Automated Design of Synthetic Gene Circuits through Linear Approximation and Mixed Integer Optimization

Linh Huynh
Department of Computer Science
University of California, Davis
huynh@ucdavis.edu

John Kececioglu
Department of Computer Science
University of Arizona, Tucson
kece@cs.arizona.edu

Ilias Tagkopoulos
Department of Computer Science & UC Davis Genome Center
University of California, Davis
itagkopoulos@ucdavis.edu

ABSTRACT

Synthetic biology aspires to revolutionize the way we construct biological circuits, as it promises fast time-to-market synthetic systems through part standardization, model abstraction, design and process automation. However, the automated design of synthetic circuits remains an unsolved problem, despite the increasing number of practitioners in the field. One reason behind that, is the absence of an efficient mathematical formulation for the combinatorial optimization problem of selecting genes and promoters when synthesizing the candidate circuits. Here, we propose an optimization framework that is based on a linear relaxation of the non-linear optimization problem, which proves to be a good approximation of the non-linear dynamics present in biological systems. Further evaluation of the proposed framework in a real non-linear synthetic circuit (a toggle switch), and with the use of a mutant promoter library, resulted in a rapid and reproducible convergence to a synthetic circuit that exhibits the desired characteristics and temporal expression profiles. This work is a step towards a unifying, realistic framework for the automated construction of biological circuits with desired temporal profiles and user-defined constraints.

1. INTRODUCTION

When it comes to automated biological circuit design, CAD tools are still in their infancy despite notable developments in the field. In this context, the use of mathematical optimization has been very limited [2] and with mixed results, while the main challenge still remains: how can we develop algorithms that cope with the combinatorial explosion and complex models that describe biological behavior? Here, we introduce a novel optimization formulation for synthetic circuit design that finds the optimal part configuration, given a library of biological parts, an objective function (e.g. the desired temporal profile of the output protein), user-defined constraints (e.g. circuit size), and an existing topology that provides connectivity (e.g. gene A must positively regulate gene B) but not individual parts (e.g. gene A, gene B, regulation strength).The optimization method translates the circuit design problem into a nonlinear integer programming formulation that it solves using spatial branch and bound techniques.

2. METHODS

Linear formulation: For the current analysis, assume a database of parts that has m promoters and n proteins. We introduce the following equation to express the concentration of protein i as a function of the available promoters and proteins:

$$\frac{df_i}{dt} = \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{k=1}^{m} a_{jk} y_{ik} f_j - (d_i + \mu) f_i + b_i \tag{1}$$

where the parameter a_{jk} is proportional to the production rate of protein i if protein j is bound at the promoter k upstream of gene i. Parameter d_i captures both the *degradation and auto-regulation* of protein i. The y_{ik} are binary variables defined as:

$$y_{ik} = \begin{cases} 1 & \text{If promoter } k \text{ is upstream} \\ & \text{of protein } i \\ 0 & \text{Otherwise} \end{cases}$$

Furthermore, we can add inducers in the system, by adding the term $-K_{inducer} f_i$ into equation 1. The solution of the linear ODE system is given by

$$\dot{F} = AF + B \tag{2}$$

where the elements of the A matrix are defined as

$$A_{ij} = \begin{cases} \sum_{k=1}^{m} a_{jk} y_{ik} & If\ i \neq j \\ -d_i - \mu & If\ i = j \end{cases} \tag{3}$$

and with F and B given as

$$F = (f_1(t), f_2(t), ..., f_n(t))^T,\ B = (b_1, b_2, ..., b_n)^T$$

Assuming that A is diagonalizable, there exists a matrix $S = (s_{ij})$, and a diagonal matrix D with its diagonal elements the eigenvalues of the system, i.e. $(\lambda_1, \lambda_2, ..., \lambda_n)$. Then $S^{-1}AS = D$ and thus $S^{-1}\dot{F} = DS^{-1}F + S^{-1}B$. Substituting $G = S^{-1}F$ and $E = S^{-1}B$, we end up with $\dot{G} = DG + E$, which has the following solution:

$$g_i(t) = C_i e^{\lambda_i t} - \frac{E_i}{\lambda_i} \tag{4}$$

As expected, the solution of a linear ODE system is nonlinear, and thus can describe dynamics of basic biological functions, since the latter are usually expressed with exponential functions. Mapping back to F, and since $F = SG$ we end up with the solution :

$$f_i(t) = \sum_{j=1}^{n} s_{ij} g_j(t) \qquad (5)$$

Objective function: Our optimization framework will try to obtain the parts set that minimize the difference between the desired temporal profile and the actual one. As such, if $f_p(t)$ and $f_p^*(t)$ are the estimated concentration and the desired concentration of protein p at the time point t, respectively, then our objective function is the following:

$$Z = \sum_{t \in T} (f_p(t) - f_p^*(t))^2 \qquad (6)$$

Linear constraints: The user may add additional constraints to the optimization system. For example, if a specific gene i should be included (or conversely, should be absent) in the design, then we can introduce a binary variable x_i, that will denote the presence or absence of gene i in the final circuit. In addition, the user may restrict the number of promoters for any given gene, limit the number of genes per promoter, or disallow large polycistronic promoters in the circuit:

$$\sum_{k=1}^{m} y_{ik} \leq M_1 x_i \quad \forall i = 1...n \qquad (7)$$

$$\sum_{i=1}^{n} y_{ik} \leq M_2 \quad \forall k = 1...m \qquad (8)$$

$$\sum_{k=1}^{m} y_{ik} \geq x_i \quad \forall i = 1...n \qquad (9)$$

Constraints are also in place due to stability issues of the resulting circuit. In order for the system to have stable dynamics, all eigenvalues must be distinct and their real parts must be negative. This leads to the following constraint:

$$Re(\lambda_i) \leq -\varepsilon \; and \; ||\lambda_i - \lambda_j|| \geq \varepsilon \quad \forall i \neq j = 1...n \qquad (10)$$

Finally, after the addition of standard diagonalization equations, normalization of eigenvalue vector space, and the addition of boundary conditions as constraints on the initial concentration of each protein i, the concentration of the desired protein is given by:

$$\sum_{j=1}^{n} (C_j e^{\lambda_j t} - E_j/\lambda_j) = f_p(t) \qquad (11)$$

Optimization problem : Now that we have defined all constraints in our system, we can formulate the optimization problem as solving for variables x_i and y_{ik} so that

$$\textbf{Minimize } Z \textit{ subject to } (4), (6) - (11)$$

This is a mixed integer non-linear programming (MINLP) problem which can be efficiently solved in practice using spatial branch & bound (e.g. Couenne [1]).

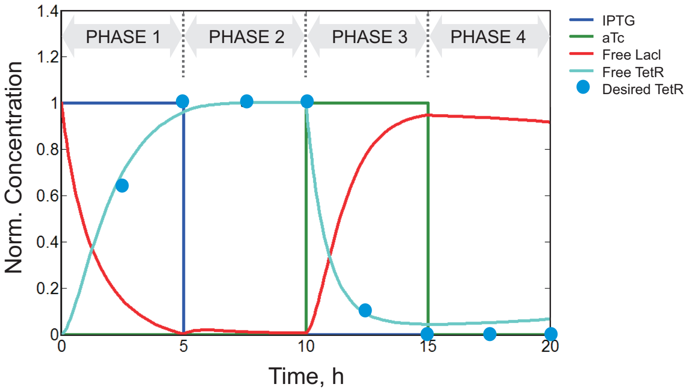

Figure 1: Expression profile of the resulting synthetic circuit, with promoters T7 (in upstream of LacI) and L3 (in upstream of TetR). The desired profile (input, depicted with blue dots) and actual profile (cyan line) for the TetR protein is shown. The temporal profile was split into four phases, based on changes in the inducer concentrations. Phase 1: IPTG high, aTc low; Phase 2: IPTG low, aTc low; Phase 3: IPTG low, aTc higt; Phase 4: IPTG low, aTc low.

3. RESULTS

To evaluate the capacity of our optimization framework, we assessed its performance in the case of a toggle switch design [4]. As an input to our optimization framework, we collected a mutant library for the TetO and LacO promoter that was experimentally characterized recently [3]. The values of a_{jk} are estimated from data in this library and other parameters such as degradation rates and association constants are chosen from [2]. As shown in figure 1, the system was able to find a set of parts (promoters T7 and L3, upstream of LacI and TetR, respectively) that led to a synthetic circuit that approximates well the desired transient dynamics.We further simulated these solutions, which resulted to circuits that also exhibit flip-flop characteristics at various degrees. From the complete solution space, less than 2% (7 sets) had this property.

4. REFERENCES

[1] P. Belotti, J. Lee, L. Liberti, F. Margot, and A. Waechter. Branching and bounds tighteningtechniques for non-convex MINLP. *Optimization Methods and Software*, 24(4):597–634, 2009.

[2] M. Dasika and C. Maranas. OptCircuit: an optimization based method for computational design of genetic circuits. *BMC Systems Biology*, 2(1):24, 2008.

[3] T. Ellis, X. Wang, and J. Collins. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature biotechnology*, 27(5):465–471, 2009.

[4] T. Gardner, C. Cantor, and J. Collins. Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403(6767):339–342, 2000.