errors as the way that we conduct the experiment (induction, starvation etc.) is not kept constant enough to receive reproducible results and that there's a need to recalibrate the protocol we use to perform the experiment.

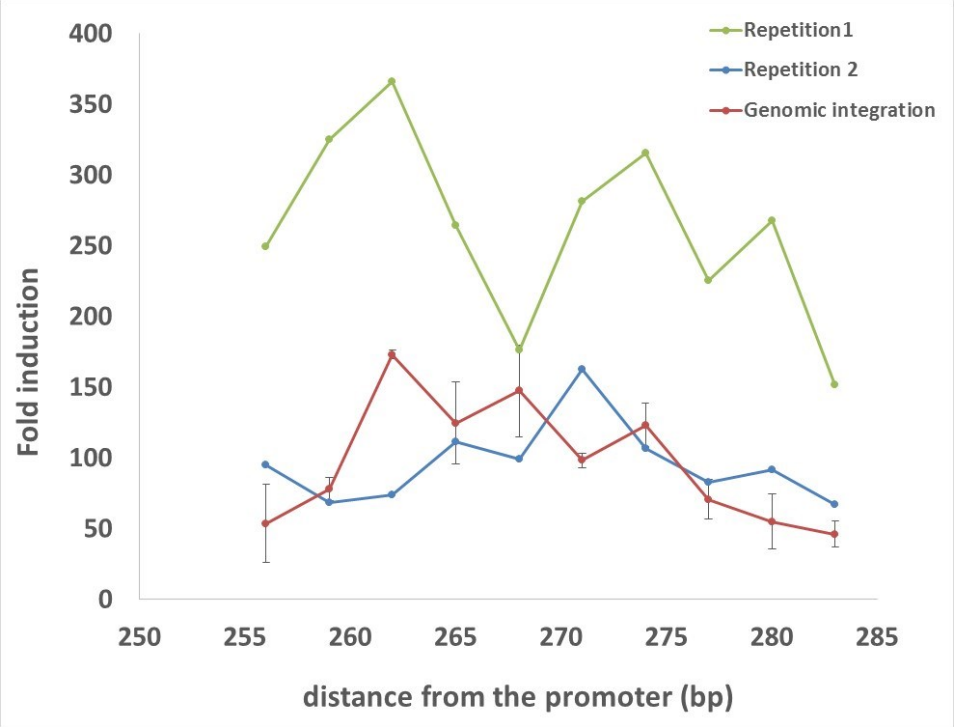


Figure 25. Expression levels library subset. *Comparison between a subset of point from the yeast library either expressed from plasmids (green/ blue) or integrated to the genome (red).*

Discussion

Transcriptional regulation is one of the major levels in controlling gene expression. Since most of the control is encoded in noncoding regions, and different regulatory DNA sequences can drive distinct levels of gene expression, unraveling the link between DNA sequence and expression levels is key for understanding transcriptional control. This requires an understanding of the set of rules that govern basic principles in transcriptional regulation. These include the effect on expression of the number of TF binding-sites, their location, orientation, affinity and interaction with different factors. Without this understanding, we cannot point how and which change in a given sequence affects transcription. Recent technological advances in genomics made it possible to design, construct and measure the effect of thousands systematically designed regulatory sequences on expression. Moreover, the decrease in cost of DNA sequencing and synthesis have led to a series of works performing high-throughput experiments aiming to decipher the regulatory code. But two questions still remain. First, can we actually decipher the code solely by increasing the number of sequences designed and measured? Indeed, utilizing thousands of regulatory sequences have gained some insights into how information is encoded in the language of DNA. Nevertheless, they lack rules that enable us to predict how a given enhancer will exert its regulatory effect. Second, can we really base ground rules for enhancers on data from synthetic constructs that will allow us to understand, predict and design expression patterns from regulatory sequences? Will it enable us to annotate regulatory sequences?

In this work, we address these questions by harnessing synthetic biology tools to study basic questions in enhancers' organization. Our approach was unique due to the fact that we utilized

a model that allowed us to focus our design. By using a model-based approach we were able not only to screen for a relatively small amount of sequences, but also to propose a mechanistic model. We presented a new mechanism for quenching-like repression in bacterial enhancers using a combined thermodynamic modelling, synthetic biology and bioinformatic approach.

Model-Based Approach for Studying Enhancers:

In the past few years, multiple papers have offered novel methods to study enhancers by incorporating reporter assays with high-throughput sequencing [49–51,107]. These methods enabled the study and quantitative measurement of thousands of synthetically designed regulatory sequences. Despite the high-throughput capability they offer, these works yielded some sequence observations but haven't provided a mechanistic or causal understanding of these observations. As a result, our understanding of enhancer grammar rules remains limited. The uniqueness of our approach is rooted in the fact that we based our experiments on a model, which we believe is missing in the current alternative approaches for studying enhancers. This enabled us to define set of structure rules for enhancers' organization and suggest a possible mechanism for quenching repression. As discussed below, it has also potentially allowed us to be able to annotate σ^{54} promoters as well. Nevertheless, can we assume that the synthetic approach taken here in this work can be applied to a much more complexed system? We believe that expression assays combined with modelling would probably not be enough to get full insights on decoding enhancers in eukaryotes due to their complexity. Most enhancers' are mediated by very large complexes of proteins, where simple binding of a transcription factor to its binding site, as we measure in the bacterial system, in most cases is not sufficient to drive gene expression. A possible approach to overcome this difficulty could be the use of synthetic transcription factors

that are fused to activator/repressor (as VP64 or KRAB) that were shown to be able to drive gene expression independently.

Main Experimental Findings:

First, we constructed a preliminary DNA looping-based mechanistic structure‑function model. The model established a direct relationship between the values of enhancer structural control parameters (for example, number of binding sites for a TF, inter-site spacing, looping length and so on) to a predicted regulatory function that was based on elastic (bending and stiffening) and entropic (excluded volume) characteristics of a thick chain with protrusions. Second, based on the numerical results we characterized synthetic enhancer libraries focused on testing the predictions of the structure‑function model, allowing us to improve the mechanistic model. Our model-guided approach has enabled us to set predictions that were experimentally tested with synthetic enhancer and define structural rules that are listed below:

· TF Excluded volume effect can produce both up-regulatory and down-regulatory effect, depending on the location of the binding site in the looping region.

· TF Bending effect can produce both up-regulatory and down-regulatory effect, depending on the location of the binding site in the looping region.

· TF Stiffening effect causes an overall shift to the down-regulatory regime.

· The overall regulatory effect caused by TF bending and/or excluded volume strongly depends on the location of the TF binding site within the looping region. This dependence results in a periodic pattern of ~10.5 bp, corresponding to the helical repeat of the dsDNA.

· The overall regulatory effect strongly depends on the size of the TF, or conversely on the number of TF binding sites oriented in-phase, with larger TFs causing a larger effect.

· When more than one TF binding site is present, the overall regulatory effect strongly depends on the relative orientation between the bound TFs. In-phase orientations augment the effect and retain the ∼10.5 bp periodic pattern, while out-of-phase orientations diminish the effect and introduce a ∼5 bp (half helical repeat) period.

· TF excluded volume and TF bending are competing effects. The oscillatory regulatory pattern is dictated by the dominating effect between the two. It is therefore possible to reverse the observed regulatory pattern by augmenting the non-dominant effect, making it dominant (e.g. by sufficiently enlarging a bending TF, whose bending effect is dominant). The competition between the effects also produces an effective stiffening effect, resulting in an overall shift to the down-regulatory regime.

Overall, our experimental results coincided very nicely with the theoretical results. To fully account for all of our experimental observations, we had to incorporate small local bending and stiffening effects into our excluded-volume model. Interestingly, these additional elastic effects, which fine-tuned our model, have made both our experiments and model applicable to past observations made on bacterial enhancers. These include studies of IHF-dependent enhancers[100,101], where IHF was shown to both upregulate and downregulate expression depending on position[19], and non-IHF dependent enhancers [17,18,102] . As a result, a novelty of our experimental and modelling results is that the conservation of TF-binding orientation within bacterial enhancers seems to be a generic phenomenon for all transcription factors and is not limited to a handful of DNA-bending proteins.

Repression based on DNA looping has been experimentally shown in several operons in E.*coli*. The best-studied examples are ara and lac operons and lambda repressor [79–82] . In the lactose operon, repression occurs by co-repressor proteins that are recruited as direct obstacles to RNA polymerase binding, and/or by bending the promoter DNA into a tight loop. Here, the oscillations emanate from the fact that the two operator binding sites are distal from one another and must face one another to directly interact for loop formation. Similarly, the lambda repressor binds cooperatively in two distal locations, thus losing cooperativity when the operators were separated by non-integer turns[103]. The periodicity of ~10 bp was also documented in a phenomenon called "allostery". Using single-molecule experiments, it was shown that binding of a protein on DNA is substantially stabilized or destabilized by another protein bound nearby due to the distortion of the double-helical DNA structure. Though this effect is very local and decays within ~25 bp [104] and thus is not suitable for our model and experimental observations. Alternatively, we claim that the oscillations observed in this work are due to the volume of the protein, which results in its ability to repress DNA looping when facing inside the loop. This assumption was supported by both the fusion protein experiment and the additive effect observed in quenching of one binding site versus two binding sites. Repression of transcriptional loop formation was reported in a work done in *Drosophila*. In this work[10], 3C experiments showed that repressors prevent the looping of distal enhancers to the promoter, but the mechanism remained unknown, and it is not clear whether indeed the volume plays a role in this system. Though, this may imply that our mechanism may be relevant in higher organisms as well.

Bioinformatic analysis to support the excluded volume effect:

In efforts to demonstrate the relevancy of our results to real genomes, we utilized standard bioinformatic tools. We performed annotation of 61 *qrr* enhancers to test the sensitivity of the *qrr* looping regions to integer multiples of the helical repeat. Indeed, the mean relative identity for the annotated *qrr* enhancers exhibits an oscillatory behavior. Interestingly, the identity was not only for cross-correlated enhancers, but also within each enhancer to itself. We then checked whether there was some underlying signature for a conserved sequence within the looping region. The AT content is enriched at positions that are integer multiples of 10.6 bp, whereas the analysis on the non-looping region results in no particular repetitive pattern of AT/GC content within the upstream sequence. This observation provides further support to the special sensitivity of *qrr* enhancer sequences to the helical periodicity as compared with non-looping sequences. The observation for enrichment for AT content within these enhancers requires deeper screening to explore whether this is unique to the *qrr* gene or a more general phenomenon, as some works support poly AT enrichment in context of DNA loops[105].

In addition to aligning the *qrr* enhncer gene, we carried out a comparative study on the annotated pAstC/AruC promoter in E. *coli*, S. *typhimurium*, and P. *aeruginosa*. We examined the structure of these bacterial enhancers as a function of their location on the genomes, and spacing between the binding sites. We found once again that the structure is remarkably conserved.

Our experimental, modelling and bioinformatic analysis suggests that sensitivity to INDELs that are integer multiples of the helical repeat could be an evolutionary fingerprint for enhancers, whereas INDELs of odd half-integer multiples of the helical repeat should be flagged as

candidates for important regulatory variation. This assumption was further tested by genomically editing E.*coli's pspG* and *Nac* genes. The experiment was done by utilizing CRISPR/cas9 technology, introducing insertions/deletions of 3, 6 and 11bp within the looping region between the activator binding site and the σ^{54} binding site. As we predicted, introducing a half helical repeat into the looping region led to a significant decrease in pspG and Nac levels, which gives further support to the sensitivity of INDELs in these regions.

Testing for Looping in Yeast

In our efforts to expand this work, and test the applicability to eukaryotes, we decided to proceed and design synthetic enhancer experiment in yeast similar to the bacterial one. A worked published by Sharon et al.[51] performed a high-throughput experiment in yeast, measuring thousands of systematically designed promoters. A ~10-bp periodic relationship between gene expression and binding-site location was shown for Gcn4 transcription factor. Though, these results were not reproducible for seven other TFs tested in this work. Thus, it is unclear whether this phenomenon plays a role in regulating only a subset of TFs but encouraged us to explore our model in yeast. Our excluded-volume model is based on DNA looping, and no direct evidence for DNA looping exists in yeast. Thus, a starting point for us was to design a library in yeast, where we ask whether we can find evidence for DNA looping. Both directions are still ongoing work and currently we still don't have enough data to draw conclusions. Our preliminary data in yeast does not support DNA looping, at least in the system that we tested. One can argue that yeast UASs are close enough to the core promoter and hence do not to require looping as part of gene activation. Other mechanisms such as the linking model and the scanning model might play a role

in long-distance transcriptional activation in yeast[33]. Alternatively, constrains such as nucleosome favoring sequences, and the surrounding sequence adjacent to binding site, were shown to have an effect on the expression regulation in yeast[51] Perhaps a more meticulous design of sequences is needed to actually observe a genuine regulatory effect.

Applicability to Eukaryotic Enhancers

In this work we choose to focus on bacterial model, as it is known to be a much simpler system in terms of enhancers' complexity (e.g. number of binding sites and distanced for their target gene). All organisms share the same DNA language. Since the physical and biochemical characteristic of TFs and DNA molecules do not change between organisms, can we assume that the rules governing transcription regulation are similar?

As discussed in the introduction, in eukaryotes, the diversity of distal regulatory regions is vastly richer than in bacteria. There are several well-known examples of 'promoter-proximal' distal regulatory regions, which are clusters of transcription factor-binding sites that are located within 100‑ 300 bp away from a core PolII promoter and are thus similar in sequence length and regulatory content to bacterial enhancers. One example is the YY1 repressor which regulates the c-*fos* promoter. This transcription factor was shown to bend DNA in an oriented dependent manner [106], similarly to the quenching repression that was observed in this work. We believe that this subset of enhancers would probably obey the set of rules that we defined in the bacterial system and it would be of great interest to test that experimentally. Naturally, applying our model-based approach will require us to modify our model's boundary condition as well.

Moreover, we will have to take into account molecular details as nucleosomes and different looping criterion.

Limitations of the Study

One limitation of the synthetic enhancer approach is the incredibly large sequence space that can be potentially explored. Since both natural enhancers and promoter proximal regions are typically 200-500 bp long, the amount of random sequence variation is huge. Although we used a model-based approach for our libraries' design, and test particular regulatory predictions of our model, it could be that sampling only a rather small number of cassettes is not enough for characterizing global effects. Moreover, changing the context of binding site from the natural architecture, could result in unintended consequences, such as creation of new binding sites (although in our design we screen them out to the best of our knowledge).

Another limitation of this study could result from the fact that our libraries were expressed from plasmids and not chromosomally integrated to the genome. Drawbacks from plasmid-based overexpression include variations in plasmid copy number. To minimize that, in bacteria we used a low-copy number plasmid. In the yeast system, using plasmids has led to very noisy results. We thus started to integrate the library into the yeast genome.

An additional limitation can result from the assay we utilize for our measurements. Since we measure fluorescence as an output to our design, it may result in missing some effects that are averaged out. Although our experimental results overall coincided with the model predictions, to actually prove elastic effects experimentally, complementary methods as cyclization assay and

atomic force microscopy (AFM) may give additional support and structural insights. Though these are *in vitro* methods and might not recapitulate the *in vivo* state.

Contribution of this work to different biological fields:

We believe that our results have relevance to a wide array of biological fields:

In synthetic biology they reveal a looping-based regulatory mechanism, which can be integrated into complex gene circuit designs. We observed in some configurations very strong repression (~80%) and medium activation (~150%). It would be interesting to try build circuits combining configurations of DNA binding proteins as those used in this work, for creating a specific regulated behavior.

In genomics and bioinformatics, our analysis proposes a genetic signature for identifying looping regions based on the existence of a periodic 11 bp signature. In addition, our data sheds light on the importance of INDEL mutations, and in particular on function altering of 5-6 bp (and odd integer multiples thereof) INDELs and their possible role in genomic variation and disease.

In transcriptional regulation, we show the existence of a looping-based mechanism for quenching repression, which may be especially important in eukaryotic enhancers. It would be interesting to examine whether this mechanism also play a role in both *Drosophila* and mammalian cells, where looping plays a major role in regulation for distal enhancers.

In developmental biology, our combined strategy may help uncover enhancer regulatory mechanisms, by providing a model as to how to reduce the size of synthetic enhancer libraries making this approach tractable in a whole organism setting.

In polymer physics, we present a new model for simulating polymer behavior. While WLC model has been long used to study DNA behavior as a polymer, we performed a simulation study of the looping for a self-avoiding worm-like chain (SAWLC) model, taking excluded volume effects into consideration for the first time.

Future Work

One direction that can be followed is to extend our method to mammalian cells, and in the process, expand and test additional features of our mode. The goal is to gain a better understanding of how cis-regulatory elements in mammalian cells may interact in either a repressive or up regulatory manner to regulate gene expression.

Recently, [49,51] have shown that using synthetic oligo libraries (OL), next generation sequencing technology, and FACs sorting, it is possible to characterize >50,000 synthetic cis-regulatory regions in yeast, mammalian cells, and mouse models simultaneously. The increase in synthetic enhancer library capacity provided by this technology can now allow us to apply the model-based synthetic enhancer approach to higher eukaryotes, as the added level of precision in design can now be channeled to probe the increased level of complexity observed in eukaryotic regulatory regions as compared with the bacterial ones.

In addition, we plan to broaden the bioinformatics analysis. Since all experiments in this work were done using plasmids, we are currently working on genomic editing of σ^{54} enhancers in *E. coli* to get further support for the relevance of our results for genomes. We've presented initial support for the *pspG* gene, and partial results for the *Nac* gene that further corroborate our results.

Lastly, we are currently developing an algorithm that will combine our INDELs and periodicity signature, for prediction and annotation of unknown σ^{54} promoters. As discussed in the introduction, enhancers are notoriously difficult to annotate. There are many reasons for this difficulty. First, regulatory regions diverge rapidly in evolution, and are thus poorly conserved. Second, the relative position, number of sites, and transcription factor type are not necessarily conserved as there seems to be redundancy in regulatory output for different binding site arrangements[108]. Third, the binding site themselves are prone to mutations either due to redundancy, or because of a specific evolutionary constraint that demands a stronger or weaker binding at that position[109]. Finally, the underlying mechanisms which lead to specific DNA binding as well as the structure-function relationship between transcription factor binding arrangement and their regulatory output function are still poorly understood. As a result, only a handful of enhancers in both eukaryotic and prokaryotic organisms have been fully annotated.

Currently, there are ~100 documented σ^{54} promoters in common databases as RegulonDB[110], only 24 of which were experimentally validated. A series of algorithms for predicting promoters have been developed in the last few decades, mostly based on identifying consensus sequences recognized by the σ^{54} holoenzyme complex within the promoter region[111,112]. Though, the σ^{54} motifs are short and not fully conserved among species, which may lead to many false positives. Other algorithms have used genome structure constraints as DNA duplex stability, and machine learning approach by using existing annotated promoters and trying to predict new ones based on common features[112,113]. Though, the data sets constructed in these methods were too small to reflect the statistical profile of σ^{54} promoters, as recent Chip-Seq experiments vastly expanded

the list of putative σ^{54} binding sites to over 200, implying a stressing need to improve σ^{54} predicting algorithms.

In continuation of the work presented here, we are currently applying bioinformatics analysis combining the excluded volume effect and the genomic signature of helical repeats documented here, for annotating previously unknown σ^{54} promoters. We believe that this can serve as an additional valuable tool mainly for identifying UASs and annotating unknown σ^{54} promoters and can potentially improve existing databases. Preliminary data suggest that indeed looking at many genomes, we are able to find the helical periodicity stamp while searching for new σ^{54} promoters. This led us to generate an algorithm for finding these promoters, based on finding these genomic stamps. Currently, we are running a RNA-seq library that is aimed to validate experimentally the predicted σ^{54} promoters generated by our algorithm. The ability to annotate new σ^{54} promoters by using regulatory rules is novel and supports the notion that such rules actually exist.

References

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).
2. GUARENTE, L. UASs and enhancers: common mechanism of transcriptional activation in yeast and mammals. *Cell* **52,** 303–305
3. Xu, H. & Hoover, T. R. Transcriptional regulation at a distance in bacteria. *Curr. Opin. Microbiol.* **4,** 138–144 (2001).
4. Huo, Y.-X. *et al.* Protein-induced DNA bending clarifies the architectural organization of the sigma54-dependent glnAp2 promoter. *Mol. Microbiol.* **59,** 168–80 (2006).
5. Ninfa, A. J., Reitzer, L. J. & Magasanik, B. Initiation of transcription at the bacterial glnAp2 promoter by purified E. coli components is facilitated by enhancers. *Cell* **50,** 1039–46 (1987).
6. Levine, M. & Manley, J. L. Transcriptional repression of eukaryotic promoters. *Cell* **59,** 405–408 (1989).
7. Perros, M., Steitz, T. A., Fried, M. G., Hudson, J. M. & Lewis, M. DNA Looping and Lac Repressor-CAP Interaction. *Science* **274,** 1929–1932 (1996).
8. Hawley, D. K., Johnson, A. D. & McClure, W. R. Functional and physical characterization of transcription initiation complexes in the bacteriophage lambda OR region. *J. Biol. Chem.* **260,** 8618–8626 (1985).
9. Bertrand-Burggraf, E., Hurstel, S., Daune, M. & Schnarr, M. Promoter properties and negative regulation of the uvrA gene by the LexA repressor and its amino-terminal DNA binding domain. *J. Mol. Biol.* **193,** 293–302 (1987).
10. Chopra, V. S., Kong, N. & Levine, M. Transcriptional repression via antilooping in the Drosophila embryo. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 9460–9464 (2012).

11. Gray, S. & Levine, M. Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila. *Genes Dev.* **10,** 700–710 (1996).
12. Small, S., Blair, A. & Levine, M. Regulation of Two Pair-Rule Stripes by a Single Enhancer in theDrosophilaEmbryo. *Dev. Biol.* **175,** 314–324 (1996).
13. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Dev. Camb. Engl.* **122,** 205–14 (1996).
14. Kim, J. & Shapiro, D. J. In Simple Synthetic Promoters YY1-Induced DNA Bending Is Important in Transcription Activation and Repression. *Nucleic Acids Res.* **24,** 4341–4348 (1996).
15. Natesan, S. & Gilman, M. Z. DNA bending and orientation-dependent function of YY1 in the c-fos promoter. *Genes Dev.* **7,** 2497–2509 (1993).
16. Redd, M. J., Stark, M. R. & Johnson, A. D. Accessibility of alpha 2-repressed promoters to the activator Gal4. *Mol. Cell. Biol.* **16,** 2865–2869 (1996).
17. Atkinson, M. R., Pattaramanon, N. & Ninfa, A. J. Governor of the glnAp2 promoter of Escherichia coli. *Mol. Microbiol.* **46,** 1247–57 (2002).
18. Feng, J., Goss, T. J., Bender, R. A. & Ninfa, A. J. Repression of the Klebsiella aerogenes nac promoter. *J. Bacteriol.* **177,** 5535–5538 (1995).
19. Wassem, R., De Souza, E. M., Yates, M. G., Pedrosa, F. de O. & Buck, M. Two roles for integration host factor at an enhancer-dependent nifA promoter. *Mol. Microbiol.* **35,** 756–764 (2000).
20. Gilmour, D. S. Promoter proximal pausing on genes in metazoans. *Chromosoma* **118,** 1–10 (2009).
21. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94,** 890–898 (2005).
22. Carey, M. The enhanceosome and transcriptional synergy. *Cell* **92,** 5–8 (1998).
23. Hong, J.-W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. *Science* **321,** 1314 (2008).

24. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* 201518552 (2015). doi:10.1073/pnas.1518552112

25. Suzuki, H. I., Young, R. A. & Sharp, P. A. Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. *Cell* **168,** 1000–1014.e15 (2017).

26. Li, G. *et al.* Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* **148,** 84–98 (2012).

27. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38,** 1341–1347 (2006).

28. Amit, R., Garcia, H. G., Phillips, R. & Fraser, S. E. Building enhancers from the ground up: a synthetic biology approach. *Cell* **146,** 105–18 (2011).

29. Petrascheck, M. *et al.* DNA looping induced by a transcriptional enhancer in vivo. *Nucleic Acids Res.* **33,** 3743–3750 (2005).

30. de Bruin, D., Zaman, Z., Liberatore, R. A. & Ptashne, M. Telomere looping permits gene activation by a downstream UAS in yeast. *Nature* **409,** 109–113 (2001).

31. Mahmoudi, T., Katsani, K. R. & Verrijzer, C. P. GAGA can mediate enhancer function in trans by linking two separate DNA molecules. *EMBO J.* **21,** 1775–1781 (2002).

32. Dobi, K. C. & Winston, F. Analysis of Transcriptional Activation at a Distance in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **27,** 5575–5586 (2007).

33. Lainé, J.-P., Singh, B. N., Krishnamurthy, S. & Hampsey, M. A physiological role for gene loops in yeast. *Genes Dev.* **23,** 2604–2609 (2009).

34. Gruber, T. M. & Gross, C. A. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* **57,** 441–66 (2003).

35. Rappas, M., Bose, D. & Zhang, X. Bacterial enhancer-binding proteins: unlocking sigma54-dependent gene transcription. *Curr. Opin. Struct. Biol.* **17,** 110–6 (2007).

36. Schumacher, J., Zhang, X., Jones, S., Bordes, P. & Buck, M. ATP-dependent transcriptional activation by bacterial PspF AAA+protein. *J. Mol. Biol.* **338,** 863–75 (2004).

37. Su, W., Porter, S., Kustu, S. & Echols, H. DNA-looping and enhancer activity: association between DNA-bound NtrC activator and RNA polymerase at the bacterial glnA promoter. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 5504–8 (1990).

38. Davidson, E. H. *Genomic Regulatory Systems: In Development and Evolution*. 261 (Academic Press, 2001).

39. Magasanik, B. The regulation of nitrogen utilization in enteric bacteria. *J. Cell. Biochem.* **51,** 34–40 (1993).

40. Francke, C. *et al.* Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12,** 385 (2011).

41. Shimkets, L. J. Intercellular signaling during fruiting-body development of Myxococcus xanthus. *Annu. Rev. Microbiol.* **53,** 525–49 (1999).

42. Cheng, C. & Sharp, P. A. RNA polymerase II accumulation in the promoter-proximal region of the dihydrofolate reductase and gamma-actin genes. *Mol. Cell. Biol.* **23,** 1961–7 (2003).

43. Core, L. J. & Lis, J. T. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319,** 1791–2 (2008).

44. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. **39,** 1507–1511 (2007).

45. Miyamoto, T., Razavi, S., DeRose, R. & Inoue, T. Synthesizing Biomolecule-Based Boolean Logic Gates. *ACS Synth. Biol.* **2,** 72–82 (2013).

46. Friedland, A. E. *et al.* Synthetic Gene Networks that Count. *Science* **324,** 1199–1202 (2009).

47. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403,** 339–342 (2000).

48. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403,** 335–338 (2000).

49. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30,** 265–70 (2012).

50. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30,** 271–277 (2012).

51. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30,** 521–30 (2012).

52. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45,** 1021–1028 (2013).

53. The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity - Google Search. Available at: https://www.google.co.il/search?q=The+cis-regulatory+logic+of+Hedgehog+gradient+responses%3A+key+roles+for+gli+binding+affinity%2C+competition%2C+and+cooperativity&rlz=1C1GGRV_enIL756IL756&oq=The+cis-regulatory+logic+of+Hedgehog+gradient+responses%3A+key+roles+for+gli+binding+affinity%2C+competition%2C+and+cooperativity&aqs=chrome..69i57.435j0j7&sourceid=chrome&ie=UTF-8. (Accessed: 13th November 2017)

54. Extensive low-affinity transcriptional interactions in the yeast genome. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/16809671. (Accessed: 13th November 2017)

55. Rogers, K. W. & Schier, A. F. Morphogen gradients: from generation to interpretation. *Annu. Rev. Cell Dev. Biol.* **27,** 377–407 (2011).

56. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. M. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462,** 65–70 (2009).

57. Fakhouri, W. D. *et al.* Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. *Mol. Syst. Biol.* **6,** 341 (2010).

58. Atkinson, M. R., Savageau, M. A., Myers, J. T. & Ninfa, A. J. Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in Escherichia coli. *Cell* **113,** 597–607 (2003).

59. Amit, R., Garcia, H. G., Phillips, R. & Fraser, S. E. Building enhancers from the ground up: a synthetic biology approach. *Cell* **146,** 105–18 (2011).

60. Amit, R. Anti-Cooperative and Cooperative Protein-Protein Interactions between TetR Isoforms on Synthetic Enhancers. *J. Comput. Biol.* **19,** 115–125 (2012).

61. Hillen, W. & Berens, C. Mechanisms underlying expression of Tn10 encoded tetracycline resistance. *Annu. Rev. Microbiol.* **48,** 345–369 (1994).

62. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. **6,** 12–16 (2009).

63. Hammar, P. *et al.* The lac repressor displays facilitated diffusion in living cells. *Science* **336,** 1595–1598 (2012).

64. Jiang, Y. *et al.* Multigene editing in the Escherichia coli genome using the CRISPR-Cas9 system. *Appl. Environ. Microbiol.* AEM.04023-14 (2015). doi:10.1128/AEM.04023-14

65. Mosberg, J. A., Lajoie, M. J. & Church, G. M. Lambda Red Recombineering in Escherichia coli Occurs Through a Fully Single-Stranded Intermediate. *Genetics* **186,** 791–799 (2010).

66. Svenningsen, S. L., Waters, C. M. & Bassler, B. L. A negative feedback loop involving small RNAs accelerates Vibrio cholerae's transition out of quorum-sensing mode. *Genes Dev.* **22,** 226–238 (2008).

67. Tu, K. C. & Bassler, B. L. Multiple small RNAs act additively to integrate sensory information and control quorum sensing in Vibrio harveyi. *Genes Dev.* **21,** 221–233 (2007).

68. pubmeddev. Home - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/. (Accessed: 4th June 2017)

69. Miyashiro, T., Wollenberg, M. S., Cao, X., Oehlert, D. & Ruby, E. G. A single qrr gene is necessary and sufficient for LuxO-mediated regulation in Vibrio fischeri. *Mol. Microbiol.* **77,** 1556–1567 (2010).

70. Barrios, H., Valderrama, B. & Morett, E. Compilation and analysis of sigma(54)-dependent promoter sequences. *Nucleic Acids Res.* **27,** 4305–4313 (1999).

71. Lenz, D. H. *et al.* The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. *Cell* **118,** 69–82 (2004).

72. Kratky, O. & Porod, G. Röntgenuntersuchung gelöster Fadenmoleküle. *Recl. Trav. Chim. Pays-Bas* **68,** 1106–1122 (1949).

73. Pollak, Y., Goldberg, S. & Amit, R. Self-avoiding wormlike chain model for double-stranded-DNA loop formation. *Phys Rev E* **90,** 052602 (2014).

74. Beck, L. L., Smith, T. G. & Hoover, T. R. Look, no hands! Unconventional transcriptional activators in bacteria. *Trends Microbiol.* **15,** 530–537 (2007).

75. Ninfa, A. J. *et al.* Using two-component systems and other bacterial regulatory factors for the fabrication of synthetic genetic devices. *Methods Enzymol.* **422,** 488–512 (2007).

76. Lewis, M. *et al.* Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. *Science* **271,** 1247–1254 (1996).

77. Ramos, J. L. *et al.* The TetR Family of Transcriptional Repressors. *Microbiol. Mol. Biol. Rev.* **69,** 326–356 (2005).

78. Qin, Y., Keenan, C. & Farrand, S. K. N- and C-terminal regions of the quorum-sensing activator TraR cooperate in interactions with the alpha and sigma-70 components of RNA polymerase. *Mol. Microbiol.* **74,** 330–46 (2009).

79. Becker, N. a, Kahn, J. D. & Maher, L. J. Bacterial repression loops require enhanced DNA flexibility. *J. Mol. Biol.* **349,** 716–30 (2005).

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

81. Lee, D. H. & Schleif, R. F. In vivo DNA loops in araCBAD: size limits and helical repeat. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 476–80 (1989).

82. Müller, J., Oehler, S. & Müller-Hill, B. Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. *J. Mol. Biol.* **257,** 21–9 (1996).

83. Wray, L. V. & Reznikoff, W. S. Identification of repressor binding sites controlling expression of tetracycline resistance encoded by Tn10. *J. Bacteriol.* **156,** 1188–1191 (1983).

84. Qin, Y. *et al.* Quorum-sensing signal binding results in dimerization of TraR and its release from membranes into the cytoplasm. *EMBO J.* **19,** 5212–5221 (2000).

85. Bassler, B. L., Wright, M. & Silverman, M. R. Sequence and function of LuxO, a negative regulator of luminescence in Vibrio harveyi. *Mol. Microbiol.* **12,** 403–412 (1994).

86. Brunwasser-Meirom, M. *et al.* Using synthetic bacterial enhancers to reveal a looping-based mechanism for quenching-like repression. *Nat. Commun.* **7,** 10407 (2016).

87. Gilbert, L. A. *et al.* CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154,** 442–451 (2013).

88. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157,** 1262–1278 (2014).

89. Lloyd, L. J. *et al.* Identification of a New Member of the Phage Shock Protein Response in Escherichia coli, the Phage Shock Protein G (PspG). *J. Biol. Chem.* **279,** 55707–55714 (2004).

90. Opalka, N. *et al.* Structure of the Filamentous Phage pIV Multimer by Cryo-electron Microscopy. *J. Mol. Biol.* **325,** 461–470 (2003).

91. Genome Browser - Storz Lab: Section on Environmental Gene Regulation - science@NICHD. Available at: https://science.nichd.nih.gov/confluence/display/segr/Genome+Browser. (Accessed: 4th June 2017)

92. Umbarger, M. A. Chromosome conformation capture assays in bacteria. *Methods San Diego Calif* **58,** 212–220 (2012).

93. Griffith, J., Hochschild, A. & Ptashne, M. DNA loops induced by cooperative binding of lambda repressor. *Nature* **322,** 750–752 (1986).

94. Jeong, J., Le, T. T. & Kim, H. D. Single-molecule fluorescence studies on DNA looping. *Methods San Diego Calif* **105,** 34–43 (2016).

95. Cournac, A. & Plumbridge, J. DNA Looping in Prokaryotes: Experimental and Theoretical Approaches. *J. Bacteriol.* **195,** 1109–1119 (2013).

96. Horowitz, D. S. & Wang, J. C. Torsional rigidity of DNA and length dependence of the free energy of DNA supercoiling. *J. Mol. Biol.* **173,** 75–91 (1984).

97. Frolova, E., Majors, J. & Johnston, M. Binding of the glucose-dependent Mig1p repressor to the GAL1 and GAL4 promoters in vivo: Regulation by glucose and chromatin structure. *Nucleic Acids Res.* **27,** 1350–1358 (1999).

98. Jensen, N. B. *et al.* EasyClone: method for iterative chromosomal integration of multiple genes in Saccharomyces cerevisiae. *FEMS Yeast Res.* **14,** 238–248 (2014).

99. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27,** 38–52 (2017).

100. Hoover, T. R., Santero, E., Porter, S. & Kustu, S. The integration host factor stimulates interaction of RNA polymerase with NIFA, the transcriptional activator for nitrogen fixation operons. *Cell* **63,** 11–22 (1990).

101. Claverie-Martin, F. & Magasanik, B. Role of integration host factor in the regulation of the glnHp2 promoter of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 1631–1635 (1991).

102. Kiupakis, A. K. & Reitzer, L. ArgR-independent induction and ArgR-dependent superinduction of the astCADBE operon in Escherichia coli. *J. Bacteriol.* **184,** 2940–50 (2002).

103. Matthews, K. S. DNA looping. *Microbiol. Rev.* **56,** 123–136 (1992).

104. Probing Allostery Through DNA | Science. Available at: http://science.sciencemag.org/content/339/6121/816. (Accessed: 10th September 2017)

105. Johnson, S., Chen, Y.-J. & Phillips, R. Poly(dA:dT)-Rich DNAs Are Highly Flexible in the Context of DNA Looping. *PLOS ONE* **8,** e75799 (2013).

106. Natesan, S. & Gilman, M. Z. DNA bending and orientation-dependent function of YY1 in the c-fos promoter. *Genes Dev.* **7,** 2497–2509 (1993).

107. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 19498–19503 (2012).

108. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. (Academic Press, 2010).

109. Rowan, S. *et al.* Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. *Genes Dev.* **24,** 980–5 (2010).

110. RegulonDB Database. Available at: http://webcache.googleusercontent.com/search?q=cache:http://regulondb.ccg.unam.mx/&gws_rd=cr&dcr=0&ei=D2i2We-4GcjHgAaIyKDoBQ. (Accessed: 11th September 2017)

111. Lin, H., Deng, E.-Z., Ding, H., Chen, W. & Chou, K.-C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **42,** 12961–12972 (2014).

112. de Avila e Silva, S. *et al.* DNA duplex stability as discriminative characteristic for Escherichia coli σ(54)- and σ(28)- dependent promoter sequences. *Biol. J. Int. Assoc. Biol. Stand.* **42,** 22–28 (2014).

113. Maleki, A., Vaezinia, V. & Fekri, A. Promoter Prediction in Bacterial DNA Sequences Using Expectation Maximization and Support Vector Machine Learning Approach. *J. Data Min. Genomics Proteomics* **6,** (2015).

Appendix

1 Supplementary Note 1: looping in the context of the wormlike chain (WLC) model

1.1 The discrete WLC model

To evaluate the probability ratio $\hat{R}_n(N,k,s,...)$ theoretically we need to model the looping probabilities $P_{looped,n}(N,k,s,...)$ of the different enhancer configurations. We begin with the simple case of DNA looping without any proteins present. DNA is typically modeled as a discrete semi-flexible chain made of individual links of length l, such that the deviation of one link from its adjacent counterpart depends solely on some elastic bending energy. The chain links do not interact with each other. This class of polymer models is based on the original work of Kratky and Porod [12] and is referred to as the class of wormlike chain models (WLC). This class includes both discrete (i.e., a chain consists of a finite number of links with a certain length) and continuous models.

A chain is described by the locations of its links, and a local coordinate system defined by three orthonormal vectors $\hat{u},\hat{v},\hat{t}$ at each link. The vector $\hat{t}$ points along the direction of the chain links. For the continuous WLC these vectors are defined continuously along the chain contour. For the discrete WLC, a chain is defined by its joint locations $\mathbf{r}_i$, along with the local coordinate systems of all links. An example of a discrete WLC of five links, with unit link length is shown in Supp. Fig. 1A.

The elastic energy of the entire chain can be broken into a sum of contributions from individual chain links. The elastic energy of a single ith link in the chain consists of two contributions. The first contribution is the the elastic energy associated with bending link $i\in\{2,...,N\}$ relative to link $i-1$ with angles θ_i,ϕ_i (zenith and azimuthal angles in local spherical coordinates of the previous link). This is the conventional bending energy, which can be written as:

$$\beta E_i^{bend}=\frac{a}{2}\left|\hat{t}_i-\hat{t}_{i-1}\right|^2=a\left(1-\cos\theta_i\right), \tag{1}$$

assuming azimuthal symmetry, where $\beta=(k_bT)^{-1}$, T is the temperature, k_b is the Boltzmann factor, and a is the bending constant of the DNA chain. The second contribution accounts for the energy associated with twisting each link in the DNA double helix by Ω_i (twist angle) with respect to the previous link (i.e. rotating $\hat{u}_i$ relative to $\hat{u}_{i-1}$ at an angle of Ω_i). This term is modeled similarly to the term used to describe the twisting energy of a torsion spring [9]:

$$\beta E_i^{twist}=c\left(\Omega_i-\Omega_0\right)^2, \tag{2}$$

where c is the twisting rigidity constant, and where we denote the relaxed twisting angle of the DNA chain (the native twist of ≈ 1.81 radian per nm) by Ω_0. Consequently, the resulting elastic energy is

$$\beta E^{el}(\theta_i,\phi_i,\Omega_i)=a\left(1-\cos\theta_i\right)+c\left(\Omega_i-\Omega_0\right)^2. \tag{3}$$

Note that all the angles θ_i,ϕ_i,Ω_i are given in the local coordinate system of the $(i-1)$th link.

We number the links of a chain in the range $1..N$ for a chain of N links. For a specific configuration of the chain we introduce a notation $\{\theta_n,\phi_n,\Omega_n\}$ to denote the set of all the links' angles of the chain, from link 1 to link n. It follows then that the energy of the entire chain consisting of N links is given by

$$E\left(\{\theta_N, \phi_N, \Omega_N\}\right) = \sum_{i=2}^{N} E^{el}\left(\theta_i, \phi_i, \Omega_i\right). \tag{4}$$

The configurational partition function for the model DNA chain consisting of N links immediately follows:

$$Z_N = \int_{-1}^{1} d\cos\theta_2 \int_0^{2\pi} d\phi_2 \int_0^{2\pi} d\Omega_2 \cdots \int_{-1}^{1} d\cos\theta_N \int_0^{2\pi} d\phi_N \int_0^{2\pi} d\Omega_N \exp\left[-\beta E\left(\{\theta_N, \phi_N, \Omega_N\}\right)\right]. \tag{5}$$

This model has been extensively studied in the past [15, 8, 9].

1.2 The discrete WLC Monte-Carlo algorithm

The probability of looping for a given choice of parameters $P_{looped,n}\left(N, k, s, ...\right)$ can be calculated using a Monte-Carlo algorithm based on the importance sampling method [7] (as described in [13]). The algorithm generates a faithful statistical ensembles consisting of $N_c \approx 10^9$ DNA chains. Any physical observable can then be computed from the generated ensemble by

$$\langle f \rangle = \frac{\sum_{j=1}^{N_c} f\left(\{\theta_N, \phi_N, \Omega_N\}_j\right)}{N_c}. \tag{6}$$

The probability of looping is computed by $P_{looped} = \langle f_{looped} \rangle$ where

$$f_{looped}\left(\{\theta_N, \phi_N, \Omega_N\}_j\right) = \begin{cases} 1 & \text{configuration } j \text{ is looped} \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

This provides the basis for the self-avoiding wormlike chain model presented in the next section.

2 Supplementary Note 2: looping in the context of the self-avoiding wormlike chain (SAWLC) model

2.1 The discrete SAWLC model

Except for a few notable exceptions [2, 16], the WLC model does not take into account energetic and entropic effects that emerge from the cross-section or "thickness" of the DNA double helix. In order to model the effects of the effective DNA cross-section size, we must take into account an additional contribution to the elastic energy. We engulf each end-point of a link (a joint in the chain) by a "hard-wall" spherical shell of diameter w. An example of such a chain with three links (four joints), and $w = 2.5l$ is shown in Supp. Fig. 1B.

This allows us to model the final contribution to the elastic energy as a set of hard-wall potentials. We denote the end-point of link i as joint i. As previously mentioned, the chain links are numbered $1..N$. Joint 0 is the beginning terminus of the chain. For the simple case in which the chain link length l is larger than the chain diameter ($l \geq w$) and therefore no two neighboring hard-wall spheres overlap, the hard-wall potential energy for the ith chain link is defined as:

$$E_i^{hw}\left(\{\theta_i, \phi_i, \Omega_i\}\right) = \begin{cases} \infty & \text{joint } i \text{ overlaps with one or more joints } 0..(i-1) \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

This allows us to write an expression for the total elastic energy associated with the chain of spheres as follows:

$$E\left(\{\theta_N,\phi_N,\Omega_N\}\right)=\sum_{i=2}^{N}E^{el}\left(\theta_i,\phi_i,\Omega_i\right)+\sum_{i=1}^{N}E_i^{hw}\left(\{\theta_i,\phi_i,\Omega_i\}\right). \tag{9}$$

We term this model the self-avoiding wormlike chain model (SAWLC). Similarly, Eq. (5) holds for the SAWLC, with the substitution of (9) for (4).

Performing the substitutions (9), (8) into (5) and opening the sums yields:

$$\begin{aligned} Z_N &= \int_{-1}^{1}d\cos\theta_1\int_{0}^{2\pi}d\phi_1\int_{0}^{2\pi}d\Omega_1\exp\left[-\beta E^{el}\left(\theta_1,\phi_1,\Omega_1\right)\right]\Theta_i^{hw}\left(\{\theta_1,\phi_1,\Omega_1\}\right)\cdots \\ &\cdots\int_{-1}^{1}d\cos\theta_N\int_{0}^{2\pi}d\phi_N\int_{0}^{2\pi}d\Omega_N\exp\left[-\beta E^{el}\left(\theta_N,\phi_N,\Omega_N\right)\right]\Theta_N^{hw}\left(\{\theta_N,\phi_N,\Omega_N\}\right), \end{aligned} \tag{10}$$

where

$$\Theta_i^{hw}\left(\{\theta_i,\phi_i,\Omega_i\}\right)=\begin{cases}0 & \text{joint } i \text{ overlaps one or more joints } 0..(i-1)\\ 1 & \text{otherwise}\end{cases}, \tag{11}$$

which in turn can be written as:

$$Z_N=\int\limits_{no\ overlap}d\cos\theta_1 d\phi_1 d\Omega_1\exp\left[-\beta E^{el}\left(\theta_1,\phi_1,\Omega_1\right)\right]\cdots\int\limits_{no\ overlap}d\cos\theta_N d\phi_N d\Omega_N\exp\left[-\beta E^{el}\left(\theta_N,\phi_N,\Omega_N\right)\right]. \tag{12}$$

In the case in which the chain link length l is smaller than the chain diameter ($l<w$), two or more consecutive spheres overlap. This is resolved by introducing Δi such that there can be an interaction between links j,k only if $|j-k|\geq\Delta i\geq\frac{w}{l}$. This changes (11) to:

$$\Theta_i^{hw}\left(\{\theta_i,\phi_i,\Omega_i\}\right)=\begin{cases}0 & \text{joint } i \text{ overlaps one or more joints } 0..(i-\Delta i)\\ 1 & \text{otherwise}\end{cases}. \tag{13}$$

Recently, we made a significant contribution to the understanding of polymer cyclization or looping by providing a detailed analysis of this phenomenon in the context of the SAWLC [13]. In this work, we first confirmed numerically previous renormalization group predictions [2, 16, 10] and scaling theory [6] models, and subsequently provided new numerical analysis that was applicable to resolving the hyper-bendability effect observed in DNA cyclization experiments for short dsDNA [3, 17]. The algorithm used for the polymer cyclization analysis forms the basis for the algorithm that we describe below.

2.2 The discrete SAWLC Monte-Carlo algorithm

To faithfully sample the configurational space of the SAWLC model we utilized a Monte-Carlo algorithm based on the weighted-biased sampling method [4] to generate statistical ensembles consisting of $N_c\approx10^7-10^9$ DNA chains. In brief, the algorithm generates chains by growing them one link at a time, while systematically checking that the new link does not cross any previous links in the chain. When such a crossing occurs, this fact is taken into account by updating the weighting

factor of the chain W. This counter-weight balances the over/under-representation of that chain in the ensemble. For a more detailed description of the method see [13].

Any physical observable f can then be estimated from the generated ensemble using the following formula:

$$\langle f \rangle = \frac{\sum_{j=1}^{N_c} f\left(\{\theta_N, \phi_N, \Omega_N\}_j\right) W\left(\{\theta_N, \phi_N, \Omega_N\}_j\right)}{\sum_{j=1}^{N_c} W\left(\{\theta_N, \phi_N, \Omega_N\}_j\right)}, \tag{14}$$

which includes the probability of looping defined by Eq. 7.

2.3 Protein-DNA interactions and looping in the context of the SAWLC model

2.3.1 Modeling a DNA chain with a bound protein

In order to expand our basic self-avoiding simulation of polymer configurations to model a nucleoprotein structure made of dsDNA and proteins, we first consider some broad geometric characteristics of dsDNA structure. dsDNA is composed of two interwound helical strands, which form a double-helix backbone linked by base-pairs. The backbone exhibits major and minor grooves. The major groove is the wider groove between the two, while the minor groove corresponds to the situation where strands are closer together on the relevant side of the double helix than on the other.

Since most DNA binding proteins bind onto the major groove of the DNA, we must continuously keep track of the location of the major groove along the DNA to properly model protein-DNA binding. To that end, we use one of the vectors defining the local coordinate system of each link: $\hat{u}_i$ (Supp. Fig. 1A). On each link, $\hat{u}_i$ points from the center of the chain's cross-section to the center of the major groove. As a result, we model a protein bound to a binding site of n basepairs, with m being the index of the first basepair of the binding site, by a protrusion from the chain shaped like a sphere with some diameter $w_{protein}$ representative of the protein's volume (Supp. Fig. 1C):

$$\mathbf{r}_{protein} = \mathbf{r}_c + \left(\frac{w + w_{protein}}{2}\right) \hat{u}_c, \tag{15}$$

where we defined the center link of the binding site as

$$c \equiv m + n/2. \tag{16}$$

Note that $\mathbf{r}_c$ is the joint at the end of link c.

2.3.2 Monte-Carlo simulation of a DNA chain with a bound protein

In order to adapt our algorithm to the case of protein-bound DNA, we have to take into account not only the growing chain but also the location of the protrusion as defined by (Eq. 15). During the application of the Monte-Carlo algorithm, upon reaching link $(m + n/2)$ in a specific chain, the simulation attempts to add a hard-wall sphere with diameter $w_{protein}$ at the location $\mathbf{r}_{protein}$. If the sphere overlaps any of the previously generated objects (chain links or other proteins) the chain is discarded, and a new chain is grown from the beginning. The excluded volume of the new sphere is taken into account, when subsequent links are generated. Only completed nucleoprotein chains are properly weighted and counted within the new nucleoprotein configurational distribution.

2.3.3 Modeling bending and stiffening of a thick chain

The excluded volume of the protein bound to a binding site of n basepairs, with m being the index of the first basepair of the binding site, is introduced via a spherical protrusion, as described in section 2.3.1 for the general case. Local DNA bending and stiffening effects induced by the bound protein are simulated via changes applied to the $[m, m+n-1]$ chain links as they are being generated.

Local stiffening of the DNA is simulated by using a different bending constant $a' > a$ (Eq. 1).

Local bending of the DNA by an angle of κ is simulated by introduction of a small rotation of the local coordinates system at each link in the protrusion binding site by κ/n around $\hat{v}_c$ (the unit vector $\hat{v}$ of the center link). This results in a cumulative bending of the chain "around" the bound protein (Supp. Fig. 1D). The energy associated with this additional rotation is zero. To properly generate the bend using our Monte-Carlo approach, the simulation estimates the value of $\hat{v}_c$ of the center link of the binding site for all binding-site links $i \in [m, m+n-1]$ by:

$$\hat{v}_c \approx R_{(c-i)2\pi/P,\hat{t}_i}\hat{v}_i, \tag{17}$$

where P is the helical repeat of the chain ($P \approx 10.5$ for DNA) and $R_{(c-i)2\pi/P,\hat{t}_i}$ is a rotation matrix by an angle $(c-i)\frac{2\pi}{P}$ around $\hat{t}_i$. This approximation is valid for the simulation of DNA since $n/2$ is smaller than the bend and twist persistence lengths of $\sim 50-100$ nm. by at least an order of magnitude (see Simulation Parameters in Materials and Methods). Thus, at these length scales the DNA is expected to be fairly straightand each successive link i is rigidly twisted by $\sim \frac{2\pi}{P}$ around $\hat{t}_i$ relative to link $i-1$.

2.3.4 Looping criterion

After numerically generating the configurational ensemble, the subset of "looped" configurations needs to be properly defined numerically. In the basic cyclization model, looping was defined [13] by having both ends of the polymer within some distance δ. However, unlike the DNA cyclization experiments in which the DNA chain segment simply closes on itself, σ^{54} transcription initiation requires the interaction of two proteins bound to the chain's ends. In particular, the driver interacts with the σ^{54} factor located directly on the DNA beneath the polymerase, thus requiring the driver to access the polymerase complex from the bottom [14]. Therefore, to faithfully model this process via looping, we must add at least two proteins to the ends of the chain - the driver and the polymerase. For simplicity, we neglect the volume of the σ^{54}. We define a looped state as one in which the driver protein interacts with the σ^{54} factor in only a subset of the possible solid angles, corresponding to the σ^{54} acceptance cone and the extent of the GAFTGA loop of NtrC [14]. Thus, the criteria for a looped driver-promoter interaction as they are shown in Supp. Fig. 1E, are as follows:

1. The driver and the σ^{54} factor must be in close proximity, defined by maximal separation of $|\mathbf{r}_{drv} - \mathbf{r}_{\sigma^{54}}| < \frac{w_{drv}}{2} + \varepsilon$.
2. The direction from the DNA to σ^{54} factor is collinear with the the direction from the σ^{54} factor to the driver, within the range $\delta\omega$. I.e. $(\mathbf{r}_{drv} - \mathbf{r}_{\sigma^{54}})$ is collinear with $-\hat{u}_1$ within the range $\delta\omega$. (Conditions 1-2 define a volume $\delta\mathbf{r}$ in which $\mathbf{r}_{drv}$ must be located, illustrated as a green cone in Supp. Fig. 1E).
3. The direction from the DNA to the driver is collinear with the the direction from the driver to the σ^{54} factor, within the range $\delta\omega$. I.e. $(\mathbf{r}_{\sigma^{54}} - \mathbf{r}_{drv})$ is collinear with $\hat{u}_N$ within the range $\delta\omega$ (Illustrated as a light blue cone with its base in $\mathbf{r}_{drv}$ centered around $\hat{u}_N$ in Supp. Fig. 1E).

Using this definition, our algorithm can compute $P_{looped,0}(N)$ for a bacterial enhancer-promoter system of looping length N.

2.3.5 Modeling looping of a thick chain with a bending or stiffening protrusion

Any configuration of bound proteins can be simulated in this way, allowing us to compute $P_{looped,n}(N,k,s,...)$ in a straightforward fashion. For example, we consider a single TetR transcription factor bound to an enhancer of length 100 bp. The TetR transcription factor is roughly spherical and can be modeled well by a spherical protrusion with diameter of $w_{TetR} = 5.44$ nm, and with a binding site size of $n = 18$ bp, on an enhancer-promoter chain that is $N = 500$ bp long. Consequently, such a protein bound to a binding site spanning links $25-42$ is modeled by a sphere located at $\mathbf{r}_{TetR} = \mathbf{r}_{34} + \left(\frac{w+w_{TetR}}{2}\right)\hat{u}_{34}$, according to Eq. (15). Additional stiffening of the DNA by a factor of two is simulated by using $a' = 2a$ during the generation of the links $25-42$. Additional bending of the DNA by $\kappa = 10^\circ$ is simulated by 18 instantaneous rotations of the local coordinate system by $\frac{10^\circ}{18}$ during generation of the links $25-42$ around the orientation of $\hat{v}_{34}$, approximated at each link $25 \le i \le 42$ as described by Eq. (17). This then allows us to generate the properly weighted nucleoprotein ensemble, and compute $P_{looped,n=1}(N=500,k=34)$ in a straightforward fashion as defined above.

3 Supplementary Note 3: modeling the experiment

3.1 Thermodynamic modeling of looping-initiated transcription

Our proposed numerical model is capable of calculating the probability of looping for a given experimental setup. The output of the *in vivo* looping experiments, however, is measured indirectly by monitoring the resultant reporter protein fluorescence level when the cells reach steady-state. We need to derive a model that connects the experimental readout to the probability of looping. Since we measure the average fluorescence for a population of cells in our experiments, we only need to provide a prediction for the mean expression levels in our simulations. While a full stochastic model (e.g., [**?**]) can provide a prediction for the second and higher moments of the distribution, it is has been shown that the stochastic model's expression for the first moment is equivalent to the mean expression level obtained from a thermodynamic equilibrium model, that is constructed using the same assumptions [**?**].

To model a minimal enhancer, which consists of only the driver binding site and the promoter, we make the following assumptions:

- We assume that the glnAp1 promoter is only active when the concentration of $NRI \sim P$ is vanishingly small (for a justification of this assumption see [1]). When a small amount of $NRI \sim P$ accumulates, the hexameric complex assembles, which simultaneously strongly represses the glnAp1 promoter while activating glnAp2. This means that all transcription in our system is a result of the glnAp2 promoter.

- There is always a bound "poised" polymerase at the promoter awaiting an activation signal. While it was recently shown that the RNAP releases from and rebinds to the promoter frequently [5], it was also previously shown that σ^{54} promoters are mostly occupied by poised polymerases both *in vivo* and *in vitro* [11**?** **?**]. This means that we do not need to include states that lack bound RNAP in our model.

- For the $NRI \sim P$ hexamerization, a cooperative process, the appropriate expression for equilibrium binding is given by:

$$\frac{\left(\frac{[NRI]}{K_{NRI}}\right)^m}{1+\left(\frac{[NRI]}{K_{NRI}}\right)^m}, \tag{18}$$

where $[NRI]$ is the concentration of phosphorylated $NRI \sim P$ dimers, K_{NRI} is the $NRI \sim P$ dissociation constant that incorporates the cooperativity of the binding interaction, and $m > 1$ is some coefficient that signifies the multimerization of $NRI \sim P$ at the two NRI sites. One can expect m to be as high as 6, but it could also be lower since $NRI \sim P$ is a dimer in solution. Hence, we expect $3 < m < 6$ [**?**]. The subsequent constant production of $NRI \sim P$ in our experiments and the cooperative binding allows us to assume that:

$$\left(\frac{[NRI]}{K_{NRI}}\right)^m \gg 1, \tag{19}$$

which allows us to posit that a driver complex [i.e. $(NRI \sim P)^m$] is always bound at the driver binding sites.

Given these assumption, we only need to model two states: the driver-and-RNAP-occupied enhancer-promoter non-looped state, and the transcriptionally active looped state.

- Finally, we assume that the rates of driver binding, looping, and unlooping are much faster than the subsequent rates involved in transcription. This means that before ATP can be hydrolyzed and an open complex be formed at the promoter, the driver-DNA-σ^{54} complex has sufficient time to equilibrate. This in turn means that the DNA-bound driver complex has sufficient time to explore its conformational space. This assumption is supported by kinetic data recently obtained by [5]. This means that the two relevant states are in thermodynamic equilibrium, and can be modeled accordingly.

Given these assumptions, we begin by writing the rate equation, which describes the kinetics of looping-initiated transcription for a single bacterial enhancer:

$$\frac{d}{dt}[mRNA] = \alpha P_{int}(N) - \beta [mRNA], \tag{20}$$

where $P_{int}(N)$ is the probability of the driver complex bound N links from the polymerase to interact with the polymerase, $[mRNA]$ is the mRNA concentration, α is the rate for transcription per unit volume after the looped structure between the polymerase and driver has formed, and β is an mRNA degradation rate constant. In steady state the fluorescence reporter level is proportional to the number of mRNA transcripts, resulting in:

$$Fl \propto [mRNA] = \frac{\alpha}{\beta} P_{int}(N), \tag{21}$$

where Fl corresponds to reporter protein concentration (or fluorescence level readout). Often the exact values of α, β and the proportion of Fl to $[mRNA]$ are not known.

The probability of the driver and the polymerase to be in close enough proximity to interact is given by:

$$P_{looped}(N) = \frac{\int\limits_{\substack{\text{looped}\\ \text{configurations}}} \exp[-\beta E_{conf}]\, d^N\xi_i}{\int\limits_{\substack{\text{all}\\ \text{configurations}}} \exp[-\beta E_{conf}]\, d^N\xi_i}, \tag{22}$$

where ξ_i are the N coordinates that define a conformation, E_{conf} is the energy of a configuration, and $\beta = (k_b T)^{-1}$, where T is the temperature and k_b is the Boltzmann constant. What actually constitutes a "looped configuration" is defined in 2.3.4. In order to write down a simpler expression for the probability of looping, we denote for convenience a general expression for a partial partition function constrained by some condition:

$$\mathcal{Z}_{condition}(N) = \int_{\substack{\text{configurations that} \\ \text{satisfy the condition}}} \exp\left[-\beta E_{conf}\right] d^N \xi_i, \tag{23}$$

where the integral is taken only over those configurations that satisfy the specific condition. This allows us to define the looping probability functions as follows:

$$P_{looped}(N) = \frac{\mathcal{Z}_{looped}(N)}{\mathcal{Z}_{all}(N)} = \frac{\mathcal{Z}_{looped}(N)}{\mathcal{Z}_{looped}(N) + \mathcal{Z}_{non-looped}(N)}. \tag{24}$$

To develop an expression for the interaction probability, we add a configuration-independent driver-polymerase interaction energy denoted by E_{nr} to the energy of a looped configuration E_{conf} . This allows us to define the interaction partition functions as follows:

$$\mathcal{Z}_{interacting}(N) = e^{\beta E_{nr}} \mathcal{Z}_{looped}(N), \tag{25}$$

leading to the following expression for the probability of interactions:

$$P_{int,0}(N) = \frac{e^{\beta E_{nr}} \mathcal{Z}_{looped}(N)}{e^{\beta E_{nr}} \mathcal{Z}_{looped}(N) + \mathcal{Z}_{non-looped}(N)}, \tag{26}$$

where the index 0 corresponds to the number of TF binding sites in the enhancer. Since $\mathcal{Z}_{non-looped}(N) \approx \mathcal{Z}_{all}(N)$, by dividing the numerator and denominator by $\mathcal{Z}_{all}(N)$ we arrive at:

$$P_{int,0}(N) \approx \frac{e^{\beta E_{nr}} P_{looped}(N)}{1 + e^{\beta E_{nr}} P_{looped}(N)} = \frac{\frac{P_{looped}(N)}{K_{nr}}}{1 + \frac{P_{looped}(N)}{K_{nr}}}, \tag{27}$$

where $K_{nr} \equiv \exp(-\beta E_{nr})$ is the dissociation constant of the protein-protein interaction in the looped conformation . For convenience we define the "looping capacity" as:

$$\chi(N) \equiv \frac{P_{looped}(N)}{K_{nr}}, \tag{28}$$

resulting in a compact result for the minimal enhancer interaction probability:

$$P_{int,0}(N) \approx \frac{\chi_0(N)}{1 + \chi_0(N)}. \tag{29}$$

Note, that Amit et al. [1] demonstrated experimentally that the model described above and summarized in Eq. (29) adequately describes the transcriptional kinetics of our $(NRI \sim P)^6 - \sigma^{54}$ system. In addition, Friedman et al. [5] used single molecule kinetics experiments to show that assumptions above are valid via careful measurements of every rate constant in the enhancer-poised promoter complex formation and subsequent transcription initiation.

3.2 Connecting thermodynamic model and experiment

By normalizing the results of one experiment by those of another experiment (see Eq. (21)) we can derive an experimentally measurable expression that both eliminates the need to experimentally determine $\frac{\alpha}{\beta}$ and corresponds to the ratio of the driver-polymerase interaction probabilities $P_{int}^{(i)}$ for the experiments $i = 1, 2$:

$$\frac{Fl^{(1)}}{Fl^{(2)}} = \frac{P_{int}^{(1)}(N)}{P_{int}^{(2)}(N)}. \tag{30}$$

Below, we demonstrate that under certain assumptions that hold for our experimental setup, $\frac{P_{int}^{(1)}(N)}{P_{int}^{(2)}(N)}$ can be replaced by the ratio of the looping probabilities for the two experiments $i = 1, 2$.

Based on the kinetic measurements that were made by [5], we can now estimate the magnitude of $P_{int,0}(N)$. We begin with the rate equations described in their model [5, Fig.7]:

$$\begin{aligned}
\frac{d}{dt}[RP_1] &= 0 = k_1[DNA][\sigma^{54}] - (k_{-1} + k_2)[RP_1] + k_{-2}[RP_2], \\
\frac{d}{dt}[RP_2] &= 0 = k_2[RP_1] - (k_{-2} + k_3)[RP_2] + k_5[RP_0], \\
\frac{d}{dt}[RP_0] &= 0 = k_3[RP_2] - (k_5 + k_4)[RP_0], \\
\frac{d}{dt}[\sigma^{54}] &= 0 = -k_1[DNA][\sigma^{54}] + k_{-1}[RP_1] + k_4[RP_0],
\end{aligned} \tag{31}$$

where $[\sigma^{54}]$ is the concentration of available σ^{54}, RP_1 and RP_2 are two types of closed complexes in which DNA remains base-paired, and RP_0 is the open complex in the looped configuration (interacting state). Solving for the fraction of DNA in interacting state, and plugging in the values found by [5], we can extract the probability for finding the driver-polymerase in a bound-looped state RP_0 as follows:

$$P_{int,0} = \frac{[RP_0]}{[RP_0] + [RP_1] + [RP_2]} = \frac{k_2 k_3}{k_2(k_3 + k_4 + k_5) + k_3 k_4 + k_{-2}(k_4 + k_5)} = 0.0101. \tag{32}$$

This implies (see Eq. (29)) that $P_{int,0}(N) \approx \frac{\chi_0(N)}{1+\chi_0(N)} \ll 1$, and therefore

$$P_{int,0}(N) \approx \frac{\chi_0(N)}{1 + \chi_0(N)} \approx \chi_0(N). \tag{33}$$

Using Eqs. (30), (33) and (28), we obtain

$$\frac{Fl^{(1)}}{Fl^{(2)}} = \frac{P_{int}^{(1)}(N)}{P_{int}^{(2)}(N)} \approx \frac{P_{looped}^{(1)}(N)}{P_{looped}^{(2)}(N)} \tag{34}$$

for the experiments i=1,2. This expression enables direct comparison of our experimental measurements to the numerical simulations. Finally, we note that since [5] used DNA segments of comparable lengths to the ones modeled by us, we will assume that Eq. (33) and the assumption

$$\chi_n(N) \ll 1 \tag{35}$$

hold for all looping capacities (all n) considered in this work.

3.3 Enhancer with transcription factor (TF) binding site

When the enhancer contains one TF binding site ($n = 1$) we can write the probability of interaction in a similar fashion to Eq. (27):

$$P_{int,1}\left(N,k,[TF]\right)=\frac{e^{\beta E_{nr}}\mathcal{Z}_{\substack{\text{looped}\\\text{no TF}}}\left(N\right)+\frac{[TF]}{K_{TD}}e^{\beta E_{nr}}\mathcal{Z}_{\substack{\text{looped}\\\text{bound TF}}}\left(N,k\right)}{e^{\beta E_{nr}}\mathcal{Z}_{\substack{\text{looped}\\\text{no TF}}}\left(N\right)+\frac{[TF]}{K_{TD}}e^{\beta E_{nr}}\mathcal{Z}_{\substack{\text{looped}\\\text{bound TF}}}\left(N,k\right)+\mathcal{Z}_{\substack{\text{non-looped}\\\text{no TF}}}\left(N\right)+\frac{[TF]}{K_{TD}}\mathcal{Z}_{\substack{\text{non-looped}\\\text{bound TF}}}\left(N,k\right)}, \tag{36}$$

where $k < N$ is the number of base pairs between the center of the driver binding site and the center of the TF binding site and K_{TD} is the binding constant of the TF to its binding site. Again, by dividing the numerator and the denominator by $\mathcal{Z}_{all}\left(N\right)$:

$$P_{int,1}\left(N,k,[TF]\right)\approx\frac{\chi_0\left(N\right)+\frac{[TF]}{K_{TD}}\chi_1\left(N,k\right)}{1+\chi_0\left(N\right)+\frac{[TF]}{K_{TD}}\left(1+\chi_1\left(N,k\right)\right)}, \tag{37}$$

where χ_0 and χ_1 denote the looping capacities for an enhancer without a bound TF (equal to the looping capacity of the minimal enhancer) and the looping capacity for an enhancer with a TF bound to it, respectively. To quantify the effect of TF binding on transcription, we divide the reporter concentration with available TF by the reporter concentration obtained without TF . Following Eqs. (30), (29) and (37) we obtain:

$$r_1\left(N,k,[TF]\right)\equiv\frac{Fl_1\left(N,k,[TF]\right)}{Fl_0\left(N\right)}=\frac{P_{int,1}\left(N,k,[TF]\right)}{P_{int,0}\left(N\right)}\approx\frac{1+\frac{[TF]}{K_{TD}}\frac{\chi_1(N,k)}{\chi_0(N)}}{1+\frac{[TF]}{K_{TD}}\left(\frac{1+\chi_1(N,k)}{1+\chi_0(N)}\right)}. \tag{38}$$

By performing the experiment with saturating concentrations of the TF we can quantify the maximal regulatory effect:

$$R_1\left(N,k\right)\equiv\lim_{[TF]\to\infty}r_1\left(N,k,[TF]\right)\approx\frac{\frac{\chi_1(N,k)}{\chi_0(N)}}{\frac{1+\chi_1(N,k)}{1+\chi_0(N)}}. \tag{39}$$

Further simplifying with Eqs. (33) and (35) yields:

$$R_1\left(N,k\right)\approx\frac{\chi_1\left(N,k\right)}{\chi_0\left(N\right)}\approx\frac{P_{looped,1}\left(N,k\right)}{P_{looped,0}\left(N\right)}\equiv\hat{R}_1\left(N,k\right), \tag{40}$$

where $P_{looped,n}$ is the probability of looping with n TFs bound. Eq. (40) thus provides the connection between the experimental expression level ratio of saturating and zero TF concentrations $R_1\left(N,k\right)$ and the theoretical looping probability ratio $\hat{R}_1(N,k)$. The latter can be obtained directly from the looping probabilities $P_{looped,1}(N,k)$ and $P_{looped,0}(N)$, which are computed using the numerical Monte-Carlo simulation.

3.4 Enhancer with two TF binding sites

In this case ($n = 2$), we define an additional parameter s to denote the number of base pairs between the center of the first TF binding site and the center of the second TF binding site. In a similar fashion to the one binding site case, we obtain:

$$P_{int,2}\left(N,k,s,[TF]\right)\approx\frac{\chi_0\left(N\right)+\frac{[TF]}{K_{TD}}\chi_1\left(N,k\right)+\frac{[TF]}{K_{TD}}\chi_1\left(N,k+s\right)+\left(\frac{[TF]}{K_{TD}}\right)^2\chi_2\left(N,k,s\right)}{1+\chi_0\left(N\right)+\frac{[TF]}{K_{TD}}\left(1+\chi_1\left(N,k\right)+\chi_1\left(N,k+s\right)\right)+\left(\frac{[TF]}{K_{TD}}\right)^2\left(1+\chi_2\left(N,k,s\right)\right)}, \tag{41}$$

$$\begin{aligned} r_2\left(N,k,s,[TF]\right) &\equiv \frac{Fl_2\left(N,k,s,[TF]\right)}{Fl_0\left(N\right)}=\frac{P_{int,2}\left(N,k,s,[TF]\right)}{P_{int,0}\left(N\right)} \\ &\approx \frac{1+\frac{[TF]}{K_{TD}}\frac{\chi_1(N,k)+\chi_1(N,k+s)}{\chi_0(N)}+\left(\frac{[TF]}{K_{TD}}\right)^2\frac{\chi_2(N,k,s)}{\chi_0(N)}}{1+\frac{[TF]}{K_{TD}}\left(\frac{1+\chi_1(N,k)+\chi_1(N,k+s)}{1+\chi_0(N)}\right)+\left(\frac{[TF]}{K_{TD}}\right)^2\left(\frac{1+\chi_2(N,k,s)}{1+\chi_0(N)}\right)}, \end{aligned} \tag{42}$$

$$R_2\left(N,k,s\right)\equiv\lim_{[TF]\to\infty}r_2\left(N,k,s,[TF]\right)\approx\frac{\frac{\chi_2(N,k,s)}{\chi_0(N)}}{\frac{1+\chi_2(N,k,s)}{1+\chi_0(N)}}\approx\frac{P_{looped,2}\left(N,k,s\right)}{P_{looped,0}\left(N\right)}\equiv\hat{R}_2\left(N,k,s\right). \tag{43}$$

Here $\chi_2(N,k,s)$ is the looping capacity for an enhancer of length N with two TFs bound at k and $(k+s)$. As with Eq. (40), Eq. (43) provides the connection between the experimental observations and the looping probabilities calculated by our simulations.

4 Supplementary Note 4: supporting data

4.1 Supplementary model data for Figure 1: A geometric model for a maximal regulatory effect at the $k = {}^{N}/_{2}$ position

In order to provide an explanation for the non-monotonic increase of the regulatory effect as k approaches $N/2$ observed both for our model (Fig. 1) and experiments with TraR (Fig. 2B, 3C), LacI-GST (Fig. 4A), and TetR (Figure 2D), we chose to analyze the most-likely looped configuration ζ_0 predicted by our model. For simplicity, we neglected the volume of the chain, i.e., we used the WLC model. Discrete minimal-energy looped configurations were found for $N = 10, 20, 50$ as the numerical solutions that minimize the total configuration energy of Eq. (4), with boundary conditions that satisfy the ideal looping criterion, namely that the chain ends coincide. For comparison, solutions were approximated to continuous curves $\tilde{\mathbf{r}}_N(uNl)$ with $u \in [0,1]$ and then renormalized to unit length $\mathbf{r}_N\left(u\right)=\tilde{\mathbf{r}}_N\left(uNl\right)/Nl$. The curves for the different N values were found to collapse onto a single curve $\mathbf{r}_{\zeta_0}\left(u\right)$, indicating that the minimal energy loop ζ_0 is independent of N.

Supp. Fig. 2A shows ζ_0 on a unit length curve, parametrized by $u \in [0,1]$, in the loop plane. Note that the minimal energy loop for the 3D problem is in fact two-dimensional. The figure shows that the minimal-energy (or most-likely-to-occur) loop is not a circle as might conventionally be thought, but rather shaped like a teardrop. The implications of this geometry can be seen in Supp. Fig. 2B, where we plot the numerically-calculated squared curvature $\left|\frac{dt_{\zeta_0}}{du}\right|^2$ of the minimal energy loop ζ_0 as a function of the normalized length u. The figure shows that the region of highest curvature is at $u=\frac{1}{2}$, precisely half-way along the contour, whereas the curvature near both ends is very small. This indicates that it is energetically preferential for loops to be more curved near the center of the loop than close to the ends, and loops whose shapes are the same as or closely resemble the minimal loop ζ_0 will be most probable.

We now calculate the curvature-dependence of the energy of link j, for a particular configuration ζ. This is certainly valid in the elastic regime, i.e. for chains whose length is $N \lesssim b$. From Eq. (1):

$$E_j^{\zeta} = \frac{a}{2}\left|\hat{t}_{j+1} - \hat{t}_j\right|_{\zeta}^2 = a\left(1 - \cos\theta_{j+1}\right). \tag{44}$$

In the limit of $N \gg 1$, the discrete curve along the points $\{\mathbf{r}_0, ..., \mathbf{r}_N\}$ can be approximated with a continuous curve $\tilde{\mathbf{r}}(s)$ parametrized by $s \in [0, Nl]$, which yields:

$$\frac{\left|\hat{t}_{j+1} - \hat{t}_j\right|^2}{l^2} \approx \left|\frac{d\tilde{\mathbf{t}}(s)}{ds}\right|_{s=jl}^2. \tag{45}$$

After renormalizing $\tilde{\mathbf{r}}(s)$ to a unit length curve $\mathbf{r}_{\zeta}(u) = \tilde{\mathbf{r}}(s)/Nl$ parametrized by $u \in [0, 1]$, we obtain:

$$\left|\hat{t}_{j+1} - \hat{t}_j\right|^2 \approx \frac{1}{N^2}\left|\frac{d\mathbf{t}_{\zeta}(u)}{du}\right|_{u=j/N}^2 = \frac{1}{N^2}\kappa_{\zeta}^2(u), \tag{46}$$

where $\kappa_{\zeta}(u) = \left|\frac{d\mathbf{t}_{\zeta}(u)}{du}\right|$ is the curvature, which depends only on the geometric shape of the specific configuration ζ. Thus the energy contribution at the jth link is approximated by:

$$E_j^{\zeta} \approx \frac{a}{2}\frac{1}{N^2}\kappa_{\zeta}^2\left(\frac{j}{N}\right). \tag{47}$$

We now ask where to expect the maximal regulatory effect for a bound TF, based on the shape of ζ_0 and on Eq. (47). Since the energy depends on the squared curvature $\kappa_{\zeta}^2\left(\frac{j}{N}\right)$ and the curvature was found numerically to be maximal at $j = N/2$, we expect a TF at the $N/2$ position to produce the greatest change in energy, and thus to produce the strongest regulatory effect. We consider the separate effects of the bound TF. Stiffening (increasing a) hinders curving in all directions (including the loop-forming direction), which increases the energy of the ζ_0 -like configurations and results in a down-regulatory effect. Bending either assists curving in the loop-forming direction (reduces energy) or hinders curving (increasing energy), depending on the direction of the bending, resulting in an up- or down-regulatory effect, respectively. A protrusion hinders curving in the loop-forming direction if it is positioned "within the loop" since curvature is limited by the presence of the protrusion. Similarly, a protrusion assists curving in the loop-forming direction if it is positioned "outside the loop" by limiting the curvature in the non-loop-forming direction.

References

[1] Roee Amit, Hernan G. Garcia, Rob Phillips, and Scott E. Fraser. Building enhancers from the ground up: A synthetic biology approach. *Cell*, 146(1):105–118, July 2011. ISSN 0092-8674. doi: 10.1016/j.cell.2011.06.024. URL `http://www.sciencedirect.com/science/article/pii/S0092867411006660`.

[2] Zheng Yu Chen and Jaan Noolandi. Renormalization-group scaling theory for flexible and wormlike polymer chains. *The Journal of Chemical Physics*, 96(2):1540–1548, January 1992. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.462138. URL `http://scitation.aip.org/content/aip/journal/jcp/96/2/10.1063/1.462138`.

[3] T. E. Cloutier and J. Widom. DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proceedings of the National Academy of Sciences*, 102(10):3645–3650, February 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0409059102. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC553319/`.

[4] Juan J. De Pablo and Fernando A. Escobedo. Monte Carlo methods for polymeric systems. In I. Prigogine and Stuart A. Rice, editors, *Advances in Chemical Physics*, volume 105, pages 337–367. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1998. ISBN 9780470141649, 9780471196303. URL http://onlinelibrary.wiley.com/doi/10.1002/9780470141649.ch11/summary.

[5] Larry J. Friedman and Jeff Gelles. Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation. *Cell*, 148(4):679–689, February 2012. ISSN 0092-8674. doi: 10.1016/j.cell.2012.01.018. URL http://www.cell.com/article/S0092867412000839/abstract.

[6] Pierre-Gilles Gennes. *Scaling Concepts in Polymer Physics*. Cornell University Press, Ithaca, NY, November 1979. ISBN 978-0-8014-1203-5.

[7] Peter Grassberger. Pruned-enriched rosenbluth method: Simulations of θ polymers of chain length up to 1 000 000. *Physical Review E*, 56(3):3682–3693, September 1997. doi: 10.1103/PhysRevE.56.3682. URL http://link.aps.org/doi/10.1103/PhysRevE.56.3682.

[8] John F. Marko and Eric D. Siggia. Stretching DNA. *Macromolecules*, 28(26):8759–8770, December 1995. ISSN 0024-9297. doi: 10.1021/ma00130a008. URL http://dx.doi.org/10.1021/ma00130a008.

[9] J. David Moroz and Philip Nelson. Torsional directed walks, entropic elasticity, and DNA twist stiffness. *Proceedings of the National Academy of Sciences*, 94(26):14418–14422, December 1997. ISSN 0027-8424, 1091-6490. URL http://www.pnas.org/content/94/26/14418.

[10] Manish Nepal, Alon Yaniv, Eyal Shafran, and Oleg Krichevsky. Structure of DNA coils in dilute and semidilute solutions. *Physical Review Letters*, 110(5):058102, January 2013. doi: 10.1103/PhysRevLett.110.058102. URL http://link.aps.org/doi/10.1103/PhysRevLett.110.058102.

[11] Alexander J. Ninfa, Lawrence J. Reitzer, and Boris Magasanik. Initiation of transcription at the bacterial glnAp2 promoter by purified e. coli components is facilitated by enhancers. *Cell*, 50(7):1039–1046, September 1987. ISSN 0092-8674. doi: 10.1016/0092-8674(87)90170-X. URL http://www.sciencedirect.com/science/article/pii/009286748790170X.

[12] O. Kratky, G. Porod. Röntgenuntersuchung gelöster fadenmoleküle. *Rec. Trav. Chim. Pays-Bas.*, 68:1106–1123, 1949.

[13] Yaroslav Pollak, Sarah Goldberg, and Roee Amit. Self-avoiding wormlike chain model for double-stranded-DNA loop formation. *Physical Review E*, 90(5):052602, November 2014. doi: 10.1103/PhysRevE.90.052602. URL http://link.aps.org/doi/10.1103/PhysRevE.90.052602.

[14] Mathieu Rappas, Daniel Bose, and Xiaodong Zhang. Bacterial enhancer-binding proteins: unlocking σ54-dependent gene transcription. *Current Opinion in Structural Biology*, 17(1):110–116, February 2007. ISSN 0959-440X. doi: 10.1016/j.sbi.2006.11.002. URL http://www.sciencedirect.com/science/article/pii/S0959440X06002077.

[15] Jiro Shimada and Hiromi Yamakawa. Ring-closure probabilities for twisted wormlike chains. application to DNA. *Macromolecules*, 17(4):689–698, April 1984. ISSN 0024-9297. doi: 10.1021/ma00134a028. URL http://dx.doi.org/10.1021/ma00134a028.

[16] Douglas R. Tree, Abhiram Muralidhar, Patrick S. Doyle, and Kevin D. Dorfman. Is DNA a good model polymer? *Macromolecules*, 46(20):8369–8382, October 2013. ISSN 0024-9297. doi: 10.1021/ma401507f. URL http://dx.doi.org/10.1021/ma401507f.

[17] Reza Vafabakhsh and Taekjip Ha. Extreme bendability of DNA less than 100 base pairs long revealed by single-molecule cyclization. *Science*, 337(6098):1097–1101, August 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1224139. URL http://www.sciencemag.org/content/337/6098/1097.

תקציר

מעצמים (אנהנסרים) הינם רצפי דנ"א שאינם מקודדים, המהווים אתרי קישור לפקטורי שעתוק ומשחקים תפקיד חיוני בבקרה על ביטוי גנים ותהליכים התפתחותיים. על אף שמעצמים נחקרו רבות במהלך העשורים האחרונים, אנו עדיין לא מבינים מהם החוקים שמכתיבים את המבנה שלהם וההשפעה של חוקים אלו על התפקיד הבקרתי שלהם. מעצמים יכולים להיות ממוקמים מאד קרוב לגן אותו מבקרים או כמה מאות בסיסים ממנו. הדרך בה בקרה מרחוק מאפשרת מגע עם המקדם (פרומוטור) באיזור הבקרה הינה ע"י כך שהדנ"א יוצר לולאה ובכך הקרבה בין המקדם לאתרי הקישור במעצם אפשרית. אנו התמקדנו בעבודה זו בחיידקים, בתת קבוצה של גנים המבוקרים על ידי פקטור סיגמא 54. פקטור זה ייחודי בכך שהוא יוצר לולאה על מנת לתקשר עם אזורי בקרה הנמצאים במרחק ממנו. פקטורי שעתוק יכולים לגרום לשפעול של גנים (אקטיבטורים) וכך לעיכוב שלהם (רפרסורים). מנגנון עיכוב שלא נחקר רבות הינו עיכוב "מדכא" (quenching) בו מעכב נקשר במרחק מה מהאקטיבטור ומסוגל לעכב אותו למרות שאין קישור ישיר בין השניים והם נקשרים על גבי רצפים שונים בדנ"א. המנגנון לעיכוב זה אינו ידוע על אף שתועד במקור בזבוב פירות וכן ישנן עדויות לכך שהוא מצוי גם בשמרים, חיידקים ותאי יונק. על מנת לחקור את מנגנון עיכוב זה נקטנו בעבודה זו בגישה סינטטית. ראשית, פיתחנו בשיתוף פעולה עם פיזיקאי במעבדה, מודל פיסיקלי המדמה את יצירת הלולאה, הנותן תחזית לכיצד יצירת הלולאה תושפע מנוכחות של פקטורי שעתוק על גבה. לאחר מכן יצרנו ספריות רבות של מעצמים סינטטיים שאת רצפיהם תכננו במחשב בבניית ארכיטקטורות שונות של פקטורי שעתוק ובחנו את התחזיות של המודל באופן ניסיוני. לבסוף, אימתנו את התוצאות שהתקבלו בגנומים של חיידקים ע"י אנליזה ביואינפורמטית.

בעבודה זו הדגמנו כיצד מודל (excluded volume) בו הנפח של פקטור השעתוק מונע יצירת לולאת דנ"א, יכול להביא לעיכוב מדכא של הגן המבוקר. יתרה מכך, רמת הבקרה מושפעת באופן משמעותי באופן סידור פקטורי השעתוק ומוכתבת ע"י מחזוריות של 11 בסיסים. אנו מדגימים שפקטור שעתוק הפונה אל תוך הלולאה מעכב משמעותית את הביטוי של הגן המדווח הנותן מדד לבקרה. בנוסף, ההשפעה השלילית מועצמת כאשר יש שני אתרים הפונים אל תוך הלולאה. ממצא זה הודגם גם כן, ע"י יצירת חלבון בעל נפח גדול יותר בעזרת איחוי של חלבון נוסף אל פקטור השעתוק.

ההשלכות של ממצאים אלו הינם שברמה הגנומית, שינויים במהלך האבולוציה של הוספה ומחיקה של בסיסים באזורי לולאה יתרחשו לרוב במכפלות שלמות של 11 בסיסים. שינוי זה מאפשר לשמור על האוריינטציה של פקטור השעתוק עם קומפלקס של הפולימראז-סיגמא 54 ובכך לא ישנה את רמות הביטוי של הגן המבוקר. לעומת זאת, הוספה ומחיקה של בסיסים באזורי לולאה במכפלות של 6 בסיסים ישנו את אוריינטציה זו ולכן זה יגרום לשינוי בביטוי ולפיכך נצפה שזה יהיה נדיר. היפותזה זאת נבחנה על שורה

של מעצמים בגן *qrr* ממינים שונים של *Vibrio* ואכן נראה שיש תמיכה ראשונית לממצאים שלנו ברמת הגנום.

על מנת לבסס את השערתנו, בצענו סדרת ניסויים נוספת בה אנו ערכנו את הגנום ושינינו את הרצף בלולאה ע"י הוספת בסיסים ובחינת ההשפעה על הגן המבוקר. כאשר הוכנסו 6 בסיסים באזור הלולאה, היתה השפעה רבה על רמות הביטוי של הגן המבוקר, אשר רמתו ירדה משמעותית, בדיוק כפי שהראו התחזיות והממצאים שלנו. ניסוי זה נותן ביסוס נוסף להשערתנו שישנה רגישות להכנסות והוצאות של בסיסים אשר מתרחשות באופן טבעי במהלך האבולוציה ואלו כנראה מוכתבות מהמחזוריות של 11 בסיסים בדנ"א.

בכדי לבדוק האם לממצאים שלנו יש חשיבות גם במערכות יותר מורכבות, החלטנו לעבוד בנוסף במערכת שמר ולבחון האם היא מודל טוב לבדיקת התחזיות שלנו. הבחירה לעבוד בשמר, נבעה מהעובדה שרצפים מאקטבים, בדומה לגנומים של חיידקים, נמצאים ביחס למקדם מרחק של עד כמאות בסיסים וכן המבנה הפנימי של רצפים אלו מזכיר את זה של מעצמים חיידקיים. מאחר והמודל שלנו הוא מבוסס על יצירת לולאה, ובשמר אין תיעוד לכך שלולאה אכן נוצרת בבקרה מרחוק, החלטנו לבדוק האם אנו יכולים לאמת שאכן בקרה מרחוק בשמר נעשית ע"י יצירה של לולאה. לשם כך יצרנו ספריה נוספת בכדי לבדוק השערה זו. במערכת בה בחרנו לעבוד לא הצלחנו להראות שיצירת לולאה בשמרים אכן מתרחשת, ויתכן שמנגנונים אחרים מאפשרים את הקישור בין הרצפים הרחוקים והמקדם של הגן אותו הם מבקרים.

מחקר זה נעשה תחת הנחייתו של פרופ' משנה רועי עמית
בפקולטה להנדסת ביוטכנולוגיה ומזון בטכניון.

אני מודה לטכניון על התמיכה הכספית הנדיבה במהלך השתלמותי.

מחקר על בקרת שעתוק המבוסס על יצירת לולאה בעזרת שימוש בכלים של ביולוגיה סינטטית

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

דוקטור לפילוסופיה

מיכל ברונוסר-מירום

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

אב תשע"ז חיפה יולי 2017

מחקר על בקרת שעתוק המבוסס על יצירת לולאה בעזרת שימוש בכלים של ביולוגיה סינטטית

מיכל ברונוסר-מירום