Precision measurement of cis-regulatory energetics in living cells

Talitha Forcier[1], **Andalus Ayaz**[1], **Manraj S. Gill**[1,§], **Daniel Jones**[1,2,¶], **Rob Phillips**[2], **Justin B. Kinney**[1]

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, USA;
[2] Department of Applied Physics, California Institute of Technology, USA

***For correspondence:**
jkinney@cshl.edu (JBK)

Present address: [§] Department of Biology, Massachusetts Institute of Technology, USA; [¶] Department of Cell and Molecular Biology, Uppsala University, Sweden

Abstract Gene expression in all organisms is controlled by cooperative interactions between DNA-bound transcription factors (TFs). However, measuring TF-TF interactions that occur at individual cis-regulatory sequences remains difficult. Here we introduce a strategy for precisely measuring the Gibbs free energy of such interactions in living cells. Our strategy uses reporter assays performed on strategically designed cis-regulatory sequences, together with a biophysical modeling approach we call “expression manifolds”. We applied this strategy in *Escherichia coli* to interactions between two paradigmatic TFs: CRP and RNA polymerase (RNAP). Doing so, we consistently obtain measurements precise to ~ 0.1 kcal/mol. Unexpectedly, CRP-RNAP interactions are seen to deviate in multiple ways from the prior literature. Moreover, the well-known RNAP binding motif is found to be a surprisingly unreliable predictor of RNAP-DNA binding energy. Our strategy is compatible with massively parallel reporter assays in both prokaryotes and eukaryotes, and should thus be highly scalable and broadly applicable.

Introduction

Cells regulate the expression of their genes in response to biological and environmental cues. A major mechanism of gene regulation in all organisms is the binding of transcription factor (TF) proteins to cis-regulatory elements encoded within genomic DNA. DNA-bound TFs interact with one another, either directly or indirectly, forming cis-regulatory complexes that modulate the rate at which nearby genes are transcribed (***Ptashne and Gann, 2002***; ***Courey, 2008***). Different arrangements of TF binding sites within cis-regulatory sequences can lead to different regulatory programs, but the rules that govern *which* arrangements lead to *which* regulatory programs remain largely unknown. Understanding these rules, which are collectively called “cis-regulatory grammar” (***Weingarten-Gabbay and Segal, 2014***), is a major challenge in modern biology.

A diverse array of high-throughput technologies have revolutionized our understanding of transcriptional regulation in recent years. It is now possible to map the genome-wide binding sites of transcription factors *in vivo* (***Ren et al., 2000***; ***Johnson et al., 2007***), sometimes to nucleotide resolution (***Rhee and Pugh, 2011***). Large collaborative efforts using such methods have been carried out to comprehensively annotate cis-regulatory elements in model organisms (***modENCODE Consortium et al., 2010***; ***Gerstein et al., 2010***) and in humans (***ENCODE Project Consortium, 2012***). Complementing such techniques are high-throughput *in vitro* methods for characterizing TF binding specificity (***Mukherjee et al., 2004***; ***Meng et al., 2005***; ***Berger et al., 2006***; ***Zhao et al., 2009***; ***Jolma et al., 2010***; ***Slattery et al., 2011***). These methods have been applied to a large fraction of the TFs in select model organisms (***Noyes et al., 2008***; ***Badis et al., 2009***) as well as in humans (***Jolma et al., 2013***). However, neither class of method addresses the critical question of what TFs do once bound to DNA. In particular, there are no systematic methods, either high-throughput or low-throughput, for characterizing the TF-TF interactions that occur within cis-regulatory complexes in living cells.

Measuring the quantitative strength of interactions between DNA-bound TFs is critical for elucidating cis-regulatory grammar. In particular, knowing the Gibbs free energy of TF-TF interactions is essential for building biophysical models ***Bintu et al.*** (***2005***); ***Sherman and Cohen*** (***2012***) that can quantitatively explain gene regulation in terms of simple protein-DNA and protein-protein interactions. Biophysical models have proven remarkably successful at quantitatively explaining regulation by a small number of well-studied cis-regulatory sequences. Arguably, the biggest successes have been achieved in the bacterium *E. coli*, particularly in the context of the *lac* promoter (***Vilar and Leibler, 2003***; ***Kuhlman et al., 2007***; ***Kinney et al., 2010***; ***Garcia and Phillips, 2011***; ***Brewster et al., 2014***) and the O_R/O_L control region of the λ phage lysogen (***Ackers et al., 1982***; ***Shea and Ackers, 1985***; ***Cui et al., 2013***). But in both cases, the biophysical level of understanding that has been achieved required decades of focused study. New approaches for dissecting cis-regulatory energetics, approaches that are both general and systematic, will be needed before this quantitative level of understanding can be obtained for any cis-regulatory sequence having any arrangement of TF binding sites.

Here we address this need by describing a systematic experimental/modeling strategy for dissecting the biophysical mechanisms of transcriptional regulation in living cells. Our strategy is based on reporter assays and is not a new experimental method per se. Rather, it shows how key biophysical quantities in transcriptional regulation can be measured to high precision by performing relatively simple experiments on strategically chosen cis-regulatory sequences, then analyzing the resulting data appropriately. Our rationale for introducing this strategy is that reporter assays can be readily performed in a wide variety of systems, making this strategy highly flexible and broadly applicable. Moreover, massively parallel reporter assays should allow this strategy to be dramatically scaled up.

Our strategy centers on the measurement and modeling of mathematical objects that we call “expression manifolds.” The underlying idea is to perform *multidimensional* measurements. If a hypothesized biophysical model is true, these measurements will collapse to a lower-dimension manifold embedded in this measurement space. If such data collapse is observed, specific values for the parameters of the hypothesized biophysical model can be inferred. On the other hand, if such collapse is not observed, the hypothesized biophysical model can be rejected and a different biophysical model is seen to be needed.

To demonstrate its utility, we applied this strategy to a regulatory paradigm in *E. coli*: activation of the σ^{70} RNA polymerase holoenzyme (RNAP) by the cAMP receptor protein (CRP). RNAP is arguably the best understood RNA polymerase in biology (***Ruff et al., 2015***), and CRP is arguably the best understood transcriptional activator (***Busby and Ebright, 1999***). CRP activates transcription when bound to DNA at various positions upstream of RNAP by forming favorable interactions with the RNAP α subunit. Such regulation is often described as "class I" or "class II", depending on the spacing between the RNAP and CRP binding sites. Both classes of interaction are known to depend strongly on the spacing between binding sites, but the *in vivo* Gibbs free energies of these interactions have been reported for only one such spacing: when the CRP site is centered -61.5 bp relative to the transcription start site (TSS), as occurs at the *E. coli lac* promoter.

By measuring and modeling expression manifolds, we systematically determined the *in vivo* Gibbs free energy (ΔG) of CRP-RNAP interactions that occur at a variety of different binding site spacings. These ΔG values were consistently measured to a precision of ~ 0.1 kcal/mol, roughly 3% of the strength of a hydrogen bond. Although our results broadly agree with the prior literature, there are key divergences. We find that class I CRP-RNAP interactions, which occur when CRP is centered upstream of $\sim$ -60.5 bp, are generally much stronger than have been suggested. Moreover, we find that the class II CRP-RNAP interaction that occurs when CRP is centered at -40.5 bp can either activate or repress transcription depending on features of the RNAP binding site that have yet to be understood.

In the course of these experiments we obtained other key biophysical information. First, we were able to distinguish between two qualitatively different mechanisms of transcriptional activation: "stabilization" of RNAP-DNA binding (also called "recruitment" (***Ptashne, 2003***)) versus "acceleration" of the transcript initiation rate by DNA-bound RNAP. Contrary to prior *in vitro* studies, we find that *in vivo* class II activation by CRP at -41.5 bp occurs exclusively through stabilization, not acceleration. Second, we were able to measure the strength with which both CRP and RNAP bind their respective sites. This strength is quantified by the grand canonical potential (denoted here by $\Delta\Psi$), which accounts for the ΔG of binding as well as the *in vivo* concentration of each protein. Importantly, we find that the actual *in vivo* $\Delta\Psi$ of RNAP-DNA binding deviates substantially from the predictions of the established RNAP binding motif. This result highlights the perils of assuming simple models for protein-DNA binding energy when modeling the biophysics of transcriptional regulation.

In what follows, we first illustrate this expression manifold strategy in the context of simple repression, which provides a general way to measure the $\Delta\Psi$ of TF-DNA binding. This strategy is then used to measure the $\Delta\Psi$ of CRP binding to a near-consensus DNA site that we use in subsequent experiments. Next we show how expression manifolds, inferred from measurements of simple activation, can be used to determine the ΔG of TF-RNAP interactions. This strategy is used to measure CRP-RNAP interactions at a variety of class I and class II positions, and the deviations of these measurements from the prior literature are discussed. Finally, we compare the values of $\Delta\Psi$ for RNAP-DNA binding, obtained in the course of the above analyses, to the predictions of the RNAP-DNA binding motif from ***Kinney et al.*** (***2010***).

Results

Strategy for measuring TF-DNA interactions *in vivo*

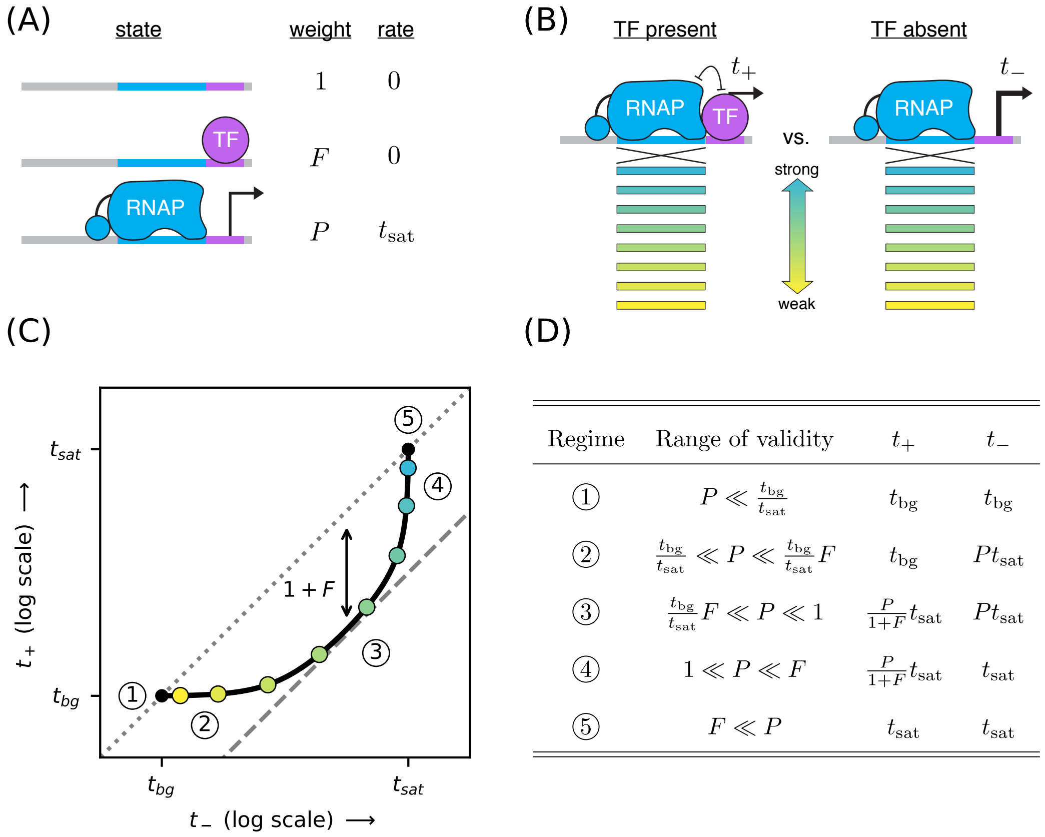
We begin by showing how expression manifolds can be used to measure the *in vivo* strength of TF binding to a specific DNA binding site. This measurement is accomplished by using the TF of interest as a transcriptional repressor. We place the TF binding site directly downstream of the RNAP binding site so that the TF, when bound to DNA, sterically occludes the binding of RNAP. We then measure the rate of transcription from a few dozen variant RNAP binding sites. Transcription from each variant site is assayed in both the presence and in the absence of the TF.

Figure 1A illustrates a thermodynamic model (***Bintu et al., 2005***; ***Sherman and Cohen, 2012***) for this type of simple repression. In this model, promoter DNA can be in one of three states: unbound, bound by the TF, or bound by RNAP. These three state are assumed to occur with a relative frequency that is consistent with thermal equilibrium, i.e., with a probability proportional to its Boltzmann weight.

The energetics of protein-DNA binding determine the Boltzmann weight for each state. By convention we set the weight of the unbound state equal to 1. The weight of the TF-bound state is then given by $F = [\mathrm{TF}]K_F$ where $[\mathrm{TF}]$ is the concentration of the TF and K_F is the affinity constant in inverse molar units. Similarly, the weight of the RNAP-bound state is $P = [\mathrm{RNAP}]K_P$. In what follows we refer to F and P as the "binding factors" for the TF-DNA and RNAP-DNA interactions, respectively. We note that these can also be written as $F = e^{-\Delta\Psi_F/k_BT}$ and $P = e^{-\Delta\Psi_P/k_BT}$ where k_B is Boltzmann's constant, T is temperature, and $\Delta\Psi_F$ and $\Delta\Psi_P$ respectively denote the grand canonical potential of binding for the TF and RNAP. Note that the grand canonical potential is equal to the Gibbs free energy of binding plus a term that accounts for the entropic cost of pulling each protein out of solution. For reference, $1\ k_BT = 1.62$ kcal/mol at 37 °C.

The overall rate of transcription is computed by summing the amount of transcription produced by each state, weighting each state by the probability with which it occurs. In this case we assume the RNAP-bound state initiates at a rate of t_{sat}, and that the other states produce no transcripts. We also add a term, t_{bg}, to account for background transcription (e.g., from an unidentified promoter further upstream). The rate of transcription in the presence of the TF is thus given by

$$t_+ = t_{\mathrm{sat}} \frac{P}{1 + F + P} + t_{\mathrm{bg}}. \quad (1)$$



Regime	Range of validity	t_+	t_-
①	$P \ll \frac{t_{\mathrm{bg}}}{t_{\mathrm{sat}}}$	t_{bg}	t_{bg}
②	$\frac{t_{\mathrm{bg}}}{t_{\mathrm{sat}}} \ll P \ll \frac{t_{\mathrm{bg}}}{t_{\mathrm{sat}}} F$	t_{bg}	$P t_{\mathrm{sat}}$
③	$\frac{t_{\mathrm{bg}}}{t_{\mathrm{sat}}} F \ll P \ll 1$	$\frac{P}{1+F} t_{\mathrm{sat}}$	$P t_{\mathrm{sat}}$
④	$1 \ll P \ll F$	$\frac{P}{1+F} t_{\mathrm{sat}}$	t_{sat}
⑤	$F \ll P$	t_{sat}	t_{sat}

Figure 1. Strategy for measuring TF-DNA interactions. (A) A thermodynamic model of simple repression. Here, promoter DNA can transition between three possible states: unbound, bound by a TF, or bound by RNAP. Each state has an associated Boltzmann weight and rate of transcript initiation. F is the TF binding factor and P is the RNAP binding factor; see text for a description of how these dimensionless binding factors relate to binding affinity and binding energy. t_{sat} is the rate of specific transcript initiation from a promoter fully occupied by RNAP. (B) Transcription is measured in the presence (t_+) and absence (t_-) of the TF. Measurements are made for promoters containing RNAP binding sites of differing binding strength (blue-yellow gradient). (C) If the model in panel A is correct, plotting t_+ vs. t_- for the promoters in panel B (colored dots) will trace out a 1D expression manifold. Mathematically, this manifold reflects Equation 1 and Equation 2 computed over all possible values of the RNAP binding factor P with the other parameters (F, t_{sat}) held fixed. Note that these equations include a background transcription term t_{bg}; it is assumed throughout that $t_{\mathrm{bg}} \ll t_{\mathrm{sat}}$ and that t_{bg} is independent of RNAP binding site sequence. The resulting manifold exhibits five distinct regimes (circled numbers), corresponding to different ranges for the value of P that allow the mathematical expressions in Equations 1 and 2 to be approximated by simplified expressions. In regime 3, for instance, $t_+ \approx t_-/(1+F)$, and thus the manifold approximately follows a line parallel to the diagonal but offset below it by a factor of $1+F$ (dashed line). Data points in this regime can therefore be used to determine the value of F. (D) The five regimes of the expression manifold, including approximate expressions for t_+ and t_- in each regime, as well as the range of validity for P.

In the absence of the TF ($F = 0$), the rate of transcription becomes

$$t_- = t_{\mathrm{sat}} \frac{P}{1+P} + t_{\mathrm{bg}}. \tag{2}$$

Our goal is to measure the TF-DNA binding factor F. To do this, we create a set of promoter sequences where the RNAP binding site is varied but the TF binding site is kept fixed. We then measure transcription from these promoters in both the presence and absence of the TF, respectively denoting the resulting quantities by t_+ and t_- (Figure 1B). Our rationale for doing this is that changing the RNAP binding site sequence should, according to our model, affect only the RNAP-DNA binding affinity K_P. All of our measurements should therefore lie along a one-dimensional "expression manifold" residing within the two-dimensional space of (t_-, t_+) values. Moreover, this expression

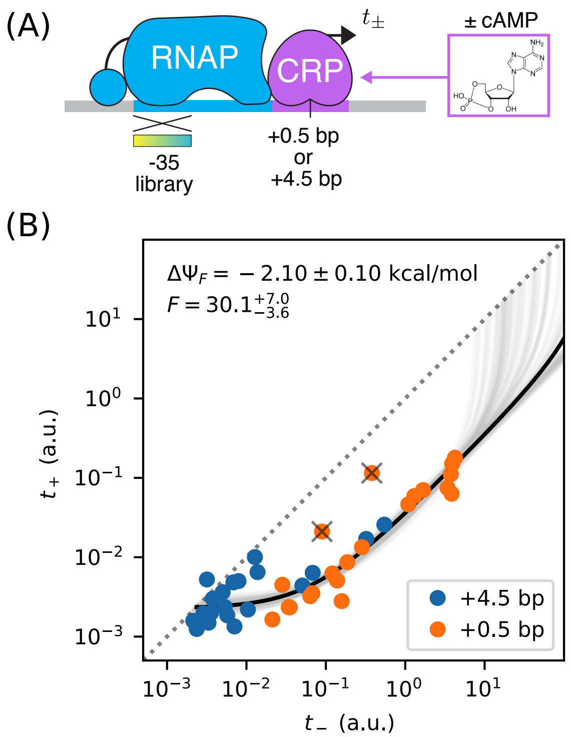


Figure 2. Precision measurement of *in vivo* CRP-DNA binding. (A) Expression measurements were performed on promoters for which CRP represses transcription by occluding RNAP. Each promoter assayed contained a near-consensus CRP binding site centered at +0.5 bp or +4.5 bp, as well as an RNAP binding site with a partially mutagenized -35 region (gradient). t_+ (alternatively, t_-) denotes measurements made in JK10 cells grown in the presence (absence) of the small molecule cAMP. (B) Dots indicate measurements for 42 such promoters. A best-fit expression manifold (black) was inferred from $n = 40$ of these data points after the exclusion of 2 outliers (gray 'X's). Gray lines indicate 100 plausible expression manifolds fit to bootstrap-resampled data points. The parameters of these manifolds were used to determine the CRP-DNA binding factor F and, equivalently, the grand canonical potential $\Delta\Psi_F = -k_B T \log F$. See Materials and Methods for more information about our curve fitting procedure and the reporting of parameter uncertainties.

manifold should follow the specific mathematical form implied by Equations 1 and 2 when P is varied and the other parameters (t_{sat}, t_{bg}, F) are held fixed. See Figure 1C.

The geometry of this expression manifold is nontrivial. In particular, when $F \gg 1$ and $t_{\text{bg}}/t_{\text{sat}} \ll 1$, there are five different regimes corresponding to different values of the RNAP binding factor P for which the expressions for t_+ and t_- approximately simplify. These regimes are listed in Figure 1D. In regime 1, P is so small that both t_+ and t_- are dominated by background transcription, i.e., $t_+ \approx t_- \approx t_{\text{bg}}$. P is somewhat larger in regime 2, causing t_- to be proportional to P while t_+ remains dominated by background. In regime 3, both t_+ and t_- are proportional to P in this regime, with $t_+/t_- \approx 1/(1+F)$. In regime 4, t_- saturates at t_{sat} while t_+ remains proportional to P. Regime 5 occurs when both t_+ and t_- are saturated, i.e., $t_+ \approx t_- \approx t_{\text{sat}}$.

Precision measurement of *in vivo* CRP-DNA binding

The placement of CRP downstream of the RNAP binding site is known to repress transcription (***Morita et al., 1988***). We therefore reasoned that placing a DNA binding site for CRP downstream of RNAP would allow us to measure the binding factor of that site. Figure 2 illustrates measurements of the expression manifold used to characterize the strength of CRP binding to the 22bp site GAA$\underline{\texttt{TGTGA}}$CCTAGA$\underline{\texttt{TCACA}}$TTT. This site contains the well-known consensus site, which comprises two dyadic pentamers (underlined) separated by a 6bp spacer (***Gunasekera et al., 1992***). We performed measurements using this CRP site centered at two different locations relative to the TSS: +0.5 bp and +4.5 bp.[1] To avoid influencing CRP binding strength, the -10 region of the RNAP site was kept fixed in the promoters we assayed while the -35 region of the RNAP binding site was varied (Figure 2A). Promoter DNA sequences are shown in Appendix 1 Figure 1.

We obtained t_- and t_+ measurements for these constructs using a modified version of the β-galactosidase assay of ***Miller*** (***1972***); see Materials and Methods for details. Our measurements are largely consistent with an expression manifold having the expected mathematical form (Figure 2B). Moreover, the measurements for CRP at the two different spacings (+0.5 bp and +4.5 bp) appear consistent with each other, although the measurements at +4.5 bp have consistently lower values for P. A small number of data points do deviate substantially from this manifold, but the presence of such outliers is not surprising from a biological perspective: introducing mutations into the RNAP binding site has the potential to create a new binding site, either for RNAP itself or for other TFs. Fortunately, outliers appear at a rate small enough for us to identify and exclude them by inspection.

We quantitatively modeled the expression manifold in Figure 2B by fitting $n+3$ parameters to our $2n$ measurements, where $n = 42$ is the number of non-outlier data points, each point corresponding to an assayed promoter. The $n + 3$ parameters were $t_{\rm sat}$, $t_{\rm bg}$, F, and P_1, P_2, …, P_n, where each P_i is the RNAP binding factor of promoter i. Nonlinear least squares optimization was then used to infer values for these parameters. Uncertainties in $t_{\rm sat}$, $t_{\rm bg}$, and F were quantified by repeating this procedure on bootstrap-resampled data points.

These results yielded highly uncertain values for $t_{\rm sat}$ because none of our measurements appear to fall within regime 4 or 5 of the expression manifold. A reasonably precise value for $t_{\rm bg}$ was obtained, but substantial scatter about our model predictions in regime 1 and 2 remain. This scatter likely reflects some variation in $t_{\rm bg}$ from promoter to promoter, variation that is to be expected since the source of background transcription is not known and the appearance of even very weak promoters could lead to such fluctuations.

These data do, however, determine a highly precise value for the strength of CRP-DNA binding: $F = 30.1^{+7.0}_{-3.6}$ or, equivalently, $\Delta\Psi_P = -2.10 \pm 0.10$ kcal/mol.[2] This expression manifold approach is thus able to measure TF-DNA binding energies to a precision of ~ 0.1 kcal/mol, about 2% of the hydroxyl-oxygen hydrogen bond (5.0 kcal/mol), the kind routinely found in liquid water. We note that CRP forms ~ 38 hydrogen bonds with DNA when it binds to a consensus DNA site (***Parkinson et al., 1996***), and that previous *in vitro* measurements of the Gibbs free energy of CRP-DNA binding to its consensus site have yielded ~ -15 kcal/mol (***Ebright et al., 1989***; ***Gunasekera et al., 1992***). Our result indicates that, in living cells, this Gibbs free energy is almost entirely canceled by the entropic cost of removing a CRP molecule from the cytoplasmic environment.

Strategy for measuring TF-RNAP interactions *in vivo*

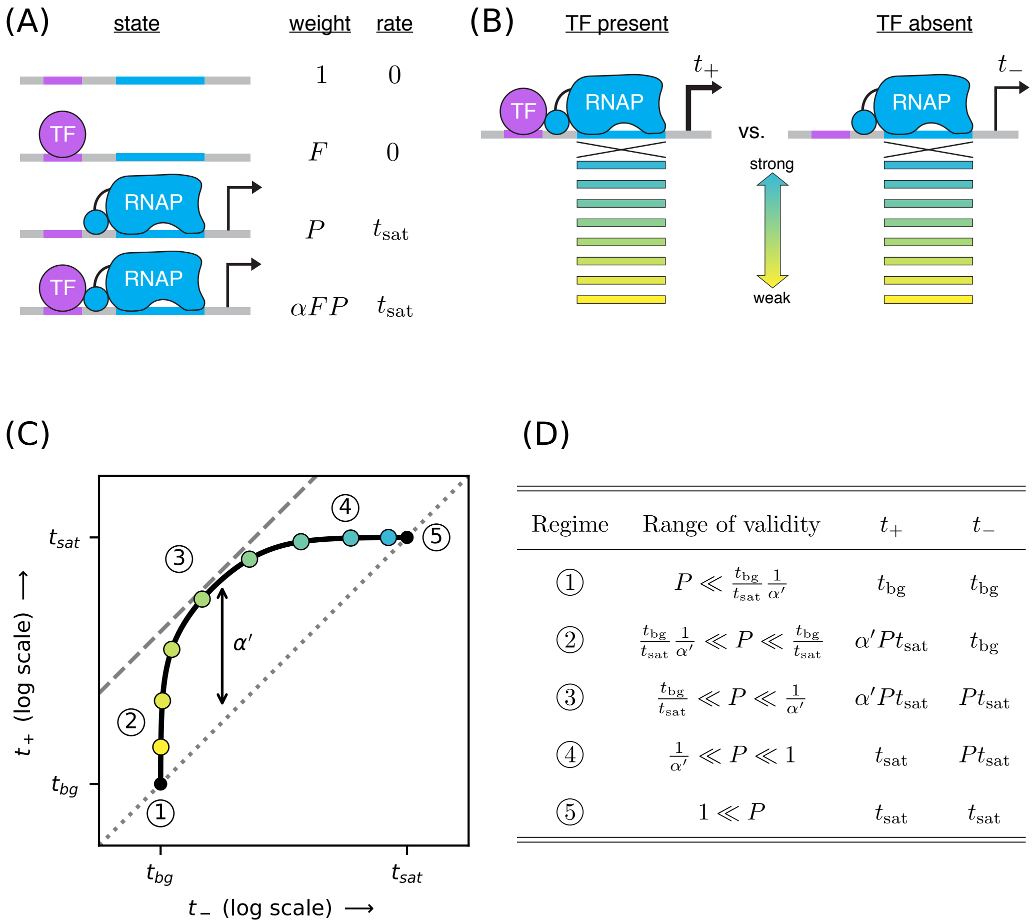
Next we discuss how to measure activating interactions between TFs and RNAP. A common mechanism of transcriptional activation is stabilization (also called recruitment (***Ptashne, 2003***)). This occurs when a DNA-bound TF stabilizes the RNAP-DNA closed complex. Stabilization effectively increases the RNAP affinity K_P, and thus the binding factor P, while not affecting the rate of transcript initiation from the RNAP-DNA closed complexes.

A thermodynamic model for activation by stabilization is illustrated in Figure 3A. Here promoter DNA can be in four states: unbound, TF-bound, RNAP-bound, or doubly bound. In the doubly bound state, a “cooperatively factor” α is included in the Boltzmann weight. This cooperatively factor is related to the TF-RNAP Gibbs free energy of interaction, ΔG_α, via $\alpha = e^{-\Delta G_\alpha / k_B T}$. Activation occurs when $\alpha > 1$ ($\Delta G_\alpha < 0$). The resulting activated transcription rate is given by

$$t_+ = t_{\rm sat} \frac{P + \alpha F P}{1 + F + P + \alpha F P} + t_{\rm bg}. \tag{3}$$

[1] The first transcribed base is, in this paper, assigned position 0 instead of the more conventional +1. Half-integer positions indicate centering between neighboring nucleotides.

[2] See Materials and Methods for a discussion of how uncertainties in these values are computed and reported.



Regime	Range of validity	t_+	t_-
①	$P \ll \frac{t_{bg}}{t_{sat}}\frac{1}{\alpha'}$	t_{bg}	t_{bg}
②	$\frac{t_{bg}}{t_{sat}}\frac{1}{\alpha'} \ll P \ll \frac{t_{bg}}{t_{sat}}$	$\alpha' P t_{sat}$	t_{bg}
③	$\frac{t_{bg}}{t_{sat}} \ll P \ll \frac{1}{\alpha'}$	$\alpha' P t_{sat}$	$P t_{sat}$
④	$\frac{1}{\alpha'} \ll P \ll 1$	t_{sat}	$P t_{sat}$
⑤	$1 \ll P$	t_{sat}	t_{sat}

Figure 3. Strategy for measuring TF-RNAP interactions. (A) A thermodynamic model of simple activation. Here, promoter DNA can transition between four different states: unbound, bound by the TF, bound by RNAP, or doubly bound. As in Figure 1, F is the TF binding factor, P is the RNAP binding factor, and t_{sat} is the rate of transcript initiation from an RNAP-saturated promoter. The cooperativity factor α quantifies the strength of the interaction between DNA-bound TF and RNAP molecules; see text for more information on this quantity. (B) As in Figure 1, expression is measured in the presence (t_+) and absence (t_-) of the TF for promoters that have RNAP binding sites of varying strength (blue-yellow gradient). (C) If the model in panel A is correct, plotting t_+ vs. t_- (colored dots) will reveal a 1D expression manifold that corresponds to Equation 4 (for t_+) and Equation 2 (for t_-) evaluated over the possible values of P. Circled numbers indicate the five regimes of this manifold. In regime 3, $t_+ \approx \alpha' t_-$ where α' is the renormalized cooperativity factor given in Equation 5; data in this regime can thus be used to measure α'. Separate measurements of F, using the strategy in Figure 1, then allow one to compute α from knowledge of α'. (D) The five regimes of the expression manifold in panel C. Note that these regimes differ from those in Figure 1D.

This can be rewritten as

$$t_+ = t_{sat}\frac{\alpha' P}{1+\alpha' P} + t_{bg}, \tag{4}$$

where

$$\alpha' = \frac{1+\alpha F}{1+F} \tag{5}$$

is a renormalized cooperatively that accounts for the strength of TF-DNA binding. As before, t_- is given by Equation 2. Note that $\alpha' \leq \alpha$ and that $\alpha' \approx \alpha$ when $F \gg 1$ and $\alpha \gg 1$.

As before, we measure both t_+ and t_- for RNAP binding sites of varying strength (Figure 3B). These measurements will, according to our model, lie along an expression manifold resembling the one shown in Figure 3C. This expression manifold exhibits five distinct regimes when $\frac{t_{sat}}{t_{bg}} \gg \alpha' \gg 1$. These regimes are listed in Figure 3D.

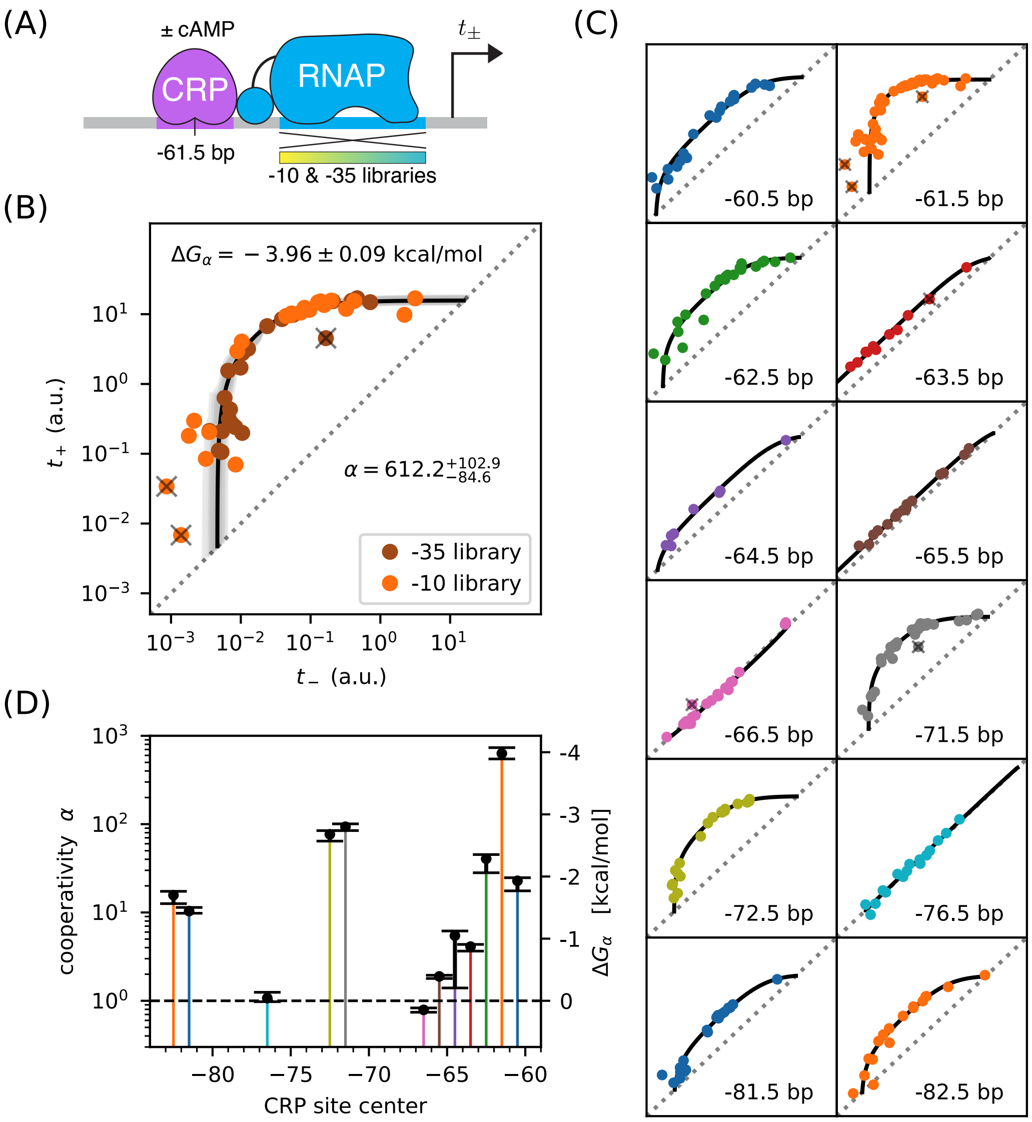


Figure 4. Precision measurement of class I CRP-RNAP interactions. (A) t_+ and t_- were measured for promoters containing a CRP binding site centered at -61.5 bp. The RNAP sites of these promoters were mutagenized in either their -10 or -35 regions (gradient). As in Figure 2, t_+ and t_- correspond to expression measurements made in the presence and absence of cAMP, respectively. (B) Data obtained for 47 variant promoters having the architecture shown in panel A. Three data points designated as outliers are indicated by 'X's. The expression manifold that best fits the 44 non-outlier points is shown in black; 100 plausible manifolds, estimated from bootstrap-resampled data points, are shown in gray. The resulting values for α and $\Delta G_\alpha = -k_BT \log \alpha$ are also provided. (C) Expression manifolds obtained for CRP binding sites centered at a variety of class I positions. (D) Inferred values for the cooperativity factor α and corresponding Gibbs free energy ΔG_α for the 12 different promoter architectures assayed in panel C. Error bars indicate the central 68% confidence interval, estimated by fitting to bootstrap-resampled data, while dots indicate the median of these estimates. Numerical values for α and ΔG_α at all of these class I positions are provided in Table 1.

Precision measurement of class I CRP-RNAP interactions

CRP activates transcription at the *lac* promoter and other promoters by binding to a 22 bp site centered at -61.5 bp relative to the TSS. This is an example of class I activation, which is mediated by an interaction between CRP and the RNAP α C-terminal domain (αCTD) (***Busby and Ebright, 1999***). *In vitro* experiments have shown this class I CRP-RNAP interaction to activate transcription by stabilizing the RNAP-DNA complex.

We measured t_+ and t_- for 47 variants of the lac* promoter (see Materials and Methods, as well as Appendix 1 Figure 1). These promoters have the same CRP binding site assayed for Figure 2, but positioned at -61.5 bp, upstream of RNAP (Figure 4A). They differ from one another in the -10 or -35 regions of their respective RNAP binding sites. Figure 4B shows the resulting measurements. With the exception of 3 outlier points, these measurements appear consistent with stabilizing activation via a Gibbs free energy of $\Delta G_\alpha = -3.96 \pm 0.09$ kcal/mol, corresponding to a cooperativity of $\alpha \sim 600$. We note that, with $F \approx 30$ determined in Figure 2, $\alpha' = \alpha$ to 3% accuracy.

This observed cooperativity is substantially stronger than suggested by previous work. Early *in vivo* experiments suggested a much lower cooperativity value, e.g. 50-fold (***Beckwith et al., 1972***), 20-fold (***Ushida and Aiba, 1990***), or even 10-fold (***Gaston et al., 1990***). These previous studies, however, only measured the ratio t_+/t_- for a specific choice of RNAP binding site. This ratio is (by Equation 4) always less than α and the differences between these quantities can be substantial.

However, even studies that have used explicit biophysical modeling have determined lower cooperativity values: ***Kuhlman et al.*** (***2007***) reported a cooperativity of $\alpha \approx 240$ ($\Delta G_\alpha \approx -3.4$ kcal/mol), while ***Kinney et al.*** (***2010***) reported $\alpha \approx 220$ ($\Delta G_\alpha \approx -3.3$ kcal/mol). Both of these studies, however, relied on the inference of complex biophysical models with many parameters. The expression manifold in Figure 3, by contrast, is characterized by only three parameters ($t_{\rm sat}$, $t_{\rm bg}$, α'), all of which can be approximately determined by visual inspection. In fact, while measuring this affinity manifold we isolated multiple specific promoters exhibiting $t_+/t_- \approx 400$, directly showing that $\alpha \gtrsim 400$.

To test the generality of this approach, we measured expression manifolds for 11 other potential class I activation positions. At every one of these positions we clearly observed the collapse of data to a 1D expression manifold of the expected shape (Figure 4C). By quantitatively modeling these manifolds, we determined the cooperativity α and the Gibbs free energy ΔG_α at each position. Uncertainties in these quantities were determined by the modeling of bootstrap-resampled data points (Materials and Methods). The resulting values for both α and ΔG_α are shown in Figure 4D. As first shown by ***Gaston et al.*** (***1990***) and ***Ushida and Aiba*** (***1990***), α depends strongly on the spacing between the CRP and RNAP binding sites, exhibiting a strong ~ 10.5 bp periodicity reflecting the helical twist of DNA. However, as with the measurement in Figure 4B, the α values we measure are far stronger than the t_+/t_- ratios previously reported by ***Gaston et al.*** (***1990***) and ***Ushida and Aiba*** (***1990***); see Table 1.

Acceleration vs. stabilization

E. coli TFs can regulate multiple different steps in the transcript initiation pathway (***Lee et al., 2012***; ***Browning and Busby, 2016***). For example, instead of stabilizing RNAP binding to DNA, TFs can activate transcription by increasing the rate at which DNA-bound RNAP initiates transcription, a process we refer to as “acceleration”. CRP, in particular, has previously been reported to activate transcription in part by acceleration when positioned appropriately with respect to RNAP (***Niu et al., 1996***; ***Rhodius et al., 1997***).

We investigated whether expression manifolds might be used to distinguish activation by acceleration from activation by stabilization. First we generalized the thermodynamic model in Figure 3A to accommodate both α-fold stabilization and β-fold acceleration (Figure 5A). This is accomplished by using the same set of states and Boltzmann weights as in the model for stabilization, but assigning a transcription rate $\beta t_{\rm sat}$ (rather than just $t_{\rm sat}$) to the TF-RNAP-DNA ternary complex. The resulting activated rate of transcription is given by

$$t_+ = t_{\rm sat}\frac{P}{1+F+P+\alpha FP} + \beta t_{\rm sat}\frac{\alpha FP}{1+F+P+\alpha FP} + t_{\rm bg}. \tag{6}$$

This simplifies to

$$t_+ = \beta' t_{\rm sat}\frac{\alpha' P}{1+\alpha' P} + t_{\rm bg} \tag{7}$$

where α' is the same as in Equation 5 and

$$\beta' = \frac{1 + \alpha\beta F}{1 + \alpha F} \tag{8}$$

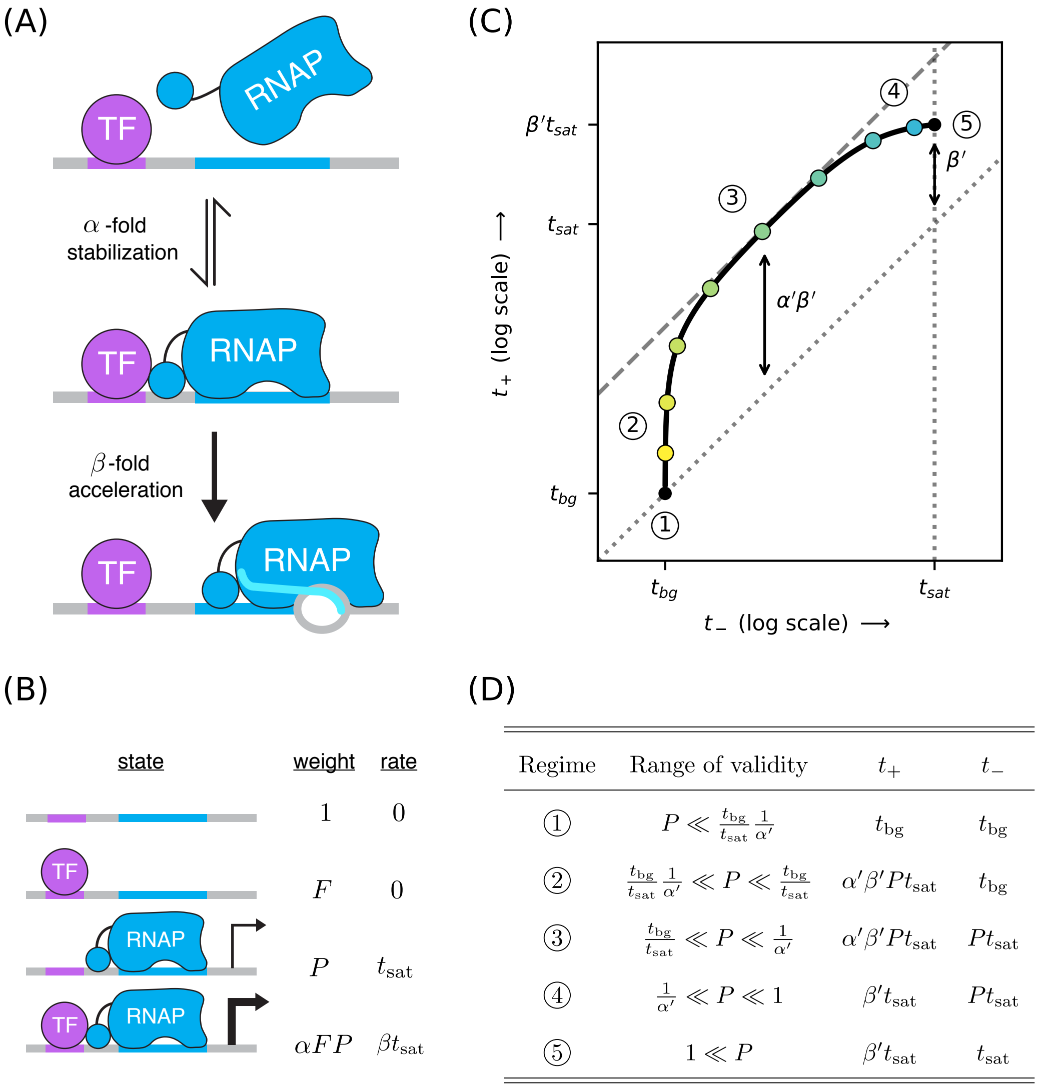
is a renormalized version of the acceleration rate β. The resulting expression manifold is illustrated in Figure 5C. Like the expression manifold for stabilization, this manifold has up to five distinct regimes corresponding to different values of P (Figure 5D). Unlike the stabilization manifold however, $t_+ \neq t_-$ in the strong RNAP binding regime (regime 5): $t_+ \approx \beta' t_{\text{sat}}$ while $t_- \approx t_{\text{sat}}$.

We next asked whether class I activation by CRP has an acceleration component. Previous *in vitro* work had suggested that the answer is 'no' (***Malan et al., 1984***; ***Busby and Ebright, 1999***), but our expression manifold approach allows us to address this question *in vivo*. We proceeded by assaying promoters containing variants of the consensus RNAP binding site, TTGACAn(17)TATAAT, that contain SNPs in their -10 or -35 regions (Figure 6A and Appendix 1 Figure 1). Note that, because the consensus RNAP binding site is 1 bp shorter than in the constructs measured for Figure 4, the CRP site at -60.5 bp in this construct corresponds to the -61.5 bp location in the constructs assayed for Figure 4B.

The resulting data (Figure 6B) are seen to largely fall along the previously measured all-stabilization expression manifold in Figure 4B. In particular, many of these data points lie at the intersection of this manifold with the $t_+ = t_-$ diagonal. We thus find that, for CRP at -61.5 bp, $\beta = 1$ to the precision of our experiments. We also identify an unambiguous value of $t_{\text{sat}} = 16.0^{+0.8}_{-1.0}$ a.u. for the transcription initiation rate of an RNAP saturated promoter. Single-cell measurements suggest that this t_{sat} value corresponds to $\sim 0.23 \pm 0.11$ transcripts per second per promoter (***So et al., 2011***). Comparing this value of t_{sat} to the t_{sat} obtained for the other manifolds in Figure 4C, we were able to estimate β for these other positions. Figure 6C shows the results: we find that $\beta \approx 1$ at all of the other class I positions for which reasonably precise estimates of β could be obtained. These results confirm that class I transcriptional activation by CRP occurs *in vivo* almost entirely through stabilization and not through acceleration.

Table 1. Summary of results for class I activation by CRP. The α and ΔG_α values listed here correspond to the values plotted in Figure 4D. n is the number of data points used to infer these values, while "outliers" is the number of data points excluded in this analysis. For comparison we show the fold-activation measurements (i.e., t_+/t_-) reported in ***Gaston et al.*** (***1990***) and ***Ushida and Aiba*** (***1990***). In these columns, n/a indicates that no measurement was reported at that CRP site spacing.

position (bp)	n	outliers	ΔG_α (kcal/mol)	α	t_+/t_- (Gaston)	t_+/t_- (Ushida)
-60.5	21	0	-1.95 ± 0.09	23.7	3.85	n/a
-61.5	47	3	-3.96 ± 0.09	612	9.05	20.6
-62.5	23	0	-2.35 ± 0.12	45.1	4.22	n/a
-63.5	11	1	-0.89 ± 0.05	4.21	n/a	n/a
-64.5	8	0	-1.10 ± 0.21	5.90	n/a	n/a
-65.5	17	0	-0.39 ± 0.03	1.90	n/a	n/a
-66.5	20	1	0.14 ± 0.03	0.79	0.78	0.84
-71.5	36	1	-2.81 ± 0.05	95.1	2.50	16.4
-72.5	19	0	-2.70 ± 0.08	79.0	3.49	n/a
-76.5	16	0	-0.03 ± 0.08	1.04	0.54	n/a
-81.5	21	0	-1.44 ± 0.05	10.3	n/a	n/a
-82.5	17	0	-1.72 ± 0.09	16.2	n/a	6.99



Regime	Range of validity	t_+	t_-
①	$P \ll \frac{t_{\text{bg}}}{t_{\text{sat}}}\frac{1}{\alpha'}$	t_{bg}	t_{bg}
②	$\frac{t_{\text{bg}}}{t_{\text{sat}}}\frac{1}{\alpha'} \ll P \ll \frac{t_{\text{bg}}}{t_{\text{sat}}}$	$\alpha'\beta' P t_{\text{sat}}$	t_{bg}
③	$\frac{t_{\text{bg}}}{t_{\text{sat}}} \ll P \ll \frac{1}{\alpha'}$	$\alpha'\beta' P t_{\text{sat}}$	$P t_{\text{sat}}$
④	$\frac{1}{\alpha'} \ll P \ll 1$	$\beta' t_{\text{sat}}$	$P t_{\text{sat}}$
⑤	$1 \ll P$	$\beta' t_{\text{sat}}$	t_{sat}

Figure 5. A strategy for distinguishing two different mechanisms of transcriptional activation. (A) A TF can activate transcription in two ways: (i) by "stabilizing" the RNAP-DNA complex or (ii) by "accelerating" the rate at which this complex initiates transcripts. (B) A thermodynamic model for the dual mechanism of transcriptional activation illustrated in panel A. Note that α multiplies the Boltzmann weight of the doubly bound complex, whereas β multiplies the transcript initiation rate of this complex. (C) Data points measured as in Figure 3C will lie along a 1D expression manifold having the form shown here. This manifold is computed using t_+ values from Equation 7 and t_- values from Equation 2, evaluated using an RNAP binding factor P ranging from 0 to ∞. Note that regime 5 occurs at a point positioned β'-fold above the diagonal, where β' is related to β through Equation 8. Measurements in or near the strong promoter regime ($P \gtrsim 1$) can thus be used to determine the value of β' and, consequently, the value of β. (D) The five regimes of this expression manifold. Note that the ranges of validity for these regimes are the same as in Figure 3D, but that the t_+ values differ.

Surprises in class II regulation

Many *E. coli* TFs participate in what is referred to as class II activation (***Browning and Busby, 2016***). This type of activation occurs when the TF binds to a site that overlaps the -35 element (often completely replacing it) and interacts directly with the main body of RNAP. CRP is known to participate in class II activation at many promoters (***Keseler et al., 2011***; ***Salgado et al., 2013***), including the galP1 promoter, where it binds to a site centered at position -41.5 bp (***Adhya, 1996***). *In vitro* studies

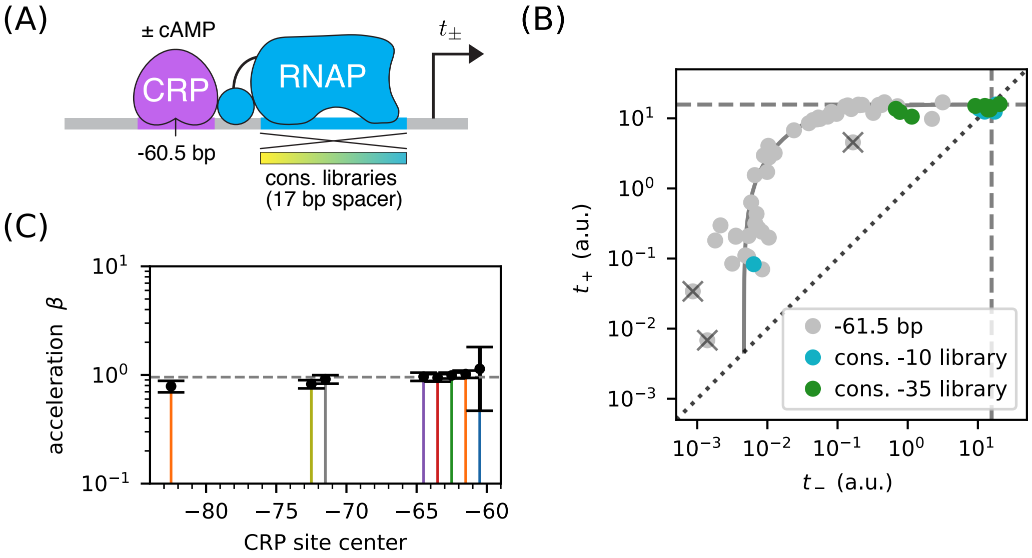


Figure 6. Class I activation by CRP occurs exclusively through stabilization. (A) t_+ and t_- were measured for promoters containing variants of the consensus (i.e., maximal strength) RNAP binding site, as well as a CRP binding site centered at -60.5 bp. Because the consensus RNAP site is 1 bp shorter than the RNAP sites assayed above, CRP at -60.5 bp here corresponds to CRP at -61.5 bp in Figure 4. (B) $n = 18$ data points obtained for the constructs in panel A, overlaid on the measurements from Figure 4B (gray). The value for t_{sat} inferred for Figure 4B is indicated by dashed lines. From these new data points we conclude that $\beta' \approx 1$, and thus $\beta \approx 1$. (C) Values for β inferred for other CRP positions using the data in Figure 4B and assuming the value of t_{sat} shown in panel B. Thus, we detect no acceleration at any class I promoter architectures. Note that β values could not be confidently determined at some CRP positions shown in Figure 4D.

have shown CRP to activate transcription at -41.5 bp relative to the TSS through a combination of stabilization and acceleration (***Niu et al., 1996***; ***Rhodius et al., 1997***).

We sought to reproduce this finding *in vivo* by measuring expression manifolds. We therefore placed a consensus CRP site at -41.5 bp, replacing much of the -35 element in the process, then varied the -10 element of the RNAP binding site (Figure 7A). Surprisingly, we observed that the resulting expression manifold saturates at the same t_{sat} value shared by all class I promoters. Thus, CRP appears to activate transcription *in vivo* solely through stabilization, and not at all through acceleration, when located at -41.5 bp relative to the TSS (Figure 7B).

The genome-wide distribution of CRP binding sites suggests that CRP also participates in class II activation at position -40.5 bp (***Keseler et al., 2011***; ***Salgado et al., 2013***). When measuring an expression manifold at this position, however, we obtained a scatter of 2D points that did not collapse to any discernible 1D expression manifold (Figure 7D). Some of these promoters exhibit activation, some exhibit repression, and some exhibit no regulation by CRP.

Our observations complicate the current understanding of class II regulation by CRP. Our *in vivo* measurements of CRP at -41.5 bp call into question the mechanism of activation previously discerned using *in vitro* techniques. The scatter observed when CRP is positioned at -40.5 bp suggests that, at this position, the -10 region of the RNAP binding site influences the values of at least two relevant biophysical parameters (not just P, as our model predicts). A potential explanation for both observations is that, because CRP and RNAP are so intimately positioned at class II promoters, even minor changes in their relative orientation caused by differences between *in vivo* and *in vitro* conditions or by changes in RNAP site sequence could have a major effect on CRP-RNAP interactions. Such sensitivity would not be expected to occur in class I activation, due to the flexibility with which the RNAP αCTDs are tethered to the main complex.

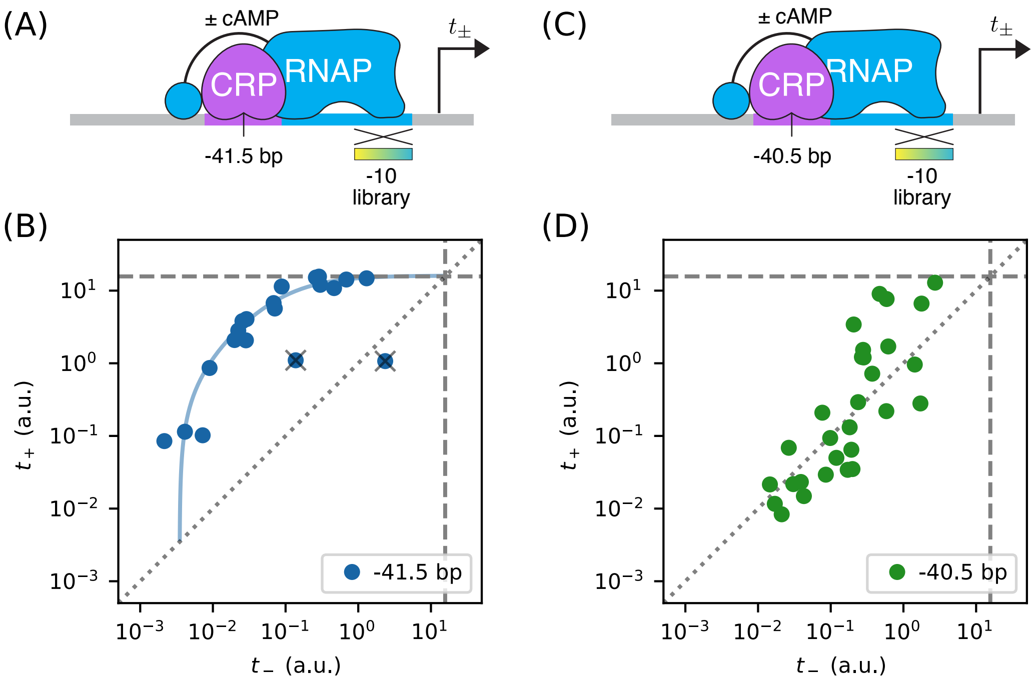


Figure 7. Surprises in class II regulation by CRP. (A) Regulation by CRP centered at -41.5 bp was assayed using RNAP binding sites that have variant -10 elements (gradient). (B) The observed expression manifold plateaus at the value of $t_{\rm sat}$ determined in Figure 6B, thus indicating no detectable acceleration by CRP. This lack of acceleration is at odds with prior *in vitro* studies (***Niu et al., 1996***; ***Rhodius et al., 1997***). (C) Regulation by CRP centered at -40.5 bp was assayed in an analogous manner. (D) Unexpectedly, data from the promoters in panel C do not collapse to a 1D expression manifold. This finding falsifies the biophysical models in Figures 3A and 5B and indicates that CRP can either activate or repress transcription from this position, depending on as-yet-unidentified features of the RNAP binding site.

Avoiding parametric models of protein-DNA binding energy

The measurement and modeling of expression manifolds has another important advantage over previous approaches for dissecting cis-regulatory sequences using massively parallel reporter assays (***Kinney et al., 2010***; ***Belliveau et al., 2018***): it sidesteps the need to parametrically model how protein-DNA binding affinity depends on DNA sequence. In modeling the expression manifolds for class I activation by CRP (Figure 4C) we obtained values for the RNAP binding factor, $P = [\mathrm{RNAP}]K_P$, for each of the variant RNAP binding sites we measured. Specifically, each inferred value for P was determined by the position of the corresponding measurement along the length of the manifold.

RNAP has a very well established sequence motif (***McClure et al., 1983***). Indeed, its DNA binding requirements were among the first characterized for any DNA-binding protein (***Pribnow, 1975***). More recently, a high-resolution model for RNAP-DNA binding energy was determined using data from a massively parallel reporter assay called Sort-Seq (***Kinney et al., 2010***). This "energy matrix model" assumes that the base pair at each position contributes additively to the overall binding energy. This model is largely consistent with previously described RNAP binding motifs but, unlike those motifs, it can predict binding energy in physically meaningful energy units (i.e., kcal/mol). In what follows we denote these binding energies as $\Delta\Delta G_P$, because they describe differences in the Gibbs free energy of binding between two DNA sites.

There is good reason to believe this matrix model to be the most accurate current model of RNAP-DNA binding. However, subsequent work has suggested that the predictions of this model might still have substantial inaccuracies (***Brewster et al., 2012***). To investigate this possibility, we compared our measured values for the grand canonical potential of RNAP-DNA binding ($\Delta\Psi_P = -k_BT \log P$) to

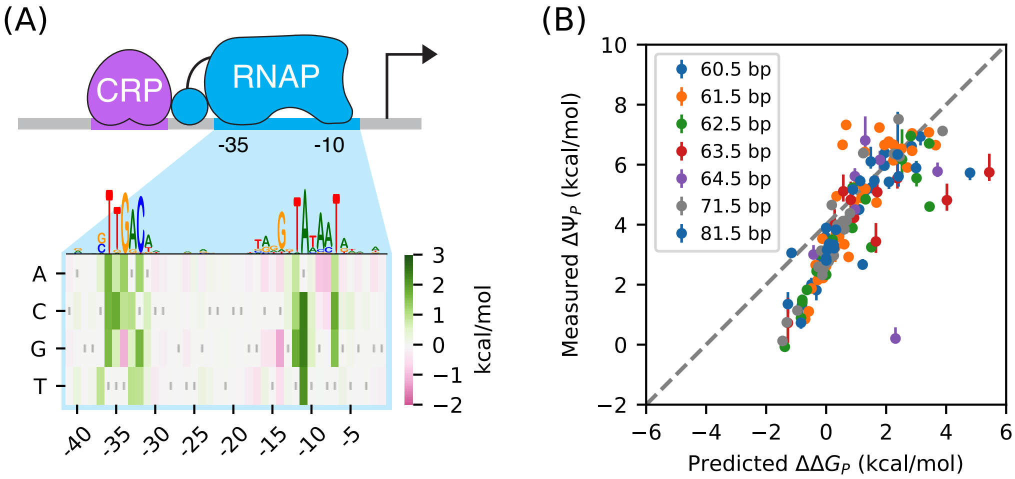


Figure 8. RNAP-DNA binding energy cannot be accurately predicted from sequence. (A) The "matrix model" for RNAP-DNA binding inferred by ***Kinney et al.*** (***2010***). This model assumes that the DNA base pair at each position in the RNAP binding site contributes additively to $\Delta\Psi_P$. Shown are the $\Delta\Delta G_P$ values assigned by this model to mutations away from the lac* RNAP site. The sequence of the lac* RNAP site is indicated by gray vertical bars. A sequence logo representation for this matrix model is provided for reference. (B) Matrix model predictions plotted against the values of $\Delta\Psi_P = -k_B T \log P$ inferred by fitting the expression manifolds in Figure 4C. Error bars on these measurements represent 64% confidence intervals computed using bootstrap resampling. Note that measured $\Delta\Psi_P$ binding energies are absolute, whereas the $\Delta\Delta G_P$ predictions of the matrix model are relative to the lac* RNAP site, which thus corresponds to $\Delta\Delta G_P = 0$ kcal/mol.

binding energies predicted from this matrix model from ***Kinney et al.*** (***2010***), which is illustrated in Figure 8A. These values are plotted against one another in Figure 8B. Although there is a strong correlation between the predictions of the model and our measurements, deviations of 1 kcal/mol or larger (corresponding to variations in P of 5-fold or greater) are not uncommon. There also appears to be systematic deviations of this model from the diagonal.

This finding is sobering: even for one of the best understood DNA-binding proteins in biology, predictions of *in vivo* protein-DNA binding energy are still quite crude. When used in conjunction with thermodynamic models, as in (***Kinney et al., 2010***), the inaccuracies of these models can have major effects on predicted transcription rates. Expression manifolds sidestep the need to parametrically model such binding energies, enabling the direct inference of grand canonical potential values for each RNAP binding site assayed.

Discussion

Expression manifolds provide a new strategy for dissecting the biophysics of transcriptional regulation in living cells. The key idea is to perform measurements of regulatory element activity that lie in a multidimensional space. These promoters are chosen so that, if a hypothesized biophysical model is correct, measurements will collapse to a lower-dimensional manifold embedded within this space. If the data collapse as expected, one can infer the parameters of the hypothesized biophysical model. If the data do not collapse, one learns that a different biophysical model is needed.

Here, we measured expression manifolds characterizing both simple repression and simple activation by CRP. Two expression measurements were made for each assayed promoter, one in the presence of cAMP (t_+) and one in the absence of cAMP (t_-). Each promoter thus corresponded to a point (t_-, t_+) in 2D. For each CRP-RNAP spacing, we assayed promoters that differed only in the DNA sequence of the RNAP binding site. Our biophysical models assumed that this site controls only one relevant biophysical quantity: the affinity of RNAP for DNA. Thus, we expected that these 2D measurements would collapse to a 1D expression manifold, with different positions along the manifold corresponding to different values of RNAP-DNA binding affinity.

Robust data collapse was observed for CRP binding sites located at all except one of the positions we assayed. In these cases, we were able to infer precise values for the energetic parameters of our models. Inferring a model for simple repression allowed us to determine the strength of CRP-DNA binding ($\Delta\Psi_F = -2.10 \pm 0.10$ kcal/mol). Inference of models for simple activation then allowed us to determine values for the CRP-RNAP interaction, as quantified by the Gibbs free energy ΔG_α; these interaction energies were consistently determined to a precision of ~ 0.1 kcal/mol.

Expression manifolds for different biophysical models often have different shapes. Measuring and modeling expression manifolds can thus allow one to distinguish between qualitatively different mechanisms of transcriptional activation. In our experiments, all transcriptional activation was seen to occur through CRP-mediated stabilization of RNAP-DNA binding, as opposed to CRP-mediated acceleration of transcript initiation. This was true even for class II activation by CRP centered at -41.5 bp, a position for which previous in vitro experiments had suggested a substantial acceleration component.

Expression manifolds also allow the measurement of protein-DNA binding energy without the need for parametric models of how this binding energy depends on DNA sequence. In the experiments described here, we obtained measurements for RNAP-DNA binding energy, as quantified by $\Delta\Psi_P$, for each of the assayed promoters. These measurements deviate substantially from the predictions of the established RNAP-DNA binding motif (***Kinney et al., 2010***). This is a cautionary tale: even for very well studied TFs, one cannot assume that published motifs accurately predict the affinity of individual DNA binding sites.

Unexpectedly, our data did not collapse to an expression manifold when CRP was centered at -40.5 bp. This result allowed us to reject our hypothesized biophysical model. We thus learned that the DNA sequence of the core RNAP binding site somehow controls how RNAP interacts with CRP in this class II configuration. Additional work will be required to understand this sequence-dependence, which to our knowledge has not been previously reported.

Our strategy has been designed to be compatible with massively parallel reporter assays (MPRAs), which use ultra-high-throughput DNA sequencing to measure the activities of thousands of transcriptional regulatory sequences simultaneously. We expect that MPRAs, performed on microarray-synthesized promoter libraries, should allow hundreds of expression manifolds to be measured in a single experiment. MPRAs will also facilitate the study of TFs that cannot be controlled by a small molecule: one can measure t_+ and t_- by assaying promoters that either do or do not have a functional TF binding site but are otherwise identical. The ease with which MPRAs can assay promoters with different combinations of sites turned “on” and “off” should enable the study of more complex regulatory architectures, beyond just simple repression and simple activation.

Based on these results, we advocate a very different approach to dissecting transcriptional regulatory grammar than has been pursued by other groups. Instead of assaying and modeling many different arrangements of transcription factor binding sites (***Gertz et al., 2009***; ***Sharon et al., 2012***; ***Mogno et al., 2013***; ***Smith et al., 2013***; ***Levo and Segal, 2014***; ***White et al., 2016***) or the activity of completely random DNA (***de Boer et al., 2017***), we suggest that more attention be paid to the interactions that occur within *specific* binding site configurations. Expression manifolds provide a useful way of interrogating individual protein-DNA and protein-protein interactions that occur in a specific promoter architecture without requiring a holistic model that aims to describe arbitrary binding site arrangements. Using MPRAs to simultaneously assay hundreds of systematically varied architectures, we expect that it should be possible to build biophysical models of transcriptional regulatory grammar from the ground up.

What would high-precision knowledge of transcriptional regulatory grammar in bacteria do for us? For one thing, it would greatly facilitate the interpretation of bacterial genome sequences. Currently, it is difficult to predict the functional consequences of TF binding sites just from their locations relative to annotated TSSs. Knowing the distance-dependent interactions between RNAP and common *E. coli* TFs would greatly illuminate how previously annotated binding sites for these

TFs actually affect expression. Such knowledge would also facilitate MPRA-based efforts to dissect previously unannotated regulatory sequences across the genome (**Belliveau et al., 2018**).

Precise knowledge of transcriptional regulatory grammar in bacteria would also have important implications for synthetic biology. Currently, complex biological computations are performed in synthetic systems by stringing simple promoter "parts" together into complex regulatory networks. By contrast, naturally occurring promoters can often perform quite complex computations themselves via the multi-protein-DNA complexes that they scaffold (**Kuhlman et al., 2007**; **Cui et al., 2013**). Such computational mechanisms have many potential advantages, including faster response times and increased robustness to stochastic fluctuations. These advantages could be particularly useful in metabolic engineering, which requires rapid and reliable control over the expression of multiple genes in a pathway (**Smanski et al., 2016**; **Nielsen and Keasling, 2016**; **Zhao et al., 2018**). But although the potential capabilities of complex promoters have been explored both theoretically (**Buchler et al., 2003**; **Bintu et al., 2005**) and experimentally (**Setty et al., 2003**; **Mayo et al., 2006**; **Segall-Shapiro et al., 2018**), there remains little capability in synthetic biology to design complex promoters with predictable quantitative behavior. High-precision knowledge of the energetics underlying transcriptional regulatory grammar could enable this capability.

Will expression manifolds be useful for understanding transcriptional regulation in eukaryotes? Both FACS-based MPRAs (**Sharon et al., 2012**; **Weingarten-Gabbay et al., 2017**) and RNA-Seq-based MPRAs (**Melnikov et al., 2012**; **Kwasnieski et al., 2012**; **Patwardhan et al., 2012**) are well established in eukaryotes so, on a technical level, experiments analogous to those described here should be feasible. The bigger question, we believe, is whether the results of such experiments would be interpretable. Eukaryotic transcriptional regulation is far more complex than transcriptional regulation in bacteria. In fact, it is not even clear what mutations to the basal promoter in eukaryotes might correspond to the mutations in the RNAP site that we relied upon here. Still, we believe that pursuing this strategy in eukaryotes is worthwhile. Despite the underlying complexities, simple “effective” models of regulatory biophysics might work surprisingly well.

Materials and Methods

Media

Expression measurements were performed on cells grown in rich defined media (RDM; purchased from Teknova) (**Neidhardt et al., 1974**) supplemented with 10 mM $NaHCO_3$, 1 mM IPTG (Sigma), and 0.2% glucose. In what follows we refer to this media as RDM’. RDM’ was further supplemented with 50 μg/ml kanamycin (Sigma) when growing cells, as well as 250 μM cAMP (Sigma) when measuring t_+.

Strains

Expression measurements were performed in *E. coli* strain JK10, which has genotype Δ*cyaA* Δ*cpdA* Δ *lacY* Δ*lacZ* Δ*dksA*. JK10 is derived from strain TK310 (**Kuhlman et al., 2007**), which is Δ*cyaA* Δ*cpdA* Δ*lacY*. The Δ*cyaA* Δ*cpdA* mutations prevent TK310 from synthesizing or degrading cAMP, thus allowing *in vivo* cAMP concentrations to be quantitatively controlled by adding cAMP to the growth media. Into TK310 we introduced the Δ*lacZ* mutation, yielding strain DJ33; this mutation allows Miller assays to be used in conjunction with plasmid-based reporters driving *lacZ* expression. In our initial experiments, we found that the growth rate of DJ33 in RDM’ varies strongly with amount of cAMP added to the media. Fortunately, we isolated a spontaneous knock-out mutation in *dksA* (thus yielding JK10), which caused the growth rate (~ 30 min doubling time) in RDM’ to be independent of cAMP concentrations below ~ 500 μM.[3] The JK10 genotype was confirmed by whole genome sequencing.

[3] Note, however, that JK10 will not grow in minimal media in the absence of cAMP.

Reporter constructs

Expression of the *lacZ* gene was driven from variants of a plasmid we call pJK48. These reporter constructs were cloned as follows. We started with the vector pJK14 from ***Kinney et al.*** (***2010***). pJK14 contains a pSC101 origin of replication (~ 5-10 copies per cell), a kanamycin resistance gene, and a *ccdB* cloning cassette positioned immediately upstream of a *gfpmut2* reporter gene and flanked by outward-facing BsmBI restriction sites. First, the *gfpmut2* gene in this vector was replaced with *lacZ*, yielding pJK47. Next, the ribosome binding site in the 5′ UTR of *lacZ* was weakened, yielding pJK47.419; this weakening prevents *lacZ* expression from a maximally active promoter from substantially slowing cell growth in RDM′. pJK47.419 was propagated in DB3.1 *E. coli* (Invitrogen), which is resistant to the CcdB toxin.

The promoters we assayed were variants of what we call the lac* promoter. The lac* promoter is similar to the endogenous *lac* promoter of *E. coli* MG1655 except for (i) it contains a CRP binding site with a consensus right pentamer and (ii) it contains mutations that were introduced in an effort to remove previously reported cryptic promoters (***Reznikoff, 1992***). Promoter-containing insertion cassettes were created through overlap-extension PCR and flanked by outward-facing BsaI restriction sites. All primers were ordered from Integrated DNA Technologies. Note that some of the primers used to create these inserts were synthesized using pre-mixed phosphoramidites at specified positions; this is how a 24% mutation rate in the -10 or -35 regions of the RNAP binding site was achieved. The resulting promoter sequences are illustrated in Appendix 1 Figure 1.

To clone variants of pJK48, we separately digested the pJK47.419 vector with BsmBI (NEB) and the appropriate insert with BsaI (NEB). Digests were then cleaned up (Qiagen PCR purification kit) and ligated together in at 1:1 molar ratio for 1 hour using T4 DNA ligase (Invitrogen). After 90 min dialysis, plasmids were transformed into electrocompetent JK10 cells. Individual clones were plated on LB supplemented with kanamycin (50 μg/ml), while libraries were grown in 50 ml LB supplemented with kanamycin. After initial cloning, each clone was re-streaked, grown in LB+kan, and stored as a catalogued glycerol stock. The promoter region of each clone was sequenced in both directions. Only plasmids with validated promoter sequences were used for the measurements presented in this paper. The promoter sequences of all constructs used in this study, as well as their measured t_+ and t_- values, are provided at https://github.com/jbkinney/18_expressionmanifolds.

Miller assays

Expression was quantified using ONPG-based β-galactosidase activity measurements adapted from the method of ***Miller*** (***1972***). Specifically, we obtained t_+ and t_- measurements for each clone as follows.

First, each clone was streaked out on LB+kan agar and grown overnight. A colony was then picked and used to inoculate a 1.5 ml overnight LB+kan liquid culture. Either 8 μl, 6 μl, or 4 μl of the overnight culture were then diluted into 200 μl RDM′+kan. 25 μl of each dilution was then added to 175 μl RDM′+kan in a 96-well optical bottom plate and supplemented with either 0 μM cAMP (for t_-) or 250 μM cAMP (for t_+). The plate was then covered with Breathe-Easier film (USA Scientific) and cells were cultured for ~ 3 hr at 37 °C, shaking at 900 RPM in a microplate shaker. During this time, 5.5 ml of lysis buffer was freshly prepared using 1.5 ml RDM′, 4.0 ml PopCulture reagent (Millipore), 114 μl of 35 mg/ml chloramphenicol (Sigma), and 44 μl of 40 U/μl rLysozyme (Sigma).

Microplate film was removed and cell density (quantified by A_{600}) was measured using an Epoch 2 Microplate Spectrophotometer (BioTek). Cells were then lysed by adding 25 μl lysis buffer to each microplate well, incubating the microplate at room temperature for 10 minutes without shaking, then cooling the microplate at 4 °C for a minimum of 15 minutes. In each well of a 96-well optical bottom plate, 50 μl of lysate was then added to 50 μl of pre-chilled Z-buffer (***Miller, 1972***) containing 1 mg/ml ONPG (Sigma). Samples were sealed with optical film and both A_{420} and A_{550} were periodically measured in the plate reader over an extended period of time (every 1.5 min for 1 hour or every 15 min for 10 hours, depending on the level of expression expected).

The expression levels t_+ and t_- were quantified from these absorbance data using the formula

$$t_\pm = \frac{\Delta A_{420} - \Delta A_{550}}{V \cdot \Delta T \cdot A_{600}}, \tag{9}$$

where V = 50 is the volume of lysate in μl added to the ONPG reaction, ΔT is the change in time from the beginning of the measurement, and ΔA_X indicates a change in absorbance at X nm over this time interval. Only data from wells with $A_{600} \lesssim 0.5$ were analyzed. Note that the A_{550} term in Equation 9 is not multiplied by 1.75 as it is in ***Miller*** (***1972***). This is because our A_{550} measurements are used to compensate for condensation on the microplate film, not for cellular debris as in (***Miller, 1972***); our lysis procedure produces no detectable cellular debris. In practice, Equation 9 was not evaluated using individual measurements, but was rather computed from the slope of a line fit to non-saturated absorbance measurements using custom Python scripts. Raw A_{420}, A_{550}, and A_{600} values, as well as our analysis scripts, are available at https://github.com/jbkinney/18_expressionmanifolds. In all the figures, median values from at least 3 independent Miller measurements were used to define each measured t_+ and t_- data point.

Parameter inference

Expression manifold parameters were fit to measured t_+ and t_- values as follows. First, outlier data points were called by eye and excluded from the parameter fitting procedure. We denote the remaining measurements using $t_+^{i,\text{data}}$ and $t_-^{i,\text{data}}$, where $i = 1, 2, \dots n$ indexes the non-outlier data points. These $2n$ measurements were used to fit $n + 3$ parameters: the saturated transcription rate (t_sat), the background transcription rate (t_bg), the renormalized cooperativity (α')[4], and the RNAP binding factors for each assayed RNAP site ($P_1, P_2, \dots, P_n$). This was accomplished using nonlinear least squares. Specifically, we minimized the loss function $\mathcal{L}(\theta)$

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \left(\left[\log \frac{t_+^{i,\text{model}}(\theta)}{t_+^{i,\text{data}}} \right]^2 + \left[\log \frac{t_-^{i,\text{model}}(\theta)}{t_-^{i,\text{data}}} \right]^2 \right) \tag{10}$$

where $\theta = \{t_\text{sat}, t_\text{bg}, \alpha', P_1, P_2, \dots, P_n\}$ are the model parameters and

$$t_+^{i,\text{model}}(\theta) = t_\text{sat} \frac{\alpha' P_i}{1 + \alpha' P_i} + t_\text{bg}, \quad t_-^{i,\text{model}}(\theta) = t_\text{sat} \frac{P_i}{1 + P_i} + t_\text{bg}. \tag{11}$$

The solid black lines in Figure 2B and Figures 4B,C show the expression manifolds fit to all n data points. The gray lines in Figure 2B and Figure 4B represent parameters fit to bootstrap-resampled data points.

The values reported for F and α, as well as for ΔG_F and ΔG_α, were computed using parameters fit to bootstrap-resampled data. For the occlusion data in Figure 2B, we reported

$$F = (F_{50})^{+(F_{84}-F_{50})}_{-(F_{50}-F_{16})}, \quad \Delta G_F = -k_B T \log F_{50} \pm k_B T \left(\frac{\log F_{84} - \log F_{16}}{2} \right), \tag{12}$$

where $1 k_B T = 1.62$ kcal/mol (corresponding to 37 °C) and where F_{84}, F_{50}, and F_{16} respectively denote the 84th, 50th, and 16th percentiles of F values obtained from bootstrap resampling. For the activation data in Figures 4B and 4C, we computed α from α' via $\alpha = \alpha' - (\alpha' - 1)/F_{50}$. We then reported

$$\alpha = (\alpha_{50})^{+(\alpha_{84}-\alpha_{50})}_{-(\alpha_{50}-\alpha_{16})}, \quad \Delta G_\alpha = -k_B T \log \alpha_{50} \pm k_B T \left(\frac{\log \alpha_{84} - \log \alpha_{16}}{2} \right), \tag{13}$$

where α_{84}, α_{50}, and α_{16} respectively denote the 84th, 50th, and 16th percentiles of α values obtained from bootstrap resampling.

By visual inspection of Figure 6B, we determined that $\beta \approx 1$ at 61.5 bp. In Figure 6C, we therefore report for each position X, an acceleration β^X given by $t_\text{sat}^X / t_\text{sat}^{-61.5}$ where $t_\text{sat}^{-61.5}$ is the saturated rate of transcription inferred for -61.5 bp in Figure 4B and, similarly, t_sat^X denotes the saturated rate of

[4] Note that $\alpha' = 1/(1 + F)$ in the case of simple repression, as in Figure 2.

transcription inferred for position X in Figure 4C. Plotted points show the median values, while error bars show the [16%, 85%] quantile interval.

Figure 8 shows $P_{i,50}$ values with error bars extending from $[P_{i,16}$ to $P_{i,84}]$. Such values were computed using P-values determined from data in which the individual replicates for each promoter were bootstrap resampled, but for which all promoters were used in the inference procedure.

Author contributions

JBK conceived of this study. TF and JBK designed this study. JBK, TF, AA, MSG, and DJ carried out the experiments. TF and JBK carried out the computational analysis. JBK wrote the manuscript with input from MSG, RP, DJ, TF, and AA. JBK funded this study.

Acknowledgments

We thank Bryce Nickels and Stirling Churchman for helpful feedback. This work was supported by a CSHL/Northwell Health Alliance grant to JBK and by NIH Cancer Center Support Grant 5P30CA045508.

References

Ackers, G., Johnson, A., and Shea, M. (1982). Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A*, 79(4):1129–1133.

Adhya, S. (1996). The lac and gal operons today. *Regulation of Gene Expression in Esherichia coli*, pages 1–20.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723.

Beckwith, J., Grodzicker, T., and Arditti, R. (1972). Evidence for two sites in the lac promoter region. *J Mol Biol*, 69(1):155–160.

Belliveau, N. M., Barnes, S. L., Ireland, W. T., Jones, D. L., Sweredoski, M. J., Moradian, A., Hess, S., Kinney, J. B., and Phillips, R. (2018). Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc Natl Acad Sci USA*, 115(21):E4796–E4805.

Berger, M., Philippakis, A., Qureshi, A., He, F., Estep, P., and Bulyk, M. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–1435.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15(2):116–124.

Brewster, R. C., Jones, D. L., and Phillips, R. (2012). Tuning promoter strength through RNA polymerase binding site design in Escherichia coli. *PLoS Comput Biol*, 8(12):e1002811.

Brewster, R. C., Weinert, F. M., Garcia, H. G., Song, D., Rydenfelt, M., and Phillips, R. (2014). The transcription factor titration effect dictates level of gene expression. *Cell*, 156(6):1312–1323.

Browning, D. F. and Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nat Rev Microbiol*, 14(10):638–650.

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proc Natl Acad Sci USA*, 100(9):5136–5141.

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

Courey, A. J. (2008). *Mechanisms in transcriptional regulation*. Blackwell, Malden, MA.

Cui, L., Murchland, I., Shearwin, K. E., and Dodd, I. B. (2013). Enhancer-like long-range transcriptional activation by λ CI-mediated DNA looping. *Proc Natl Acad Sci USA*, 110(8):2922–2927.

de Boer, C., Sadeh, R., Friedman, N., and Regev, A. (2017). Deciphering cis-regulatory logic with 100 million synthetic promoters. *bioRxiv*, pages 1–31.

Ebright, R. H., Ebright, Y. W., and Gunasekera, A. (1989). Consensus DNA site for the Escherichia coli catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the E. coli lac DNA site. *Nucl Acids Res*, 17(24):10295–10305.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Garcia, H. G. and Phillips, R. (2011). Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci USA*, 108(29):12173–12178.

Gaston, K., Bell, A., Kolb, A., Buc, H., and Busby, S. (1990). Stringent spacing requirements for transcription activation by CRP. *Cell*, 62(4):733–743.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dosé, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Rätsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., modENCODE Consortium, Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., and Waterston, R. H. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, 330(6012):1775–1787.

Gertz, J., Siggia, E. D., and Cohen, B. A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–218.

Gunasekera, A., Ebright, Y., and Ebright, R. (1992). DNA sequence determinants for binding of the Escherichia coli catabolite gene activator protein. *J Biol Chem*, 267(21):14713–14720.

Johnson, D., Mortazavi, A., Myers, R., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, 20(6):861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339.

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A. G., Mackie, A., Paulsen, I., Gunsalus, R. P., and Karp, P. D. (2011). EcoCyc: a comprehensive database of Escherichia coli biology. *Nucl Acids Res*, 39(Database issue):D583–90.

Kinney, J. B., Murugan, A., Callan, C. G., and Cox, E. C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci USA*, 107(20):9158–9163.

Kuhlman, T., Zhang, Z., Saier, M. H., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of Escherichia coli. *PNAS*, 104(14):6043–6048.

Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., and Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci USA*, 109(47):19498–19503.

Lee, D. J., Minchin, S. D., and Busby, S. J. W. (2012). Activating transcription in bacteria. *Annu Rev Microbiol*, 66(1):125–152.

Levo, M. and Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nat Rev Genet*, 15(7):453–468.

Malan, T., Kolb, A., Buc, H., and McClure, W. (1984). Mechanism of CRP-cAMP activation of lac operon transcription initiation activation of the P1 promoter. *J Mol Biol*, 180(4):881–909.

Mayo, A. E., Setty, Y., Shavit, S., Zaslaver, A., and Alon, U. (2006). Plasticity of the cis-regulatory input function of a gene. *PLoS Biol*, 4(4):e45.

McClure, W. R., Hawley, D. K., Youderian, P., and Susskind, M. M. (1983). DNA determinants of promoter selectivity in Escherichia coli. *Cold Spring Harb Symp Quant Biol*, 47 Pt 1:477–481.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Kinney, J. B., Kellis, M., Lander, E. S., and Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*, 30(3):271–277.

Meng, X., Brodsky, M. H., and Wolfe, S. A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat Rev Microbiol*, 23(8):988–994.

Miller, J. (1972). *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M. Y., Will, S., Alekseyenko, A. A., Artieri, C., Booth, B. W., Brooks, A. N., Dai, Q., Davis, C. A., Duff, M. O., Feng, X., Gorchakov, A. A., Gu, T., Henikoff, J. G., Kapranov, P., Li, R., MacAlpine, H. K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S. K., Riddle, N. C., Sakai, A., Samsonova, A., Sandler, J. E., Schwartz, Y. B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K. H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S. E., Brent, M. R., Cherbas, L., Elgin, S. C. R., Gingeras, T. R., Grossman, R., Hoskins, R. A., Kaufman, T. C., Kent, W., Kuroda, M. I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J. W., Ren, B., Russell, S., Cherbas, P., Graveley, B. R., Lewis, S., Micklem, G., Oliver, B., Park, P. J., Celniker, S. E., Henikoff, S., Karpen, G. H., Lai, E. C., MacAlpine, D. M., Stein, L. D., White, K. P., and Kellis, M. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012):1787–1797.

Mogno, I., Kwasnieski, J. C., and Cohen, B. A. (2013). Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res*, 23(11):1908–1915.

Morita, T., Shigesada, K., Kimizuka, F., and Aiba, H. (1988). Regulatory effect of a synthetic CRP recognition sequence placed downstream of a promoter. *Nucl Acids Res*, 16(15):7315–7332.

Mukherjee, S., Berger, M., Jona, G., Wang, X., Muzzey, D., Snyder, M., Young, R., and Bulyk, M. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36(12):1331–1339.

Neidhardt, F. C., Bloch, P. L., and Smith, D. F. (1974). Culture medium for enterobacteria. *J Bacteriol*, 119(3):736–747.

Nielsen, J. and Keasling, J. D. (2016). Engineering Cellular Metabolism. *Cell*, 164(6):1185–1197.

Niu, W., Kim, Y., Tau, G., Heyduk, T., and Ebright, R. H. (1996). Transcription activation at class II CAP-dependent promoters: two interactions between CAP and RNA polymerase. *Cell*, 87(6):1123–1134.

Noyes, M., Christensen, R., Wakabayashi, A., Stormo, G., Brodsky, M., and Wolfe, S. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, 133(7):1277–1289.

Parkinson, G., Wilson, C., Gunasekera, A., Ebright, Y. W., Ebright, R. H., Ebright, R. E., and Berman, H. M. (1996). Structure of the CAP-DNA complex at 2.5 angstroms resolution: a complete picture of the protein-DNA interface. *J Mol Biol*, 260(3):395–408.

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S.-I., Cooper, G. M., Ahituv, N., Pennacchio, L. A., and Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol*, 30(3):265–270.

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc Natl Acad Sci USA*, 72(3):784–788.

Ptashne, M. (2003). Regulated recruitment and cooperativity in the design of biological regulatory systems. *Philos Transact A Math Phys Eng Sci*, 361(1807):1223–1234.

Ptashne, M. and Gann, A. (2002). *Genes and signals*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Ren, B., Robert, F., Wyrick, J., Aparicio, O., Jennings, E., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T., Wilson, C., Bell, S., and Young, R. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309.

Reznikoff, W. S. (1992). The lactose operon-controlling elements: a complex paradigm. *Mol Microbiol*, 6(17):2419–2422.

Rhee, H. S. and Pugh, B. F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, 147(6):1408–1419.

Rhodius, V. A., West, D. M., Webster, C. L., Busby, S. J., and Savery, N. J. (1997). Transcription activation at class II CRP-dependent promoters: the role of different activating regions. *Nucl Acids Res*, 25(2):326–332.

Ruff, E. F., Record, M. T., and Artsimovitch, I. (2015). Initial events in bacterial transcription initiation. *Biomolecules*, 5(2):1035–1062.

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernández, S., Alquicira-Hernández, K., López-Fuentes, A., Porrón-Sotelo, L., Huerta, A. M., Bonavides-Martinez, C., Balderas-Martínez, Y. I., Pannier, L., Olvera, M., Labastida, A., Jiménez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chávez, V., Hernández-Alvarez, A., Morett, E., and Collado-Vides, J. (2013). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucl Acids Res*, 41(Database issue):D203–13.

Segall-Shapiro, T. H., Sontag, E. D., and Voigt, C. A. (2018). Engineered promoters enable constant gene expression at any copy number in bacteria. *Nat Rev Microbiol*, 36(4):352–358.

Setty, Y., Mayo, A., Surette, M., and Alon, U. (2003). Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci U S A*, 100(13):7702–7707.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*, 30(6):521–530.

Shea, M. A. and Ackers, G. K. (1985). The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol*, 181(2):211–230.

Sherman, M. S. and Cohen, B. A. (2012). Thermodynamic state ensemble models of cis-regulation. *PLoS Comput Biol*, 8(3):e1002407.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H. J., and Mann, R. S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, 147(6):1270–1282.

Smanski, M. J., Zhou, H., Claesen, J., Shen, B., Fischbach, M. A., and Voigt, C. A. (2016). Synthetic biology to access and expand nature's chemical diversity. *Nat Rev Microbiol*, 14(3):135–149.

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet*, 45(9):1021–1028.

So, L.-h., Ghosh, A., Zong, C., Sepúlveda, L. A., Segev, R., and Golding, I. (2011). General properties of transcriptional time series in Escherichia coli. *Nature Genetics*, 43(6):554–560.

Ushida, C. and Aiba, H. (1990). Helical phase dependent action of CRP: effect of the distance between the CRP site and the -35 region on promoter activity. *Nucl Acids Res*, 18(21):6325–6330.

Vilar, J. M. G. and Leibler, S. (2003). DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–989.

Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A., and Segal, E. (2017). Deciphering Transcriptional Regulation of Human Core Promoters. *bioRxiv*, pages 1–27.

Weingarten-Gabbay, S. and Segal, E. (2014). The grammar of transcriptional regulation. *Hum. Genet.*, 133(6):701–711.

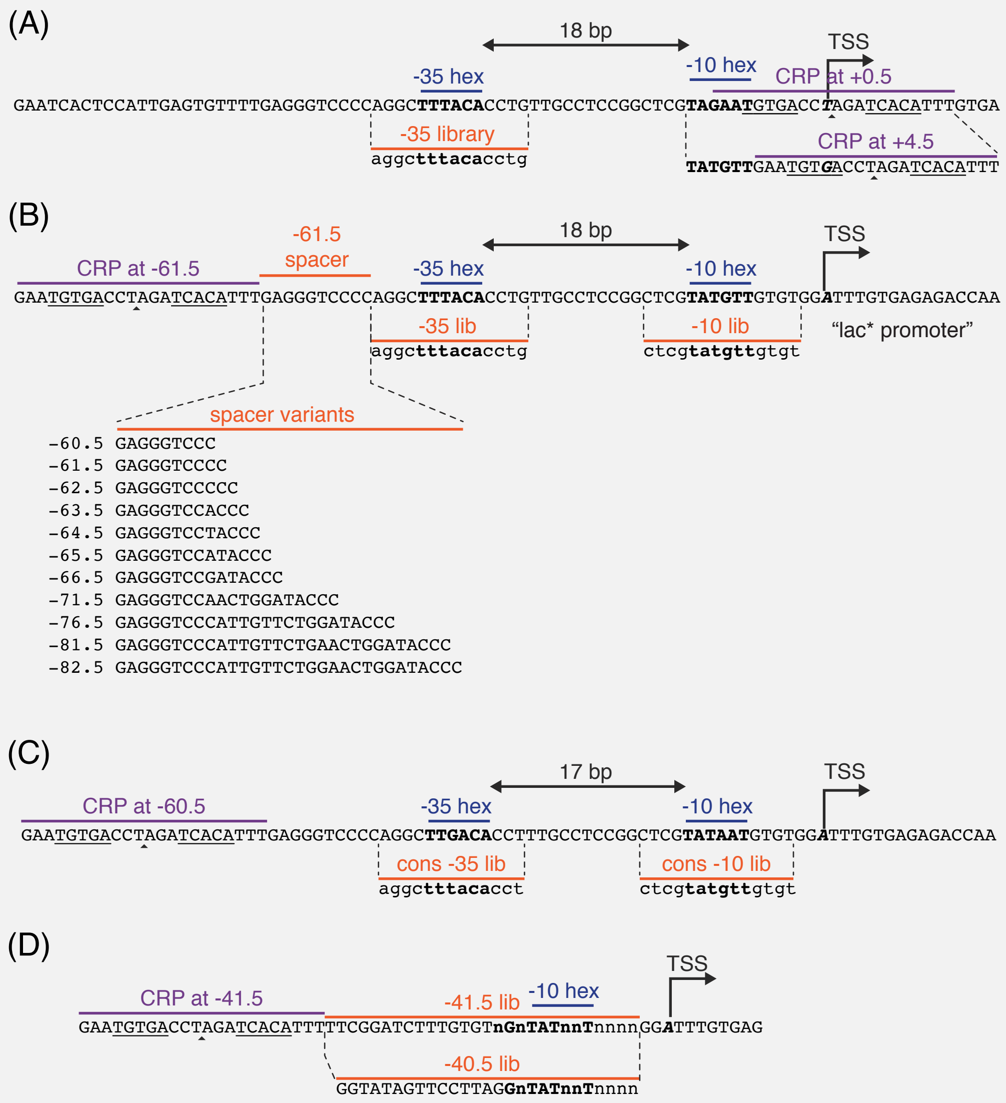
White, M. A., Kwasnieski, J. C., Myers, C. A., Shen, S. Q., Corbo, J. C., and Cohen, B. A. (2016). A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. *Cell Rep*, 17(5):1247–1254.

Zhao, E. M., Zhang, Y., Mehl, J., Park, H., Lalwani, M. A., Toettcher, J. E., and Avalos, J. L. (2018). Optogenetic regulation of engineered cellular metabolism for microbial chemical production. *Nature*, 555(7698):683–687.

Zhao, Y., Granas, D., and Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12):e1000590.

Appendix 1

Promoter variants


(A)
18 bp
TSS
-35 hex
-10 hex
CRP at +0.5
GAATCACTCCATTGAGTGTTTTGAGGGTCCCCAGGCTTTACACCTGTTGCCTCCGGCTCGTAGAATGTGACCTAGATCACATTTGTGA
-35 library
aggctttacacctg
CRP at +4.5
TATGTTGAATGTGACCTAGATCACATTT
(B)
-61.5
spacer
18 bp
TSS
CRP at -61.5
-35 hex
-10 hex
GAATGTGACCTAGATCACATTTGAGGGTCCCCAGGCTTTACACCTGTTGCCTCCGGCTCGTATGTTGTGTGGATTTGTGAGAGACCAA
-35 lib
aggctttacacctg
-10 lib
ctcgtatgttgtgt
“lac* promoter”
spacer variants
-60.5 GAGGGTCCC
-61.5 GAGGGTCCCC
-62.5 GAGGGTCCCCC
-63.5 GAGGGTCCACCC
-64.5 GAGGGTCCTACCC
-65.5 GAGGGTCCATACCC
-66.5 GAGGGTCCGATACCC
-71.5 GAGGGTCCAACTGGATACCC
-76.5 GAGGGTCCCATTGTTCTGGATACCC
-81.5 GAGGGTCCCATTGTTCTGAACTGGATACCC
-82.5 GAGGGTCCCATTGTTCTGGAACTGGATACCC
(C)
17 bp
TSS
CRP at -60.5
-35 hex
-10 hex
GAATGTGACCTAGATCACATTTGAGGGTCCCCAGGCTTGACACCTTTGCCTCCGGCTCGTATAATGTGTGGATTTGTGAGAGACCAA
cons -35 lib
aggctttacacct
cons -10 lib
ctcgtatgttgtgt
(D)
TSS
CRP at -41.5
-41.5 lib
-10 hex
GAATGTGACCTAGATCACATTTTTCGGATCTTTGTGTnGnTATnnTnnnnGGATTTGTGAG
-40.5 lib
GGTATAGTTCCTTAGGnTATnnTnnnn

Appendix 1 Figure 1. Promoter sequences used in this study. In all panels, the -35 and -10 hexamers of the RNAP binding site are in bold. CRP binding site centers are indicated by small triangles. The dyadic pentamers of the core CRP binding site in each construct are underlined. The transcription start site (TSS) is bold and italicized. Lowercase bases (‘a’,‘c’,‘g’, and ‘t’) indicate positions synthesized with a 24% mutation rate. The lowercase character ‘n’ indicates completely randomized positions. (A) Occlusion promoters assayed for Figure 2. (B) Class I promoters assayed for Figure 4. In the main text we refer to the wild-type promoter with CRP at -61.5 bp as the “lac* promoter”. The lac* promoter served as the template for all of the promoters shown here. (C) Strong class I promoters assayed for Figure 6. (D) Class II promoters assayed for Figure 7.