How the avidity of polymerase binding to the –35/–10 promoter sites affects gene expression

Tal Einav[a,1] and Rob Phillips[a,b,c,1]

[a]Department of Physics, California Institute of Technology, Pasadena, CA 91125; [b]Department of Applied Physics, California Institute of Technology, Pasadena, CA 91125; and [c]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125

Edited by Ken A. Dill, Stony Brook University, Stony Brook, NY, and approved May 16, 2019 (received for review April 3, 2019)

Although the key promoter elements necessary to drive transcription in *Escherichia coli* have long been understood, we still cannot predict the behavior of arbitrary novel promoters, hampering our ability to characterize the myriad sequenced regulatory architectures as well as to design new synthetic circuits. This work builds upon a beautiful recent experiment by Urtecho et al. [G. Urtecho, *et al.*, *Biochemistry*, 68, 1539–1551 (2019)] who measured the gene expression of over 10,000 promoters spanning all possible combinations of a small set of regulatory elements. Using these data, we demonstrate that a central claim in energy matrix models of gene expression—that each promoter element contributes independently and additively to gene expression—contradicts experimental measurements. We propose that a key missing ingredient from such models is the avidity between the –35 and –10 RNA polymerase binding sites and develop what we call a multivalent model that incorporates this effect and can successfully characterize the full suite of gene expression data. We explore several applications of this framework, namely, how multivalent binding at the –35 and –10 sites can buffer RNA polymerase (RNAP) kinetics against mutations and how promoters that bind overly tightly to RNA polymerase can inhibit gene expression. The success of our approach suggests that avidity represents a key physical principle governing the interaction of RNA polymerase to its promoter.

transcription regulation | avidity | statistical mechanics

Promoters modulate the complex interplay of RNA polymerase (RNAP) and transcription factor binding that ultimately regulates gene expression. While our knowledge of the molecular players that mediate these processes constantly improves, more than half of all promoters in *Escherichia coli* still have no annotated transcription factors in RegulonDB (1) and our ability to design novel promoters that elicit a target level of gene expression remains limited.

As a step toward taming the vastness and complexity of sequence space, the recent development of massively parallel reporter assays has enabled entire libraries of promoter mutants to be simultaneously measured (2–4). Given this surge in experimental prowess, the time is ripe to reexamine how well our models of gene expression can extrapolate the response of a general promoter.

A common approach to quantifying gene expression, called the energy matrix model, assumes that every promoter element contributes additively and independently to the total RNAP (or transcription factor) binding energy (3). This model treats all base pairs on an equal footing and does not incorporate mechanistic details of RNAP–promoter interactions such as its strong binding primarily at the −35 and −10 binding motifs (Fig. 1*A*). A newer method recently took the opposite viewpoint, designing an RNAP energy matrix that includes only the −35 element, the −10 element, and the length of the spacer separating them (5), neglecting the sequence composition of the spacer or the surrounding promoter region.

Although these methods have been successfully used to identify important regulatory elements in unannotated promoters (6) and predict evolutionary trajectories (5), it is clear that there is more to the story. Even in the simple case of the highly studied *lac* promoter, such energy matrices show systematic deviations from measured levels of gene expressions, indicating that some fundamental component of transcriptional regulation is still missing (7).

We propose that one failure of current models lies in their tacit assumption that every promoter element contributes independently to the RNAP binding energy. By naturally relaxing this assumption to include the important effects of avidity, we can push beyond the traditional energy matrix analysis in several key ways, including the following: (*i*) We can identify which promoter elements contribute independently or cooperatively without recourse to fitting, thereby building an unbiased mechanistic model for systems that bind at multiple sites. (*ii*) Applying this approach to RNAP–promoter binding reveals that the −35 and −10 motifs bind cooperatively, a feature that we attribute to avidity. Moreover, we show that models that instead assume the −35 and −10 elements contribute additively and independently sharply contradict the available data. (*iii*) We show that the remaining promoter elements (the spacer, upstream [UP], and background shown in Fig. 1*A*) do contribute independently and additively to the RNAP binding energy and formulate the corresponding model for transcriptional regulation that we call a multivalent model. (*iv*) We use this model to explore how the interactions between the −35 and −10 elements can buffer RNAP kinetics against mutations. (*v*) We analyze the surprising phenomenon that overly tight RNAP–promoter binding leads to decreased gene expression. (*vi*) We validate our model by analyzing the expression of over 10,000 promoters in *E. coli* recently published by Urtecho et al. (8) and demonstrate that

Significance

Cellular behavior is ultimately governed by the genetic program encoded in the cell's DNA and through the arsenal of molecular machines that actively transcribe its genes, yet we lack the ability to predict how an arbitrary DNA sequence will perform. To that end, we analyze the performance of over 10,000 regulatory sequences and develop a model that can predict the behavior of any sequence based on its composition. By considering promoters that vary only by one or two elements, we can characterize how different components interact, providing fundamental insights into the mechanisms of transcription.

Author contributions: T.E. designed research; T.E. and R.P. performed research; T.E. analyzed data; and T.E. and R.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: phillips@pboc.caltech.edu or tal.einav@alumni.caltech.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1905615116/-/DCSupplemental.

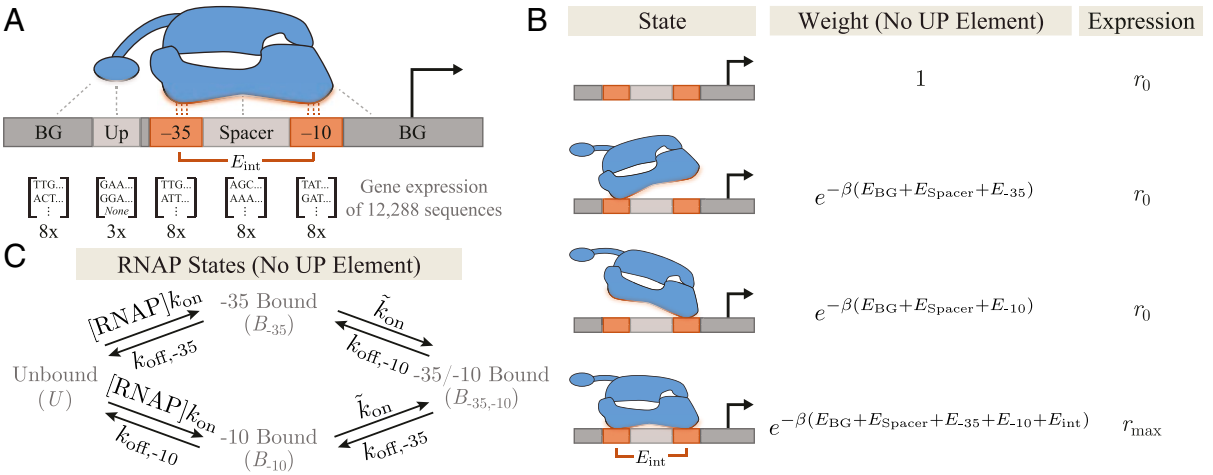

Fig. 1. The bivalent nature of RNAP–promoter binding. (*A*) Expression was measured for promoters comprising any combination of −35, −10, spacer, UP, and background (BG) elements. (*B*) Without an UP element, RNAP contacts the promoter at the −35 and −10 sites, giving rise to expression r_0 when unbound or partially bound and $r_{\max}$ when fully bound. (*C*) Having two binding sites alters the dynamics of RNAP binding. k_{on} represents the on-rate from unbound to partially bound RNAP, $\tilde{k}_{\text{on}}$ the analogous rate from partially to fully bound RNAP, and $k_{\text{off},j}$ denotes the unbinding rate from site j.

our framework markedly improves upon the traditional energy matrix analysis.

While this work focuses on RNAP–promoter binding, its implications extend to general regulatory architectures involving multiple tight-binding elements, including transcriptional activators that make contact with RNAP (CRP in the *lac* promoter) (9), transcription factors that oligomerize (as recently identified for the *xylE* promoter) (6), and transcription factors that bind to multiple sites on the promoter (DNA looping mediated by the Lac repressor) (10). More generally, this approach of categorizing which binding elements behave independently (without resorting to fitting) can be applied to multivalent interactions in other biological contexts, including novel materials, scaffolds, and synthetic switches (11, 12).

Results

The –35 and –10 Binding Sites Give Rise to Gene Expression That Defies Characterization as Independent and Additive Components. Decades of research have shed light upon the exquisite biomolecular details involved in bacterial transcriptional regulation via the family of RNAP σ factors (13). In this work, we restrict our attention to the σ^{70} holoenzyme (8), the most active form under standard *E. coli* growth conditions, whose interaction with a promoter includes direct contact with the −35 and −10 motifs (two hexamers centered roughly 10 and 35 bases upstream of the transcription start site), a spacer region separating these two motifs, an UP element just upstream of the −35 motif that anchors the C-terminal domain (αCTD) of RNAP, and the background promoter sequence surrounding these elements.

Urtecho et al. (8) constructed a library of promoters composed of every combination of eight −35 motifs, eight −10 motifs, eight spacers, eight backgrounds (BGs), and three UP elements (Fig. 1*A*). Each sequence was integrated at the same locus within the *E. coli* genome and transcription was quantified via DNA barcoding and RNA sequencing. One of the three UP elements considered was the absence of an UP binding motif, and this case serves as the starting point for our analysis.

The energy matrix approach used by Urtecho et al. (8) posits that every base pair of the promoter will contribute additively and independently to the RNAP binding energy, which by appropriately grouping base pairs is equivalent to stating that the free energy of RNAP binding will be the sum of its contributions from the background, spacer, −35, and −10 elements (*SI Appendix*, section A). Hence, the gene expression (GE) is given by the Boltzmann factor

$$\text{GE} \propto e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{-35}+E_{-10})}. \quad [1]$$

Note that all E_js represent free energies (with an energetic and entropic component); to see the explicit dependence on RNAP copy number, refer to *SI Appendix*, section A. Fitting the 32 free energies (one for each background, spacer, −35, and −10 element) and the constant of proportionality in Eq. **1** on 25% of the data enables us to predict the expression on the remainder of the $8 \times 8 \times 8 \times 8 = 4{,}096$ promoters (*Materials and Methods*).

Fig. 2*A* demonstrates that Eq. **1** leads to a poor characterization of these promoters ($R^2 = 0.57$; parameter values listed in *SI Appendix*, section B), suggesting that critical features of gene expression are missing from this model. One possible resolution is to assume that the level of gene expression saturates for very strong promoters at $r_{\max}$ and for very weak promoters at r_0 (caused by background noise or spurious transcription; *SI Appendix*, section B), namely,

$$\text{GE} = \frac{r_0 + r_{\max} e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{-35}+E_{-10})}}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{-35}+E_{-10})}}. \quad [2]$$

Since Eq. **2** still assumes that each promoter element contributes additively and independently to the total RNAP binding energy, it also makes sharp predictions that markedly disagree with the data (*SI Appendix*, section C). Inspired by these inconsistencies, we postulated that certain promoter elements, likely the −35 and −10 sites, may contribute synergistically to RNAP binding.

To that end, we consider a model for gene expression shown in Fig. 1*B* where RNAP can separately bind to the −35 and −10 sites. RNAP is assumed to elicit a large level of gene expression $r_{\max}$ when fully bound but the smaller level r_0 when unbound or partially bound. Importantly, the Boltzmann weight of the fully bound state contains the free energy E_{int} representing

the avidity of RNAP binding to the −35 and −10 sites. Physically, avidity arises because unbound RNAP binding to either the −35 or the −10 sites gains energy but loses entropy, while this singly bound RNAP attaching at the other (−10 or −35) site again gains energy but loses much less entropy, as it was tethered in place rather than floating in solution. Hence we expect $e^{-\beta E_{\text{int}}} \gg 1$, and including this avidity term implies that RNAP no longer binds independently to the −35 and −10 sites.

Our coarse-grained model of gene expression neglects the kinetic details of transcription whereby RNAP transitions from the closed to the open complex before initiating transcription. Instead, we assume that there is a separation of timescales between the fast process of RNAP binding/unbinding to the promoter and the other processes that constitute transcription. In the quasi-equilibrium framework shown in Fig. 1*B*, gene expression is given by the weighted average over all states,

$$\text{GE} = r_0 \frac{1 + e^{-\beta(E_{\text{BG}} + E_{\text{Spacer}})} \left(e^{-\beta E_{-35}} + e^{-\beta E_{-10}} + \frac{r_{\text{max}}}{r_0} e^{-\beta\left(E_{-35} + E_{-10} + E_{\text{int}}\right)} \right)}{1 + e^{-\beta(E_{\text{BG}} + E_{\text{Spacer}})} \left(e^{-\beta E_{-35}} + e^{-\beta E_{-10}} + e^{-\beta(E_{-35} + E_{-10} + E_{\text{int}})} \right)}. \quad [3]$$

We call this expression a multivalent model since it reduces to the energy matrix Eq. **1** (with constant of proportionality $r_{\text{max}} E^{-\beta E_{\text{int}}}$) in the limit where gene expression is negligible when the RNAP is not bound ($r_0 \approx 0$) and the promoter is sufficiently weak or the RNAP concentration is sufficiently small that polymerase is most often in the unbound state (so that the denominator ≈ 1). The background and spacer are assumed to contribute to RNAP binding in both the partially and fully bound states, an assumption that we rigorously justify in *SI Appendix*, section D.

Fig. 2*B* demonstrates that the multivalent model Eq. **3** is better able to capture the system's behavior ($R^2 = 0.91$) while requiring only two more parameters (r_0 and E_{int}) than the energy matrix model Eq. **1**. The sharp boundaries on the left and right represent the minimum and maximum levels of gene expression, $r_0 = 0.18$ and $r_{\text{max}} = 8.6$, respectively (*SI Appendix*, section E). The multivalent model predicts that the top 5% of promoters will exhibit expression levels of 7.6 (compared with 8.5 measured experimentally) while the weakest 5% of promoters should express at 0.2 (compared with the experimentally measured 0.1). In addition, this model quickly gains predictive power, as its coefficient of determination diminishes only slightly ($R^2 = 0.86$) if the model is trained on only 10% of the data and used to predict the remaining 90%.

Epistasis-Free Models of Gene Expression Lead to Sharp Predictions That Disagree with the Data. To further validate that the lower coefficient of determination of the energy matrix approach (Eq. **1**) was not an artifact of the fitting, we can use the epistasis-free nature of this model to predict the gene expression of double mutants from that of single mutants. More precisely, denote the gene expression $\text{GE}^{(0,0)}$ of a promoter with the consensus −35 and −10 sequences (and any background or spacer sequence). Let $\text{GE}^{(1,0)}$, $\text{GE}^{(0,1)}$, and $\text{GE}^{(1,1)}$ represent promoters (with this same background and spacer) whose −35/−10 sequences are mutated/consensus, consensus/mutated, and mutated/mutated, respectively, where "mutated" stands for any nonconsensus sequence. As derived in *SI Appendix*, section D, the gene expression of these three later sequences can predict the gene expression of the promoter with the consensus −35 and −10 without recourse to fitting, namely,

$$\text{GE}^{(0,0)} = \text{GE}^{(1,1)} \frac{\text{GE}^{(0,1)}}{\text{GE}^{(1,1)}} \frac{\text{GE}^{(1,0)}}{\text{GE}^{(1,1)}}. \quad [4]$$

Fig. 2*A*, *Inset* compares the epistasis-free predictions (x axis, right-hand side of Eq. **4**) with the measured gene expression (y axis, left-hand side of Eq. **4**). These results demonstrate that the simple energy matrix formulation fails to capture the interaction between the −35 and −10 binding sites. While this calculation cannot readily generalize to the multivalent model since it exhibits epistasis, it is analytically tractable for weak promoters where the multivalent model displays a marked improvement over the energy matrix model (*SI Appendix*, section C).

RNAP Binding to the UP Element Occurs Independently of the Other Promoter Elements. Having seen that the multivalent model (Eq. **3**) can outperform the traditional energy matrix analysis on promoters with no UP element, we next extend the former model to promoters containing an UP element. Given the importance of the RNAP interactions with the −35 and −10 sites seen above, Fig. 3*A* shows three possible mechanisms for how the UP element could mediate RNAP binding. For example, the C terminus could bind strongly and independently so that RNAP has three distinct binding sites. Another possibility is that the RNAP αCTD binds if and only if the −35 binding site is bound. A third alternative is that the UP element contributes independently to RNAP binding (analogous to the spacer and background).

To distinguish between these possibilities, we analyze the correlations in gene expression between every pair of promoter elements (UP and −35, spacer and background, etc.) to determine the strength of their interaction. Each model in Fig. 3*A* will have a different signature: Fig. 3*A*, *Top* schematic predicts strong interactions between the −35 and −10, between the UP and −35, and between the UP and −10; Fig. 3*A*, *Middle* schematic would give rise to strong dependence between the −35 and −10 as well as between the UP and −10, while the UP and −35 elements would be perfectly correlated; and in the Fig. 3*A*, *Bottom* schematic the UP elements contributes independently of the other promoter elements.

This analysis, which we relegate to *SI Appendix*, section D, demonstrates that the UP element is approximately independent of all other promoter elements ($R^2 \gtrsim 0.6$) as are the background and spacer, indicating that Fig. 3*A*, *Bottom* schematic characterizes the binding of the UP element. This leads us to the general form of transcriptional regulation by RNAP, shown in Eq. **5**:

$$\text{GE} = r_0 \frac{1 + e^{-\beta(E_{\text{BG}} + E_{\text{Spacer}} + E_{\text{UP}})} \left(e^{-\beta E_{-35}} + e^{-\beta E_{-10}} + \frac{r_{\text{max}}}{r_0} e^{-\beta\left(E_{-35} + E_{-10} + E_{\text{int}}\right)} \right)}{1 + e^{-\beta(E_{\text{BG}} + E_{\text{Spacer}} + E_{\text{UP}})} \left(e^{-\beta E_{-35}} + e^{-\beta E_{-10}} + e^{-\beta(E_{-35} + E_{-10} + E_{\text{int}})} \right)}. \quad [5]$$

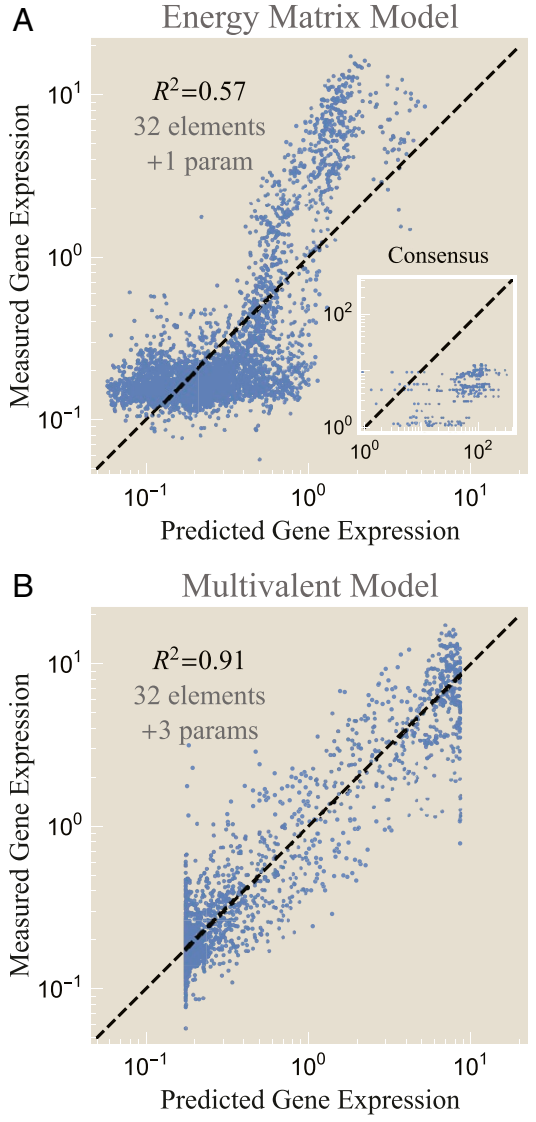

Fig. 2. Expression of promoters with no UP element. Model predictions using (*A*) an energy matrix (Eq. **1**) where the −35 and −10 elements independently contribute to RNAP binding and (*B*) a multivalent model (Eq. **3**) where the two sites contribute cooperatively. (*A*, *Inset*) The epistasis-free nature of the energy matrix model makes sharp predictions about the gene expression of the consensus −35 and −10 sequences that markedly disagree with the data. Parameter values are given in *SI Appendix*, section B.

Fig. 3*B* demonstrates how the expression of all promoters containing one of the two UP elements combined with each of the eight background, spacer, −35, and −10 sequences ($2 \times 8^4 = 8{,}192$ promoters) closely matches the model predictions ($R^2 = 0.88$). We note that the large number of outliers on the left edge of the data may be attributable to noise, since more than half of all promoters have predicted gene expression <0.2 (*SI Appendix*, section E). Remarkably, since we used the same free energies and gene expression rates from Fig. 2*B*, characterizing these 8,192 promoters required only two additional parameters (the free energies of the two UP elements). This result emphasizes how understanding each modular component of gene expression can enable us to harness the combinatorial complexity of sequence space.

Sufficiently Strong RNAP–Promoter Binding Energy Can Decrease Gene Expression. Although the 12,288 promoters considered above are well characterized by Eq. **5** on average, the data demonstrate that the full mechanistic picture is more nuanced. For example, Urtecho et al. (8) found that gene expression (averaged over all backgrounds and spacers) generally increases for −35/−10 elements closer to the consensus sequences. In terms of the gene expression models studied above (Eqs. **1**–**3**), promoters with fewer −35/−10 mutations have more negative free energies E_{-35} and E_{-10} leading to larger expression. Yet the strongest promoters with the consensus −35/−10 violated this trend, exhibiting less expression than promoters one mutation away. Thus, Urtecho et al. (8) postulated that past a certain point, promoters that bind RNAP too tightly inhibit transcription initiation and decrease gene expression.

The promoters with a consensus −35/−10 are shown as red points in Fig. 3*B*, and indeed these promoters are all predicted to bind tightly to RNAP and hence express at the maximum level $r_{\max} = 8.6$, placing them on the right edge of the data. Yet depending on their UP, background, and spacer, many of these promoters exhibit significantly less gene expression than expected. Motivated by this trend, we posit that the state of transcription initiation can be characterized by a free energy ΔE_{trans} relative to unbound RNAP that competes with the free energy ΔE_{RNAP} between fully bound and unbound RNAP (*SI Appendix*, section E), analogous to a nonequilibrium boundary-crossing problem with an effective barrier height ΔE_{trans} (14).

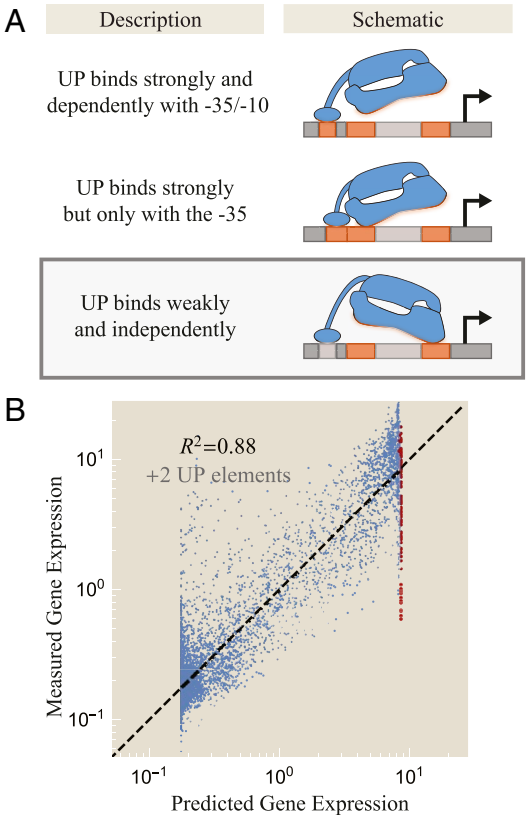

Fig. 3. The interaction between RNAP and the UP element. (*A*) Possible mechanisms by which the RNAP C terminus can bind to the UP element (orange segments represent strong binding comparable to the −35 and −10 motifs; gray segments represent weak binding comparable to the spacer and background). The data support the *Bottom* schematic (*SI Appendix*, section D). (*B*) The corresponding characterization of 8,192 promoters identical to those shown in Fig. 2 but with one of two UP binding motifs. Red points represent promoters with a consensus −35 and −10. Data were fitted using the same parameters as in Fig. 2*B* and fitting the binding energies of the two UP elements (parameter values in *SI Appendix*, section B).

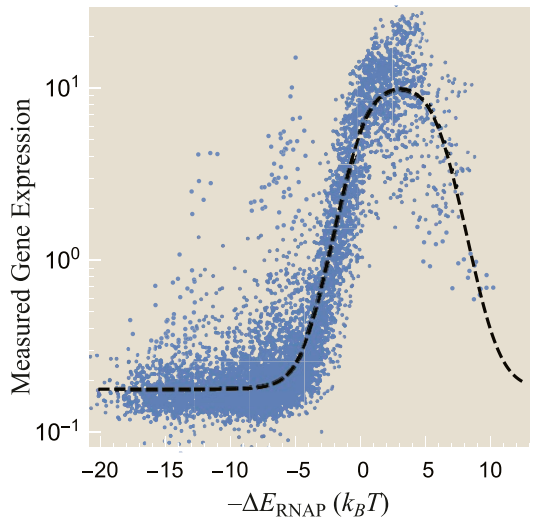

Fig. 4. Gene expression is reduced when RNAP binds a promoter too tightly. Shown is measured expression vs. inferred promoter strength $-\Delta E_{\mathrm{RNAP}}$ (stronger promoters on the right). Expression decreases once the RNAP binding becomes comparable to the free energy of the transcription initiation state $-\Delta E_{\mathrm{trans}} = 6.2\ k_BT$. The dashed line shows the prediction of the multivalent model.

Assuming the rate of transcription initiation is proportional to the relative Boltzmann weights of these two states, the level of gene expression r_{max} in Eq. **5** will be modified to

$$\frac{r_{\mathrm{max}} + r_0\, e^{-\beta(\Delta E_{\mathrm{RNAP}} - \Delta E_{\mathrm{trans}})}}{1 + e^{-\beta(\Delta E_{\mathrm{RNAP}} - \Delta E_{\mathrm{trans}})}}. \qquad [6]$$

As expected, this expression reduces to r_{max} for promoters that weakly bind RNAP ($e^{-\beta(\Delta E_{\mathrm{RNAP}} - \Delta E_{\mathrm{trans}})} \ll 1$) but decreases for strong promoters until it reaches the background level r_0 when RNAP is glued to the promoter and unable to initiate transcription. Upon reanalyzing the gene expression data with the inferred value $\Delta E_{\mathrm{trans}} = -6.2\ k_B\,T$ (*SI Appendix*, section E), we find that gene expression diminishes for the strongest RNAP–promoter free energies ΔE_{RNAP} as shown in Fig. 4 (stronger promoters to the right). This suggests that for sufficiently strong promoters, the rate-limiting step in transcription initiation changes from RNAP dissociation to promoter escape.

The Bivalent Binding of RNAP Buffers Its Interaction with DNA Against Promoter Mutations. In this final section, we investigate how the avidity between the −35 and −10 sites changes the dynamics of RNAP binding. More specifically, we consider the effective dissociation constant governing RNAP binding when both the −35 and −10 sites are intact and compare it to the case where only one site is capable of binding. To simplify this discussion, we focus exclusively on RNAP binding to the −35 and −10 motifs as shown in the rates diagram Fig. 1*C*, absorbing the effects of the background, spacer, and UP elements into these rates.

At equilibrium, there is no flux between the four RNAP states. We define the effective dissociation constant

$$K_D^{\mathrm{eff}} = \frac{K_{-35}K_{-10}}{c_0 + K_{-35} + K_{-10}} \qquad [7]$$

which represents the concentration of RNAP at which there is a 50% likelihood that the promoter is bound (*SI Appendix*, section F). $K_j = \frac{k_{\mathrm{off},j}}{k_{\mathrm{on}}}$ stands for the dissociation constant of free RNAP binding to the site j and $c_0 = \frac{\tilde{k}_{\mathrm{on}}}{k_{\mathrm{on}}} = [\mathrm{RNAP}]\, e^{-\beta E_{\mathrm{int}}}$ represents the increased local concentration of singly bound RNAP transitioning to the fully bound state (i.e., E_{int} and c_0 are the embodiments of avidity in the language of statistical mechanics and thermodynamics, respectively). Note that K_D^{eff} is a sigmoidal function of K_{-10} with height K_{-35} and midpoint at $K_{-10} = c_0 + K_{-35}$.

Fig. 5 demonstrates how the effective RNAP dissociation constant K_D^{eff} changes when mutations to the −10 binding motif alter its dissociation constant K_{-10}. When the −35 sequence is weak (dashed lines, $k_{\mathrm{off},-35} \to \infty$), $K_D^{\mathrm{eff}} \approx K_{-10}$, signifying that RNAP binding relies solely on the strength of the −10 site. In the opposite limit where RNAP tightly binds to the −35 sequence (solid lines), the cooperativity c_0 and the dissociation constant K_{-35} shift the curve horizontally and bound the effective dissociation constant to $K_D^{\mathrm{eff}} \leq K_{-35}$. This upper bound may buffer promoters against mutations, since achieving a larger effective dissociation constant would require not only wiping out the −35 site but in addition mutating the −10 site. Finally, in the case where the cooperativity c_0 is large, $K_D^{\mathrm{eff}} \approx \frac{K_{-10}K_{-35}}{c_0}$, indicating that as soon as one site of the RNAP binds, the other is very likely to also bind, thereby giving rise to the multiplicative dependence on the two K_Ds.

To get a sense for how these numbers translate into physiological RNAP dwell times on the promoter, we note that the lifetime of bound RNAP is given by $\tau = \frac{1}{K_D^{\mathrm{eff}} k_{\mathrm{on}}}$ (*SI Appendix*, section F). Using $K_D^{\mathrm{eff}} \approx 550$ nM for the *lac* promoter (15) and assuming a diffusion-limited on-rate $10^7 \frac{1}{\mathrm{M \cdot s}}$ leads to a dwell time of 5 s, comparable to the measured dwell time of RNAP–promoter in the closed complex (16). It would be fascinating if recently developed methods that visualize real-time single-RNAP binding events probed the dwell time of the promoter constructed by Urtecho et al. (8) to see how well the predictions of the multivalent model match experiments (16).

Discussion

While high-throughput methods have enabled us to measure the gene expression of tens of thousands of promoters, they nevertheless only scratch the surface of the full sequence space. A typical promoter composed of 200 bp has 4^{200} variants (more than the number of atoms in the universe). Nevertheless, by understanding the principles governing transcriptional regulation, we can begin to cut away at this daunting complexity to design better promoters.

In this work, we analyzed a recent experiment by Urtecho et al. (8) measuring gene expression of over 10,000 promoters in *E. coli* using the σ^{70} RNAP holoenzyme. These sequences comprised all combinations of a small set of promoter elements, namely, eight −10s, eight −35s, eight spacers, eight backgrounds, and three

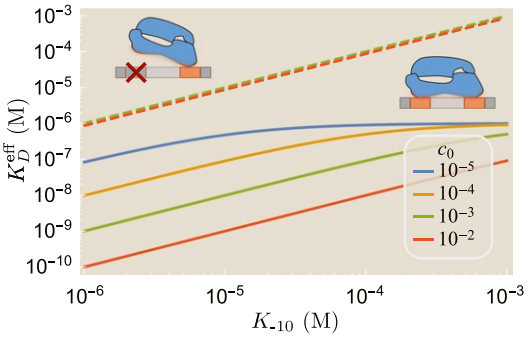

Fig. 5. The dissociation between RNAP and the promoter. Shown is RNAP binding to a promoter with a strong (solid lines, $K_{-35} = 1$ μM) or weak (dashed lines, $K_{-35} \to \infty$) −35 sequence. c_0 represents the local concentration of singly bound RNAP.

UPs depicted in Fig. 1*A*, providing an opportunity to deepen our understanding of how these elements interact and to compare different quantitative models of gene expression.

We first analyzed these data using classic energy matrix models which posit that each promoter element contributes independently to the RNAP–promoter binding energy. As emphasized by Urtecho et al. (8) and others, such energy matrices poorly characterize gene expression (Fig. 2*A*, $R^2 = 0.57$) and offer testable predictions that do not match the data (*SI Appendix*, section C), mandating the need for other approaches (7, 8).

To meet this challenge, we first determined which promoter elements contribute independently to RNAP binding (*SI Appendix*, section D). This process, which was done without recourse to fitting, demonstrated that the −35 and −10 elements bind in a concerted manner that we postulated is caused by avidity. In this context, avidity implies that when RNAP is singly bound to either the −35 or the −10 sites, it is much more likely (compared with unbound RNAP) to bind to the other site, similar to the boost in binding seen in bivalent antibodies (17) or multivalent systems (12, 18, 19). Surprisingly, we found that outside the −35/−10 pair, the other components of the promoter contributed independently to RNAP binding.

Using these findings, we developed a multivalent model of gene expression (Eq. **5**) that incorporates the avidity between the −35/−10 sites as well as the independence of the UP/spacer/background interactions. This model was able to characterize the 4,096 promoters with no UP element (Fig. 2*B*, $R^2 = 0.91$) and the 8,192 promoters containing an UP element (Fig. 3*B*, $R^2 = 0.88$). These results surpass those of the traditional energy matrix model (Fig. 2*A*, $R^2 = 0.57$), requiring only two additional parameters that could be experimentally determined (e.g., the interaction energy E_{int} arising from the −35/−10 avidity and the level of gene expression r_0 of a promoter with a scrambled −10 motif, a scrambled −35 motif, or both motifs scrambled).

These promising findings suggest that determining which components bind independently is crucial to properly characterize multivalent systems. It would be fascinating to extend this study to RNAP with other σ factors (13) as well as to RNAP mutants with no α CTD or that do not bind at the −35 site (20, 21). Our model predicts that polymerase in this last category with one strong binding site should conform to an energy matrix approach.

Quantitative frameworks such as the multivalent model explored here can deepen our understanding of the underlying mechanisms governing a system's behavior. For example, while searching for systematic discrepancies between our model prediction and the gene expression measurements, we found that promoters predicted to have the strongest RNAP affinity did not exhibit the largest levels of gene expression (thus violating a core assumption of nearly all models of gene expression that we know of). This led us to posit a characteristic energy for transcription initiation that reduces the expression of overly strong promoters (Fig. 4). In addition, we explored how having separate binding sites at the −35 and −10 elements buffers RNAP kinetics against mutations; for example, no single mutation can completely eliminate gene expression of a strong promoter with the consensus −35 and −10 sequence, since at least one mutation in both the −35 and −10 motifs would be needed (Fig. 5).

Finally, we end by zooming out from the particular context of transcription regulation and note that multivalent interactions are prevalent in all fields of biology (22), and our work suggests that differentiating between independent and dependent interactions may be key to not only characterizing overall binding affinities but also understanding the dynamics of a system (23). Such formulations may be essential when dissecting the much more complicated interactions in eukaryotic transcription where large complexes bind at multiple DNA loci (24, 25) and more broadly in multivalent scaffolds and materials (11, 12).

Materials and Methods

Gene expression was measured as the ratio of RNA to DNA barcodes (8). We fitted both the energy matrix and multivalent models on 75% of the data and characterized the predictive power on the remaining 25%, repeating the procedure 10 times. The coefficient of determination R^2 was calculated for $y_{\text{data}} = \log_{10}(\text{gene expression})$ to prevent the largest gene expression values from dominating the result (*SI Appendix*, section B). The Mathematica notebook (doi: 10.22002/D1.1242) contains the data analyzed in this work and can recreate all plots.

ACKNOWLEDGMENTS. We thank Suzy Beeler, Vahe Galstyan, Peng (Brian) He, and Zofii Kaczmarek for helpful discussions. This work was supported by the Rosen Center at the California Institute of Technology and the National Institutes of Health through Grant 1R35 GM118043-01.

1. S. Gama-Castro *et al.*, RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* **44**, D133–D143 (2016).
2. R. P. Patwardhan *et al.*, High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
3. J. B. Kinney, A. Murugan, C. G. Callan, E. C. Cox, Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9158–9163 (2010).
4. F. Inoue, N. Ahituv, Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
5. A. H. Yona, E. J. Alm, J. Gore, Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, 1530 (2018).
6. N. M. Belliveau *et al.*, Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4796–E4805 (2018).
7. T. L. Forcier *et al.*, Measuring cis-regulatory energetics in living cells using allelic manifolds. *eLife* **7**, e40618 (2018).
8. G. Urtecho, A. D. Tripp, K. Insigne, H. Kim, S. Kosuri, Systematic dissection of sequence elements controlling σ70 promoters using a genomically-encoded multiplexed reporter assay in *E. coli*. *Biochemistry* **58**, 1539–1551 (2019).
9. T. Kuhlman, Z. Zhang, M. H. Saier, T. Hwa, Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6043–6048 (2007).
10. J. Q. Boedicker, H. G. Garcia, S. Johnson, R. Phillips, DNA sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation. *Phys. Biol.* **10**, 066005 (2013).
11. C. T. Varner, T. Rosen, J. T. Martin, R. S. Kane, Recent advances in engineering polyvalent biological interactions. *Biomacromolecules* **16**, 43–55 (2015).
12. G.-H. Yan *et al.*, Artificial antibody created by conformational reconstruction of the complementary-determining region on gold nanoparticles. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E34–E43 (2018).
13. A. Feklístov, B. D. Sharon, S. A. Darst, C. A. Gross, Bacterial sigma factors: A historical, structural, and genomic perspective. *Annu. Rev. Microbiol.* **68**, 357–376 (2014).
14. S. Roy, S. Garges, S. Adhya, Activation and repression of transcription by differential contact: Two sides of a coin. *J. Biol. Chem.* **273**, 14059–14062 (1998).
15. L. Bintu *et al.*, Transcriptional regulation by the numbers: Models. *Curr. Opin. Genet. Dev.* **15**, 116–124 (2005).
16. F. Wang *et al.*, The promoter-search mechanism of *Escherichia coli* RNA polymerase is dominated by three-dimensional diffusion. *Nat. Struct. Mol. Biol.* **20**, 174–181 (2013).
17. J. S. Klein, P. J. Bjorkman, Few and far between: How HIV may be evading antibody avidity. *PLoS Pathog.* **6**, e1000908 (2010).
18. S. Banjade, M. K. Rosen, Phase transitions of multivalent proteins can promote clustering of membrane receptors. *eLife* **3**, e04123 (2014).
19. J. Huang *et al.*, Detection, phenotyping, and quantification of antigen-specific T cells using a peptide-MHC dodecamer. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1890–E1897 (2016).
20. A. Kumar *et al.*, The minus 35-recognition region of *Escherichia coli* sigma 70 is inessential for initiation of transcription at an extended minus 10 promoter. *J. Mol. Biol.* **232**, 406–418 (1993).
21. L. Minakhin, K. Severinov, On the role of the *Escherichia coli* RNA polymerase sigma 70 region 4.2 and alpha-subunit C-terminal domains in promoter complex formation on the extended −10 galP1 promoter. *J. Biol. Chem.* **278**, 29710–29718 (2003).
22. A. Gao *et al.*, Evolution of weak cooperative interactions for biological specificity. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11053–E11060 (2018).
23. J. D. Stone *et al.*, Interaction of streptavidin-based peptide-MHC oligomers (tetramers) with cell-surface TCRs. *J. Immunol.* **187**, 6281–6290 (2011).
24. N. Goardon *et al.*, ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. *EMBO J.* **25**, 357–366 (2006).
25. M. Levine, C. Cattoglio, R. Tjian, Looping back to leap forward: Transcription enters a new era. *Cell* **157**, 13–25 (2014).