ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

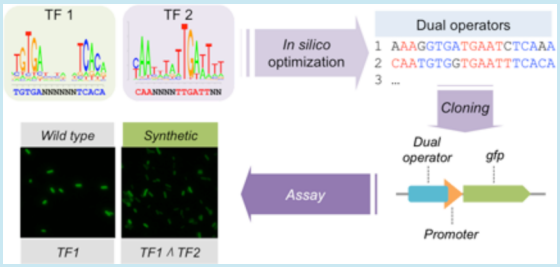
Expanding the Logic of Bacterial Promoters Using Engineered Overlapping Operators for Global Regulators

María-Eugenia Guazzaroni[†] and Rafael Silva-Rocha*[,‡]

[†]FFCLRP and [‡]FMRP, University of São Paulo, Ribeirão Preto, SP Brazil

Supporting Information

ABSTRACT: The understanding of how the architecture of *cis*-regulatory elements at bacterial promoters determines their final output is of central interest in modern biology. In this work, we attempt to gain insight into this process by analyzing complex promoter architectures in the model organism *Escherichia coli*. By focusing on the relationship between different TFs at the genomic scale in terms of their binding site arrangement and their effect on the target promoters, we found no strong constraint limiting the combinatorial assembly of TF pairs in *E. coli*. More strikingly, overlapping binding sites were found equally associated with both equivalent (both TFs have the same effect on the promoter) and opposite (one TF activates while the other repress the promoter) effects on gene expression. With this information on hand, we set an *in silico* approach to design overlapping sites for three global regulators (GRs) of *E. coli*, specifically CRP, Fis, and IHF. Using random sequence assembly and an evolutionary algorithm, we were able to identify potential overlapping operators for all TF pairs. In order to validate our prediction, we constructed two *lac* promoter variants containing overlapping sites for CRP and IHF designed *in silico*. By assaying the synthetic promoters using a GFP reporter system, we demonstrated that these variants were functional and activated by CRP and IHF *in vivo*. Taken together, presented results add new information on the mechanisms of signal integration in bacterial promoters and provide new strategies for the engineering of synthetic regulatory circuits in bacteria.



KEYWORDS: *regulatory networks, cis-regulatory elements, synthetic biology*

Transcriptional regulation in living cells takes place mainly through the integration of multiple signals at the promoter of target genes.[1] Specifically in bacteria, a number of transcriptional factors (TFs) are dedicated to sensing external and internal stimuli in order to control the expression of their target genes.[2,3] In general terms, transcriptional regulation is achieved through the interplay between different proteins at gene promoters, each protein binding to a well-defined *cis*-regulatory element or operator.[3,4] In this context, sigma factors are at the front line of the signal integration process since they confer sequence specificity to the RNA polymerase (RNAP) holoenzyme.[5] Therefore, the interplay between different sigma factors present in the cell mediates the first level of transcriptional regulation.[5,6] While the tuning of the abundance/activity of sigma factors defines which set of target promoters could be expressed, the second level of regulation is mediated mainly through TFs that could assist (activators) or block (repressors) the binding of the RNAP at the target sites.[2,4] Depending on the number of target genes regulated by a given TF, it can be classified as a *local* (a TF with one or few targets) or *global* (a TF with several targets) regulator, but the boundaries for each category might be somewhat arbitrary.[2−4] Usually, genes encoding local TFs are located adjacent to their targets.[7,8] Furthermore, as occur in the model bacteria *Escherichia coli*, global regulators (GRs) are dedicated to sense high priority signals that are crucial for the life style of the organism[2,9] and act in many cases in association with local TFs to reach a composite response.[2,3,9] Thus, signal integration through GRs imposes a hierarchical regulatory network to coordinate and optimize the transcriptional response to changing conditions.[10,11] However, many of the GRs of *E. coli* are also nucleoid-associated proteins (NAPs) that play a crucial role in the compaction of genomic DNA and on the organization of the chromosomal structure.[12−15] This is especially true for the H-NS protein that seems to be the NAP with major contribution to the chromosomal organization in *E. coli*.[14] While an important portion of the regulatory effect of NAPs on gene expression could be associated with their role on chromatin organization,[12,13,16] the promoter-specific action of these proteins during signal integration has been extensively investigated.[15]

The last decades of research on the molecular aspects of transcriptional regulation in bacteria have revealed that the interplay of TFs at the promoter region obeys a number of rules that start to be deciphered.[17] For instance, transcriptional activators usually bind to some preferred positions in the upstream region relative to the core promoter (i.e., the sequence recognized by the RNAP) and are categorized into different classes depending on the mode of interaction with the

Received: February 24, 2014
Published: July 18, 2014

RNAP.[17,18] In the same way, transcriptional repressors binding sites usually overlap with the core promoters to block the access of the RNAP, but additional mechanisms also exist.[17,18] In addition, it has long been known that the binding of a TF to some specific positions may influence the binding of another TF (or another molecule of the same TF) to a different operator, a process known as cooperativity.[19] This process can be either due to the existence of protein−protein interactions between TFs[20,21] or to changes in the local topology of the DNA molecule triggered by the binding of a first TF that is transmitted to neighbor sites, as occurs in the recently characterized DNA allosteric mechanism.[22] The sum of all those factors would ultimately generate the final outcome of the transcription regulation process, in the sense that it will determine when and at what rate the target gene will be expressed.[19,23,24]

According to the current knowledge regarding gene regulation, it is reasonable to infer that the architecture of a given promoter region (in terms of the diversity and arrangement of its *cis*-regulatory elements) should be the key element (at least in prokaryotes) determining the regulatory logic displayed by the target gene in response to a set of stimuli.[19,25] The role of the *cis*-regulatory architecture might be more controversial but still recognized as relevant in gene regulation in multicellular organisms.[26] Thus, the understanding of the rules behind the codification of the regulatory logic into the promoter architecture is crucial for many applications, including the reverse engineering of regulatory circuits in synthetic biology.[27] In this regard, several experimental and computational works have investigated the process of combinatorial gene regulation in the past few years.[19,23−25,28,29] However, most of the gained information on *cis*-regulatory logic has been derived for isolated model promoters[28] or have been performed using local regulators (such as LacI, TetR, and AraC), which are usually not involved in complex signal integration processes in natural systems.[25,28,30−32] In this sense, the investigation of signal integration mechanisms mediated by GRs could provide some important clues on the architecture/logic relationship of *cis*-regulatory elements at a global scale. In this report, we aimed to investigate *in silico* some features of complex promoters regulated by GRs in the model organism *E. coli*. We focused our analysis first on GRs by investigating the relationship between their binding sites at the genomic scale.[33] Our results highlight the absence of constraints limiting the combination of multiple binding sites either in tandem (adjacent) or overlapping, indicating that different global TFs could be easily combined in either arrangements. We further explored these features using *in silico* binding site evolution for three GRs (CRP, Fis, and IHF) and designed dual CRP-IHF binding sites that were shown to be functional *in vivo*. The results presented here provide important clues for the understanding of bacterial promoter architectures subjected to combinatorial regulation, and implemented a new approach for the engineering of regulatory circuits in synthetic biology.

RESULTS AND DISCUSSION

Plasticity of TF Binding Sites in the Genome of *E. coli*.

The available experimental and computational evidence on the mechanisms of gene regulation in several organisms suggest that the arrangement of TF binding sites in promoter regions are not occasional but indeed reflect some mechanistic constraints related to transcriptional control.[25,34−36] While most evidence for this has been generated for TF studied in isolation,[25,35,36] new reports have reinforced the notion that these arrangements also reflect constraints related to some specific requirements for TF−TF interaction during gene regulation.[34] In this sense, the systematic analysis of TF-binding site arrangements at a genomic scale has the potential to reveal such "hidden" rules related to combinatorial control of transcription. For example, in eukaryotes, Kazemian and coworkers recently reported a strong bias in the organization of TF binding sites in developmental regulatory regions in *Drosophila* due to the cooperative action of some specific pairs of TF *in vivo*.[37] In prokaryotes, Ezer and collaborators[34] used computer simulations to suggest that the organization of TF-binding sites in *E. coli* reflect physical constraints related to the search of the target sequence by the TF through facilitated diffusion. Thus, the disposition of closely spaced binding sites would determine if the first TF enhance or difficult the binding of the other to the second operator, with important consequences to the final promoter logic.[34] Although very promising, these proposed physical constraints are still lacking experimental validation.

In order to get further knowledge on the mechanisms of combinatorial control in *E. coli*, we analyzed the available information on the location of TF-binding sites in the promoters of this organism. We seek bias in pairs of binding sites that could indicate some sort of preferential action mode between global TFs. For this, we used the information on the localization of binding sites available at the RegulonDB.[33] We only used the data set for binding sites for which experimental evidence were available. For the analysis presented below, we used information on (i) the disposition of the binding sites respect to each other (i.e., adjacent or overlapping; Figure 1)

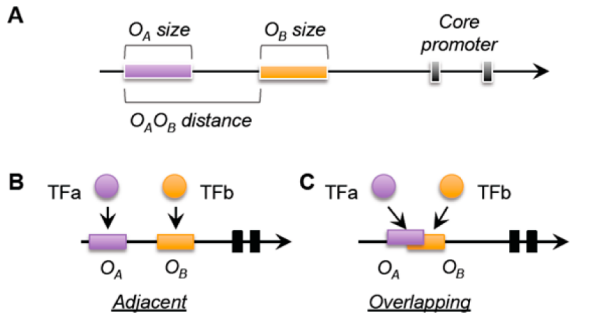


Figure 1. Schematic representation of the analyzed binding site arrangements. (A) For distance calculation, the left-most positions of the *cis*-regulatory operators of two TFs (exemplified as O_A and O_B) were taken. According to the calculated distances, the pairs of sites where classified as adjacent (B) or overlapping (C). Only sites with a distance shorter than 100 bp were analyzed. Abbreviations: TFa, transcriptional factor "a"; TFb, transcriptional factor "b"; O_A, operator "a"; O_B, operator "b".

and (ii) the effect of each site on the target promoter (i.e., positive or negative). With this information on hand, we inspect the relationship between binding sites for the seven TFs classified as GRs, namely CRP, IHF, Fis, FNR, ArcA, Lrp, and H-NS[2]. In order to first demonstrate that the data set indeed provide biologically relevant information on the *cis*-regulatory logic of *E. coli* promoters, we analyzed the arrangement of pairs of sites for the same GR (e.g., two binding sites for CRP). With this analysis, we would expected to observe that, if the binding sites for these global regulators were randomly distributed, half

dx.doi.org/10.1021/sb500084f | ACS Synth. Biol. 2014, 3, 666−675

Figure 2. Analysis of overlapping and adjacent sites for seven global regulators in *E. coli*. Using the annotated binding sites for TFs available at the RegulonDB,[33] we mapped pairs of binding sites located within a distance less than 100 bp where at least one element is targeted by a GR. For each pair, the effect of the regulators on the target promoter was mapped, and the pairs were classified as having the same (e.g., both are activators) or opposite effects. In all plots, the asterisks highlight significant differences (p-value < 0.05) in terms of the site effects on the target promoters. (A) Relationship between two binding sites belonging to the same GR. (B) The relationship between binding sites for a given GRs (e.g., CRP) and any other GR (except the one at stake) or local regulators found either overlapping or adjacent to the first site. (C) Relationship between pairs of GRs.

of the pairs of binding sites would perform the same function while the other half would perform opposite functions. However, as shown in Figure 2A, pairs of sites belonging to the same GR are strongly biased to perform the same function on the target promoter, both in overlapping and adjacent pairs, which is in accordance with the cooperative action of TFs during gene regulation.[24] Next, we analyzed pairs of sites where one operator is recognized by a first GR while the other is targeted by any TF of *E. coli* (local or global) except the first regulator. The results in Figure 2B revealed some bias for pairs of sites for certain GRs, yet not as strong as in the previous analysis. For instance, pairs of overlapping sites that include CRP, Fis, or H-NS operators showed a bias toward performing opposite effects on the target promoters. On the other hand, pairs of sites that include an operator for IHF displayed a bias toward performing the same effect on gene expression (Figure 2B). Additionally, adjacent pairs including sites for CRP, Lrp, or H-NS were biased for presenting opposite effects, while pairs including one ArcA site preferentially displayed the same effect on the promoter.

Despite the biases observed in the cases presented above, it is worth noticing that sites with "equivalent" and "opposite" effects were observed in all the pairs analyzed (Figure 2B), indicating that different combinations of binding sites are allowed for the seven studied GRs. Since the previous analysis takes into account different pairs of TFs, we decided to inspect the relationship between pairs of GRs (i.e., each site is recognized by one GR), to see if any preferential arrangement emerged. As shown in Figure 2C, only overlapping pairs for CRP/H-NS and adjacent pairs for CRP-Lrp sites presented a statistically significant bias for opposite effects, while no clear tendencies were observed for the remaining comparisons. Although the number of pairs in each comparison is much smaller than before (Figure 2B), the results presented in Figure 2C reinforce the notion that overlapping and adjacent binding sites for GRs are compatible with both equivalent (both regulators have the same influence) or opposite effects at the target promoters. Two well-known examples of this concept are the *arcB* promoter, where CRP and AraC are activators, and the *lac* promoter, which is activated by CRP and repressed by LacI.[38] On the basis of the evidence presented above, one could argue that a great number of complex regulatory behaviors could be generated through the combination of most of the GRs of *E. coli* (and possibly from many other bacteria), in the sense that the arrangement of *cis*-regulatory elements would not be constrained by the identity of the TFs involved. Taken together, this evidence indicates that no strong constraint limits the association of different GRs at target promoters in *E. coli*. In order to get more evidence for this supposition, we used *in silico* tools and *in vivo* assays to engineer combinatorial control in bacteria, as described in the following sections.

In Silico Optimization of Overlapping Binding Sites.

The molecular characterization of the binding sites for many TFs has provided valuable information on the sequences preferentially recognized by them.[4] Since many characterized TFs have well-defined requirements for specific nucleotide positions in the target sequences,[4] it is reasonable to expect that overlapping binding sites would be constrained by the binding specificity of the TFs involved. However, these elements would be of high interest for the engineering of synthetic circuit,[27,31,39] since it would allow a high compression of signal integrating elements in short DNA segments. As could be observed in the analysis presented above, the disposition of binding sites for global regulators in *E. coli* reveals that many different

dx.doi.org/10.1021/sb500084f | *ACS Synth. Biol.* 2014, 3, 666−675

combinations of bindings sites (even those overlapping each other) are indeed possible and thus functional operators could be efficiently recognized by different global regulators. In this sense, it would be useful to explore *in silico* which type of constraints would limit the appearance/engineering of dual binding sites for global regulators. For the purpose of the analysis presented here, we consider only complete overlapping, which we defined as a situation where the shorter binding site is contained inside the sequence of the larger one. This situation imposes a more severe constraint for the appearance of overlapping binding sites but would be of fundamental interest for applications where the operator position relative to the core promoter is important (as in the case of transcriptional activation). For the analysis, we used the position weight matrices (PWMs) available for the CRP, Fis, and IHF as shown in Figure 3.[40,41] In this sense, we first

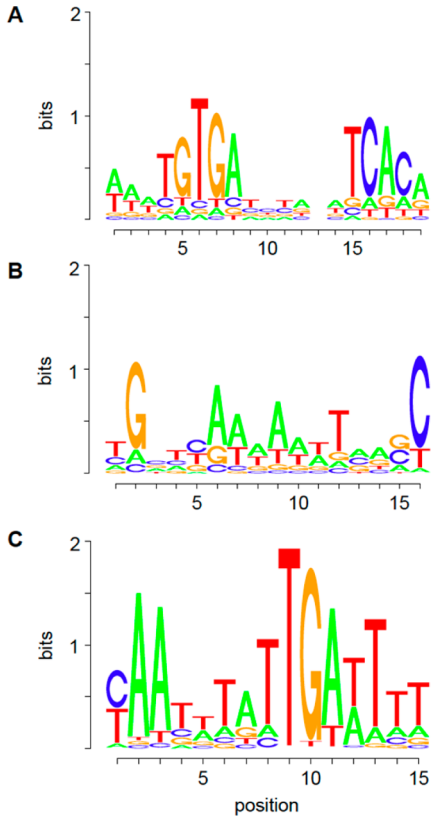


Figure 3. Sequence logos representing the binding site motifs for analyzed GRs. Logo representations were generated using WebLogo[53] with sequences extracted from Prodoric Database of TF-binding sites[40] or from previous works.[41] Figures represent the binding motifs for CRP (A), Fis (B), and IHF (C).

generated 10^5 random DNA sequences of 19 nt in length and analyzed the binding site score for each GR using the PWMs. This analysis allowed us to explore a portion of the DNA *sequence space* and to identify potential dual binding sites for each pairs of regulators. The comparison of the scores obtained for each pair is presented in Figure 4. As shown in the figure, this sequence sampling procedure allowed the identification of a number of sequences with high score for all the three pairs analyzed. In fact, the analysis of the data in Figure 4 demonstrates a statistically significant positive correlation for CRP−Fis (correlation index = 0.120, *p*-value < 0.001) and Fis−IHF (correlation index = 0.083, *p*-value < 0.001) sites, while CRP−IHF sites displayed a negative correlation (correlation index = −0.058, *p* value <0.001). The evidence of these correlations can be observed in Figure S1 (Supporting Information (SI)), where the score of the top ten sites for each TF are plotted to show the score for the other TFs. As shown in SI Figure S1A, the best site found for CRP (site "A") showed also a high score for Fis (80% of the matrix maxima). In the same way, from the top ten sites for Fis (SI Figure S1B), site "F" also displayed a good score for CRP (72% of the maxima) while site "J" was also high with the IHF matrix (71%). Finally, from the best sites for IHF (SI Figure S1C), site "I" showed a good score with the Fis matrix (74%). While the scores of secondary sites would perhaps not be enough to allow efficient TF binding *in vivo*, they indicate that few additional mutations would be sufficient to generate a fully functional site (yet evidently reducing the binding affinity of the primary site).

While the random approach presented above would allow exploring the generation of dual binding sites, it only allow the exploration of a limited portion of the potential sequence space in a computationally feasible way. One possibility to overcome this limitation would be to perform a random walking into the sequence space to select the best sequences containing potential dual sites for the target regulators. In this sense, in order to design high-core sites for pairs of TFs that in fact would be functional *in vivo*, we implemented a genetic algorithm aimed to allow the *in silico* evolution of binding sites for multiple regulators. The general strategy of the algorithm is depicted in SI Figure S2. Briefly, the algorithm performs several cycles of mutation of DNA sequences, then proceeds to calculate the sequence scores using the PWMs for the three GRs and finally performs a random selection of elements between the best candidates to use them in additional rounds of mutation. Using this procedure, several variants of optimized sequences could be obtained. The results of these simulations are shown in Figure 5. When DNA sequences were subjected to the evolution of binding sites with improved score for all three GRs, the system was able to optimize many sequences with reasonable good scores after few generations (Figure 5A). In most cases, the obtained scores were in agreement with the scores from real binding sites found in the genome of *E. coli*.[40,42] In addition, when the system was used to evolve binding sites for selected pairs of GRs, the final observed scores were even higher than those found for the three GRs (Figures 5B, C), which is expected since the sequences were optimized for only two PWMs at a time. All together, the evolutionary algorithm implemented here was capable to efficiently optimize dual overlapping binding sites for GRs, some of which were selected for further *in vivo* experimental validation.

Construction and *In Vivo* Experimental Validation of Synthetic *lac* Promoters. The utilization of a genetic algorithm to optimize overlapping binding sites for pairs of GRs raised the possibility to test if in fact the resulting DNA sequences would be functional under physiological conditions. In this sense, we selected two different sequences for CRP−IHF designed using our approach to engineer synthetic promoter variants regulated by these GRs. We focused on the CRP−IHF pair because they presented well-characterized binding consensus (Figure 3A, C), thus the optimized sequences generated here would be more reliable. For the validation, we selected the well-characterized *lac* promoter as a start point. At the wild type *Plac* sequence, a CRP site is located at the region −61.5 relative to the transcriptional start site and

dx.doi.org/10.1021/sb500084f | *ACS Synth. Biol.* 2014, 3, 666−675

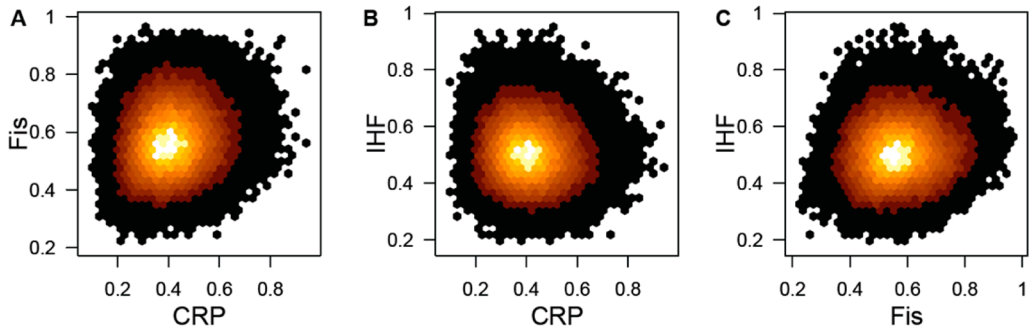


Figure 4. Distribution of the binding site scores found in random data sets. For the analysis, 10^5 random short sequences (19 nt in length) were assembled and analyzed for the binding site score using the PWMs for the three GRs of interest. Each individual plot represents the pair-to-pair comparison between the obtained binding scores. The *x*-axis and *y*-axis represent the scores of the binding sites. Colors indicate the density of sequences obtained in each particular region, from high (light yellow) to low density (back). Plots were generated using *hexbin* function of R package. Comparisons between CRP-Fis (A), CRP-IHF (B), and Fis-IHF (C) are shown.

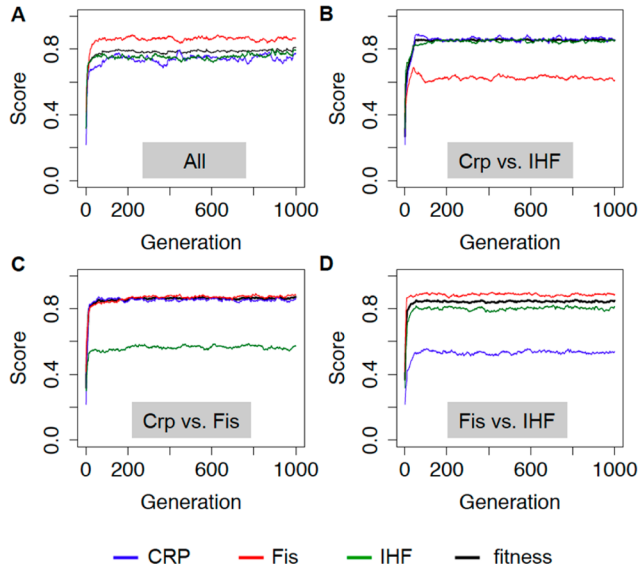


Figure 5. Optimization of overlapping binding sites for GRs. Plots represent the average PWMs scores in the populations of each generation. For comparison purposes, scores are normalized for the maximal values of the PWMs. Black line represents the average fitness of the population, which is calculated as the average of the PWMs score values in the population. The other lines represent the average score values for CRP (blue), Fis (red), and IHF (green), and the same color code is followed in all plots. (A) Optimization of overlapping binding sites for CRP, Fis, and IHF. (B) Optimization of overlapping binding sites for CRP and IHF. (C) Optimization of overlapping binding sites for CRP and Fis. (D) Optimization of overlapping binding sites for Fis and IHF. In plots B–D, the black line represents the fitness value for the pairs of GRs analyzed.

binding of CRP to this site is required for full promoter activity.[43] The structure of the *Plac* and the sequences used to construct the synthetic promoters are shown in Figure 6A. Three variants of the promoter were constructed by replacing a 19 bp sequence centered at position −61.5 by a random sequence with low score for both CRP and IHF (*Plac*Δ*CRP*) and two sequences optimized for both regulators (*Plac-CI1* and *Plac-CI2*). All four *Plac* variants were cloned upstream of a unstable version of GFP (GFPlva) in a midcopy number vector, and the resulting constructions were introduced into the wild type and mutant *E. coli* strains lacking *crp* or *ihfA* genes (Table 1). As shown in Figure 6B, the analysis of the recombinant strains harboring the promoter fusions showed by fluorescence microscopy that the two new variants of *Plac* (*Plac-CI1* and *Plac-CI2*) were indeed functional in the wild type *E. coli* strain. However, in a mutant strain of *E. coli* lacking the *ihfA* gene, no fluorescence was observed for both synthetic promoter variants, in contrast to the fully functional wild type *Plac*. Finally, when the promoters were assayed in an *E. coli* strain presenting a chromosomal deletion in the *crp* gene,[44] no fluorescence could be detected in the recombinant cells (Figure 6B). Figure 6C presents a quantitative analysis of all promoter variants in the three strains of *E. coli*, showing that the synthetic promoters were functional in the wild type strain with an activity level about 7-times lower than the parental *Plac*. However, these activities were completely abolished in strains lacking either *crp* or *ihfA* genes, indicating that the two synthetic promoters *Plac-CI1* and *Plac-CI2* were dependent on CRP and IHF, thus resulting in an AND gate performance.[45] In fact, this AND gate behavior was expect from a biological point of view but could not be predicted using the optimization method since an OR gate could also be possible. In this sense, future work would be necessary to address the relationship between the promoter architecture generated and the final logic displayed by the system. Taken together, these results strongly suggest that functional synthetic overlapping sites could be designed *in silico* using available information on the binding consensus of TFs.

Conclusions. The understanding of the architecture of *cis*-regulatory elements in complex promoters is crucial not only to allow the easy decoding of the regulatory logic of natural promoters[19,35] but also to facilitate the engineering of synthetic circuits for biotechnology.[25,35,36] The work presented here contributes to this understanding by showing that no strong constraint limits the association between different global TFs on overlapping or adjacent binding sites on the overall effect that these TFs could exert at the target promoter. Since such type of constraints would not limit the combination of multiple binding sites at synthetic promoters, the challenge would be to know how efficiently combine different binding sites into a single DNA sequence (as in the extreme case of fully overlapped sites addressed here). As showed in this work, the implementation of evolutionary algorithms could be a powerful tool to facilitate such task, requiring only the availability of good consensus sequences for the TFs of interest (as has been demonstrated for the case of CRP and IHF binding sites). In fact, our experimental results for *Plac-CI1* and *Plac-CI2* synthetic promoters evidenced that dual binding sites located at a permissive position (in this example, −61.5) result in dual

dx.doi.org/10.1021/sb500084f | *ACS Synth. Biol.* 2014, 3, 666–675

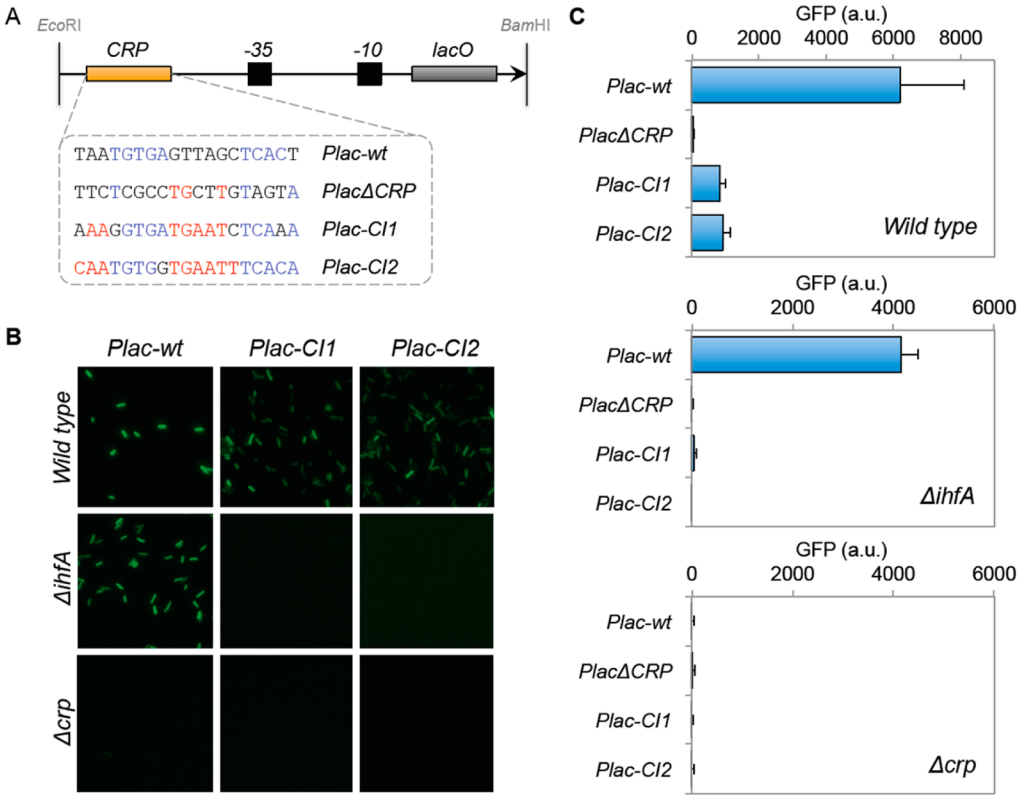


Figure 6. *In vivo* validation of dual binding sites for CRP and IHF. (A) Architecture of the lac promoter used for binding site validation. A 157 bp DNA fragment containing the lac promoter was cloned as an *Eco*RI/*Bam*HI insert upstream of an unstable variant of the GFP (GFPlva) reporter gene. The CRP and *lacO* operators are represented, as well as the RNAP binding site (−10/−35 boxes). Below the chart, the four variants assayed are shown. A 19 bp DNA sequence containing the wild type CRP operator is shown (*Plac-wt*). Another variant, which lacks the CRP binding site, was used as a control (*Plac-ΔCRP*). The two variants containing overlapping CRP-IHF binding sites are tagged as *Plac-CI1* and *Plac-CI2*. For the sequences, the nucleotides matching the CRP binding consensus are highlighted in blue while those for the IHF consensus are pointed out in red. (B) Fluorescence microscopy of *E. coli* strains (wild type, Δ*ihfA*, and Δ*crp*) harboring the wild type (*Plac-wt*) and the dual promoter variants (*Plac-CI1* and *Plac-CI2*). (C) Quantification of promoter activity in cells growing in LB media. Overnight grown strains were diluted 1:20 in fresh media and incubated to late exponential/stationary phase (OD_{600} ∼ 1.5). Cells were then washed and diluted in PBS 1X buffer. Fluorescence of cells was quantified and the GFP levels were calculated. For this, the recorded fluorescence was normalized by the cell density after subtracting the background levels of *E. coli* strains harboring promoter-less vectors. Horizontal bars indicate the standard deviation from three independent experiments.

regulation by these GRs, highlighting the effectiveness of the evolutionary algorithm implemented here. Finally, we proposed that the exploration of the *sequence space* to identify highly specific or dual (overlapping) binding sites for TFs could be an efficient strategy to design new synthetic promoters for the combinatorial integration of different signals during gene regulation in bacteria (Figure 7). Nevertheless, the integration of this approach with the increasing existing knowledge on the dynamic of reorganization of the chromatin structure in bacteria[13] will be crucial to generate new tools to decipher the regulatory logic of complex promoters and to use it in several synthetic biology applications.

METHODS

***cis*-Regulatory Data Set.** Information relative to the identified operators in the genome of *E. coli* was retrieved from the RegulonDB database,[33] version 8.1 from December 2012. From this data set, only TF-binding sites with experimental support were used. Specifically, we used the information related to (i) the start and end positions of the binding sites relative to the genomic coordinates, (ii) the effect of the binding site on the activity of the target promoter, and (iii) the operator distances relative to the transcription start site (TSS).

Analysis of Binding Site Arrangements. Using the data set extracted from RegulonDB, we analyzed the distribution of the binding sites for each of the target GRs spread over the genome. For each pair of regulators, we computed the distance between their binding sites. For this, we considered only binding sites that were within a distance lower than 100 bp from each other. For the binding site analysis, we considered as reference the leftmost base position of each binding site, ignoring if the sites were in the sense or antisense strand. In the cases where the distance between the sites were lower than the size of the first operator site, the pairs were classified as "overlapping", while in the remain cases they were called "adjacent" (Figure 1). Additionally, for each pair identified, the regulatory effects of each individual site at the target promoter were analyzed to determine if they presented the same effects (i.e., both were acting as activators or repressors) or opposite (i.e., one was acting as an activator while the other as a repressor). All analyses were performed using *ad hoc* Perl scripts and scripts written in R (http://www.r-project.org/).

Analysis of *cis*-Regulatory Elements in Random Sequences. In order to analyze the appearance of binding

dx.doi.org/10.1021/sb500084f | *ACS Synth. Biol.* 2014, 3, 666−675

Table 1. Bacterial Strains, Plasmids, and Primers Used in This Study

strain	description	ref
E. coli DH5α	*F⁻ endA1 gln V44 thi-1 recA1 relA1 gyrA96 deoR nupG Φ80dlacZΔM15 Δ(lacZYA-argF)U169, hsdR17(r_K^- m_K^+), λ⁻*	52
E. coli BW25113	*Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ⁻, rph-1, Δ(rhaD-rhaB)568, hsdR514.* Wild type strain.	50
E. coli JW5702-4	*Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ⁻, Δcrp-765::kan, rph-1, Δ(rhaD-rhaB)568, hsdR514. Δcrp* mutant strain.	50
E. coli JW1702-1	*Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ⁻, ΔihfA786::kan, rph-1, Δ(rhaD-rhaB)568, hsdR514. ΔihfA* mutant strain.	50
E. coli JW3229-1	b*Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ⁻, Δfis-779::kan, rph-1, Δ(rhaD-rhaB)568, hsdR514. Δfis* mutant strain.	50
	Plasmids	
pUC19	ApR, *oripColE1, lacZα*; standard cloning vector.	49
pPSSUB-101	*Cm*^{R}, *oripColE1*; GFPlva promoter probe vector.	48
pRV2	Km^R, *orip15a*; dual mCherry GFPlva promoter probe vector.	47
pMR1	*Cm*^{R}, *orip15a*; variant of pRV2 with *Cm* resistance marker.	this study
pMR1-*Plac-wt*	*Cm*^{R}, *orip15a*; *Plac-wt-GFPlva* transcriptional fusion.	this study
pMR1-*Plac-Δcrp*	*Cm*^{R}, *orip15a*; *Plac-Δcrp-GFPlva* transcriptional fusion.	this study
pMR1-*Plac-CI1*	*Cm*^{R}, *orip15a*; *Plac-CI1-GFPlva* transcriptional fusion.	this study
pMR1-*Plac-CI2*	*Cm*^{R}, *orip15a*; *Plac-CI2-GFPlva* transcriptional fusion.	this study
	Primers	
5-CmR-*Bgl*II	5′-GCGGAGATCTTGAGACGTTGATCGGCAC-3′	this study
3-CmR-SacI	5′-CCAAGCGAGCTCGATATC-3′	this study
5-PlacWT	5′-GCGCGAATTCTAATGTGAGTTAGCTCACTC-3′	this study
5-PlacCT	5′-GCGCGAATTCTTCTCGCCTGCTTGTAGTACATTAGGCACCCCAGG-3′	this study
5-PlacCI1	5′-GCGCGAATTCAAAGGTGATGAATCTCAAACATTAGGCACCCCAGG-3′	this study
5-PlacCI2	5′-GCGCGAATTCCAATGTGGTGAATTTCACACATTAGGCACCCCAGG-3′	this study
M13/F-20	5′-GTAAAACGACGGCCAGT-3′	49

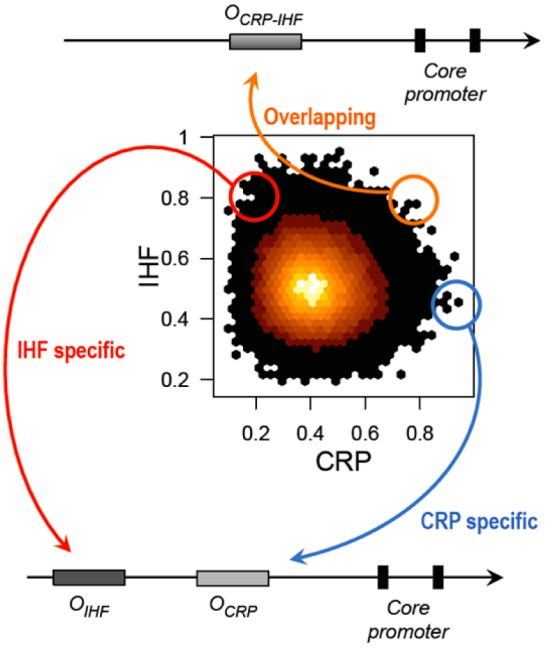


Figure 7. Exploring the sequence space for dual binding sites for GRs. The plot represents the binding score distribution for CRP and IHF found in the 10^5 random sequences data set (Figure 4B). As shown in the plot, selection of sequences at the top-leftmost (red circle) or bottom-mid right (blue) regions provides highly specific binding sites for IHF or CRP, respectively. These high specific sequences would allow the combinatorial control of target promoters through adjacent sites (scheme at the bottom of the chart). On the other hand, the region at the top-right part of the plot (orange circle) encompass sequences with high binding score for both regulators, which could be used for combinatorial control mediated by overlapping sites (scheme at the top part of the chart).

sites for three GRs (CRP, Fis, and IHF) in random data sets of short sequences, we used the position weight matrices (PWM) found in the PRODORIC database for CRP[40] and the described by Ussery and collaborators for Fis and IHF.[41] For the analysis, we used a data set formed by 10^5 randomly assembled DNA sequences with 19 nt in length. For each of the random sequences, we computed the sequence score according to each PWM relative to CRP, Fis, or IHF binding sites. PWMs were implemented in *Perl* and results were processed using R (http://www.r-project.org/).

***In Silico* Evolution of *cis*-Regulatory Elements.** In order to perform the *in silico* evolution of binding sites, we implemented a genetic algorithm in *Perl* for the evolution and selection of short DNA sequences (19 nt in length) with optimized binding sites for multiple GRs. First, we defined as a start point a sequence (control sequence: 5′-CGGCGGC-CGGGGGTAGGGA-3′) that had the lowest binding site score for the three PWMs retrieved form the 10^5 set of random sequences. The algorithm then constructs an initial population with N elements by copying the initial sequence N times. It is worth mentioning that each time that the original sequence is copied, a point mutation can occur with a probability of $P = 0.5$. For the generation of mutants, a position in the sequence is randomly selected and replaced by a randomly selected DNA base (A, G, T, or C), in such a way that if the selected base is the same as the present in the original position, no mutation occurs. Once the population is assembled, the new sequences are scored using the PMWs for the GRs. In order to create the next generation, a fitness factor is calculated based on the PWMs scores obtained (it is important to note that for this analysis the maximal score possible for each PWMs are normalized to 1.0). Fitness is calculated basically by averaging the score for the selected GRs for which a binding site should be optimized. By using the obtained fitness, the sequence population is then ranked and only the elements with fitness higher than or equal to 90% of the maximal value found in the current population are suitable to be copied to the next generation. Once the elements with the best fitness are identified using the above criteria, a new population is assembled subsequently through the random selection of

these elements and copied to the next generation. The whole process is repeated to a final number of generations G. For our simulations, we set both the populational size N and the number of generations G to 1000. The schematic representation of the genetic algorithm is shown in Figure S2 at the Supporting Information.

Bacterial Strains and Growth Conditions. All strains, plasmids and primers used in this study are listed in Table 1. *E. coli* DH5α was used as host for DNA cloning and plasmid maintaining. Chemically competent cells and transformation of cells with plasmidial DNA were performed as described previously.[46] Unless otherwise indicated, *E. coli* cells were grown in LB media[46] at 37 °C with aeration and constant shacking at 170 rpm When required, chloramphenicol (*Cm*, 30 μg mL^{-1}), kanamycin (Km, 50 μg mL^{-1}), or isopropyl β-D-1-thiogalactopyranoside (IPTG, 100 μM) were supplemented to the growth media to ensure plasmid maintenance.

Construction of GFP Probe Vector and Promoter Cloning. For the assay of synthetic promoters with dual binding sites, a variant of the pRV2 plasmid[47] was constructed by replacing the kanamycin resistance marker with a *Cm* resistance gene. For this, a 780 bp DNA fragment captaining the *Cm* resistance marker was PCR amplified from the pPSSUB-101 vector[48] using primers 5-CmR-*Bgl*II and 3-CmR-SacI. The resulting DNA fragment was digested with *Bgl*II and SacI restriction enzymes and cloned into the pRV2 vector digested with the same enzymes. The recombinant vector was named pMR1 and used to clone the different promoter variants. The cloning of the synthetic promoters was performed as follows. First, all promoter variants were constructed using the pUC19[49] vector as template DNA. The wild type *lac* promoter was amplified using primers 5-PlacWT and M13/F20. The resulting PCR product was cloned as a 170 bp *Eco*RI/*Bam*HI fragment into the pMR1 vector prepared using the same enzymes, generating the vector pMR1-*Plac-wt*. For the generation of a *Plac* variant with a nonfunctional CRP binding site, primer 5-PlacCT was used in combination with primer M13/F20 to amplify a promoter were the CRP recognition sequence is replaced by a control sequence (5′-TTCTCGCCTGCTTGTAGTA-3′). This mutated version of *Plac* was cloned in pMR1 as before, generating the vector pMR1-*Plac*-Δ*crp*. The two dual CRP-IHF *Plac* variants were constructed using primers 5-PlacCI1-M13/F20 (*Plac-CI1*) and 5-PlacCI2-M13/F20 (*Plac-CI2*), generating the vectors pMR1-*Plac-CI1* and pMR1-*Plac-CI2*, respectively. After verification of all constructions, the plasmids were used to transform chemically competent cells of *E. coli* wild type or harboring deletions of the *crp* or *ihfA* regulators.[50] The final strains were used to assay the promoter activity using the GFP reporter.

***In Vivo* Promoter Assay.** For the quantification of promoter activity *in vivo*, the wild type and mutant strains of *E. coli* harboring the different plasmids (pMR1-*Plac-wt*, pMR1-*Plac*-Δ*crp*, pMR1-*Plac-CI1* and pMR1-*Plac-CI2*) were used. As controls to calculate the basal expression of GFP, strains harboring the promoter-less pMR1 vector were also assayed. For the experiments, single colonies of each strain were inoculated in LB media supplemented with *Cm* and incubated overnight at 37 °C with air shaking. After pregrowth, 0.5 mL of the cultures were washed twice with 10 mM $MgSO_4$ buffer and resuspended in equal volume of the buffer. Cells were then diluted 1:20 in fresh LB-*Cm* media supplemented with 100 μM of IPTG and incubated for 4 h to reach late exponential/stationary phase. At this point, the optical density of cells at 600 nm (OD_{600}) was measured and samples were taken, washed in 10 mM $MgSO_4$ buffer and resuspended in the same buffer. Cell suspensions were used for fluorescence assays with excitation of 488 nm and emission of 510 nm in a Shimadzu RF-5301PC Fluorimeter. The GFP levels were calculated by normalizing the fluorescence by the OD_{600} of the cells and were expressed in arbitrary units. For each promoter, the background levels of GFP from an *E. coli* strain harboring the empty vector (pMR1) were subtracted to give the specific promoter activity. Experiments were performed with three biological replicas.

Confocal Fluorescence Microscopy. For confocal fluorescence microscopy experiments, 5 μL of cells at late exponential/stationary phase (pregrown for 4 h in rich media and processed as described above) were spotted in 2% low-melting agarose pads prepared as described previously.[51] Agarose pads were then loaded into microscopy chambers and cells were visualized using a Zeiss epifluorescence microscope with excitation of 498 nm and emissions of 516 nm. Phase contrast bright-field and fluorescent images were captured with an AxioCam camera (Carl Zeiss) and processed using the AxioVision software version 3.1.

■ ASSOCIATED CONTENT

Supporting Information

Figures S1 and S2. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +55 16 3602 3107. Fax: +55 16 3633 6840. Email: silvarochar@gmail.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are thankful to Dr. Isis do Carmo Kettelhut and Dr. Gustavo Henrique Goldman for providing access to their laboratory facilities. The study was supported by the National Council of Technological and Scientific Development of Brazil (project CNPq-Universal 14/2013, Ref. No. 472893/2013-0). M-EG benefits from a Young Talent Fellowship of the National Council of Technological and Scientific Development of Brazil (CNPq, Ref. No. 370630/2013-0). R.S.-R. is supported by a Post-Doctoral Fellowship from the São Paulo Science Foundation (FAPESP, Ref. No. 2013/04125-2).

■ REFERENCES

(1) Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol. 14*, 283−291.

(2) Martinez-Antonio, A., and Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol. 6*, 482−489.

(3) Balleza, E., Lopez-Bojorquez, L. N., Martinez-Antonio, A., Resendis-Antonio, O., Lozada-Chavez, I., Balderas-Martinez, Y. I., Encarnacion, S., and Collado-Vides, J. (2009) Regulation by transcription factors in bacteria: Beyond description. *FEMS Microbiol. Rev. 33*, 133−151.

(4) Ishihama, A. (2010) Prokaryotic genome regulation: Multifactor promoters, multitarget regulators, and hierarchic networks. *FEMS Microbiol. Rev. 34*, 628−645.

(5) Osterberg, S., del Peso-Santos, T., and Shingler, V. (2011) Regulation of alternative sigma factor use. *Annu. Rev. Microbiol. 65*, 37−55.

(6) Helmann, J. D., and Chamberlin, M. J. (1988) Structure and function of bacterial sigma factors. *Annu. Rev. Biochem. 57*, 839−872.
(7) Janga, S. C., Salgado, H., and Martinez-Antonio, A. (2009) Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucleic Acids Res. 37*, 3680−3688.
(8) Kolesov, G., Wunderlich, Z., Laikova, O. N., Gelfand, M. S., and Mirny, L. A. (2007) How gene order is influenced by the biophysics of transcription regulation. *Proc. Natl. Acad. Sci. U.S.A. 104*, 13948−13953.
(9) Thieffry, D., Huerta, A. M., Perez-Rueda, E., and Collado-Vides, J. (1998) From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays 20*, 433−440.
(10) Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature 431*, 308−312.
(11) Alon, U. (2007) Network motifs: Theory and experimental approaches. *Nat. Rev. Genet. 8*, 450−461.
(12) Ma, Q., Yin, Y., Schell, M. A., Zhang, H., Li, G., and Xu, Y. (2013) Computational analyses of transcriptomic data reveal the dynamic organization of the *Escherichia coli* chromosome under different conditions. *Nucleic Acids Res. 41*, 5594−603.
(13) Dorman, C. J. (2013) Genome architecture and global gene regulation in bacteria: Making progress towards a unified model? *Nat. Rev. Microbiol. 11*, 349−355.
(14) Wang, W., Li, G. W., Chen, C., Xie, X. S., and Zhuang, X. (2011) Chromosome organization by a nucleoid-associated protein in live bacteria. *Science 333*, 1445−1449.
(15) Browning, D. F., Grainger, D. C., and Busby, S. J. (2010) Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. *Curr. Opin. Microbiol. 13*, 773−780.
(16) Sobetzko, P., Travers, A., and Muskhelishvili, G. (2012) Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc. Natl. Acad. Sci. U.S.A. 109*, E42−50.
(17) Browning, D. F., and Busby, S. J. (2004) The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol. 2*, 57−65.
(18) Barnard, A., Wolfe, A., and Busby, S. (2004) Regulation at complex bacterial promoters: How bacteria use different promoter organizations to produce different regulatory outcomes. *Curr. Opin. Microbiol. 7*, 102−108.
(19) Hermsen, R., Tans, S., and ten Wolde, P. R. (2006) Transcriptional regulation by competing transcription factor modules. *PLoS Comput. Biol. 2*, e164.
(20) Dove, S. L., Joung, J. K., and Hochschild, A. (1997) Activation of prokaryotic transcription through arbitrary protein−protein contacts. *Nature 386*, 627−630.
(21) Berg, J., Willmann, S., and Lassig, M. (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol. 4*, 42.
(22) Kim, S., Brostromer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y. Q., Su, X. D., Sun, Y., and Xie, X. S. (2013) Probing allostery through DNA. *Science 339*, 816−819.
(23) Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., and Phillips, R. (2005) Transcriptional regulation by the numbers: Applications. *Curr. Opin. Genet. Dev. 15*, 125−135.
(24) 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*, 116−124.
(25) Kinkhabwala, A., and Guet, C. C. (2008) Uncovering cis regulatory codes using synthetic promoter shuffling. *PLoS One 3*, e2030.
(26) Paris, M., Kaplan, T., Li, X. Y., Villalta, J. E., Lott, S. E., and Eisen, M. B. (2013) Extensive divergence of transcription factor binding in drosophila embryos with highly conserved gene expression. *PLoS Genet. 9*, e1003748.
(27) Purnick, P. E., and Weiss, R. (2009) The second wave of synthetic biology: From modules to systems. *Nat. Rev. Mol. Cell Biol. 10*, 410−422.
(28) 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*, e45.
(29) Hermsen, R., Ursem, B., and ten Wolde, P. R. (2010) Combinatorial gene regulation using auto-regulation. *PLoS Comput. Biol. 6*, e1000813.
(30) Setty, Y., Mayo, A. E., Surette, M. G., and Alon, U. (2003) Detailed map of a cis-regulatory input function. *Proc. Natl. Acad. Sci. U.S.A. 100*, 7702−7707.
(31) Sprinzak, D., and Elowitz, M. B. (2005) Reconstruction of genetic circuits. *Nature 438*, 443−448.
(32) Siuti, P., Yazbek, J., and Lu, T. K. (2013) Synthetic circuits integrating logic and memory in living cells. *Nat. Biotechnol. 31*, 448−452.
(33) Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muniz-Rascado, L., Garcia-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martinez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernandez, S., Alquicira-Hernandez, K., Lopez-Fuentes, A., Porron-Sotelo, L., Huerta, A. M., Bonavides-Martinez, C., Balderas-Martinez, Y. I., Pannier, L., Olvera, M., Labastida, A., Jimenez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chavez, V., Hernandez-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. *Nucleic Acids Res. 41*, D203−213.
(34) Ezer, D., Zabet, N. R., and Adryan, B. (2014) Physical constraints determine the logic of bacterial promoter architectures. *Nucleic Acids Res.*, 1−12, DOI: 10.1093/nar/gku078.
(35) Hunziker, A., Tuboly, C., Horvath, P., Krishna, S., and Semsey, S. (2010) Genetic flexibility of regulatory networks. *Proc. Natl. Acad. Sci. U.S.A. 107*, 12998−13003.
(36) Cox, R. S., 3rd, Surette, M. G., and Elowitz, M. B. (2007) Programming gene expression with combinatorial promoters. *Mol. Syst. Biol. 3*, 145.
(37) Kazemian, M., Pham, H., Wolfe, S. A., Brodsky, M. H., and Sinha, S. (2013) Widespread evidence of cooperative DNA binding by transcription factors in drosophila development. *Nucleic Acids Res. 41*, 8237−8252.
(38) Collado-Vides, J., Magasanik, B., and Gralla, J. D. (1991) Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol Rev. 55*, 371−394.
(39) Elowitz, M., and Lim, W. A. (2010) Build life to understand it. *Nature 468*, 889−890.
(40) Munch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., and Jahn, D. (2003) PRODORIC: Prokaryotic database of gene regulation. *Nucleic Acids Res. 31*, 266−269.
(41) Ussery, D., Larsen, T. S., Wilkes, K. T., Friis, C., Worning, P., Krogh, A., and Brunak, S. (2001) Genome organisation and chromatin structure in *Escherichia coli*. *Biochimie 83*, 201−212.
(42) Shimada, T., Fujita, N., Yamamoto, K., and Ishihama, A. (2011) Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. *PLoS One 6*, e20081.
(43) Schmitz, A. (1981) Cyclic AMP receptor protein interacts with lactose operator DNA. *Nucleic Acids Res. 9*, 277−292.
(44) Berthoumieux, S., de Jong, H., Baptist, G., Pinel, C., Ranquet, C., Ropers, D., and Geiselmann, J. (2013) Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol. Syst. Biol. 9*, 634.
(45) Silva-Rocha, R., and de Lorenzo, V. (2008) Mining logic gates in prokaryotic transcriptional regulation networks. *FEBS Lett. 582*, 1237−1244.
(46) Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor, New York.
(47) Silva-Rocha, R., and de Lorenzo, V. (2011) Implementing an OR-NOT (ORN) logic gate with components of the SOS regulatory network of *Escherichia coli*. *Mol. Biosyst. 7*, 2389−2396.
(48) Basu, S., Mehreja, R., Thiberge, S., Chen, M. T., and Weiss, R. (2004) Spatiotemporal control of gene expression with pulse-generating networks. *Proc. Natl. Acad. Sci. U.S.A. 101*, 6355−6360.

(49) Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) Improved M13 phage cloning vectors and host strains: Nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene 33*, 103−119.
(50) Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol. Syst. Biol. 2*, 2006 0008.
(51) Young, J. W., Locke, J. C., Altinok, A., Rosenfeld, N., Bacarian, T., Swain, P. S., Mjolsness, E., and Elowitz, M. B. (2012) Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nat. Protoc. 7*, 80−88.
(52) Grant, S. G., Jessee, J., Bloom, F. R., and Hanahan, D. (1990) Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants. *Proc. Natl. Acad. Sci. U.S.A. 87*, 4645−4649.
(53) Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: A sequence logo generator. *Genome Res. 14*, 1188−1190.