ACS SyntheticBiology
Cite This: ACS Synth. Biol. 2018, 7, 1834−1843
Letter
pubs.acs.org/synthbio

Predicting Transcriptional Output of Synthetic Multi-input Promoters

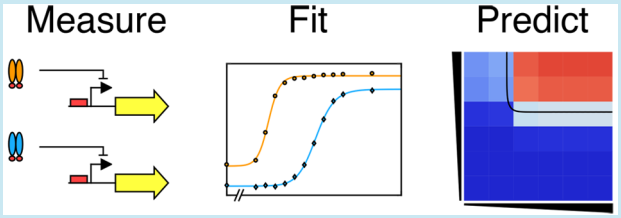
David M. Zong,[†] Selahittin Cinar,[‡] David L. Shis,[§] Krešimir Josić,[‡,§,⊥] William Ott,[*,‡] and Matthew R. Bennett[*,§,∥]

[†]Graduate Program in Systems, Synthetic, and Physical Biology, [§]Department of Biosciences, [∥]Department of Bioengineering, Rice University, Houston, Texas 77005, United States

[‡]Department of Mathematics, [⊥]Department of Biology and Biochemistry, University of Houston, Houston, Texas 77004, United States

Supporting Information

ABSTRACT: Recent advances in synthetic biology have led to a wealth of well-characterized genetic parts. As parts libraries grow, so too does the potential to create novel multi-input promoters that integrate disparate signals to determine transcriptional output. Our ability to construct such promoters will outpace our ability to characterize promoter performance, due to the vast number of input combinations. In this study, we examine the input−output relations of recently developed synthetic multi-input promoters and describe two methods for predicting their behavior. The first method uses 1-dimensional induction data obtained from experiments on single-input systems to predict the *n*-dimensional induction responses of systems with *n* inputs. We demonstrate that this approach accurately predicts Boolean (on/off) responses of multi-input systems consisting of novel chimeric transcription factors and hybrid promoters in *Escherichia coli*. The second method uses only a small amount of multi-input response data to accurately predict analog system response over the entire landscape of input combinations. Taken together, these methods facilitate the design of synthetic circuits that utilize multi-input promoters.



KEYWORDS: *transcriptional logic gates, chimeric transcription factors, multi-input synthetic hybrid promoters, transcriptional noise*

The majority of synthetic gene circuits have been built with a small set of parts (*i.e.*, genetic elements that can affect gene expression), and this has consequently limited the complexity of circuit designs.[1] Called "the component problem," synthetic biologists have encountered a dearth of well-characterized and orthogonal parts needed to build larger circuits.[2] One approach directed at solving the parts problem involves identifying and characterizing naturally occurring parts, a method called parts mining.[3] Another approach utilizes concepts developed for protein engineering to create novel parts through modification of existing transcription factors.[4] Still other ventures have focused on the forward engineering of regulatory control sequences, including the construction of synthetic promoters,[5] ribosome binding sites,[6] and terminators.[7]

Recent advances toward solving the component problem have led to a growing repository of orthogonal genetic parts. Yet, the number of possible part combinations grows much more quickly than the size of parts libraries. To complicate matters, the assembly of parts can produce unpredictable results, as the behavior of each part can change with the context in which it operates.[8−10] Hence, the design and construction of biological circuits is often based on trial and error, resulting in a laborious and inefficient process. To address this issue, some researchers have developed computational models of synthetic circuits to simulate and evaluate circuit designs before building them in the lab, accelerating the process of genetic circuit development.[11−14]

Of the many types of parts being designed for synthetic biology, one of the most important classes is the multi-input promoter. Such promoters allow transcriptional output at one locus to depend on two or more transcription factors, enabling the efficient construction of complex regulatory architectures. Indeed, many synthetic gene circuits have relied on multi-input promoters, including oscillators,[13,15] logic gates,[16−19] an image edge detector,[20] a synthetic predator−prey ecosystem,[21] and a pattern formation circuit.[22] However, because multi-input promoters are not common in bacteria, synthetic biologists have had to resort to engineering novel promoters that respond to multiple transcription factors. This has led to various types of multi-input promoters that largely fall into one of three classes: tandem promoters, hybrid promoters, and chimera-responsive promoters.

Tandem promoters are constructed by placing two promoters one after the other.[16,17] Therefore, transcription can be initiated at either of two places, each regulated by a different transcription factor. One drawback to this approach is that the 5′ UTRs generated differ depending on the initiation

Received: April 16, 2018
Published: July 24, 2018

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843

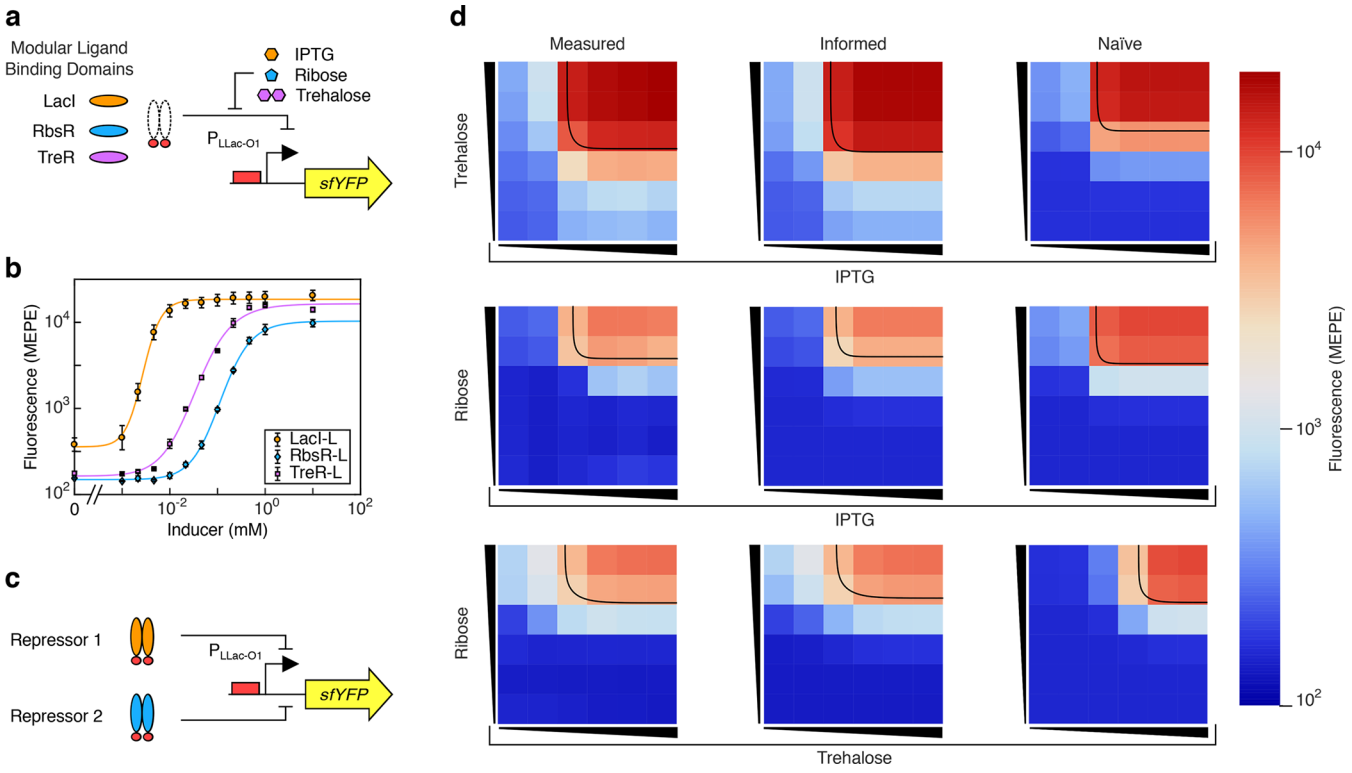


Figure 1. Predicting the output of two-input ligand-inducible transcriptional repressor systems. (a) Single-input systems consist of one of three chimeric repressors, each with a LacO DBD. These chimeric repressors regulate the expression of sfYFP as a function of their cognate inducers. (b) Single-input response profiles. Points depict mean fluorescence of the flow cytometer distribution, normalized to beads. Error bars represent standard deviation from three experimental replicates. We fit sigmoidal functions to the fluorescence data (curves). (c) Two-input systems. For each system, we coexpressed two chimeric transcription factors that simultaneously regulate the $P_{LLacO-1}$ promoter driving the expression of sfYFP. (d) Heatmaps representing the expression of sfYFP for the two-input systems. From left to right and bottom to top, boxes represent 0, 0.001, 0.01, 0.1, 1, and 10 mM of the indicated inducer. First column: Experimental data. Second column: Informed model prediction. Third column: Naïve model prediction. Black curves indicate location of half-maximal induction.

site, possibly resulting in different translational activities.[8] This problem can be alleviated by using enzymatic cleavage of the UTR, thus standardizing the structure of the mRNA and normalizing translation initiation.[8,11]

Hybrid promoters (also known as combinatorial promoters) are engineered to contain more than one operator site at the core promoter site, which allows different species of transcription factors to bind.[15,23−27] This enables these promoters to respond to different transcription factors simultaneously and always generate the same mRNA independent of how the promoter was activated. One drawback of hybrid promoters, however, is that they can be hard to engineer, as operator sites must be precisely placed within the promoter.

Chimera-responsive promoters are not engineered promoters. Instead, multiple transcription factors are engineered to bind the same operator site within the promoter.[18,19] To do this, chimeric transcription factors are constructed by appending the DNA binding domain (DBD) of one transcription factor to the ligand binding domain (LBD) of another.[28] If multiple transcription factors, each containing the same DBD, are present in a cell, each will regulate the target promoter based on the presence or absence of its ligand. The drawback of this method is that one swaps a DNA engineering problem for a protein engineering problem. Provided that the correct operator sites are present, chimeric transcription factors can be used in conjunction with tandem promoters and hybrid promoters, increasing the number of potential input combinations.

In light of these recent developments, it has become feasible to construct novel multi-input promoters with a high number of inputs. But, as the number of potential input combinations grows, conventional expression characterization techniques will become untenable—fully characterizing every input combination will be unworkable. For example, if one possesses a library of x transcription factors, then there exist $x(x-1)/2$ possible pairwise combinations. To further complicate matters, if one wishes to characterize the response curve of an inducible promoter system with only one input requiring n measurements at different concentrations, then a similar characterization for a two-input system would require n^2 measurements, a three-input system n^3 measurements, and so on.[29]

Here, we describe and evaluate two methods for predicting the output of synthetic multi-input promoters responding to varying concentrations of multiple inducing ligands. Our first method uses data only from single-input systems to predict multi-input system output. We first characterized each of a set of chimeric transcription factors acting upon a target promoter in isolation. We then developed naïve models that use only this single-input information to predict the behavior of two-input systems consisting of two chimeras operating simultaneously. With the naïve modeling in place, we constructed and characterized three two-input chimeric repressor systems *in vivo*. Our naïve models accurately predicted the Boolean (on/off) responses of these chimeric repressor systems. To demonstrate the flexibility of the method, we successfully

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843

extended it to a hybrid promoter system compatible with the set of characterized chimeric transcription factors.

Our second method is based on the following idea: To predict the response of a multi-input system for every combination of inducing ligand concentrations, one has only to construct the multi-input system *in vivo* and then test it with a small number of ligand concentration combinations. Importantly, the number of measurements needed increases only linearly with the number of inputs. These measurements serve as training data for our model, that then predicts multi-input system response at all inducer concentration combinations. We found this second method to be more accurate than the first for all of the two-input systems we built, and extremely accurate for a three-input hybrid promoter system.

Overall, this work provides methods for accurately predicting the transcriptional output of synthetic multi-input promoters. This ability is critical to the forward design of complex gene circuits relying on multiple environmental or intercellular inputs.

RESULTS AND DISCUSSION

Multi-input Chimeric Repressor Systems. We first developed a "naïve" approach that uses data from single-input systems to predict the behavior of a multi-input system, assuming we can factorize input interactions. To do this, we examined several versions of chimeric AND gates previously reported by Shis *et al.*[18] These transcriptional AND gates are built using LacI/GalR family transcription factor chimeras, each containing the DNA binding domain of LacI and the ligand binding domain of another family member (Figure 1a): LacI-L (which responds to isopropyl β-D-1-thiogalactopyranoside (IPTG)), RbsR-L (ribose), or TreR-L (trehalose).[18] When two different chimeras are present, each will independently regulate the $P_{LLacO-1}$ promoter, meaning that both ligands will be required for transcription.

We began by determining the system response for each chimera acting in isolation. To do this, we used flow cytometry to measure the response of the $P_{LLacO-1}$ promoter driving expression of sfYFP in the presence of each of the three chimeric repressors individually. As shown in Figure 1b, each single-input system behaves as expected, with sfYFP expression monotonically increasing as inducer concentration increases.

We modeled each induction profile with a sigmoidal function of the form

$$F(I) = CH^+(I) + \varepsilon \quad \text{with} \quad H^+(I) = \frac{I^p}{I_*^p + I^p} \tag{1}$$

Here F denotes fluorescence, I denotes inducer concentration, ε represents fluorescence in the absence of inducer, C is the maximum fluorescence increase in response to inducer, p is the Hill coefficient, and I_* is the inducer concentration at which the Hill function H^+ attains its half-maximal value. We fit the model given in eq 1 to the measured steady-state fluorescence data for each of the three chimeric repressors (Figure 1b). In each case, the model tightly fits the data with low error (see Supporting Information (SI) for fitting details).

Our naïve model uses these single-input induction curves to predict the two-input expression of systems consisting of two chimeric repressor species operating simultaneously (Figure 1c): Let $F_1(I_1)$ and $F_2(I_2)$ denote the single-input induction curves (obtained above) for any two of the three inducers (IPTG, ribose, or trehalose). Using eq 1, we have

$$F_1(I_1) = C_1\frac{I_1^{p_1}}{I_{*,1}^{p_1} + I_1^{p_1}} + \varepsilon_1 = C_1H_1^+(I_1) + \varepsilon_1$$

$$F_2(I_2) = C_2\frac{I_2^{p_2}}{I_{*,2}^{p_2} + I_2^{p_2}} + \varepsilon_2 = C_2H_2^+(I_2) + \varepsilon_2$$

The naïve prediction for the two-input system is then given by

$$F(I_1, I_2) = \alpha_0 + \alpha_1H_1^+(I_1) + \alpha_2H_2^+(I_2) + \alpha_3H_1^+(I_1)H_2^+(I_2) \tag{2}$$

That is, we descriptively model two-input system response as a quadratic polynomial function of H_1^+ and H_2^+. The first three terms in eq 2 capture the "leaky" expression that may occur when at most one inducer is present. Coefficient α_3 quantifies the strength of the interaction term. A two-input system functions well as an (analog) AND gate when α_3 is much larger than α_0, α_1, and α_2.

The task is now to estimate the parameters α_0, α_1, α_2, and α_3, using only the parameters of F_1 and F_2. In the absence of either inducer (*i.e.*, setting $H_1^+ = H_2^+ = 0$), we assume the repressor that binds more tightly in the uninduced state controls AND gate fluorescence, and thereby obtain

$$\alpha_0 = \min(\varepsilon_1, \varepsilon_2) \tag{3}$$

Similarly, when both inducers are present at saturating concentrations, we set $H_1^+ = H_2^+ = 1$ and obtain

$$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 = \min(C_1 + \varepsilon_1, C_2 + \varepsilon_2) \tag{4}$$

When one inducer is present at saturating concentration and the other is absent, we assume that although repression by the induced repressor has been relieved, AND gate fluorescence cannot be greater than single-input fluorescence for the other repressor in the uninduced state. This assumption is equivalent to

$$\alpha_0 + \alpha_1 = \varepsilon_2, \quad \alpha_0 + \alpha_2 = \varepsilon_1 \tag{5}$$

obtained by setting $H_1^+ = 1$, $H_2^+ = 0$ and $H_1^+ = 0$, $H_2^+ = 1$, respectively. Combining eqs 3, 4, and 5 yields a system of equations for the α_j:

$$\alpha_0 = \min(\varepsilon_1, \varepsilon_2) \tag{6a}$$

$$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 = \min(C_1 + \varepsilon_1, C_2 + \varepsilon_2) \tag{6b}$$

$$\alpha_0 + \alpha_1 = \varepsilon_2 \tag{6c}$$

$$\alpha_0 + \alpha_2 = \varepsilon_1 \tag{6d}$$

We solve eqs 6 to complete the derivation of our naïve prediction for AND gate fluorescence. Since no two of the three single-input induction curves intersect one another (see Figure 1b), for the sake of clarity we solve eqs 6 assuming $F_1(I) \geq F_2(I)$ for every inducer concentration I, obtaining

$$\alpha_0 = \varepsilon_2, \quad \alpha_1 = 0, \quad \alpha_2 = \varepsilon_1 - \varepsilon_2, \quad \alpha_3 = (C_2 + \varepsilon_2) - \varepsilon_1$$

Our naïve prediction for AND gate fluorescence is therefore given by

$$F(I_1, I_2) = \varepsilon_2 + (\varepsilon_1 - \varepsilon_2)H_2^+(I_2) + ((C_2 + \varepsilon_2) - \varepsilon_1) \times H_1^+(I_1)H_2^+(I_2) \tag{7}$$

We emphasize that this model depends only on single-input information.

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843

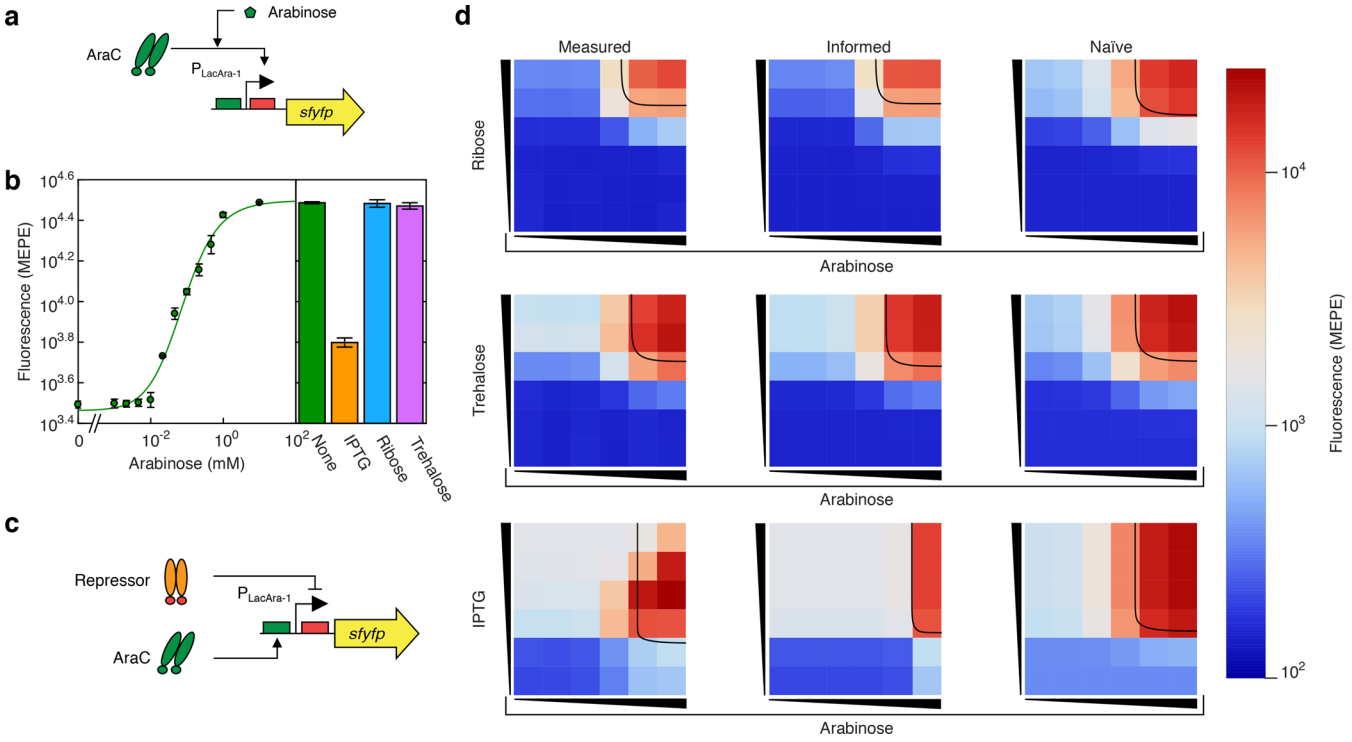


Figure 2. Predicting the output of two-input hybrid promoter systems. (a) Single-input AraC circuit. The $P_{LacAra-1}$ promoter drives sfYFP production in the presence of arabinose. (b) Production of sfYFP after 2 h of induction as a function of arabinose concentration. Points: mean fluorescence of the flow cytometer distribution. Error bars: standard deviation from three experimental replicates. We fit a sigmoidal function to the data (green curve). Bars represent sfYFP production at 10 mM arabinose and 10 mM of the other inducer indicated. (c) The two-input case for the $P_{LacAra-1}$ promoter. A chimeric repressor and AraC operate simultaneously. (d) Heatmaps representing the expression of sfYFP for the two-input systems. From left to right and bottom to top, boxes represent 0, 0.001, 0.01, 0.1, 1, 10 mM of the indicated inducer. First column: Experimental data. Second column: Informed model prediction. Third column: Naïve hybrid promoter prediction. Black curves indicate location of half-maximal induction.

To test the naïve model, we built three two-input expression systems by pairwise combining the three single-input systems. For each two-input system, we constructed plasmids that express a pair of chimeric repressors that regulate expression of sfYFP driven by the $P_{LLacO-1}$ promoter. We measured the mean production of sfYFP at 6 inducer concentrations for each inducer, ranging from 0 to 10 mM, producing a total of 36 different induction conditions for each system (Figure 1d, left column). All three two-input systems exhibited AND gate behavior: High levels of both inducers were needed to see significant expression.

The naïve model accurately predicts the Boolean (on/off) response profiles of the two-input chimeric repressor systems (Figure 1d, right column; see SI for detailed quantitative error analysis). This finding supports one of the primary claims of our paper: The naïve model enables efficient Boolean characterization of multi-input systems and requires only single-input data for model specification. We will show that this claim holds for multi-input synthetic hybrid promoter systems as well. Although the naïve model succeeds in the Boolean context, its (analog) predictions are not highly accurate for some pairs of inducer concentrations. In particular, the model performs well for all three two-input chimeric repressor systems when both inducer concentrations are low or high, but nontrivial prediction errors occur in the TreR-L systems at low to medium concentrations of trehalose and medium to high concentrations of the second inducer (ribose or IPTG). Further, the predicted position for the isocline of half-maximal induction does not always match the position obtained from the experimental response profile (black lines in Figure 1d).

In order to accurately predict analog system response over the entire landscape of input combinations, we developed a second predictive method that uses the polynomial-Hill model described by eq 2 in a different way. Rather than use single-input information to infer model parameters, our second predictive method calls for training the model on a *small subset* of the data obtained from the experimentally constructed two-input system. We call the second method "informed modeling".

For each pair of chimeric repressors, we fit the 8 parameters for eq 2 using only measured fluorescence data from just 12 of the 36 inducer pairs. In particular, we used the 11 inducer pairs with at least one inducer at 10 mM and the pair with both inducers at zero (see SI). The informed model then predicts fluorescence response for any pair of inducer concentrations. The predictions of the informed model outperform those of its naïve counterpart for all three of the chimeric repressor pairs (Figure 1d, center column; see SI for error quantification). In particular, the informed model accurately predicts the positions of the three isoclines corresponding to half-maximal induction.

We based our choice of training data set on scalability considerations. Thinking of the two-input inducer space as a 6-by-6 grid, the set of inducer pairs with at least one inducer at 10 mM consists of two one-dimensional "slices". As we explain in the Discussion, training a *D*-input system would then require *D* one-dimensional slices, so the size of the training data set scales only linearly with number of inputs. We based

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843

our selection of slices on the idea that the two-input system responds most sensitively to changes in the concentration of a given inducer when saturated with the other inducer. The inducer pair with both inducers at zero calibrates the fluorescence floor. We do not claim that our choice of training data set is optimal, only that it works well. In particular, errors associated with the predictions of the informed model are nearly as low as those obtained by fitting the polynomial-Hill model to all of the measured two-input data (see SI).

Multi-input Hybrid Promoter Systems. We conjectured that the predictive models we developed for the chimeric repressor systems would also work with another type of multi-input promoter: the synthetic hybrid promoter. To test this conjecture, we constructed, observed, and predictively modeled two-input systems utilizing the $P_{LacAra-1}$ promoter, a hybrid promoter containing an operator site for the activator AraC and one for the repressor LacI. This promoter (again driving sfYFP) is compatible with the previously characterized chimeric repressors, and may be used in the presence of constitutively expressed AraC and the gene(s) encoding chimeric repressor(s).

Single-Input Experiments. We first expanded our analysis of single-input systems to the AraC-regulating $P_{LacAra-1}$ single-input case. We characterized the response of the $P_{LacAra-1}$ promoter in the presence of constitutive AraC and no LacI chimeras (Figure 2a). In particular, we measured mean steady-state fluorescence of sfYFP in a range of L-arabinose from 0 to 10 mM at 12 concentrations (Figure 2b). As with the chimeric repressors, we fit a sigmoidal function of the form (eq 1) to the arabinose induction data (Figure 2b, green curve).

To observe any potential crosstalk, we measured the response of $P_{LacAra-1}$ to 10 mM of IPTG, ribose, and trehalose at 0 mM and 10 mM of arabinose (Figure 2b, right). The $P_{LacAra-1}$ promoter produced significantly reduced sfYFP expression with the addition of IPTG. This is consistent with previous results demonstrating that IPTG inhibits the activity of native AraC.[30] Ribose and trehalose did not have any nonspecific activatory or inhibitory effects.

Naïve Modeling. As for the two-input chimeric repressor systems, we developed a naïve model that predicted, based solely on single-input information, the Boolean behavior of the activator-repressor hybrid systems. We conjectured that the model (eq 7) would not work well for the hybrid systems because of the architectural differences between the $P_{LLacO-1}$ and $P_{LacAra-1}$ promoters (see SI for verification of this conjecture). We therefore developed a new naïve model for the hybrid systems, described as follows.

Start with the single-input induction curves for the chimeric repressors (R) and the activator (A), now denoted

$$F_R(I_R) = C_R H_R^+(I_R) + \varepsilon_R$$

$$F_A(I_A) = C_A H_A^+(I_A) + \varepsilon_A$$

respectively. As before, the new naïve model has the structure

$$F(I_A, I_R) = \alpha_0 + \alpha_1 H_A^+(I_A) + \alpha_2 H_R^+(I_R) + \alpha_3 H_A^+(I_A) H_R^+(I_R) \quad (8)$$

To specify the α_j, we set (H_A^+, H_R^+) equal to (0, 0), (1, 1), (1, 0), and (0, 1), respectively, and model as follows:

$$\alpha_0 = \varepsilon_R \quad (9a)$$

$$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 = \sqrt{(C_A + \varepsilon_A)(C_R + \varepsilon_R)} \quad (9b)$$

$$\alpha_0 + \alpha_1 = \varepsilon_R \quad (9c)$$

$$\alpha_0 + \alpha_2 = \sqrt{\varepsilon_A \varepsilon_R} \quad (9d)$$

The system (eqs 9) follows from two assumptions. First, when the repressor is uninduced, AND gate fluorescence is ε_R regardless of arabinose concentration. This assumption is based on the relationship $F_A^+(I_A) \geqslant \varepsilon_A > \varepsilon_R$ between the single-input induction curves. Second, we assume that when the repressor is fully induced, AND gate fluorescence values are given by geometrically averaging corresponding single-input fluorescence values. This assumption is based on the idea that repressor may bind the hybrid promoter with low probability even at saturating repressor cognate inducer concentrations.

A Caveat. We constructed the naïve model for the hybrid promoter systems using single-input chimeric repressor data from another promoter: the $P_{LLacO-1}$ promoter. This was necessary given our experimental methodology because the $P_{LacAra-1}$ promoter does not produce transcription without AraC bound, so we could not generate repressor-only induction curves for the hybrid promoter.

In general, such a scheme could be problematic, especially for promoters with substantially different expression ranges. However, we believe our methodology was sound because the expression ranges of the $P_{LLacO-1}$ and $P_{LacAra-1}$ promoters are comparable in our context. Regardless, our naive model for the hybrid promoter accurately predicts *in vivo* two-input Boolean system responses, as described below.

Implementation and Prediction of Two-Input Hybrid Promoter Systems. To test the new naïve model, we coexpressed AraC and one of the chimeric repressors, which simultaneously regulate the $P_{LacAra-1}$ promoter (Figure 2c). We measured mean sfYFP production at 6 inducer concentrations for each inducer, ranging from 0 mM to 10 mM, producing a total of 36 different induction conditions (Figure 2d, left column).

As we conjectured, the hybrid promoter naïve model (Figure 2d, right column) accurately predicts digital (on/off) system responses (see SI for error quantification). Further, this new naïve model outperforms the model (eq 7) for the ribose-arabinose and trehalose-arabinose systems. As expected, the hybrid promoter naïve model fails for the IPTG-arabinose system, due to the crosstalk between IPTG and the activation of the promoter by AraC. In future work, we intend to develop systematic theoretical and experimental frameworks for treating various types of crosstalk, including ligand-transcription factor interference and signaling molecule (HSL) crosstalk. Capturing the latter is important for the bioengineering of synthetic microbial consortia.

As for the chimeric repressor systems, we developed an informed model for the hybrid promoter systems that accurately predicts analog response over the landscape of induction conditions, once trained with a small amount of the two-input fluorescence data. For each activator-repressor pair, we fit the 8 parameters for the model described by eq 8 using only measured fluorescence data from just 12 of the 36 inducer pairs. In particular, we used the 11 inducer pairs with at least one inducer at 10 mM and the pair with both inducers at zero. For the ribose-arabinose and trehalose-arabinose systems, the informed model accurately predicts fluorescence response for the remaining 24 inducer pairs and outperforms the hybrid

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843

promoter naïve model (Figure 2d, center; see SI for error quantification). The informed model fails to capture IPTG-arabinose system response due to the aforementioned crosstalk between IPTG and AraC.

Three-Input Prediction and Measurement. To test the utility of our methodology in higher dimensions, we constructed a three-input system that responds to a combination of arabinose, ribose, and trehalose (Figure 3a). We constructed the system by coexpressing RbsR-L and TreR-L and cotransforming the $P_{LacAra-1}$ promoter.

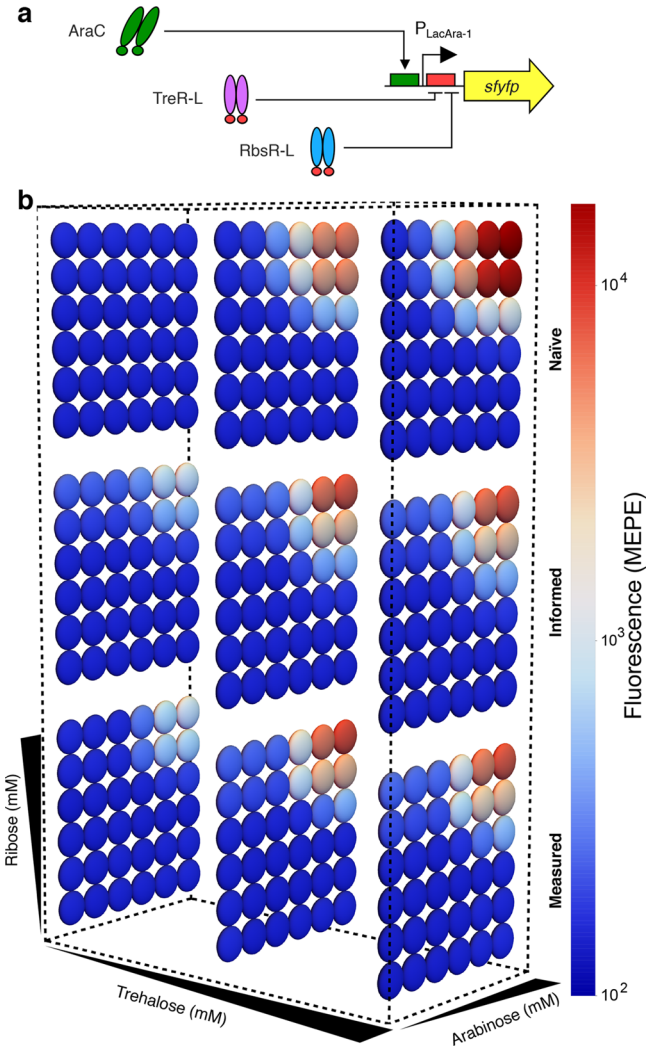


Figure 3. Predicting the output of a three-input hybrid promoter system. (a) Circuit diagram. (b) Mean fluorescence for various inducer combinations. Arabinose and ribose concentrations: 0, 0.001, 0.01, 0.1, 1, 10 mM. Trehalose concentrations: 0, 0.1, 10 mM. Top row: Naïve hybrid promoter prediction. Middle row: Informed model prediction. Bottom row: Measured data.

By combining modeling techniques used for the chimeric repressor systems and the hybrid promoter systems, we developed a naïve model that predicts three-input system Boolean behavior as a function of the arabinose, ribose, and trehalose single-input induction curves. The three-input naïve model naturally generalizes the two-input version in eq 2, and has the form

$$F(I_1, I_2, I_3) = \alpha_0 + \alpha_1 H_1^+(I_1) + \alpha_2 H_2^+(I_2) + \alpha_3 H_3^+(I_3) + \alpha_4 H_1^+(I_1) H_2^+(I_2) H_3^+(I_3) \quad (10)$$

See SI for information on how we deduce the α_i from the single-input induction curves. To test the naïve model, we measured fluorescence at 108 inducer combinations: 6 levels of arabinose and ribose (0, 0.001, 0.01, 0.1, 1, 10 mM of each), and 3 levels of trehalose (0, 0.1, 10 mM). The naïve model (Figure 3b, top) accurately predicts digital (on/off) measured system response (Figure 3b, bottom; see SI for error quantification), but quantitative discrepancies exist. In particular, measured system behaviors at 0.1 mM and 10 mM trehalose essentially match, a prediction that the naïve model does not completely capture. The naïve model falters here because the inflection point for the trehalose single-input induction curve is located at 0.15 mM, while the three-input system behaved experimentally as if 0.1 mM trehalose already saturates.

As with the two-input systems, we conjectured that an informed model would yield superior performance. We specified such a model by fitting the 11 parameters in eq 10, using a small subset of the 108 inducer combinations for which measurements were taken. In particular, we used the 13 inducer triples with at least two inducers at 10 mM and the triple with all inducers at zero (see Discussion for justification). The informed model predicts (analog) system response for the remaining 108 − 14 = 94 inducer combinations with a high level of accuracy (Figure 3b, middle; see SI for error quantification).

Noise in Multi-input Promoters. In addition to predicting mean multi-input system responses, we sought to understand how noise in multi-input systems depends on the number of inputs, input types, and promoter architecture. This is especially vital given that multi-input promoters may be used as biosensors. Noise in sensor output will propagate forward through the circuit, potentially affecting downstream circuit performance.[31]

For each single-input system and each inducer concentration, we measured the distribution of sfYFP production in the population and computed the robust coefficient of variation (RCV) of this distribution. The RCV is obtained by dividing the interquartile range by the median of the distribution. We began our analysis by looking for a functional relationship between the square of RCV and fluorescence, as previous findings show that the relationship between noise and overall expression should be linear above a threshold of protein copy number.[32,33] However, our results are inconsistent with previous findings, as both high- and low-noise regimes emerged at intermediate expression levels (Figure 4a).

To investigate the reason, we more closely examined examples of high and low noise for both the $P_{LLacO-1}$ and $P_{LacAra-1}$ promoters. In the case of high noise for the $P_{LacAra-1}$ promoter, the fluorescence distribution is bimodal (Figure 4b). Indeed, in all cases where arabinose was an inducing molecule, intermediate amounts caused bimodality in the population. The bimodality observed in the arabinose-inducible case is consistent with previous studies,[34] as the ara operon is under native regulatory control in our system and this configuration causes an all-or-nothing response in the population at intermediate amounts of arabinose.[35,36] We found that fluorescence distribution, and therefore RCV behavior, changed depending on the type of promoter used. For

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843

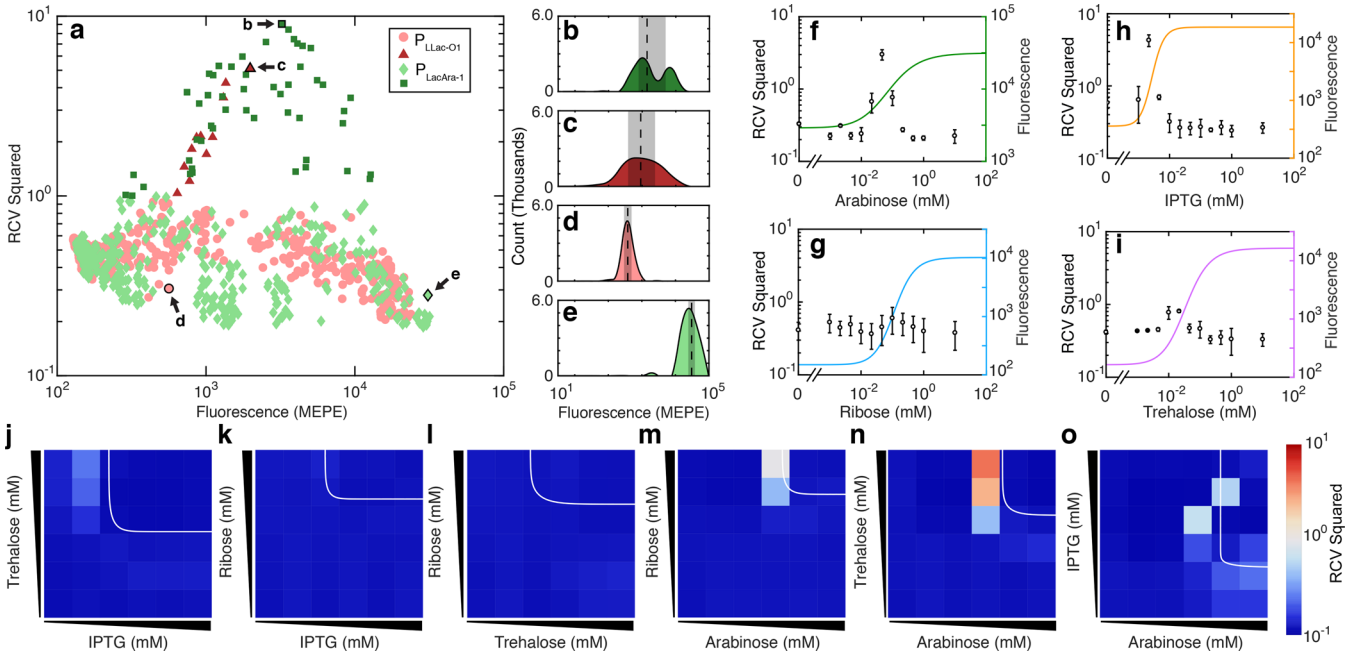


Figure 4. Noise in multi-input systems. (a) Noise (square of RCV) as a function of overall fluorescence. Each point is a flow cytometer measurement of one induction level for one of the systems mentioned previously. Measurements from $P_{LLacO-1}$ are in red and measurements from $P_{LacAra-1}$ are in green. Points with squared RCV value greater than one are darker. There appears to be no overall pattern that relates fluorescence to noise. (b, c, d, e) Fluorescence distributions associated with points marked by arrows in (a). These smoothed fluorescence distributions represent cases of high noise in the $P_{LacAra-1}$ system, high noise in the $P_{LLacO-1}$ system, low noise in the $P_{LLacO-1}$ system, and low noise in the $P_{LacAra-1}$ system, respectively. Note that the distribution is bimodal in (b). Dotted line represents median and gray region represents the interquartile range. (f, g, h, i) Square of RCV as a function of inducer concentration in the single-input cases. Hill functions expressing fluorescence as a function of inducer concentration are drawn in the background. Noise is generally highest at the inflection point of each induction curve, and is more pronounced in the AraC and LacI-L systems. (j, k, l, m, n, o) Square of RCV as a function of inducer concentrations for the two-input systems. Isoclines representing the midpoint between minimal and maximal production are drawn over the heatmaps. Noise is generally largest near the isoclines. The arabinose-IPTG two-input system is an exception, due to inducer crosstalk.

instance, we found that when high noise occurs with the $P_{LLacO-1}$ promoter, the fluorescence distribution is unimodal (Figure 4c). We suspect this might have been caused by the different mechanisms by which each of the small molecule inducers enter the cell. For instance, in our systems, IPTG both diffuses and is actively transported,[37,38] while arabinose, ribose, and trehalose are all transported by their respective transporters. We did not modify the native transporters of these sugars in our strains.

We also examined how noise magnitude varies as a function of inducer concentrations. For the single-input systems, noise magnitude peaked at the inflection points of the induction curves (Figure 4f–i). This behavior is consistent with the theory that a noisy input would produce the noisiest output at the inflection point of the system.[39] The spikes in noise level at the inflection points were substantial for arabinose and IPTG, but only modest for ribose and trehalose. RCV spikes for the ribose and trehalose cases may have been only modest because the transporters of these sugars may have been active at the tested concentrations.

Extending this thinking to the two-input systems, we observed that noise magnitude peaked near the isoclines corresponding to production halfway between minimal and maximal expression (Figure 4j–o). Further, noise characteristics depended on which pair of inducers were involved. For example, systems involving arabinose exhibited noise spikes along the arabinose axis at intermediate arabinose concentrations. When pairing two inducers that individually produced low noise, the resultant two-input system also produced low noise (*e.g.*, with ribose and trehalose). Therefore, at least with the systems we examined, the multi-input promoters inherited the noise characteristics of their constituent parts.

Discussion. As our ability to design and construct sophisticated synthetic circuits continues to grow, so too must our ability to predict the performance of such circuits *in silico*.[18] Here, we have developed and tested two predictive methods. The naïve method is philosophically ideal in that our naïve models accurately predict multi-input system Boolean responses using only single-input data (and knowledge of promoter architecture). The informed method provides a high level of (analog) accuracy, but requires a small amount of multi-input data.

Importantly, both predictive methods scale to systems with large numbers of inputs. In particular, the amount of multi-input data needed to train our informed models scales linearly with the number of inputs, while the number of inducer combinations that such models predict scales exponentially: Suppose we wish to predict a D-input system, where each input takes V possible values. Our informed method would accurately predict system response for all V^D input combinations, while requiring only $D(V - 1) + 2$ values for model specification.

When using a small amount of multi-input data to train our informed models, one natural question arises: Why have we chosen the particular "one-dimensional" subset of the multi-input data as we have? Would not another subset work just as well, or perhaps better? Answering the first question, by selecting the "one-dimensional" subset of multi-input data by

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834–1843

varying one inducer at a time while holding the others at full induction, we probe the full dynamic range of each individual inducer. To validate our choice, we compared the error associated with our informed model prediction to the error produced by using all of the multi-input data to fit the underlying model. Importantly, these error values essentially match for all of the systems we tested (see SI). This observation leaves open the possibility that other training data sets may work just as well.

We have presented two methods for predicting the output of multi-input synthetic promoter/transcription factor systems. Practitioners should consider the following when selecting a method. The naïve method is ideal when the user has a vast library of well-characterized single-input devices and wishes to evaluate the digital (on/off) behavior of potential designs without having to perform additional lab work. The naïve method provides predictions that will aid the user in narrowing down the large design space of potential multi-input combinations and in selecting candidates to build or analyze further. The informed method excels when the user has already constructed a multi-input system and wishes to probe the entire induction space. By collecting a small set of induction data, the rest of the induction space can be predicted to a high degree of accuracy.

Our informed method of prediction captures the analog nature of the inputs/output signal transduction of multi-input promoters. Such promoters are often treated as digital devices, because it can be too resource-intensive to test the entire input space. Digital inputs/output approximation works in certain situations. However, when designing microbes for complex environments such as the gut microbiome or soil, relevant signals may be in constant flux. An analog predictive approach is therefore necessary, as it facilitates the design of circuits that can accommodate a range of signals.[40] The analog approach assists with the parts problem as well: Analog circuits can require fewer parts than their digital counterparts to compute a given function.[41]

The polynomial-Hill modeling framework that we use for our naïve and informed methods of prediction is descriptive rather than mechanistic. For those who favor mechanistic modeling, we have developed and tested a statistical mechanical model for operator site binding (see SI). This energy modeling framework describes binding to DNA by transcription factors and σ-factor using the Boltzmann−Gibbs distribution.[16,25,42] For all of the multi-input systems we built, the energy model performs essentially as well as the original informed model, when trained on the same data set (see SI).

Overall, our predictive methodologies facilitate the design of synthetic microbes that rely on multiple environmental inputs.

MATERIALS AND METHODS

Strains and Plasmids. All experiments were performed in the *E. coli* strain CY15 (MG1655 Δ*lacI*Δ*sdiA*::*araC*::*lasR*::*rhlR*). The plasmid pH6 was used as the reporter plasmid for experiments concerning the $P_{LLacO-1}$ promoter. The $P_{LacAra-1}$ sfYFP reporter plasmid pDZ041 was generated by PCR amplification of the $P_{LacAra-1}$ promoter with sfYFP expression cassette and the backbone of the pH6 reporter plasmid using primers with *BsaI* sites on the overhang. The two fragments were assembled in a Golden Gate reaction.[43] Plasmids that express the chimera repressors were previously described in the methods of Shis *et al.*[18] As a control, an empty plasmid was generated by PCR amplification of the backbone of the chimera plasmid with *BsaI* sites on the overhang. Strains for characterization were made by cotransformation of a plasmid containing either a single or a pair of chimeric repressors or an empty plasmid with a reporter plasmid. A colony from the transformation was inoculated in LB media containing ampicillin and kanamycin and grown overnight. The culture was then used to make a −80 °C stock by mixing 1:1 with 50% (v/v) glycerol.

Growth and Induction Measurements. Frozen stocks of strains cotransformed with the appropriate plasmid were streaked onto LB agar plates containing ampicillin and kanamycin and incubated overnight. A colony from the plate is inoculated in 3 mL of LB media containing ampicillin and kanamycin and grown for 14−16 h. The LB culture is diluted 100:1 into M9 media containing 0.4% (v/v) glycerol and 0.2% (w/v) CAS amino acids with antibiotics in a round-bottom 96 well plate and incubated at 37 C in a microplate shaker at 800 rpm for 2 h. At this time, cultures were induced and diluted by the addition of and equivalent volume of M9 media containing 2× inducer at a 1:1 ratio. The induced cultures were returned to the plate shaker and grown for 2 h and then placed on ice. The final OD600 of these cultures ranged from 0.3 to 0.4 as measured in a Tecan Infinite M1000.

Flow Cytometry. Flow cytometry was performed using a Millipore Guava HT Cyte Flow Cytometer with a custom blue and green laser setup and an auto sampling tray. Before the start of data collection each day, Spherotech Rainbow Calibration Particles (8 peaks) (Spherotech, RCP-30−5A, Lot# AH01) were measured for calibration. Cultures were diluted 100:1 into phosphate buffered saline (Invitrogen) containing 34 ug/mL chloramphenicol on ice in a 96 well plate. The plate was transferred to a 37 C incubator for 1 h to allow the sfYFP fluorophore to mature. The plate is then placed on ice for 15 min. The plate is then placed inside the Guava HT Cyte. The machine gain settings were FSC: 128, SSC: 64, YEL: 64, with an SSC threshold of 20. All other settings were set to default. 15 000 events were collected for each sample. Data files were exported using the Guava InCyte software to FCS3.0 files. These files were opened using the FlowCal software package[44] with a customized I/O handling script specific to to InCyte's FCS3.0 export files. Samples were first gated by density in forward and side scatter, taking 50% the densest region in the forward and side scatter plane using FlowCal's built in function. The readout of the fluorescence channel in arbitrary units was normalized to units of molecules of equivalent R-phycoerythrin (MEPE) using a calibration curve generated by the beads run on the same experimental day. The mean and robust coefficient of variance (RCV) were calculated using FlowCal's built in functions. RCV is defined as the interquartile range divided by the median of the distribution. The RCV was used to reject spurious signals that would affect the calculation of the variance.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.8b00165.

Polynomial-Hill modeling framework (modeling and model fitting for single-input gates, polynomial-Hill modeling for multi-input systems); Model training using only single-input induction curves: The naïve approach

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843

(model derivation, model parameter values, and quantification of model prediction error for the two-input chimeric repressor systems and the two- and three-input hybrid promoter systems); Model training using multi-input induction data: The informed approach (model fitting techniques, model parameter values, and quantification of model prediction error for the multi-input systems); Boltzmann−Gibbs energy modeling (model derivation, model fitting, model parameter values, and quantification of model prediction error for the multi-input systems); Supporting experimental methods (flow cytometry, plasmids used) (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: ott@math.uh.edu.

*E-mail: matthew.bennett@rice.edu.

ORCID

Matthew R. Bennett: 0000-0002-4975-8854

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Ye Chen for useful discussions related to this work. This work was supported by the National Institutes of Health grant R01GM117138 (to M.R.B., K.J., W.O.), the Robert A. Welch Foundation grant C-1729 (to M.R.B.), the National Science Foundation grants DMS-1662290 (to M.R.B., K.J.), and DMS-1413437 (to W.O.), and the National Science Foundation Graduate Research Fellowship Program grant 1450681 (to D.M.Z.).

REFERENCES

(1) 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.

(2) Bennett, M., and Hasty, J. (2009) Overpowering the component problem. *Nat. Biotechnol. 27*, 450−451.

(3) Stanton, B. C., Nielsen, A. A., Tamsir, A., Clancy, K., Peterson, T., and Voigt, C. A. (2014) Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nat. Chem. Biol. 10*, 99−105.

(4) Taylor, N. D., Garruss, A. S., Moretti, R., Chan, S., Arbing, M. A., Cascio, D., Rogers, J. K., Isaacs, F. J., Kosuri, S., Baker, D., Fields, S., Church, G. M., and Raman, S. (2016) Engineering an allosteric transcription factor to respond to new ligands. *Nat. Methods 13*, 177−183.

(5) Brewster, R., Jones, D., and Phillips, R. (2012) Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. *PLoS Comput. Biol. 8*, e1002811.

(6) Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol. 27*, 946−950.

(7) Chen, Y., Liu, P., Nielsen, A. A., Brophy, J. A., Clancy, K., Peterson, T., and Voigt, C. A. (2013) Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat. Methods 10*, 659−664.

(8) Brophy, J. A., and Voigt, C. A. (2014) Principles of genetic circuit design. *Nat. Methods 11*, 508−520.

(9) Kim, K. H., and Sauro, H. M. (2010) Fan-out in gene regulatory networks. *J. Biol. Eng. 4*, 16.

(10) Del Vecchio, D., Ninfa, A. J., and Sontag, E. D. (2008) Modular cell biology: retroactivity and insulation. *Mol. Syst. Biol. 4*, 161.

(11) Nielsen, A., Der, B., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E., Ross, D., Densmore, D., and Voigt, C. (2016) Genetic circuit design automation. *Science 352*, aac7341.

(12) Yordanov, B., Dalchau, N., Grant, P. K., Pedersen, M., Emmott, S., Haseloff, J., and Phillips, A. (2014) A computational method for automated characterization of genetic components. *ACS Synth. Biol. 3*, 578−588.

(13) Stricker, J., Cookson, S., Bennett, M., and Mather, W. (2008) A fast, robust and tunable synthetic gene oscillator. *Nature 456*, 516−519.

(14) O'Brien, E. L., Van Itallie, E., and Bennett, M. R. (2012) Modeling synthetic gene oscillators. *Math. Biosci. 236*, 1−15.

(15) Chen, Y., Kim, J., Hirning, A., Josić, K., and Bennett, M. (2015) Emergent genetic oscillations in a synthetic microbial consortium. *Science 349*, 986−989.

(16) Tamsir, A., Tabor, J. J., and Voigt, C. A. (2011) Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'. *Nature 469*, 212−215.

(17) Nielsen, A. A., and Voigt, C. A. (2014) Multi-input CRISPR/Cas genetic circuits that interface host regulatory networks. *Mol. Syst. Biol. 10*, 763.

(18) Shis, D., Hussain, F., Meinhardt, S., Liskin, S., and Bennett, M. (2014) Modular, Multi-Input Transcriptional Logic Gating with Orthogonal LacI/GalR Family Chimeras. *ACS Synth. Biol. 3*, 645−651.

(19) Chan, C. T., Lee, J. W., Cameron, D., Bashor, C. J., and Collins, J. J. (2016) 'Deadman' and 'Passcode' microbial kill switches for bacterial containment. *Nat. Chem. Biol. 12*, 82−86.

(20) Tabor, J. J., Salis, H. M., Simpson, Z. B., Chevalier, A. A., Levskaya, A., Marcotte, E. M., Voigt, C. A., and Ellington, A. D. (2009) A synthetic genetic edge detection program. *Cell 137*, 1272−1281.

(21) Balagaddé, F. K., Song, H., Ozaki, J., Collins, C. H., Barnet, M., Arnold, F. H., Quake, S. R., and You, L. (2008) A synthetic *Escherichia coli* predator-prey ecosystem. *Mol. Syst. Biol. 4*, 187.

(22) Basu, S., Gerchman, Y., Collins, C. H., Arnold, F. H., and Weiss, R. (2005) A synthetic multicellular system for programmed pattern formation. *Nature 434*, 1130−1134.

(23) Cox, R., Surette, M., and Elowitz, M. (2007) Programming gene expression with combinatorial promoters. *Mol. Syst. Biol. 3*, 145.

(24) Lutz, R., and Bujard, H. (1997) Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res. 25*, 1203−1210.

(25) Chen, Y., Ho, J. M. L., Shis, D. L., Gupta, C., Long, J., Wagner, D. S., Ott, W., Josić, K., and Bennett, M. R. (2018) Tuning the dynamic range of bacterial promoters regulated by ligand-inducible transcription factors. *Nat. Commun. 9*, 64.

(26) Wu, F., Zhang, Q., and Wang, X. (2018) Design of Adjacent Transcriptional Regions to Tune Gene Expression and Facilitate Circuit Construction. *Cell Syst. 6*, 206−215.e6.

(27) Wu, F., Su, R., Lai, Y., and Wang, X. (2017) Engineering of a synthetic quadrastable gene network to approach Waddington landscape and cell fate determination. *eLife 6*, e23702.

(28) Meinhardt, S., Manley, M. W., Jr., Becker, N. A., Hessman, J. A., Maher, L. J., III, and Swint-Kruse, L. (2012) Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression. *Nucleic Acids Res. 40*, 11139−11154.

(29) Kaplan, S., Bren, A., Zaslaver, A., Dekel, E., and Alon, U. (2008) Diverse Two-Dimensional Input Functions Control Bacterial Sugar Genes. *Mol. Cell 29*, 786−792.

(30) Lee, S. K., Chou, H. H., Pfleger, B. F., Newman, J. D., Yoshikuni, Y., and Keasling, J. D. (2007) Directed evolution of AraC for improved compatibility of arabinose- and lactose-inducible promoters. *Appl. Environ. Microbiol. 73*, 5711−5715.

(31) Hooshangi, S., Thiberge, S., and Weiss, R. (2005) Ultrasensitivity and noise propogation in a synthetic transcriptional cascade. *Proc. Natl. Acad. Sci. U. S. A. 102*, 3581−3586.

(32) Arren, B., Paulsson, J., Maheshri, N., Carmi, M., Erin, O., Pilpel, Y., and Barkai, N. (2006) Noise in protein expression scales with natural protein abundance. *Nat. Genet. 38*, 636−643.

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843

(33) Taniguchi, Y., Choi, P. J., Li, G., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, S. X. (2010) Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science 329*, 533−538.
(34) Siegele, D., and Hu, J. (1997) Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proc. Natl. Acad. Sci. U. S. A. 94*, 8168−8172.
(35) Afroz, T., Biliouris, K., Kaznessis, Y., and Beisel, C. L. (2014) Bacterial sugar utilization gives rise to distinct singlecell behaviours. *Mol. Microbiol. 93*, 1093−1103.
(36) Afroz, T., Biliouris, K., Boykin, K. E., Kaznessis, Y., and Beisel, C. L. (2015) Trade-offs in Engineering Sugar Utilization Pathways for Titratable Control. *ACS Synth. Biol. 4*, 141−149.
(37) Alfred, F., Vine, C. E., Caminal, G., and Josep, L. (2012) Evidencing the role of lactose permease in IPTG uptake by Escherichia coli in fed-batch high cell density cultures. *J. Biotechnol. 157*, 391−398.
(38) Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I., and Oudenaarden, A. v. (2004) Multistability in the lactose utilization network of Escherichia coli. *Nature 427*, 737−740.
(39) Elowitz, M., Levine, A., Siggia, E., and Swain, P. (2002) Stochastic Gene Expression in a Single Cell. *Science 297*, 1183−1186.
(40) Venturelli, O. S., Egbert, R. G., and Arkin, A. P. (2016) Towards Engineering Biological Systems in a Broader Context. *J. Mol. Biol. 428*, 928−944.
(41) Daniel, R., Rubens, J. R., Sarpeshkar, R., and Lu, T. K. (2013) Synthetic analog computation in living cells. *Nature 497*, 619−623.
(42) 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.
(43) Engler, C., Kandzia, R., and Marillonnet, S. (2008) A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS One 3*, e3647.
(44) Castillo-Hair, Sebastian, M., Sexton, J. T., Landry, B. P., Olson, E. J., Igoshin, O. A., and Tabor, J. J. (2016) FlowCal: A User-Friendly, Open Source Software Tool for Automatically Converting Flow Cytometry Data from Arbitrary to Calibrated Units. *ACS Synth. Biol. 5*, 774−780.

DOI: 10.1021/acssynbio.8b00165
ACS Synth. Biol. 2018, 7, 1834−1843