Predictions of Genetic Circuit Behaviors Based on Modular Composition in Transiently Transfected Mammalian Cells

Junmin Wang
Graduate Program in
Bioinformatics
Boston University
Boston, Massachusetts 02215
Email: dawang@bu.edu

Samuel A. Isaacson
Department of Mathematics
and Statistics
Boston University
Boston, Massachusetts USA 02215
Email: isaacsas@bu.edu

Calin Belta, *Fellow*, *IEEE*
Graduate Program in
Bioinformatics
Boston University
Boston, Massachusetts 02215
Email: cbelta@bu.edu

***Abstract*—Transient transfection of cells can be highly stochastic, resulting in large variations in plasmid counts across a population. The resulting dynamics of the cells can then also be highly variable, so predicting the behaviors of transfected circuits can be a major challenge. In this work, we provide a precise definition of genetic modules, from which we then develop a method of composition that allows model-based design of circuits in transiently transfected mammalian cells. For validation, we apply our method to cascades consisting of two regulatory switches. Predictions of the mathematical models compare well with the experimental data. Our findings suggest reducing batch effects and selecting a proper model both contribute to improving model predictions.**

I. Introduction

Experimental approaches combined with modeling are an increasingly popular strategy taken by the research community to study synthetic biology. Models can be used to simulate the temporal behaviors of circuits, analyze critical features of circuits such as bi-stability [1], [2] as well as guide circuit construction [3], [4].

One challenge of synthetic biology is the problem of predicting the behaviors of genetic circuits based on the behaviors of modules [5]–[9]. For a given circuit topology, a large variety of transcriptional factors (TF) can be chosen from to compose the circuit, resulting in a combinatorial explosion of circuits that can be built. Building and testing all possible circuit designs directly via experimental approaches is infeasible, especially since the numbers of successfully constructed promoters, switch genes, terminators, etc. are rising rapidly [10]. On the other hand, building and simulating predictive models for circuits can often be completed within a reasonable time thanks to today's computational power. Hence, there is a growing demand for a dependable tool that can facilitate the prediction of circuit behaviors from modules.

In this work, we present a novel method of composition that enables forward design of complex circuits in transiently transfected mammalian cells (TTMC). Transient transfection is a widely adopted technique for delivering foreign genetic materials into cells. The transfected genetic materials utilize the cells' innate transcriptional and translational machinery to get expressed. As the name suggests, transiently transfected genes are only expressed temporarily and do not become integrated into the host's genome. Compared with stable transfection, transient transfection has many benefits including a simple procedure and minimal side-effects to host cells [11]. It is increasingly being explored in mammalian cells because many biomedical related proteins only become biologically active in mammalian cells [12], [13].

Hill-function-based models have long been used to describe gene regulation and study the time evolution of gene expression profiles in synthetic biology [14]. Unfortunately, the same protocol for modeling is less successful in the context of TTMC. In TTMC, plasmid copy number varies significantly across the population (Fig. 2b). Binning cells by plasmid copy number is a standard method to analyze experimental data. Wang et al. construct a bin-dependent model that is compatible with the process of binning and describes the experimental data more accurately than the Hill-function-based model [15]. In this work, our goal is to develop a method of modular composition that can be applied to the bin-dependent model. In the rest of the paper, we will provide a precise definition of a module, present our method of modular composition, and validate our method via experimental data. Notation-wise, Roman text is used for species and italicized text for concentrations.

II. Modules and Models

A. Module

A transcriptional regulatory module is defined as a switch gene and the promoter it regulates (Fig. 1a). The input of the module is the switch gene, and the output, the regulated promoter. The strength of the regulated promoter is often

This work was supported by the NSF under CPS Frontier grant CNS-1446607. SAI was partially supported by National Science Foundation award DMS-1255408. We thank Brian Teague at Massachusetts Institute of Technology for providing us the experimental data.

978-1-5386-6709-5/18/$31.00 ©2018 IEEE

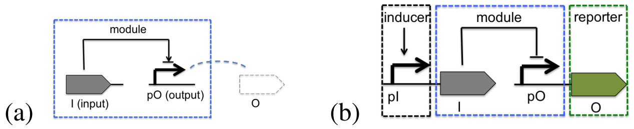

Fig. 1. (a) Graphical representation of a genetic module. The input of a module is the transcriptional factor I, while its output is the regulated promoter pO. O is the protein that is expressed by pO. (b) Graphical representation of a regulatory switch. The green dotted box stands for the reporter. The black dotted box stands for the promoter regulated by an external inducer.

indicated by the expression level of the downstream gene. The promoter can be regulated either positively or negatively, depending on whether the regulator is an activator or an inhibitor. Mathematically, a module M is expressed as: $\mathrm{M} = \{\mathrm{I}, \mathrm{pO}\}$, where I and pO stand for the TF and the promoter, respectively. We assume I is an inhibitor, but similar results can be derived if I is an activator.

The definition we choose is widely used in the community [3], [16] and has a distinct advantage. Another definition of a module in the community is a transcriptional unit, i.e., the coding sequence for a gene along with the sequences necessary for its transcription [17]. In comparison, the definition we choose captures the interaction between a TF and a promoter. It maps a module to a transcriptional regulatory model, whose parameters can be directly inferred from experimental data. Based on this definition, models for modules contain all the information needed to quantify signal propagation in a circuit.

B. Reporter

In a circuit, some proteins do not carry regulatory functions. One such example are proteins used as markers for the states of the cells, e.g., fluorescent proteins, antibodies, etc. We refer to these proteins as reporters (Fig. 1b).

C. External Inducer

Besides modules and reporters, a circuit often contains promoters regulated by external inducers (Fig. 1b). The connection of TF to these promoters makes it possible to control circuit behaviors via external inducers.

D. Model

Defining a module allows us to establish a framework for the composition of mathematical models. The simplest circuit containing at least one module, one reporter, and one external-inducer-regulated promoter is a regulatory switch (Fig. 2a). I and O can be measured via direct expression or co-expression of a fluorescent gene. In TTMC, expression levels are largely determined by the number of plasmids transfected in individual cells [16], which cannot be controlled and are extremely variant across a population. Hence, there is a need to estimate the plasmid counts, which can be achieved by co-transfecting another constitutive fluorescent protein, known as the transfection marker (TM) (Fig. 2). Data of I and O are then binned by the TM so that subpopulations of cells with similar plasmid counts can be compared.

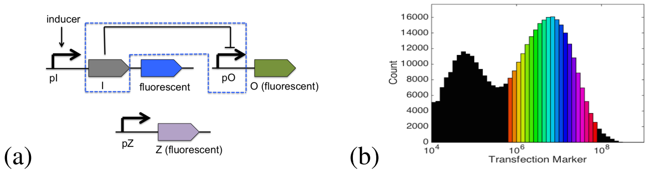

Fig. 2. Characterization of a regulatory switch. (a) Z, the TM, is used to estimate plasmid copy number in cells. (b) Distribution of the TM. The black bins are ignored because they represent untransfected cells (we plot this figure using data from [16]).

Davidsohn et al. develop a Hill-function-based ODE model which describes the time evolution of the average concentration of the input protein I and the downstream protein O [16]:

$$\begin{aligned}
\frac{dI_i}{dt} &= \alpha_i \cdot \phi(t) - \lambda_I \cdot I_i \\
\frac{dO_i}{dt} &= \beta \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^f \cdot \left((1-\gamma) \cdot \frac{1}{1+\left(\frac{I_i}{d}\right)^h} + \gamma\right) - \lambda_O \cdot O_i \\
\phi(t) &= \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor}
\end{aligned} \tag{1}$$

In (1), i represents the i-th plasmid count bin. I_i and O_i are the average concentrations of the induced and the regulated proteins in the i-th bin. α_i is the production rate of the induced protein. α_i is assumed time-invariant because I is assumably induced by a constant concentration of inducer. $\phi(t)$ captures that the population-average plasmid counts decrease over time as with transient transfection. T, the length of the cell cycle, is measured to be approximately 20 hours [16]. λ_I and λ_O are dilution rates. β is the maximal production rate of the regulated protein for cells in the 1st bin, i.e., cells with minimal plasmid counts, P_1. P_i is the mid-point of the i-th plasmid count bin. f maps the ratios of the concentrations of TMs to the ratios of plasmid counts [16]. Transfected genes in TTMC are not expressed until plasmids enter the nucleus during mitosis. It has been estimated that the average delay in the initiation of expression is 25 hours across a population of cells [16]. While in this work we are focused on the average behavior of the population, more sophisiticated models could be developed to accout for stochasticity in the initiaion of gene expression

A fundamental assumption of the model above is that the log of the maximal production rate of the regulated protein is a linear function of the log of the TM. However, this assumption does not hold at high plasmid copy number as protein production rates may slow down due to enzyme saturation [15]. In [15], Wang et al. show why this assumption may be violated in TTMC and develop an alternative bin-dependent model [15]:

$$\begin{aligned}
\frac{dI_i}{dt} &= \alpha_i \cdot \phi(t) - \lambda_I \cdot I_i \\
\frac{dO_i}{dt} &= \begin{cases} \beta \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^f \cdot \left(\frac{1-\gamma}{1+\left(\frac{I_i}{d}\right)^h} + \gamma\right) - \lambda_O \cdot O_i, & \text{if } i <= i^* \\ \beta \cdot \phi(t) \cdot \left(\frac{P_{i'}}{P_1}\right)^f \cdot \left(\frac{P_i}{P_{i'}}\right)^g \cdot \frac{1-\gamma}{1+\left(\frac{I_i}{d}\right)^h} & \\ +\beta \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^f \cdot \gamma - \lambda_O \cdot O_i, & \text{if } i > i^* \end{cases}
\end{aligned} \tag{2}$$

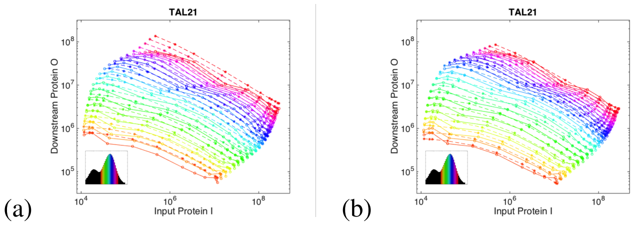

Fig. 3. Comparison between the experimental data and (a) the Hill-function-based models (b) the bin-dependent models fit to the data. Plasmid copy number is shown by color. Solid curves are experimental data, and dashed curves are model fits. Each dot on a curve represents one inducer level. Both axes are log scaled. The best-fit values of the parameters are (a): $\beta = 6.96 \times 10^4$ MEFL/hr, $f = 1.28$, $d = 2.13 \times 10^5$ MEFL, $h = 0.68$, $\gamma = 1.91 \times 10^{-5}$; (b): $\beta = 4.68 \times 10^4$ MEFL/hr, $f = 1.58$, $d = 2.90 \times 10^5$ MEFL, $h = 0.72$, $\gamma = 1.10 \times 10^{-3}$, $g = 0.83$. We plot both (a) and (b) using data from [16].

In (2), i^* is the bin that separates high plasmid copy number from the rest. For high plasmid copy number, the log of the plasmid copy number is again approximated as a linear function of the log of the TM, but with a different slope. The rest of the notations are the same as in (1).

Example 1: We fit the Hill-function-based model (1) and the bin-dependent model (2) to the TAL14, TAL21, and LmrA datasets from [16] using the method of least squares (TAL14, TAL21, and LmrA are names of the repressors in the regulatory switches shown in Fig. 2a). Each switch is induced at twelve dosages. For each dosage, I and O are measured via a flow cytometer 72 hr post-transfection. These data are segmented by the concentrations of the TM into bins of width 0.1 on a log scale. The geometric means of I and O are then calculated within each bin. The fit model values versus the experimental values of the geometric means are shown in Fig. 3. Compared to the Hill-function-based model, the bin-dependent model fits the data well at all plasmid copy numbers.

III. Circuits and Models

A. Modular Connection

Within a set of modules, two are connected if the promoter of one module expresses the TF of the other. Mathematically, the connection between modules M and M* can be represented by a tuple $(\mathrm{M}, \mathrm{M}^*)$, where $\mathrm{M} = \{\mathrm{I}, \mathrm{pO}\}$, $\mathrm{M}^* = \{\mathrm{I}^*, \mathrm{pO}^*\}$, and pO expresses I^*.

Similarly, the connection between a module M and a reporter R can be represented by a tuple (M, R), where $\mathrm{M} = \{\mathrm{I}, \mathrm{pO}\}$, and pO expresses R. The connection between an external-inducer-regulated promoter E and a module M can be represented by a tuple (E, M), where $\mathrm{M} = \{\mathrm{I}, \mathrm{pO}\}$, and E expresses I. The connection between E and R can be represented by (E, R), where E expresses R.

B. Composition of models

Based on the models for modules, we can develop models for general circuit topologies in which each promoter is either constitutively expressed or regulated by one and only one unique TF. We name the circuit to be built the target circuit. Assume the target circuit consists of m modules and n external-inducer-regulated promoters. Let $\{\mathrm{pO}_k\}_{k=1}^m$ denote the set of regulated promoters in the target circuit. Because each promoter is regulated by one unique TF, we know for all $k = 1, 2, ..., m$, there exists a unique gene, also known as the input of the module I_k such that I_k regulates pO_k. Similarly, because the strength of the promoter is indicated by the expression level of the downstream gene, we know for all $k = 1, 2, ..., m$, there exists a unique downstream gene O_k such that expression of O_k initiates at pO_k. It is worth mentioning that through the composition of modules, some TF may be regulated by others, i.e., $\{\mathrm{I}_k\}_{k=1}^m \cap \{\mathrm{O}_k\}_{k=1}^m \neq \emptyset$.

The model for the target circuit is a collection of the models for all modules and external-inducer-regulated promoters. However, models of different modules cannot be directly connected into a chain. Like most biological data, flow cytometry measurements are subject to noise. This noise may originate from imperfect experimental conditions as well as data calibration. In order to make accurate quantitative predictions of circuit behaviors, we need to reduce batch effects by bringing different batches to the same scale, based on the approach taken in [16]. The scaling factors among batches can be calculated by comparing the means and the tightness of the data of different batches (details can found in [16]). Once the scaling factors are calculated, we compensate the batch effects by dividing the parameters by the scaling factors. Mathematically, we can develop a bin-dependent model based on (2) for the k-th module for each $k = 1, 2, ..., m$:

$$\frac{dO'_{ki}}{dt} = \begin{cases} \beta'_k \cdot \phi(t) \cdot \left(\frac{P'_i}{P'_1}\right)^{f_k} \cdot \left(\frac{1-\gamma_k}{1+\left(\frac{I'_{ki}}{d'_k}\right)^{h_k}} + \gamma_k\right) \\ \qquad - \lambda_{Ik} \cdot O'_{ki}, & \text{if } i <= i^*_k \\ \beta'_k \cdot \phi(t) \cdot \left(\frac{P'_{i^*_k}}{P'_1}\right)^{f_k} \cdot \left(\frac{P'_i}{P'_{i^*_k}}\right)^{g_k} \cdot \frac{1-\gamma_k}{1+\left(\frac{I'_{ki}}{d'_k}\right)^{h_k}} \\ \qquad + \beta'_k \cdot \phi(t) \cdot \left(\frac{P'_i}{P'_1}\right)^{f_k} \cdot \gamma_k - \lambda_{Ok} \cdot O'_{ki}, & \text{if } i > i^*_k \end{cases} \quad (3)$$

where

$$I'_{ki} = \frac{I_{ki}}{c_{Ik}} \qquad \beta'_k = \frac{\beta_k}{c_{Ok} \cdot c_{Pk}} \qquad P'_{ki} = \frac{P_{ki}}{c_{Pk}}$$
$$O'_{ki} = \frac{O_{ki}}{c_{Ok} \cdot c_{Pk}} \qquad d'_k = \frac{d_k}{c_{Ik}} \quad (4)$$

In (3) and (4), I_{ki}, O_{ki}, β_k, H_k, λ_{Ik}, λ_{Ok}, i^*_k, f_k, and g_k are counterparts of I_i, O_i, β, H, λ_I, λ_O, i^*, f, and g from (2) for the k-th module. In (4), the prime variables represent the variables without batch effects. c_{Ik}, c_{Ok}, and c_{Pk} represent the scaling factors of the input I, the downstream protein O, and the TM in the k-th module. Similarly, we can develop a model for the j-th external-inducer-regulated promoter for each $j = 1, 2, ..., n$:

$$\frac{dI'_{ji}}{dt} = \alpha'_{ji} \cdot \phi(t) - \lambda_{I_j} \cdot I'_{ji}, \quad (5)$$

where

$$I'_{ji} = \frac{I_{ji}}{c_{Ij}} \qquad \alpha'_{ji} = \frac{\alpha_{ji}}{c_{Ij}} \quad (6)$$

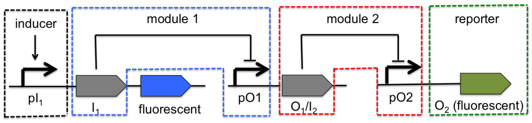

Fig. 4. Graphical representation of a two-transcriptional-repressor cascade. The circuit consists of two modules: the I$_1$-pO$_1$ module and the I$_2$/O$_1$-pO$_2$ module. I$_1$ and I$_2$ can be different combinations of TAL14, TAL21, and LmrA.

In (6), I_{ji} and α_{ji} are the concentration and the production rate of the protein expressed by the j-th external-inducer-regulated promoter in the i-th bin. I'_{ji} and α'_{ji} are counterparts of I_{ji} and α_{ji} without batch effects. Notice the same method of composition can be applied to the Hill-function-based model (1).

Example 2: Using the above method, we construct Hill-function-based models with and without rescaling, and a bin-dependent model with rescaling for each of the six possible two-repressor cascades: LmrA-TAL14, LmrA-TAL21, TAL14-LmrA, TAL14-TAL21, TAL21-LmrA, and TAL21-TAL14. The structure of a two-transcriptional-repressor cascade is illustrated in Fig. 4. Each cascade can be regarded as a composition of two modules. The scaling factors for I, O, and the TM are TAL14: 0.29, 0.93, 0.89; TAL21: 0.20, 1, 1.12; LmrA: 1, 0.41, 1 [16]. We then simulate the dynamics of the two-repressor cascades by inserting the parameter estimates into the models.

To validate the bin-dependent model and our rescaling method, we measure the differences between the simulated and the observed concentrations of EYFP 72 hours post-transfection. The error metric we adopt is the mean-fold error for all induction levels and plasmid copy numbers of each cascade. The mean-fold error is defined as $e^{\frac{\sum_{u=1}^{M}\sum_{i=1}^{N}\left|\log\left(\frac{O'_{ui}}{\hat{O}'_{ui}}\right)\right|}{MN}}$, where $\hat{O}'_{ui}$ and O'_{ui} denote the predicted and the observed concentrations of EYFP at hour 72, M the number of inducer levels, and N the number of bins. We compare the observed experimental concentrations against the concentrations simulated by the models (Fig. 5a). The bin-dependent model is shown to outperform the Hill-function-based model in almost all aspects. The Hill-function-based model with rescaling gives an average error of 1.8 fold compared to an error of 2.2 fold for the model without rescaling (Fig. 5b). This suggests that inconsistent scales between modules account for a significant portion of the error. With rescaling, the bin-dependent model makes an average error of 1.7 fold and produces smaller errors than the Hill-function-based model for five of the six cascades (Fig. 5b).

IV. Conclusion

In this work, we present a method of modular composition that addresses the issue of batch-effects in TTMC. By validating our method with real experimental data, we find that the rescaled bin-dependent model makes the most accurate predictions among the models that are tested. Our work shows promises in improving circuit designs in TTMC.

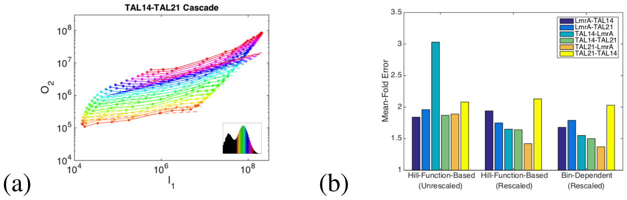

Fig. 5. (a) Comparison between the experimental data and the predictions of O_2 made by the bin-dependent model. Plasmid copy number is shown by color. Solid curves are experimental data, and dashed curves are model predictions (we plot the histogram and experimental data in this figure using data from [16]). (b) Comparison of the mean-fold errors of the Hill-function-based models, with and without rescaling, and the bin-dependent model with rescaling.

Code of Ethics

Our work does not require any ethics approval as the research is aimed at advancing the contributions of science and technology.

References

[1] S. Basu, Y. Gerchman, C. Collins, F. Arnold, and R. Weiss, "A synthetic multicellular system for programmed pattern formation," *Nature*, vol. 434, pp. 1130–34, 2005.

[2] T. Gardner, C. Cantor, and J. Collins, "Construction of a genetic toggle switch in escherichia coli," *Nature*, vol. 403, pp. 339–42, January 2000.

[3] T. Ellis, X. Wang, and J. Collins, "Diversity-based, model-guided construction of synthetic gene networks with predicted functions," *Nat. Biotechnol.*, vol. 27, no. 5, May 2009.

[4] D. D. Vecchio, "Design and analysis of an activator-repressor clock in e. coli," in *Proc. of American Control Conference*, New York, July 2007.

[5] "Synthetic biology: back to the basics," *Nat. Methods*, vol. 11, no. 5, p. 463, 05 2014.

[6] A. Gyorgy and D. D. Vecchio, "Modular composition of gene transcription networks," *PLOS Comput. Biol.*, March 2014.

[7] D. D. Vecchio and E. D. Sontag, "Dynamics and control of synthetic bio-molecular networks," in *Proc. of American Control Conference*, New York, 2007.

[8] D. D. Vecchio, Y. Qian, and A. Dy, "Control theory meets synthetic biology," *J. R. Soc. Interface*, 2016.

[9] H. Sivakumar and J. Hespanha, "Towards modularity in biological networks while avoiding retroactivity," in *Proc. of American Control Conference*, June 2013.

[10] B. Canton, A. Labno, and D. Endy, "Refinement and standardization of synthetic biological parts and devices," *Nat. Biotechnol.*, vol. 26, pp. 787–93, July 2008.

[11] T. Kim and J. Eberwine, "Mammalian cell transfection: the present and the future," *Anal. Bioanal. Chem.*, vol. 397, no. 8, pp. 3173–78, Aug 2010.

[12] A. Dalton and W. Barton, "Over-expression of secreted proteins from mammalian cell lines," *Protein Sci.*, vol. 23, no. 5, pp. 517–25, May 2014.

[13] K. Khan, "Gene expression in mammalian cells and its applications," *Adv. Pharm. Bull.*, vol. 3, no. 2, pp. 257–63, Dec 2013.

[14] U. Alon, *An Introduction to Systems Biology - Design Principles of Biological Circuits*. Chapman and Hall, 2007.

[15] J. Wang, S. A. Isaacson, and C. Belta, "Modeling genetic circuit behavior in transiently transfected mammalian cells," April 2018, manuscript submitted for publication.

[16] N. Davidsohn, J. Beal, S. Kiani, A. Adler, F. Yaman, Y. Li, Z. Xie, and R. Weiss, "Accurate predictions of genetic circuit behavior from part characterization and modular composition." *ACS Synth. Biol.*, vol. 4, no. 6, pp. 673–81, 2015.

[17] B. Pierce, *Genetics: A Conceptual Approach.*, 2nd ed. W. H. Freeman and Company, 2005.