Title

Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence

Authors

Justin B. Kinney[1,2,†], Anand Murugan[1], Curtis G. Callan Jr. [1,3], Edward C. Cox[4]

Affiliations

1. Department of Physics, Princeton University, Princeton, NJ 08544, USA
2. Lewis-Sigler Institute, Princeton University, Princeton, NJ 08544, USA
3. Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544 USA
4. Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

Corresponding Author Information

Please email correspondence to jkinney@cshl.edu.

Additional title page footnotes

[†] Current address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA

Abstract

Cells use protein-DNA and protein-protein interactions to regulate transcription. A biophysical understanding of this process has, however, been limited by the lack of methods for quantitatively characterizing the interactions that occur at specific promoters and enhancers in living cells. Here we show how such biophysical information can be revealed by a simple experiment in which a library of partially mutated regulatory sequences are partitioned according to their *in vivo* transcriptional activities and then sequenced *en masse*. Computational analysis of the sequence data produced by this experiment can provide precise quantitative information about how the regulatory proteins at a specific arrangement of binding sites work together to regulate transcription. This ability to reliably extract precise information about regulatory biophysics in the face of experimental noise is made possible by a recently identified relationship between likelihood and mutual information (1). Applying our experimental and computational techniques to the *E. coli lac* promoter, we demonstrate the ability to identify regulatory protein binding sites *de novo*, determine the sequence-dependent binding energy of the proteins that bind these sites and, importantly, measure the *in vivo* interaction energy between RNA polymerase and a DNA-bound transcription factor. Our approach provides a generally applicable method for characterizing the biophysical basis of transcriptional regulation by a specified regulatory sequence. The principles of our method can also be applied to a wide range of other problems in molecular biology.

"\body"

Introduction

Cells regulate transcription primarily through the binding of proteins to DNA binding sites within cis-regulatory modules (CRMs). Understanding how CRMs use different arrangements of binding sites to encode regulatory programs remains a major challenge for molecular biology. High-throughput methods have spurred great progress in cataloging the genome-wide distribution of binding sites (2, 3), and many techniques exist for characterizing the sequence-specificity of individual regulatory proteins (4-7). However, determining how a specific CRM integrates information from multiple DNA-bound proteins still requires a laborious series of biochemical experiments that typically provide only qualitative information (reviewed in ref. 8).

The *E. coli lac* promoter (9,10) is one of the few CRMs whose function is well understood at the biophysical level (11,12). Kuhlman et al. (12) were the first to prove that a certain aspect of this system — the up-regulation of transcription by the protein CRP (13) — could be quantitatively explained by a simple energetic interaction between CRP and the σ^{70}-dependent RNA polymerase holoenzyme (henceforth RNAP). To do this, Kuhlman et al. measured transcriptional activity resulting from different *in vivo* concentrations of active CRP and showed that the resulting functional form of this activity was consistent with a simple thermodynamic model (14). By fitting the defining parameters of this model to their data, Kuhlman et al. were then able to measure the *in vivo* interaction energy between CRP and RNAP.

Despite its success, Kuhlman et al.'s approach is not feasible as a general method for studying CRMs. First, it requires quantitative control over the *in vivo* concentrations of all of the proteins that bind the CRM of interest. Secondly, CRMs typically contain multiple binding sites for each operative regulatory protein, making it difficult to determine the specific role of each site simply by varying the concentration of the regulator. This latter fact has created difficulty for similar studies of eukaryotic enhancers (15).

We hypothesized that measuring the activities of a large number of CRMs containing scattered point mutations could provide information similar to that produced by Kuhlman et al.'s method. This approach would be feasible for studying arbitrary CRMs, and would allow the effect of each individual binding site to be characterized. Because point mutations tend to preserve the spatial arrangement of binding sites, such measurements would allow one to interrogate the same protein-DNA complexes that allow the wild-type CRM to function. A similar approach had been tried by Schneider and Stormo in 1989 (16), but the recent advent of ultra-high-throughput sequencing, together with new techniques in machine learning (1) led us to believe that this approach could be much more powerful than had previously been realized.

In this paper we report the application of this mutagenesis-based approach to the *E. coli lac* promoter. Fluorescence-activated cell sorting (FACS; ref. 17) and 454

pyrosequencing (18) was used to characterize the activities of ~200,000 *lac* promoters mutagenized in a 75 bp region containing the CRP and RNAP binding sites (Fig. 1A). The resulting sequence data allowed us to identify these binding sites *de novo,* determine the sequence-dependent binding energy of both CRP and RNAP, and measure the *in vivo* interaction energy between these two proteins in their native DNA-bound configuration. We note that previous attempts to determine *in vivo* protein-protein interaction energies from sequence data (19,20) required unproven assumptions about how arbitrary arrangements of DNA-bound proteins interact; our approach does not.

In this way, we demonstrate how deep sequencing can be used to measure protein-DNA and protein-protein interaction energies in living cells. This ability should be useful for addressing many different questions in molecular biology.

Interrogating a CRM with flow cytometry and deep sequencing

We performed six experiments on region [-75:-1] of the *E. coli lac* promoter (Figs. 1A and 1B). These experiments, summarized in Table 1, differed in the positions within region [-75:-1] that were mutagenized, the strain of *E. coli* used, and the physiological conditions under which *lac* promoter function was characterized.

The full-wt experiment used a library of reporter constructs, derived from pUA66-lacZ (Fig. 1C; ref. 21), in which region [-75:-1] of the *lac* promoter was mutagenized at 12% per nucleotide, yielding 9 ± 3 substitution mutations per sequence. Wild-type *E. coli* (strain MG1655) were transformed with this plasmid library, after which GFP expression was induced during exponential growth in minimal media supplemented with glucose, cAMP, and 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). FACS (17) was used to sort induced cells into 10 different batches, each cell according to its measured fluorescence (Fig. 1D). PCR was then used to tag mutant CRMs according to the batch in which each CRM was found (Fig. 1E). 454 pyrosequencing (18) of the resulting PCR amplicons yielded a list of 51,835 mutant CRMs and corresponding batches. The batch μ associated with each CRM σ serves as a noisy and qualitative measurement of that CRM's *in vivo* transcriptional activity. Experiments crp-wt and rnap-wt were performed the same way, but using mutant CRMs in which only the CRP or RNAP binding site was mutagenized (Table 1).

The full-500, full-150, and full-0 experiments were performed using the same plasmid library as in full-wt, but with transcriptional activity assayed in *E. coli* strain TK310 (12) grown in 500 μM, 150 μM, or 0 μM cAMP, respectively. Cells were also sorted into 5 batches instead of 10. Strain TK310 lacks adenlyate cyclase (*cyaA*, needed for cAMP synthesis) and phosphodiesterase (*cpdA*, which degrades cAMP) and is therefore unable to control its intracellular cAMP levels (12). Growing TK310 cells in media supplemented with different concentrations of cAMP thus allowed us to control the active intracellular concentration of CRP (just as Kuhlman et al. did). Importantly, though, the mechanism of *lac* promoter function is the same in both MG1655 and TK310 cells.

In total we obtained 220,591 mutant CRMs, with each CRM σ assigned a noisy, qualitative measurement μ of its transcriptional activity (Fig. 1B). These sequences σ and measurements μ comprise the only data used in the analysis that follows.

Information footprints reveal functional binding sites

Our first goal was to identify nucleotide positions that affect expression, thereby locating all functional binding sites within the probed region of the *lac* promoter. For this we used mutual information, a fundamental quantity from information theory that provides the most general measure of dependence between any two variables (22). For each nucleotide position i, we computed the number of sequences in each batch μ having each of the four possible bases b_i. We then used this information to compute the mutual information $I(b_i;\mu)$ between b_i and μ (see SI Appendix Sec. 2e for details), thereby quantifying the effect of the base at position i on our measurements. Plotting the mutual information $I(b_i;\mu)$ for each position i produced what we call an "information footprint".

Fig. 2A shows the information footprint produced by the full-wt experiment. The known binding sites of CRP and RNAP are clearly visible and each has the expected bipartite structure: CRP binds DNA as a homodimer at [-72:-51] with a 6 bp gap between its two DNA-binding domains, while RNAP binding to [-41:-1] results from the recognition of separate sequence elements centered roughly on positions -10 and -35 (9,13).

The information values displayed in Fig. 2A are small, ranging from ~0.05 bits down to values indistinguishable from zero. This is not surprising, since each $I(b_i;\mu)$ measures the effect on transcription of just one out of 75 positions. But because our data set is large (N = 51,856), these information values are determined very precisely — typically to within $\sim 4\times10^{-4}$ bits. Fig. 2A thus displays meaningful information values ranging over two orders of magnitude. This high level of sensitivity can reveal aspects of sequence function not detectable by other methods, e.g. (23). For instance, small but significant effects on expression were observed between the -10 and -35 elements and between the CRP and RNAP sites, regions not previously thought to influence transcription at the *lac* promoter. Indeed, only 10 positions ([-75, -70, -48, -42, -41, -23, -22, -21, -19, -3]) out of 75 show an insignificant effect on expression (i.e. $P > 0.05$).

Information footprints can further allow one to determine which of the identified sites are involved in the response to a specific biochemical signal or genetic perturbation. Fig. 2B shows the information footprint from the full-0 experiment, in which TK310 cells were induced in the absence of cAMP, thereby inactivating intracellular CRP. The lack of active CRP is reflected in the near-total loss of information at positions [-72:-51]: information values at all but three positions within this 22 bp site differ insignificantly from zero (i.e. $P > 0.05$). An intermediate reduction in information occurs in the full-150 footprint, while the full-500 footprint closely resembles that of full-wt (SI Appendix Fig. S3). We note that the small but significant information values at positions [-64, -57, -52] in the full-0 footprint might result from interactions between DNA and the α subunits of RNAP (24), or from alternative RNAP binding sites (25).

An important caveat to this analysis is that the quantitative features of an information footprint ultimately depend on the details of one's experiment, including the level of mutagenesis used in the initial CRM library and the specific fluorescence gates used for sorting cells. So while qualitative differences between footprints from different experiments can be revealing, it is difficult to draw conclusions from more subtle quantitative differences, such as the different shapes of the RNAP footprint in Figs. 2A and 2B. But as we show in the next section, explicit biophysical models can be fit to data in a way that does not depend on such experimental details. Quantitative differences between models inferred from different experiments can, as a result, be revealing about underlying molecular mechanisms.

Model fitting in the presence of uncharacterized noise

Next we used our data to infer quantitative models for how *in vivo* protein-DNA and protein-protein interactions modulate transcription. By 'model' we mean a mathematical function that assigns to each sequence σ a predicted value x for some quantity of interest, such as the DNA binding energy of a regulatory protein, or the rate of transcription resulting from the interactions between multiple proteins. To infer a given model, we first assumed a specific mathematical formula for the model, then determined the values of model parameters by matching the sequence-dependent predictions x to our sequence-specific measurements μ.

Such model fitting could be done in the standard Bayesian way if we knew the experimental 'error model' $p(\mu|x)$ — the probability of obtaining a measurement μ given an underlying quantity of interest x (such as binding energy or transcription rate). But in our experiments it was virtually impossible to accurately determine $p(\mu|x)$ *a priori*. Many difficult-to-characterize noise processes, including stochastic transcription, variations in cell size, and noise in FACS measurements, contribute to the quantitative form of $p(\mu|x)$. Also, in the case where x represents the DNA-binding energy of a protein, we do not know *a priori* how the binding of that protein affects transcription; indeed, this is something we want to learn from the data. Kinney et al. (1) overcame this problem by computing likelihood in the presence of an explicitly uncertain error model. They showed that regardless of what $p(\mu|x)$ actually is, the likelihood of a candidate model is well approximated by

[1] $$p(\text{data}\,|\,\text{model}) \approx \text{const} \times 2^{NI(x;\mu)}$$

in the limit where the number N of independently measured sequences is sufficiently large. Here $I(x;\mu)$ is the 'predictive information' of the model — the mutual information between model predictions x and measurements μ. Although Kinney et al. focused on the analysis of microarray data, Eq. 1 is applicable to any data set consisting of a large number of sequences and corresponding measurements. Kinney et al.'s approach therefore provides a practical substitute for standard likelihood-based inference when the experimental error model is either difficult to characterize or is unknowable *a priori* (see SI Appendix Sec. 2b for more discussion).

In the analysis that follows, we used a custom parallel tempering Monte Carlo algorithm to sample model parameters according to the right-hand-side of Eq. 1 (see SI Appendix Sec. 2f). This allowed us to determine not just the best values for model parameters, i.e. which values maximize predictive information $I(x;\mu)$, but also the uncertainty in each parameter due to finite data. Because $N > 10^4$ for all of the experiments described in this paper, even changes as small as 10^{-4} bits in the value of $I(x;\mu)$ led to substantial changes in model likelihood. The large amount of data produced by our experiments thus allowed us to determine precise quantitative values for model parameters even though our measurements were noisy and qualitative.

In vivo **interaction energies from sequence data**

Having located the binding sites of both CRP and RNAP *de novo* using information footprints, we sought an explicit model for each protein's sequence-dependent binding energy. For this we used 'energy matrix' models: each base within a protein's binding site was assumed to contribute additively to the overall binding energy. These simple models have been shown to accurately describe a number of transcription factors, e.g. (26,27), though there are known exceptions (28,29).

We fit an energy matrix for CRP to positions [-74:-49] using the full-wt data set. Energy matrix elements were sampled, according to Eq. 1, using the predictive information $I(\varepsilon_c;\mu)$ where ε_c is CRP's predicted binding energy. The resulting optimal matrix is shown in Fig. 3A. We similarly fit an energy matrix to positions [-41:-1] to model RNAP's binding energy ε_r (Fig. 3B). CRP and RNAP energy matrices were also inferred from our five other data sets (SI Appendix Fig. S4A). We note that at this stage of our analysis we were able to determine each matrix only up to an unknown multiplicative constant, not in physical units such as kcal/mol (see SI Appendix Sec. 2c).

Unlike information footprints, these energy matrices are meant to capture intrinsic properties of the regulatory proteins, properties that should not depend the on specific ways cells were sorted or on the level of mutagenesis used in the CRM library. The optimal matrices inferred from our six experiments (4 matrices for CRP, 5 for RNAP) are nearly identical, supporting this interpretation: CRP matrix elements derived from different experiments correlate by > 95% (SI Appendix Fig. S4B), while RNAP matrix elements exhibit > 92% correlation (SI Appendix Fig. S4C). Furthermore, each of these matrix models performs better on every one of our data sets than do any of the models for either CRP or RNAP currently in the literature (with two minor exceptions; see SI Appendix Sec. 2g). We find this level of quantitative agreement between experiments remarkable, considering that our six experiments used different promoter libraries, different *E. coli* strains (MG1655 or TK310), different inducing conditions (500 μM, 150 μM, or 0 μM cAMP), and different fluorescence gates for sorting cells. This close agreement in the face of important experimental differences attests to both the usefulness and correctness of Eq. 1.

Our inferred CRP and RNAP energy matrices recapitulate much of what is known about the sequence specificities of these two proteins. The known consensus sequences —

TGTGA(N)$_6$TCACA for CRP (4) and TTGACA(N)$_{18}$TATAAT for RNAP (30) — exactly match the lowest energy sequences predicted by nearly every one of our matrix models. The one exception is the RNAP matrix fit to full-0 data, which predicts that TTGATA will have slightly lower energy than TTGACA in the -35 region. We note that every one of our RNAP matrices also predicts that having a 'G' at position -14 increases RNAP binding strength. In the literature this 'G' is said to create an "extended -10 promoter", and such promoters are known to have increased transcriptional activity. Our CRP matrices are also in qualitative agreement with previous *in vitro* measurements (31), though there are some quantitative discrepancies.

Next we sought a quantitative understanding of how the interaction of CRP with RNAP affects transcription. Kuhlman et al. (12) previously showed that a simple biophysical model, based on equilibrium statistical mechanics (reviewed in ref. 14) accounted well for the effect of cAMP on *lacZ* expression in TK310 cells. We hypothesized that using energy matrices to describe the binding energies of CRP and RNAP within Kuhlman et al.'s model, then fitting all model parameters to our data *de novo*, would allow us to recover Kuhlman et al.'s results, including their measurement of the interaction energy between CRP and RNAP.

Following Kuhlman et al., we assumed that the rate of transcription τ at the *lac* promoter is proportional to the occupancy of RNAP at its binding site in thermal equilibrium. This model is quantitatively expressed as

[2]
$$\tau = \tau_{\max} \frac{C_r e^{-\varepsilon_r/RT} + C_c C_r e^{-(\varepsilon_c+\varepsilon_r+\varepsilon_i)/RT}}{1 + C_c e^{-\varepsilon_c/RT} + C_r e^{-\varepsilon_r/RT} + C_c C_r e^{-(\varepsilon_c+\varepsilon_r+\varepsilon_i)/RT}}$$

where RNAP occupancy is given by the sum of Boltzmann weights corresponding to physical states in which RNAP is bound, divided by the sum of weights for all possible states of the system. These Boltzmann weights depend on (i) the CRP and RNAP binding energies ε_c and ε_r, which we express in kcal/mol and normalize to be zero at each wild-type *lac* promoter site, (ii) the concentrations C_c and C_r of CRP and RNAP, expressed in units of each wild-type site's dissociation constant and (iii) the CRP-RNAP interaction energy ε_i, expressed in kcal/mol. τ_{max} is the transcription rate resulting from full RNAP occupancy. R = 1.98×10^{-3} kcal/mol °K is the gas constant and T = 310 °K (37 °C) is the temperature at which cells were induced.

Using $I(\tau;\mu)$ evaluated on full-wt data, we fit all of the parameters defining τ, including ε_i, C_c, and the elements of the energy matrices used to compute ε_c and ε_r. Doing so we inferred a CRP-RNAP interaction energy ε_i = -3.26 ± 0.41 kcal/mol. This value is consistent with Kuhlman et al.'s measurement of -3.4 kcal/mol (12) and represents the first time (to our knowledge) that the analysis of sequence data has been shown to successfully yield the *in vivo* interaction energy between two proteins. This procedure also yielded an *in vivo* CRP concentration of C_c = [CRP]/K_d^{wt} = $10^{-1.2 \pm 0.2}$. Fig. 3C shows these values for ε_i and C_c, as well as the optimal energy matrices for ε_c and ε_r inferred by fitting τ. These matrices closely resemble those in Fig. 3A and 3B, but, unlike the

matrices we inferred by separately fitting ε_c and ε_r, their elements are determined explicitly in physical units of kcal/mol. We note that fitting τ to full-wt data provided no information about the value of either τ_{max} or C_r (see SI Appendix Sec. 2c).

Testing biochemical mechanisms by fitting a single model to multiple data sets

cAMP is known to alter *lac* promoter activity by affecting CRP's ability to bind DNA, not CRP's interaction with RNAP. Both of these possibilities, though, are consistent with the information footprints shown in Fig. 2. By contrast, the former hypothesis predicts that the CRP concentration C_c in our model for τ should vary from experiment to experiment, while the latter predicts an experiment-dependent interaction energy ε_i. To further test the validity of our approach, we fit a single model for τ to all six of our data sets (see SI Appendix Sec. 2d). This multi-data-set model employed a single CRP energy matrix and a single RNAP energy matrix, but allowed for data-set-specific values for both ε_i and C_c.

The experiment-specific values we inferred for ε_i and C_c are shown in Figs. 4A and 4B. The ε_i values determined for all six experiments are mutually consistent ($P > 0.05$, χ^2 test), whereas the six C_c values differ very significantly: C_c values for experiments full-0 and full-150, which were performed in reduced concentrations of cAMP, were found to be much lower than the C_c values determined for the other four experiments. These results verify the well-known fact that cAMP affects the concentration of active CRP, not the strength of CRP's interaction with RNAP (9,13).

Repeating this inference using only one ε_i value for all six experiments yielded the final inferred model of this paper (Fig. 4 and SI Appendix Fig. S5). The CRP-RNAP interaction energy inferred from all six data sets is $\varepsilon_i = -2.82 \pm 0.13$ kcal/mol. This differs from Kuhlman et al.'s measurement, but only by ~ 20%. Also, the ratio of CRP concentrations inferred for full-150 and full-500 ($C_c^{full\text{-}150}/C_c^{full\text{-}500} = 10^{-0.56 \pm 0.02}$) closely agrees with the ratio of cAMP (150 μM/500 μM = $10^{-0.52}$) used in these two experiments. This agreement is consistent with Kuhlman et al.'s observation that, for TK310 cells, the concentration of active CRP is proportional to the exogenous concentration of cAMP (12). Indeed, even the expected trace amount of cAMP present in the full-0 induction media (~50 nM, a result of carry-over from the starter culture inoculum) is fully consistent with the much-reduced yet significantly nonzero value for $C_c^{full\text{-}0}/C_c^{full\text{-}500} = 10^{-3.7 \pm 0.6}$.

We note, however, that there are also puzzling quantitative oddities in our results. Our CRP concentration ratios are in good agreement with expectations, but the absolute values we inferred for C_c (full-500: $10^{-1.13 \pm 0.06}$; full-150: $10^{-1.70 \pm 0.07}$) are about 20-fold lower than the corresponding values claimed by Kuhlman et al. (SI Appendix Sec. 2g). Also, in the six-ε_i fit, the average ε_i is -2.34 ± 0.09 kcal/mol, substantially less than the -2.82 ± 0.13 kcal/mol we inferred using a common ε_i for all six experiments. Finally, all of our inferred CRP energy matrices are asymmetric and predict that CRP binds in the energetically unfavorable orientation at the wild-type CRP site. We think these strange results probably result from our models for ε_c, ε_r, and τ being too simplistic. The issue of model selection is complicated, however, and is beyond the scope of this paper.

Model validation using predictive information

Predictive information can be used to determine how well a proposed thermodynamic model integrates binding energies into a single transcriptional output. The predictive information $I(\tau;\mu)$ quantifies how well the biophysical model in Eq. 2 accounts for our measurements μ. However, we can also directly compute the predictive information $I(\varepsilon_c,\varepsilon_r;\mu)$ of the pair of binding energies, ε_c and ε_r , without any model for τ (SI Appendix Sec. 2e). The mere fact that τ is a function of ε_c and ε_r (in Eq. 2) means that $I(\tau;\mu) \leq I(\varepsilon_c,\varepsilon_r;\mu)$, i.e. the predictive information of τ is bounded above by the predictive information of the pair of energies ε_c and ε_r. This is a direct consequence of the Data Processing Inequality, a basic result in information theory (22). In our case, equality between $I(\tau;\mu)$ and $I(\varepsilon_c,\varepsilon_r;\mu)$ can be achieved only if τ preserves all of the transcriptionally relevant information encoded in the predicted values for ε_c and ε_r. We emphasize that there is no *a priori* guarantee that any quantity τ can do this, let alone a quantity derived from a simple biophysical model.

Remarkably we find (on full-wt data) that $I(\tau;\mu) = 0.732 \pm 0.007$ bits, which is identical within error bars to $I(\varepsilon_c,\varepsilon_r;\mu) = 0.732 \pm 0.006$ bits. We believe this equality is an important validation of the specific thermodynamic formula (Eq. 2) used to represent τ. This agreement also argues that our energy matrix models for ε_c and ε_r provide a valid representation of physical binding energy; if they did not, their use in the Boltzmann exponents in Eq. 2 would be unlikely to yield sensible results. We note that a simpler model in which transcription depends only linearly on ε_c and ε_r — which would be appropriate if CRP and RNAP bound DNA only as a complex — achieves only $I(\varepsilon_c+\varepsilon_r;\mu) = 0.647 \pm 0.005$ bits when all parameters are fit *ab initio* to full-wt data. This is significantly less than $I(\varepsilon_c,\varepsilon_r;\mu)$. Our specific thermodynamic model for τ, with its physically motivated functional form, therefore provides a marked improvement over the more naive linear model.

Discussion

The approach we present here can be applied to a wide variety of CRMs in a number of different organisms. No prior knowledge of a CRM's sequence architecture is needed. All that is required is that (i) the CRM of interest function on a reporter construct and (ii) a large library of reporter constructs be introduced into cells so that each cell receives a single mutant CRM. After the activity-based-partitioning and sequencing of mutant CRMs, information footprints can be used to identify all functionally relevant positions within the probed sequence. Footprints from experiments performed in growth conditions or genetic backgrounds that are known to affect expression can further help one identify which 'clumps' of informative positions correspond to discrete binding sites, as well as which of these sites are involved in transducing specific intracellular signals. Eq. 1 then allows one to infer mathematical models describing the sequence-dependent binding energy of each site's cognate protein. The same fitting procedure can also be used to build biophysical models of the *in vivo* interactions between multiple DNA-bound

proteins. Such inference requires no quantitative model of experimental noise, thus allowing experiments performed in very different ways to produce nearly identical results.

GFP reporter plasmids have been constructed for almost all *E. coli* promoters (21). Starting from these plasmids and using our protocols, it should be possible to biophysically characterize the vast majority of promoters in *E. coli*. Similar experiments can likely be performed for most CRMs in yeast. While we demonstrated our technique on a CRM containing only two protein-binding sites, the great sensitivity of this approach should allow the simultaneous effects of many DNA-binding proteins (including nucleosomes) to be discerned from a single experiment. Ultimately, our method should be useful for characterizing the detailed functional architecture of CRMs active in a wide variety of culturable cells, including stem cells and cancer cell lines. The biophysical characterization of CRMs in living animals, however, will likely require significant changes to the experimental approach described here.

The underlying principles of our approach should also be useful for studying many different systems in which sequence-dependent interactions (e.g. protein-DNA, protein-RNA, protein-protein or protein-ligand interactions) play a central role in establishing some activity of interest. Experiments like ours can be performed both *in vivo* and *in vitro*; one simply needs to partition a large number of mutant sequences according to each one's activity. Our analysis method should then allow quantitative models of sequence-dependent function to be rigorously inferred from the resulting sequence data, regardless how the activity-dependent partitioning of sequences is accomplished.

Experimental Procedures

See SI Appendix Sec. 1 for an expanded explanation of our experimental procedures.

Strains: *E. coli* strains MG1655 (wild-type) and TK310 (Δ*cyaA* Δ*cpdA* Δ*lacY*) were kindly provided by Thomas Kuhlman. Except where noted, TK310 cells were maintained in media supplemented with 500 μM cAMP in order to prevent *crp*-mediated suppression of Δ*cyaA* (32).

Library construction: Mutant *lac* promoter libraries were synthesized by IDT using defined mixtures of nucleoside phoshporamidites. Plasmid libraries consisted of ~2×10^6 independently cloned plasmids in which region [-75:-1] was exactly replaced without the introduction of artificial restriction sites (SI Appendix Sec. 1a)

Sorting: Cells were grown in exponential phase for ≥ 10 generations, diluted into buffer, and stored on ice for 0-24 hr prior to sorting. A BD Biosciences FACSVantage SE with DiVa was then used to sort 100,000 cells into each batch based solely on GFP fluorescence. Plating revealed ~70,000 viable cells per batch.

Amplicon generation and sequencing: Miniprepped plasmid from each FACS batch was used as template for amplicon-generating PCR. Two control sets of amplicons

were also generated from pUA66-lacZ plasmid. The 47 resulting amplicon libraries (45 FACS batches + 2 wild-type controls) were collated and sequenced by Roche using the Genome Sequencer FLX platform. This yielded 448,416 sequences, 308,309 of which passed our quality filters. Unfortunately, an analysis of these sequences indicated a large post-sort reduction in sequence diversity (SI Appendix Sec. 1f). To guarantee that each sequence was independently sorted, we discarded all but one copy of each sequence in each batch, leaving a total of 220,591 sequences across our six experiments.

Sequence data and analysis results: Our 454 sequence data is available on the NCBI website under accession number SRA012345. Additional information, including processed sequence reads and inferred model parameters, is available at http://www.princeton.edu/~ccallan/sortseq09/.

Acknowledgements

We are deeply grateful to Christina DeCoste, who assisted with all flow cytometry instrumentation. Amy Caudy, Thomas Kuhlman, and Paul Wiggins provided critical advice on various experimental issues. We further benefited from discussions with William Bialek, Thomas Gregor, Stanislav Shvartsman, Antoinette Sutto, Gasper Tkačik, and Michael Zhang. The work of CGC, JBK and AM was supported in part by National Science Foundation grant PHY-0650617. The work of JBK and EC was supported in part by National Institute of Health grants GM078591 and GM071508. The work of JBK was supported in part by the Simons Foundation. The work of CGC was supported in part by US Department of Energy grant DE-FG02-91ER40671. The authors declare no conflicts of interest.

References

1. Kinney JB, Tkačik G, Callan C (2007) Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA* 104:501-506.

2. Ren B, et al. (2000) Genome-wide location and function of DNA binding proteins. *Science* 290:2306-2309.

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

4. Berg O, von Hippel P (1988) Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J Mol Biol* 200:709-723.

5. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505-510.

6. Meng X, Brodsky MH, Wolfe SA (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature Biotechnol* 23:988-994.

7. Berger M, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24:1429-1435.

8. Carey MF, Peterson CL, Smale ST (2009) *Transcriptional regulation in eukaryotes: concepts, strategies, and techniques* (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

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

10. Müller-Hill B (1996) *The* lac *operon: a short history of a genetic paradigm* (Berlin: Walter de Gruyter).

11. Vilar JMG, Leibler S (2003) DNA looping and physical constraints on transcription regulation. *J Mol Biol* 331:981-989.

12. Kuhlman T, Zhang Z, Saier MH, Hwa T (2007) Combinatorial transcriptional control of the lactose operon of Escherichia coli. *Proc Natl Acad Sci USA* 104:6043-6048.

13. Busby S, Ebright RH (1999) Transcription activation by catabolite activator protein (CAP). *J Mol Biol* 293:199-213.

14. Bintu L, et al. (2005) Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* 15:116-124.

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

16. Schneider T, Stormo G (1989) Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucleic Acids Res* 17:659-674.

17. Herzenberg L, Sweet R, Herzenberg L (1976) Fluorescence-activated cell sorting. *Sci Am* 234:108-117.

18. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

19. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* 451:535-540.

20. Gertz J, Siggia ED, Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 457:215-218.

21. Zaslaver A, et al. (2006) A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. *Nat Methods* 3:623-628.

22. Cover TM, Thomas JA (1991) *Elements of information theory* (New York, NY: Wiley).

23. Patwardhan RP, et al. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnol* 27:1173-1175.

24. Ross W, et al. (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* 262:1407-1413.

25. Reznikoff WS (1992) The lactose operon-controlling elements: a complex paradigm. *Mol Microbiol* 6:2419-2422.

26. Takeda Y, Sarai A, Rivera VM (1989) Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc Natl Acad Sci USA* 86:439-443.

27. Sarai A, Takeda Y (1989) Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc Natl Acad Sci USA* 86:6513-6517.

28. Man TK, Stormo GD (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res* 29:2471-2478.

29. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315:233-237.

30. Harley CB, Reynolds RP (1987) Analysis of E. coli promoter sequences. *Nucl Acids Res* 15:2343-2361.

31. Gunasekera A, Ebright YW, Ebright RH (1992) DNA sequence determinants for binding of the Escherichia coli catabolite gene activator protein. *J Biol Chem* 267:14713-14720.

32. Harman JG, Dobrogosz WJ (1983) Mechanism of CRP-mediated cya suppression in Escherichia coli. *J Bacteriol* 153:191-199.

Main text figures and tables

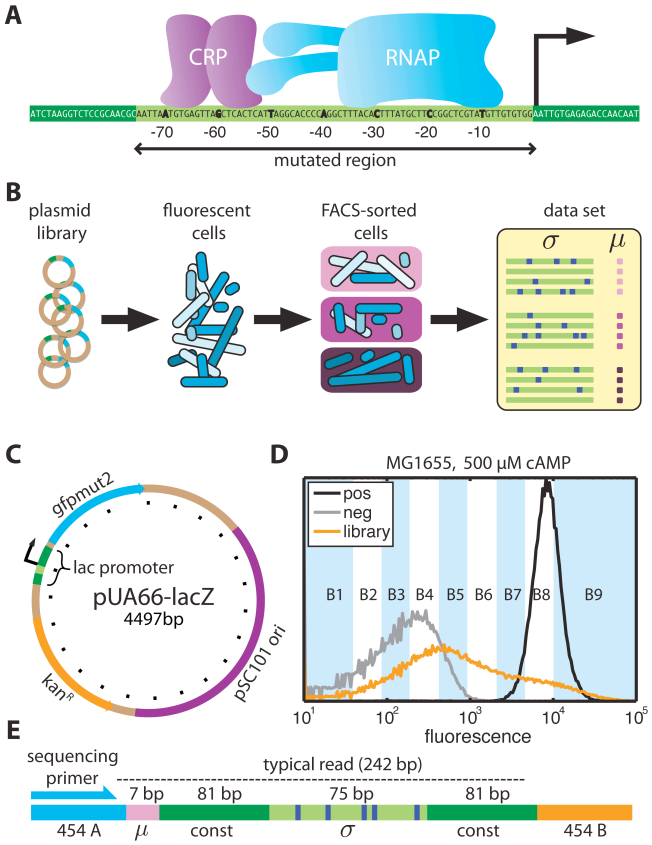

Figure 1. Overview of the experiments A) Our experiments used *lac* promoters mutagenized in region [-75:-1], which binds both CRP and RNAP. B) Plasmids containing mutant *lac* promoters driving GFP expression were transformed into *E. coli*. Induced cells were then partitioned using FACS. Deep sequencing of the mutant promoters in each FACS batch yielded a long list of sequences σ with corresponding measurements μ. C) Plasmid pUA66-lacZ (21), a very-low-copy-number plasmid on which the wild-type *lac* promoter drives the expression of GFP; tick mark spacing is 200 bp. D) Fluorescence distributions of MG1655 cells containing the full-wt plasmid library (orange), the pUA66-lacZ plasmid (black), or a negative control plasmid pJK10 (SI Appendix Fig. S1) in which region [-75:-1] of the *lac* promoter was deleted (gray). In the full-wt experiment, batches B1-B9 received cells from the indicated fluorescence ranges, while batch B0 received cells randomly sampled from the initial library. E) Each PCR amplicon contained a 7 bp DNA barcode indicating the batch μ in which each sequence σ was found. 454 pyrosequencing (18) yielded reads of about 242 bp covering the indicated regions.

607

Table 1. Summary of our six experiments

Data Set	Mut. Region	Mut. Rate	Strain	cAMP (μM)	No. μ	No. Reads
full-wt	[-75:-1]	12%	MG1655	500	10	51,835
crp-wt	[-74:-49]	24%	MG1655	500	10	46,986
rnap-wt	[-39:-4]	15%	MG1655	500	10	45,461
full-500	[-75:-1]	12%	TK310	500	5	23,431
full-150	[-75:-1]	12%	TK310	150	5	24,334
full-0	[-75:-1]	12%	TK310	0	5	28,544

608
609 **Table 1.** Shown for each of our six experiments are the mutagenized region of the *lac* promoter, the per-
610 position substitution rate, the *E. coli* strain used, the cAMP concentration used for induction, the number
611 of batches into which cells were sorted, and the final number of filtered, non-redundant reads.
612
613

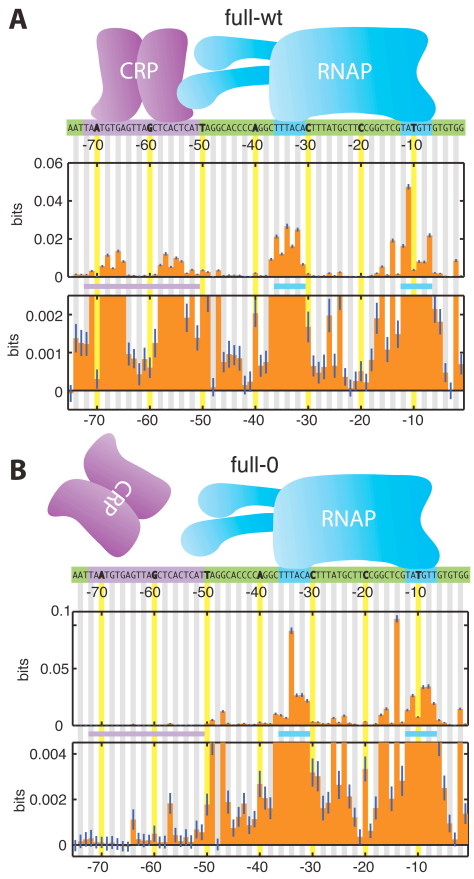

Figure 2. Information footprints. A) Footprint from full-wt data, aligned with known protein-DNA contact positions (highlighted). The lower plot is a 20X magnification of the upper plot. Blue lines indicate uncertainties due to finite sample effects (SI Appendix Sec. 2e). B) Footprint of the full-0 data set, in which intracellular CRP was inactive. SI Appendix Fig. S3 shows information footprints from all six experiments.

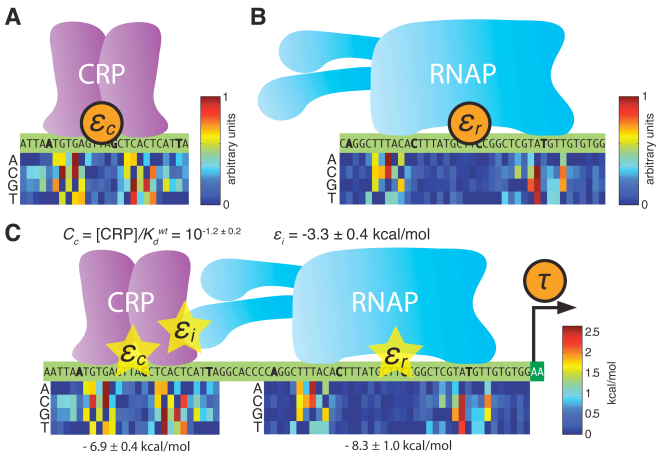

Figure 3. Models fit to full-wt data. A) The CRP energy matrix fit to [-74:-49] by maximizing $I(\varepsilon_c;\mu)$ on full-wt data. B) The RNAP energy matrix fit to [-41:-1] by maximizing $I(\varepsilon_r;\mu)$ on full-wt data. In panels A and B, each matrix column lists the energy contributions of the four possible bases at the aligned position within the site. Matrix elements range from 0 to 1 (in arbitrary units) with the lowest element in each column set to zero by convention. SI Appendix Fig. S4 shows the CRP and RNAP matrices derived from all six of our data sets. C) The thermodynamic model for τ inferred using $I(\tau;\mu)$ in Eq. 1. Optimal CRP and RNAP energy matrices are shown with elements expressed in kcal/mol (1 kcal/mol = 1.62 k_bT at T = 310 °K). It is useful to define each wild-type *lac* promoter site as having zero energy. We therefore add an energy shift, shown below each matrix, when computing ε_c and ε_r. Doing this means that C_c is the intracellular CRP concentration in units of the dissociation constant of the wild-type (zero energy) site. Values quoted for ε_i and C_c are mean ± RMSD values determined from the parameter ensembles sampled using parallel tempering Monte Carlo.

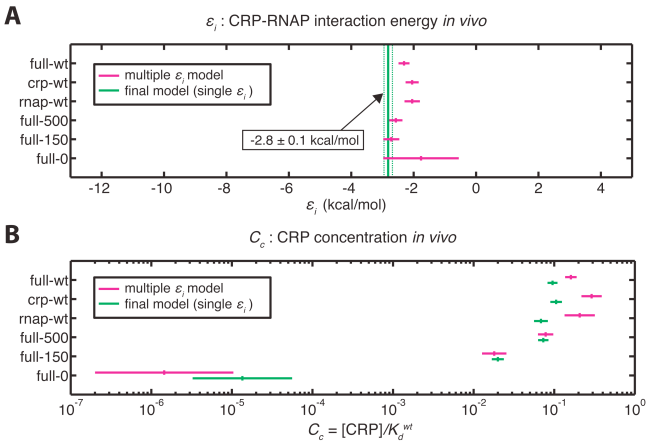

Figure 4. Parameters fit to all six data sets. A) CRP-RNAP interaction energies ε_i (mean ± RMSD) inferred by fitting τ to all six data sets, using either data-set-specific values for ε_i (magenta) or a single ε_i for all six data sets (green). B) CRP concentrations C_c inferred for these same multi-data-set models. SI Appendix Fig. S5 shows full ensemble distributions for ε_i and the six C_c parameters of the final model, together with mean and RMSD values for all the CRP and RNAP matrix elements.