ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

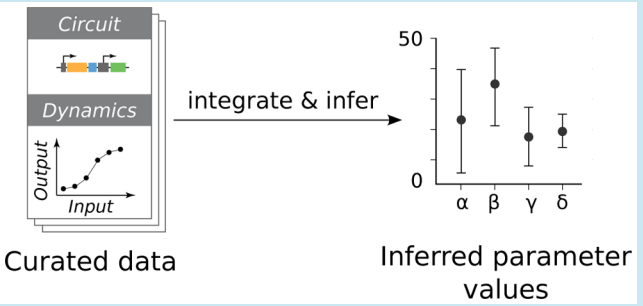
A Parts Database with Consensus Parameter Estimation for Synthetic Circuit Design

Linh Huynh and Ilias Tagkopoulos*

Department of Computer Science & UC Davis Genome Center, University of California Davis, Davis, California 95616 United States

Supporting Information

ABSTRACT: Mathematical modeling and numerical simulation are crucial to support design decisions in synthetic biology. Accurate estimation of parameter values is key, as direct experimental measurements are difficult and time-consuming. Insufficient data, incompatible measurements, and specialized models that lack universal parameters make this task challenging. Here, we have created a database (PAMDB) that integrates data from 135 publications that contain 118 circuits and 165 genetic parts of the bacterium *Escherichia coli*. We used a succinct, universal model formulation to describe the part behavior in each circuit. We introduce a constrained consensus inference method that was used to infer the value of the model parameters and evaluated its performance through cross-validation in a benchmark of 23 circuits. We discuss these results and summarize the challenges in data integration and parameter inference. This work provides a resource and a methodology that can be used as a point of reference for synthetic circuit modeling.



KEYWORDS: *parameter estimation, data integration, parameter inference, gene circuit, mathematical model*

Mathematical modeling and numerical simulation are crucial to support design decisions in synthetic biology.[1−3] For the models to be predictive rather than descriptive, it is important that the key processes are captured in the model and the modeling parameters are accurate enough so that the simulation outcome reflects reality. Currently, there is a plethora of models, most of them built to support experimental work, which are fitted to describe the observed data. Similarly, although some parameters are common, many are not, which makes interoperability difficult or impossible. Given the cost in time and expenses to experimentally measure them, parameter inference and easy access to a parameter repository for parts is important to move the field forward.[4]

Table 1 summarizes the current databases and inventories related to synthetic biology that are relevant for computational modeling. In addition to these efforts, there is a current need for a resource that stores information regarding parts, modules, and circuits together with categorized metadata.[5] Incorporating parameter values in a way that can be imported to a computational model and used for simulation is both useful and desired.[6] Normalization and standarization of the circuit, part, and parameter measurements is a formidable challenge due to the diversity of instruments, techniques, reporting guidelines, and units used.[7,8]

For the parameter inference problem, numerous approaches have been proposed over the years,[9−12] with reviews covering methods from all relevant fields.[13−16] Common approaches include fitting a single system to estimate the model parameters, or fitting parameters on multiple systems sequentially by propagating parameter values inferred at each step.[17−20] However, in most cases this estimation is inaccurate due to the many degrees of freedom, the lack of training data, and the existence of many local optima in the solution space. As an example, consider the system shown in Figure 1. In that example, two different parameter value sets have identical likelihood with respect to explaining the observed experimental data (Figure 1E), despite the fact that their inferred parameter values have an order of magnitude difference. In contrast, when the same parameter value combinations are used to predict the behavior of another circuit, as in Figure 1F, the respective simulations yield substantially different results. Model sloppiness, a phenomenon where a change in one parameter can be compensated by changes in other parameters with the output remaining the same,[21,22] is one of the main reasons behind this behavior. Noise and measurement errors can further reduce the accuracy of inferred parameter values, hence decreasing the predictive performance of the model.

To address these challenges, we resort to a data-driven approach in which available published data from circuits with shared parts are used to further constrain the inference procedure. Figure 2 illustrates our approach. First, we performed literature curation to build a database that contains the architecture and the dynamics (i.e., input−output data set) of each published circuit. Then, we built a model for each circuit, by using a consistent minimal set of universal functions that can

Received: October 19, 2015
Published: July 25, 2016

© 2016 American Chemical Society

DOI: 10.1021/acssynbio.5b00205
ACS Synth. Biol. 2016, 5, 1412−1420

Table 1. A Comparison between Our Database PAMDB and Related Inventories in Synthetic Biology[a]

	system-level experimental data		measured or inferred parameters				
database	available	machine readable format[b]	available	machine readable format[b]	universal model	no. of genes/parts	no. of publications
part registry	yes	no	yes	no	no	>20000	–
JBEI-ICE	yes	no	no	–	no	2302	83
virtual part	no	–	no	–	no	3015	N/A
SEVA	no	–	no	–	no	185	N/A
Addgene	no	–	no	–	no	>45000	N/A
BioNumber	no	–	yes	yes	no	–	>1000
BioModels	no	–	yes	yes	no	–	1483[c]
PAMDB	yes	yes	yes	yes	yes	165	45

[a]These inventories include part registry,[43] JBEI-ICE,[44] virtual part,[45] SEVA,[46] Addgene,[47] BioNumber,[48] BioModels,[49] and PAMDB. System-level experimental data from experiments do not measure any parameter directly (illustrated in Figure S1). [b]Data that can be directly downloaded and serve as an input to modeling tools (not the case for which there is no API, numerical values are mixed with text in a nonstructured format, or data are stored in images). "–" corresponds to "not applicable"; "N/A" corresponds to "not available". [c]Number of curated models.

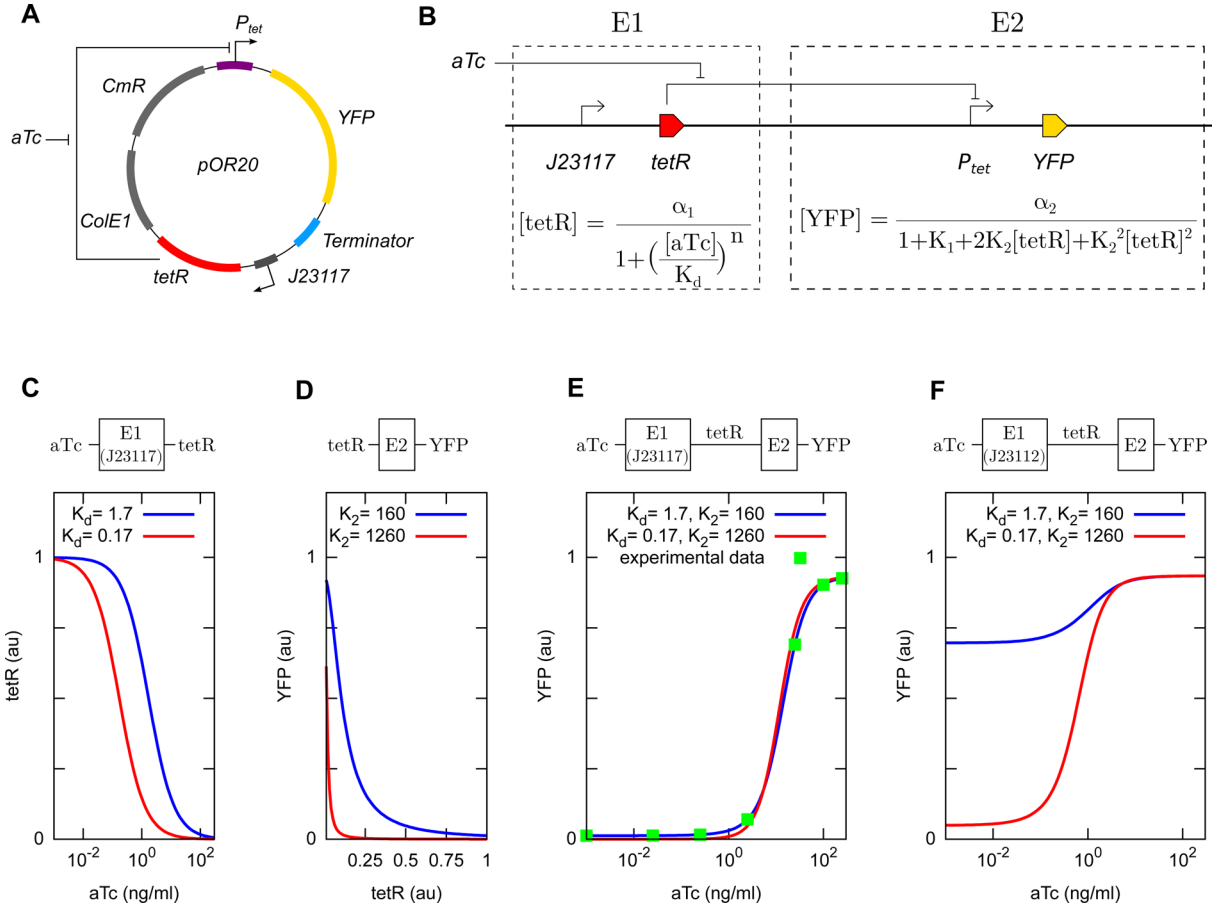


Figure 1. An illustrative example on a sloppy model. (A) The architecture of the circuit pOR20 from ref 20 that was used to characterize the tetR-Ptet system. (B) A model from ref 20 with two equations that capture the dynamics of the tetR-P_{tet} system. (C) The relationship between aTc and tetR (eq E1) with two different values of K_d, with the other parameters having the same value ($\alpha_1 = 1$, $n = 1$) in both cases. (D) Relationship between tetR and YFP (eq E2) with two different values of K_2, with the other parameters having the same value ($\alpha_2 = K_1 = 350$) in both cases. (E) Two different parameter value combinations that can fit with the experimental data from ref 20 although they have different intermediate relationships as depicted in panels C and D. (F) Two parameter value combinations from panel E lead to two different predictions for the circuit pOR20* that is the same as pOR20 with the exception of the constitutive promoter J23117 being replaced by a weaker promoter J23112.

describe regulatory and other biological behavior (see Methods). If a part is used in two or more circuits, then its parameters will appear in their respective models. This overlap of parameters with circuit models and the simultaneous parameter optimization during the training phase leads to additional constraints, which reduces the parameter space and hence the likelihood of overfitting and sloppiness effects.

To make available our curation and parameter inference results, we constructed the Parts and Modules Database (PAMDB), a database for quantitative characterization of parts and modules in synthetic biology. In this first version, PAMDB contains 135 publications, 165 parts, and the characterization data of 118 circuits. We also evaluated our parameter inference approach with a benchmark that combines the data of three

DOI: 10.1021/acssynbio.5b00205
ACS Synth. Biol. 2016, 5, 1412–1420

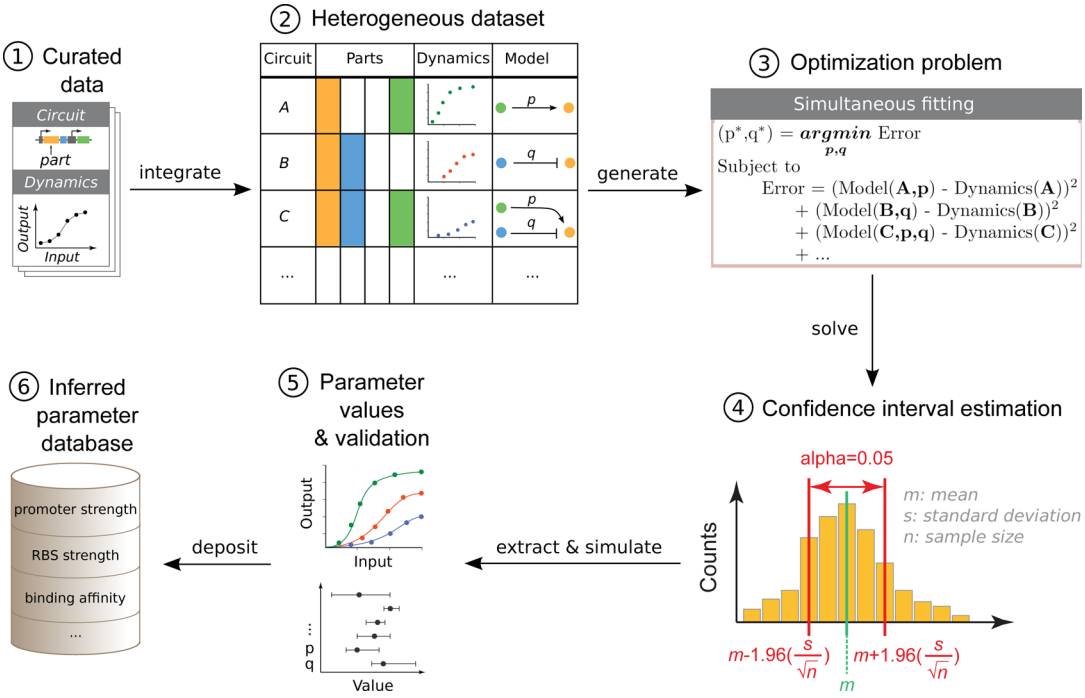


Figure 2. An overview of our proposed approach.

common regulation systems, namely pLAC, pTET, and pBAD. We discuss the challenges of the problem of integrating and mining quantitative data for synthetic biology and propose approaches to address these challenges.

RESULTS

PAMDB: A Quantitative Database for Parts and Modules. *Database Organization.* At the conceptual level, PAMDB contains seven entity classes or tables (Figure 3A). The *publication* class is the central class that organizes information for any given specific publication. Each publication entity contains a list of parts, plasmids, strains, media, and general experiment settings. Each of those elements has its own class in the database, where pertinent information is stored and queried. For instance, the media class contains the supplemental chemical components for each medium label. In addition to the curated data and metadata, PAMDB stores structured information related to model parameters and simulations. An example of a database entry is shown in Figure 3B.

Database Content. We curated 135 publications (see Methods) that contained gene circuits for *E. coli* with complete information about both the sequence and the characterization data. For parameter inference, to avoid introducing a complex model with many not well-constrained parameters, we limited our model to apply for all circuits with steady state characterization and without feed-back loops, protein−protein interactions, RNA−RNA interactions, metabolic interactions, recombination, or cell−cell communication; 45 of the 135 publications contained circuits that adhere to those specifications. In total, the current PAMDB version contains 165 parts, 118 experiments, and 26 strains. These experiments contain 538 data points. For the inferred data, 239 parameter values were included. The content of the database was summarized in Table 2.

Parameter Inference with PAMDB. There are two methodological advances in PAMDB that allow parameter inference to be performed in a cohesive manner. First, we adopted a succinct *universal model* for all circuit components. The model does not capture each process in detail, but it allows model training with small sample sizes. Second, we developed a new parameter inference method that can best be described as *constrained simultaneous fitting* tailored for synthetic gene circuits. In contrast to methods that train a model for each circuit independently (*independent fitting*) or sequentially, by fitting independently each circuit and propagating parameter values to circuits with overlapping parameters (*sequential fitting*), this method solves the optimization problem of fitting all parameters simultaneously (see Methods). To create a benchmark for these methods, we extracted circuits with promoters that were either constitutive or variants of the pLAC, pTET, and pBAD promoters. This resulted to a collection of 35 circuits, 38 genetic parts (promoter, ribosomal binding site, coding region), and 169

Table 2. A Summary of the Content of PAMDB

description	quantity
number of publications	135
with curated data	45
number of plasmids	136
number of parts	165
promoter	38
RBS	42
CDS	34
terminator	8
origin	11
antibiotic resistance	5
composite part	27
number of strains	26
number of experiments	118
curve	94
single data point	24
number of data points	538
number of parameters	239

DOI: 10.1021/acssynbio.5b00205
ACS Synth. Biol. 2016, 5, 1412−1420

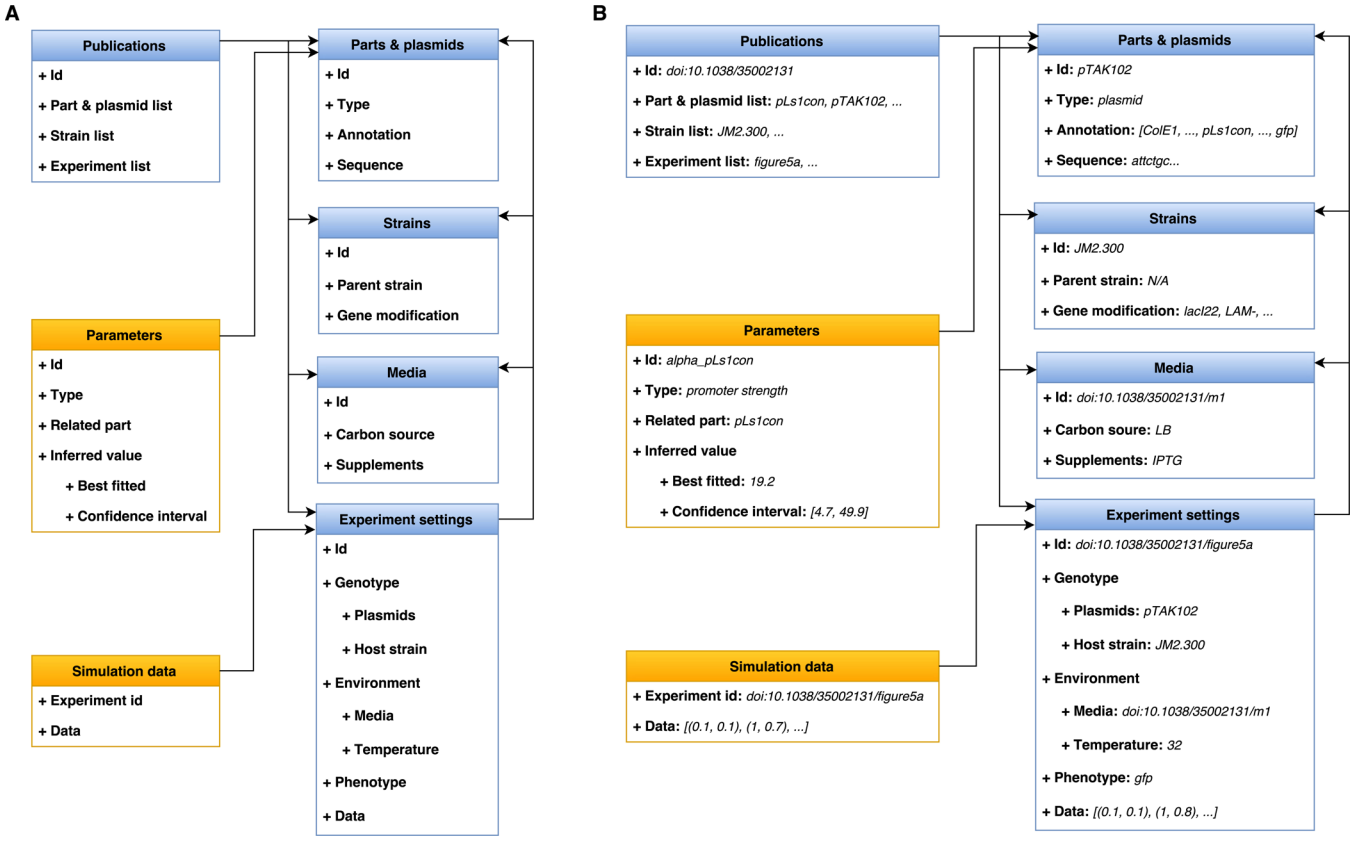


Figure 3. (A) An overview on the design of PAMDB where the blue color represents the curated data and the orange color represents the inferred data. (B) An example of a representative publication.[50]

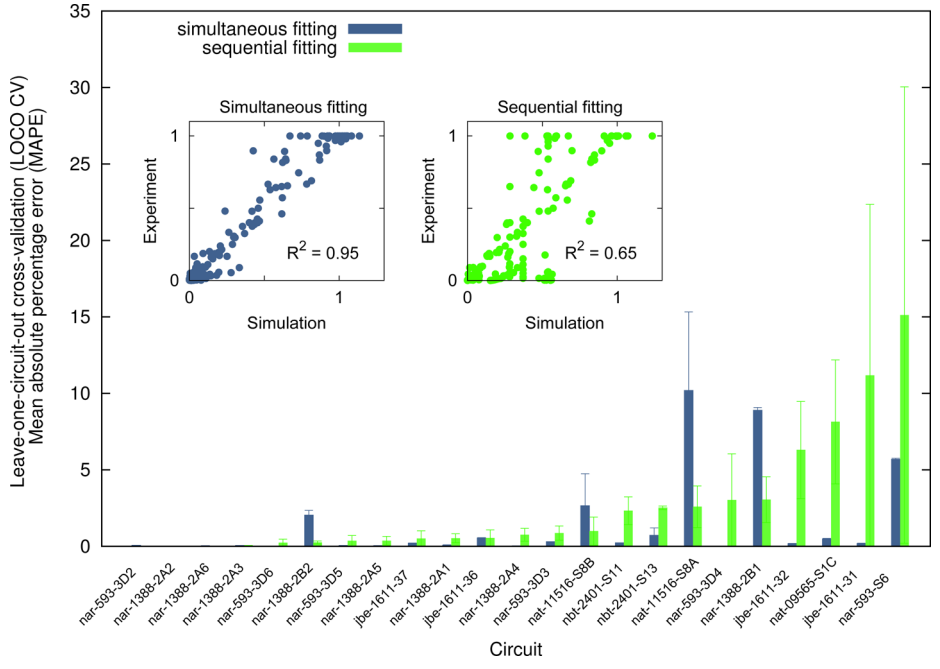


Figure 4. A comparison between the simultaneous fitting approach and the sequential fitting approach on the leave-one-circuit-out cross-validation (LOCO CV) mean absolute percentage error (MAPE) of the 23 circuits in the benchmark. Inset plots: simulated vs experimental data for all 35 circuits. Note that for the inset plots, all data sets were used.

data points. We applied the universal modeling framework with 58 parameters for all circuits and their parts.

Uncertainty Reduction through Data Integration. First, we assess what benefits, if any, are conferred by data integration. To evaluate improvements due to the data integration, we calculated the confidence intervals (CI) of the predicted parameters for all circuits by either with(simultaneous fitting) or without (independent fitting) data integration (Figure S3). From the

DOI: 10.1021/acssynbio.5b00205
ACS Synth. Biol. 2016, 5, 1412−1420

58 parameters, 21 of them (36%) had a confidence interval more than 1 order of magnitude around their value, hence could not be adequately constrained. The largest uncertainty on parameter values was observed in basal promoter activity, followed by TF binding activity, RBS, and promoter strength. In 47 out of the 58 cases measured (81%), the CI of the prediction for the parameter values was reduced by an average of 1 order of magnitude (0.97 ± 0.18; Figure S3). The uncertainty reduction was most profound in the parameters that were associated in promoter strength, which also had the largest coverage in the data set. Additionally, there was a positive correlation between the number of models covering a specific parameter and its uncertainty reduction as measured in its CI (Figure S3 inset).

Simultaneous vs Sequential Fitting. In synthetic biology, sequential fitting is the method of choice.[19,23] We evaluated the sequential and simultaneous fitting on the benchmark data set applying both on the universal model. Strikingly, there is a 46% increase on the correlation coefficient between simulated and actual circuit measurements on all 35 circuits of the benchmark when simultaneous fitting is used (Figure 4, inset plot; $R^2 = 0.95$ and 0.65 for simultaneous and sequential fitting, respectively). We also performed leave-one-circuit-out cross-validation (LOCO CV), where we estimated the parameter values of a single circuit by using the data from all other circuits. The method could only be applied in 23 out of the 35 circuits in our benchmark, since all inferred parameters should be present in the training data set. As shown in Figure 4, simultaneous fitting had a better performance in 17 out of the 23 circuits measured by mean absolute percent error (MAPE). We also used another two metrics, the normalized RMSD and the RMSPE (see Methods), which are all in agreement (Figures S3 and S5).

Performance Impact of the Universal Model. A universal model is less complex, usually more phenomenological than mechanistic and potentially less accurate than a specialized model that is tailored for any given circuit. If experimental measurements allow for the latter, then it is likely that it will perform better. Of course parts are rarely well-characterized and parameter data are scarce, a fact that necessitates the adoption of an universal model if we aspire to perform parameter inference from a collection of circuits. To evaluate its impact on performance, we applied simultaneous fitting with either the universal model or with the specialized models of Moon et al.[19] or Ellis et al.[23], respectively. The comparison on cross-validation results with the MAPE metric (Figure S6) showed that the universal model was slightly better than the Ellis specialized model (16 of 23 cases) and performed as well as the Moon specialized model (each of them outperformed 11 of 23 cases). The results are similar when the normalized RMSD metric is used (Figure S7), with the universal model outperforming the Ellis model (15 out of 23 cases) and being slightly worse than the Moon model (14 out of 23 cases). To summarize, our results show that any decrease in the performance due to using a universal model is either absent or small. As the database grows, re-examining this point with a larger benchmark and more specialized models can provide a baseline for model improvement.

DISCUSSION

Model parameter estimation is a well-known problem in engineering[24,25] and computational biology.[9,13,14] So far, parameter inference has been applied in a circuit-by-circuit analysis which has been more descriptive than predictive. In this work, we propose a method which is based on three critical points: the adoption of a universal, succinct model for describing key circuit dynamics; the integration and mining of all published data sets for circuits with overlapping parameters for restricting the solution space; and the application of simultaneous parameter inference from all available data. Our results show that by doing so, we achieve a reduction in both the parameter estimation uncertainty and the resulting error in circuit behavior prediction.

We are still facing several challenges when it comes to extracting parameter values and training models from published data. In our analysis we struggled with lack of consistent data reporting methodology and data types. In most cases, gene expression was reported in arbitrary units without a reference, which forced us to use a scaling factor for each data set that we estimated together with other model parameters. The assumption here is that arbitrary units are a linear function of relative units and that the scaling factor should have a value that maximizes the likelihood of the fitted model. While these are reasonable assumptions necessary to move forward with data integration, it remains to be seen if they hold. In many cases, the associated metadata were missing, which creates training issues during parameter inference. Even when the data accompanying a publication are consistent, they reflect measurements for a specific environmental setting (strain, medium, etc.). Given that the number of possible settings are infinite and changing even one abiotic factor in the environment can have serious repercussions to gene expression and cellular behavior, the challenge is how to extrapolate from one condition into the other in a way that is predictive. Table 3 summarizes these and other challenges and lists our recommendations on how to move the field forward in this direction.

Given the biological complexity and the degrees of freedom that are present in biological design, it is practically impossible to avoid model sloppiness and overfitting. We can, however, reduce their undesired effects by leveraging data-driven techniques as the one presented in this paper. The synthetic biology community has yet to adopt and follow standards on reporting experimental values and depositing the resulting data sets into the public domain in a coherent, consistent format (although some exceptions exist[7,8,26−28]). In contrast, circuit design standards have been significantly improved over the years, largely because of the efforts by the SBOL consortium.[29] We argue that if we are to realize the potential of data-driven techniques for automated design[30−36] and predictive synthetic biology, we have to develop equivalent standards for data repositories, measurements, methodologies, and controls, so that they can be efficiently mined and used in pertinent computational methods. Although we are far from exploiting "Big Data" techniques in this area, what we propose are steps toward building the foundations that can enable this vision in the future.

METHODS

Data Curation. We limited our curation for novel publications that describe experimentally validated synthetic circuits for *E. coli*. As such, we did not include review papers or papers with only computational results. We started with the top 10 most cited publications that were results from searching with a combination of three keywords "synthetic biology", "gene circuit" and "E coli". Then we extended this set recursively by adding publications that cited (forward) or were cited (backward) by publications that were already in the set. More specifically, at each forward step, we added all publications that cited one or more publications in the current set. Conversely, at

DOI: 10.1021/acssynbio.5b00205
ACS Synth. Biol. 2016, 5, 1412−1420

Table 3. Challenges of Integrating and Mining Quantitative Data in Synthetic Biology

challenge	why it exists	ways to address it
lack of sufficient data and meta-data	Many factors such as the host strain, medium, genetic context play an important role on part dynamics and circuit behavior. Yet, we lack characterization data for different combinations, which makes prediction challenging	Create a "report card" for each part to standarize its behavior under common conditions. Apply transfer learning methods, so data from one environment can be translated in another.
lack of a standard reporting format	Circuit characterization data are often reported in different formats (figures, tables) that need additional processing	All characterization data should be reported in a standard format that will be machine-readable and easy to integrate.
lack of an integrated database	Characterization data are only available from the original publications. There is no database (Table1) that combines and normalizes these data sources together	There should be a community effort to support an integrated database so that each publication will be accompanied by a concomitant submission of its data set.
incompatible data types	Usual measurement data are reported with arbitrary units and thus it is difficult to integrate them together to infer useful information	Relative units (such as RPU or REU) should be used. Absolute unit measurement (molecules) together with controls would provide the most information, wherever possible.
lack of universal models	Multiple models exist to describe the same biological processes. The lack of standarized parameters for each part/module creates a challenge for parameter inference.	Universal models should be used to infer parameter values and simulate the circuit behavior. Specialized models can be used in conjunction for additional accuracy.
computational complexity and local optimality	When more data are integrated, the complexity of the inference problem increases exponentially. Heuristics lead to many local optimal solutions.	Adopt a hybrid approach that balances the computational cost and optimization result.

each backward step, we added all publications that were cited by one or more publications in the current set. By repeating forward steps and backward steps alternatively up to four times, we curated a total of 135 publications that published experimental circuit characterization for *E. coli*. Since we limited the model to the one without feed-back loops, protein−protein interactions, RNA−RNA interactions, metabolic interactions, recombination or cell−cell communication, we only manually curated the characterized data of 45 publications.

Data Integration. Characterization data are usually reported with arbitrary units, so it is important to normalize their values across the entire data set. In the case of fluorescent measurements, we chose as our standard the relative expression unit (REU[26]). Ideally, the conversion of an arbitrary unit to REU is performed through a reference system (a constitutive promoter, e.g. J23101, and a strong RBS, e.g. B0032 from the part registry). However, in most publications (111 out of 118) this information was not available. For this reason, we introduced a scaling parameter, or "scale factor", ω that linearly translates REUs to arbitrary units (see Methods). The scaling factor was estimated simultaneously with all other model parameters.

Database Design and Implementation. At the physical level, PAMDB was organized as a document-based database (MongoDB) in which each curated publication was represented by a pair of documents (Figure S2), one that contained curated data and another one that contained all the inferred data. The organization of the database and a summary of its contents were shown in Figure 3.

Model. To avoid adding ill-constrained parameters, we used a simple model based on Hill and linear functions. This model captured both the processes of transcription, translation and ligand-protein binding. More specifically, when a transcription factor TF is bound to a promoter Pr, the expression level ν_g of a gene g at the downstream of Pr is modeled by

$$\nu_g = \begin{cases} N_g\left(\beta_{Pr} + \dfrac{\alpha_{Pr} - \beta_{Pr}}{1 + \left(\dfrac{K_{Pr}}{\mu'_{TF}}\right)^{n_{Pr}}}\right) & \text{TF is an activator} \\ N_g\left(\beta_{Pr} + \dfrac{\alpha_{Pr} - \beta_{Pr}}{1 + \left(\dfrac{\mu'_{TF}}{K_{Pr}}\right)^{n_{Pr}}}\right) & \text{TF is a repressor} \end{cases} \quad (1)$$

where N_g is the copy number of gene g (N_g is the plasmid copy number if g is on a plasmid, otherwise, $N_g = 1$ if g is in the chromosome). Parameters β_{Pr}, α_{Pr}, K_{Pr}, and n_{Pr} represent the basal level, the promoter strength, the binding affinity and its cooperativity, respectively, with respect to promoter Pr and its transcription factor TF. Here, μ'_{TF} is the concentration of transcription factor TF and is modeled as

$$\mu'_{TF} = \begin{cases} \mu_{TF} & \text{TF is non-inducible} \\ \dfrac{\mu_{TF}}{1 + \left(\dfrac{[L_{TF}]}{K_{L_{TF}}}\right)^{n_{L_{TF}}}} & \text{TF binds with } L_{TF} \text{ and TF binds to Pr} \\ \dfrac{\mu_{TF}}{1 + \left(\dfrac{K_{L_{TF}}}{[L_{TF}]}\right)^{n_{L_{TF}}}} & \text{TF binds with } L_{TF} \text{ and } L_{TF} - \text{TF binds to Pr} \end{cases} \quad (2)$$

DOI: 10.1021/acssynbio.5b00205
ACS Synth. Biol. 2016, 5, 1412−1420

where μ_{TF} is the protein expression level (from eq 3 below). $[L_{TF}]$is the ligand concentration when a ligand binds to the TF. Parameters $K_{L_{TF}}$ and $n_{L_{TF}}$ correspond to the dissociation constant and the Hill coefficient of the ligand, respectively. For the translation, the protein expression level μ_g of a gene g is modeled by

$$\mu_g = \alpha_r \nu_g \tag{3}$$

where α_r is the strength of the ribosome binding site r in the upstream of g. In the case where g is a reporter protein (e.g., gfp, yfp), the protein expression level μ_g in relative units is converted to the protein expression level μ_g^{au} in arbitrary units to fit with the experimental data. This is achieved by applying the following conversion formula:

$$\mu_g^{au} = \omega_g^{ref} \mu_g \tag{4}$$

where ω_g^{ref} is the scale factor for the reporter protein g as presented in publication ref. The value of this scale factor will be estimated simultaneously with the value of other parameters to fit the experimental data with the model through the parameter inference process. The units of all parameters are summarized in Table 4.

Table 4. Description and Units for Model Parameters

parameter	description	unit
α_r	RBS strength	relative RBS unit (RRU), normalized by BBa_B0032
α_{Pr}, β_{Pr}	promoter strength and basal level	relative promoter unit (RPU), normalized by BBa_J23101[7]
ν_g	gene expression level of g	RPU
μ_g, μ_{TF}	protein expression level of g, TF	REU (= RRU × RPU)
ω_g^{ref}	scale factor of g from publication ref	au/REU
K_{Pr}	binding affinity	REU
$[L_{TF}]$, $K_{L_{TF}}$	ligand concentration and dissociation constant	mM
$n_{L_{TF}}$, n_{Pr}, N_g	Hill coefficient, cooperativity and copy number	N/A

Parameter Estimation. Suppose that we need to fit n circuits, each circuit C_i is modeled by

$$y_i = M_i(x_i, \theta_i) \quad i = 1, \ldots, n \tag{5}$$

where x_i, y_i, θ_i represent the input, the output, and the set of model parameters, respectively, all for circuit C_i. Each parameter can appear in more than a single model, so we denote the set of all parameters as

$$\theta = \bigcup_{i=1}^{n} \theta_i \tag{6}$$

Let $D_i = \{(x_i^{(1)}, y_i^{(1)}, \sigma_i^{(1)}), \ldots, (x_i^{(d_i)}, y_i^{(d_i)}, \sigma_i^{(d_i)})\}$ be an experimental characterization data set with d_i data points of the circuit C_i and $\sigma_i^{(j)}$ capturing the standard deviation of the output $y_i^{(j)}$. If we assume that all data points are independent and the output value y_i of circuit C_i has a Gaussian distribution then the log-likelihood[37] is given by

$$\mathrm{LL}(D_i|\theta_i) = -\frac{1}{2}\sum_{j=1}^{d_i}\left(\frac{M_i(x_i^{(j)}, \theta_i) - y_i^{(j)}}{\sigma_i^{(j)}}\right)^2 + \text{const} \tag{7}$$

We evaluated three parameter inference methods:

(i) Independent fitting: The value θ_i^* of parameters of each circuit C_i was fitted independently by maximum log-likelihood estimation:

$$\theta_i^* = \underset{\theta_i}{\operatorname{argmax}}\ \mathrm{LL}(D_i|\theta_i) \tag{8}$$

(ii) Sequential fitting: Circuits were fitted one by one; the best-fitted solution of former circuits were propagated to fit latter circuits. This sequential fitting method is common in practice in which we fix some model parameters with values inferred in past work and estimate the value of other parameters by fitting the model with experimental data.

(iii) Simultaneous fitting: The value θ^* of all parameters was fitted at the same time by solving the optimization problem

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ \sum_{i=1}^{n} \mathrm{LL}(D_i|\theta_i) \tag{9}$$

To solve the optimization problem, we used the trust region method[38] with multiple starting values. This method has been shown to outperform others in reliability and efficiency.[39] To improve computational performance, we also converted all circuit models (represented by graphs as in ref 40) to computational functions (represented by inline code), to avoid repeating this conversion during the optimization phase. To estimate the parameter value uncertainty, we calculate the confidence interval, which is based on the bootstrapping method.[41] As such, the 95% confidence interval of a parameter $p \in \theta$ is an interval $CI(p)$, and the confidence interval length (log scale) in the comparison (Figure S3) was calculated as

$$\log_{10}\left(\frac{\max(CI(p))}{\min(CI(p))}\right) \tag{10}$$

The simulation (Figure 4 inset) and prediction (Figures 4 and S4−S7) were calculated from the best-fitted parameter values of both approaches since we estimated the deviation between the best prediction of each approach and the experimental data.

For the cross validation, we needed to compare the prediction error for different data sets, which can vary widely on their values. Therefore, we used three different normalization error metrics. Let n be the number of data points of a circuit and s_i and d_i be the simulated and the desired value at each data point, respectively. The prediction error of each circuit was calculated with the following three metrics:

(i) Mean absolute percentage error (MAPE[42])

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|s_i - d_i|}{d_i} \tag{11}$$

(ii) Root mean square percentage error (RMSPE[42])

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{s_i - d_i}{d_i}\right)^2} \tag{12}$$

(iii) These above metrics may be not applicable when $d_i = 0$ or when d_i was very small. Therefore, we normalized the root-mean-square-deviation by the mean of experimental data.

DOI: 10.1021/acssynbio.5b00205
ACS Synth. Biol. 2016, 5, 1412−1420

We called the new metric normalized root-mean-square-deviation (normalized RMSD) that was calculated as

$$\frac{\sqrt{(\sum_{i=1}^{n}(s_i - d_i)^2)/n}}{(\sum_{i=1}^{n} d_i)/n} \quad (13)$$

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.5b00205. The PAMDB database is available at http://www.pamDB.com.

Physical design of PAMDB; graphical comparisons among the different approaches studied (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: itagkopoulos@ucdavis.edu.

Author Contributions

L.H. wrote the code and performed the experiments. I.T. conceived the project and supervised all development and analysis. L.H. and I.T. analyzed the data and wrote the paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge support from the NSF CAREER Grant No. 1254205 to I.T.

■ REFERENCES

(1) Church, G. M., Elowitz, M. B., Smolke, C. D., Voigt, C. A., and Weiss, R. (2014) Realizing the potential of synthetic biology. *Nat. Rev. Mol. Cell Biol. 15*, 289−294.

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

(3) Slusarczyk, A. L., Lin, A., and Weiss, R. (2012) Foundations for the design and implementation of synthetic genetic circuits. *Nat. Rev. Genet. 13*, 406−420.

(4) Andrianantoandro, E., Basu, S., Karig, D. K., and Weiss, R. (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol. 2*, 0028.

(5) Misirli, G., Hallinan, J., Pocock, M., Lord, P., McLaughlin, J. A., Sauro, H., and Wipat, A. (2016) Data Integration and Mining for Synthetic Biology Design. *ACS Synth. Biol.*, DOI: 10.1021/acssynbio.5b00295.

(6) Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., Algaa, E., Weidemann, A., Sauer-Danzwith, H., Mir, S., et al. (2012) SABIO-RK-database for biochemical reaction kinetics. *Nucleic Acids Res. 40*, D790−D796.

(7) Kelly, J. R., Rubin, A. J., Davis, J. H., Ajo-Franklin, C. M., Cumbers, J., Czar, M. J., de Mora, K., Glieberman, A. L., Monie, D. D., and Endy, D. (2009) Measuring the activity of BioBrick promoters using an in vivo reference standard. *J. Biol. Eng. 3*, 4.

(8) Davidsohn, N., Beal, J., Kiani, S., Adler, A., Yaman, F., Li, Y., Xie, Z., and Weiss, R. (2015) Accurate predictions of genetic circuit behavior from part characterization and modular composition. *ACS Synth. Biol. 4*, 673−681.

(9) Lillacci, G., and Khammash, M. (2010) Parameter estimation and model selection in computational biology. *PLoS Comput. Biol. 6*, e1000696.

(10) Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc., Interface 6*, 187−202.

(11) Chen, W. W., Schoeberl, B., Jasper, P. J., Niepel, M., Nielsen, U. B., Lauffenburger, D. A., and Sorger, P. K. (2009) Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol. 5*, 239.

(12) Schaber, J., Baltanas, R., Bush, A., Klipp, E., and Colman-Lerner, A. (2012) Modelling reveals novel roles of two parallel signalling pathways and homeostatic feedbacks in yeast. *Mol. Syst. Biol. 8*, 622.

(13) Moles, C. G., Mendes, P., and Banga, J. R. (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res. 13*, 2467−2474.

(14) Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A., and Blom, J. G. (2009) Systems biology: parameter estimation for biochemical models. *FEBS J. 276*, 886−902.

(15) Chou, I.-C., and Voit, E. O. (2009) Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math. Biosci. 219*, 57−83.

(16) Sun, J., Garibaldi, J. M., and Hodgman, C. (2012) Parameter estimation using metaheuristics in systems biology: a comprehensive review. *IEEE/ACM Trans. Comput. Biol. Bioinf. 9*, 185−202.

(17) Mishra, D., Rivera, P. M., Lin, A., Del Vecchio, D., and Weiss, R. (2014) A load driver device for engineering modularity in biological networks. *Nat. Biotechnol. 32*, 1268−1275.

(18) Daniel, R., Rubens, J. R., Sarpeshkar, R., and Lu, T. K. (2013) Synthetic analog computation in living cells. *Nature 497*, 619−623.

(19) Moon, T. S., Lou, C., Tamsir, A., Stanton, B. C., and Voigt, C. A. (2012) Genetic programs constructed from layered logic gates in single cells. *Nature 491*, 249−253.

(20) 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.

(21) Brown, K. S., and Sethna, J. P. (2003) Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top. 68*, 021904.

(22) Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol. 3*, e189.

(23) Ellis, T., Wang, X., and Collins, J. J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol. 27*, 465−471.

(24) Beck, J. V., and Arnold, K. J. (1977) *Parameter estimation in engineering and science*, James Beck.

(25) Tarantola, A. (2005) *Inverse problem theory and methods for model parameter estimation*, Siam.

(26) Temme, K., Zhao, D., and Voigt, C. A. (2012) Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc. Natl. Acad. Sci. U. S. A. 109*, 7085−7090.

(27) Sainz de Murieta, I., Bultelle, M., and Kitney, R. I. (2016) Towards the first data acquisition standard in Synthetic Biology. *ACS Synth. Biol.*, DOI: 10.1021/acssynbio.5b00222.

(28) Wilson, E. H., Sagawa, S., Weis, J. W., Schubert, M. G., Bissell, M., Hawthorne, B., Reeves, C. D., Dean, J., and Platt, D. (2016) Genotype Specification Language. *ACS Synth. Biol. 5*, 471.

(29) Galdzicki, M., et al. (2014) The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nat. Biotechnol. 32*, 545−550.

(30) Myers, C. J., Barker, N., Jones, K., Kuwahara, H., Madsen, C., and Nguyen, N.-P. D. (2009) iBioSim: a tool for the analysis and design of genetic circuits. *Bioinformatics 25*, 2848−2849.

(31) Pedersen, M., and Phillips, A. (2009) Towards programming languages for genetic engineering of living cells. *J. R. Soc., Interface 6*, S437−S450.

(32) Chandran, D., Bergmann, F. T., and Sauro, H. M. (2009) TinkerCell: modular CAD tool for synthetic biology. *J. Biol. Eng. 3*, 19.

(33) Beal, J., Weiss, R., Densmore, D., Adler, A., Appleton, E., Babb, J., Bhatia, S., Davidsohn, N., Haddock, T., Loyall, J., Schantz, R., Vasilev, V., and Yaman, F. (2012) An End-to-End Workflow for Engineering of Biological Networks from High-Level Specifications. *ACS Synth. Biol. 1*, 317−331.

(34) Huynh, L., Kececioglu, J., Köppe, M., and Tagkopoulos, I. (2012) Automatic design of synthetic gene circuits through mixed integer non-linear programming. *PLoS One 7*, e35529.
(35) Huynh, L., Tsoukalas, A., Köppe, M., and Tagkopoulos, I. (2013) SBROME: A scalable optimization and module matching framework for automated biosystems design. *ACS Synth. Biol. 2*, 263−273.
(36) Huynh, L., and Tagkopoulos, I. (2015) Fast and accurate circuit design automation through hierarchical model switching. *ACS Synth. Biol. 4*, 890−897.
(37) Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics 25*, 1923−1929.
(38) Coleman, T. F., and Li, Y. (1996) An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim. 6*, 418−445.
(39) Raue, A., Steiert, B., Schelker, M., Kreutz, C., Maiwald, T., Hass, H., Vanlier, J., Tönsing, C., Adlung, L., and Engesser, R. (2015) Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, btv405.
(40) Huynh, L., and Tagkopoulos, I. (2014) Optimal Part and Module Selection for Synthetic Gene Circuit Design Automation. *ACS Synth. Biol. 3*, 556−564.
(41) Efron, B., and Tibshirani, R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci. 1*, 54−75.
(42) Hyndman, R. J., and Koehler, A. B. (2006) Another look at measures of forecast accuracy. *Int. J. Forecast. 22*, 679−688.
(43) Part registry. http://parts.igem.org/ (accessed 05/26/2016).
(44) Ham, T. S., Dmytriv, Z., Plahar, H., Chen, J., Hillson, N. J., and Keasling, J. D. (2012) Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. *Nucleic Acids Res. 40*, e141.
(45) Cooling, M. T., Rouilly, V., Misirli, G., Lawson, J., Yu, T., Hallinan, J., and Wipat, A. (2010) Standard virtual biological parts: a repository of modular modeling components for synthetic biology. *Bioinformatics 26*, 925−931.
(46) Martínez-García, E., Aparicio, T., Goñi-Moreno, A., Fraile, S., and de Lorenzo, V. (2015) SEVA 2.0: an update of the Standard European Vector Architecture for de-/re-construction of bacterial functionalities. *Nucleic Acids Res. 43*, D1183.
(47) Addgene. https://www.addgene.org/ (accessed: 05/26/2016).
(48) Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010) BioNumbers-the database of key numbers in molecular and cell biology. *Nucleic Acids Res. 38*, D750−D753.
(49) Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., and Shapiro, B. (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res. 34*, D689−D691.
(50) Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000) Construction of a genetic toggle switch in Escherichia coli. *Nature 403*, 339−342.

DOI: 10.1021/acssynbio.5b00205
ACS Synth. Biol. 2016, 5, 1412−1420