

PARAMETER ESTIMATION FOR TWO SYNTHETIC GENE NETWORKS: A CASE STUDY

David Braun*

Princeton University
Department of Molecular Biology
Lewis Thomas Laboratory
Princeton, NJ 08544

Subhayu Basu[†], Ron Weiss[‡]

Princeton University
Department of Electrical Engineering
J-319, E-Quad
Princeton, NJ 08544

ABSTRACT

The ability to correlate mathematical models with experimental data is fundamental for a wide range of quantitative biology disciplines. Modelling typically requires accurate knowledge of kinetic rate constants, which may be extracted by parameter estimation using physical observations of overall system behaviors. Synthetic gene networks have well characterized connectivity and are easily manipulated for validation purposes, making them ideal for studying parameter estimation techniques. Here we use two synthetic gene networks, a transcriptional cascade and a pulse generating network, to study the efficacy of a simple statistical parameter fitting algorithm. The fitting was performed on experimental data and computer-generated data (to test how well the algorithm works under ideal conditions with perfect information). Most of the experimental parameter estimations yielded tight ranges of kinetic values for both gene networks. However, the results using simulated data indicate that the algorithm was able to provide better parameter estimates for the pulse generating network than for the transcriptional cascade. This is likely a result of the larger amount of time-series data available for the pulse generator and its greater level of phenotypical complexity, leading to tighter constraints for optimization. The variation in the magnitudes of the standard deviations between parameter estimates may give an indication of system sensitivity to specific kinetic rate constants. In the future, we also plan to verify the experimental parameter estimation results by constructing network variants and attempting to predict behaviors using values obtained in this study.

1. INTRODUCTION

Parameter estimation has applications in almost any field that involves quantitative analysis of biological systems, as the need to fit mathematical models to experimental data is ubiquitous. However, direct determination of *in vivo* values for biochemical parameters is difficult and often inaccurate, and thus a computational approach to parameter estimation has been employed for a number of biological systems [1, 2]. Here we explore the use of synthetic gene networks for studying parameter estimation algorithms. These networks are especially suitable for this task because presumably their topology is completely known and they can be easily manipulated and reconfigured, allowing for validation of any obtained parameter estimates.

This paper applies a simple statistical parameter fitting technique to two synthetic gene networks. Initially, a cost function is created that measures the deviation between the experimentally-determined system data and the computer-generated simulation data. This cost function is then minimized using a global optimization algorithm, Adaptive Simulated Annealing [3]. The computed cost is dependent on the set of kinetic parameters for the system. Thus the minimum cost function provides the parameter set which fits the model most closely with the experimental data. For well-constrained systems, as the value of the cost function approaches zero, the kinetic parameter estimations should ideally approach the actual biological parameters.

Two synthetic gene networks were selected as test systems for the parameter estimation. The first is a transcriptional cascade [4], which is a system composed of a series of repressors. The second is a pulse generating network [5], which employs a feed-forward motif to produce a transient pulse of reporter gene expression in response to a permanent increase in the concentration of an activating signal.

The efficacy of the parameter estimation technique was tested using simulation and experimental data to determine if this method is useful both theoretically and experimentally. The ability to obtain accurate parameter estimates using this method depends on the constraints established by the data. The system should have relatively complex output behavior (phenotypic complexity) with respect to the network architecture and a sufficient amount of experimental data available to optimize fitting accuracy. This parameter fitting approach was first applied to experimental data from the two synthetic gene networks in order to attempt to determine the true kinetic parameters of the systems. The experimental data was first interpolated and smoothed in preparation for the optimization. To analyze the potential accuracy of the fitting algorithm on these networks and the accompanying data, a model system with a chosen set of parameters was used to generate simulated, ideal, noiseless data. The kinetic parameters for the system were then randomized, and the parameter fitting technique was used in conjunction with the simulated time-series data to attempt to recover the original kinetic values.

In section 2, we introduce the fitting algorithm. In section 3, we describe the two gene networks under study. In section 4, we report and analyze the fitting results. Finally, in section 5, we summarize the findings and point to future work.

*Partially supported by the Howard Hughes Medical Institute.

[†]Partially supported by an NSF EMT grant CCF-0432094

[‡]Partially supported by a DOE grant DE-FG02-02ER15355

2. PARAMETER FITTING ALGORITHM

The algorithm consists of three components. The first component processes the experimental data in an attempt to remove errors and reduce noise. The second component consists of a cost function that computes the deviation of simulated data from the experimental data. The third component performs global optimization using the cost function in order to find the best possible fit to the experimental data, and consequently the set of optimal kinetic parameters.

2.1. Data Preprocessing

In order to increase the number of data points that can be used for the parameter fitting, and to compensate for any missing data points, normalized fluorescence data is interpolated using piecewise cubic interpolation (pchip, MATLAB 6.5, Mathworks, Natick, MA) [6]. The data is then smoothed using a hybrid Gaussian-median filter [7]. For the experimental data reported in Section 3, a window size of five to nine measurements (determined by trial-and-error) was able to filter out much of the noise without losing any relevant features of the data.

2.2. Cost Function

The deviation of the simulation data from the experimental data (the model error) is given by the cost function:

$$E = \sum_{i=1}^n (\log Y_i^{experimental} - \log Y_i^{simulation})^2$$

In this study we used time-series fluorescence data for the Y_i values. Since biological processes tend to have lognormal distributions [8], the logarithm of each data point was used to calculate the error. Using logarithms also reduces the large errors typically associated with measurements that have high values. Overall, this least-squares approach provides a maximum-likelihood estimation of the kinetic parameters [9].

2.3. Optimization

The cost function was minimized using Adaptive Simulated Annealing (ASA) [3]. ASA is an efficient simulated annealing algorithm that uses an annealing schedule with exponentially decreasing temperatures. In this study, the rate of annealing was decreased manually in order to compensate for the large number of parameters in the system. The exact rate of annealing was ultimately determined by educated guesses and trial-and-error for each system. The optimization was performed multiple times using a different, randomized set of initial estimates for each iteration. The results were then used to create a distribution for each kinetic parameter.

3. SYNTHETIC GENE NETWORKS

3.1. Transcriptional Cascades

The transcriptional cascades are networks of repressors (i.e. genetic logic inverters) connected in series [4]. The simplest cascade consists of only one inverter (Figure 1a). In this system, the *tet* repressor (TetR) is constitutively expressed and regulates the enhanced yellow fluorescence protein (EYFP), which is under control of the $P_{Ltet-O1}$ promoter. TetR repression of $P_{Ltet-O1}$ can be

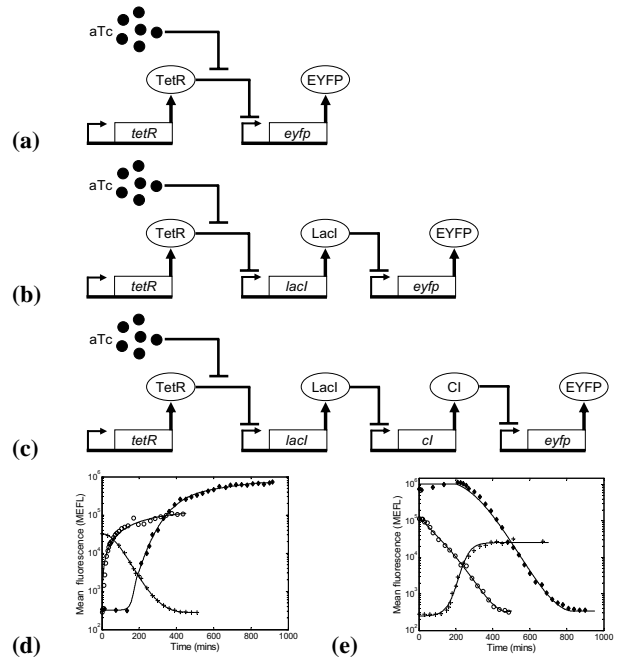


Fig. 1. (a-c) Schematic diagrams for the three transcriptional cascades. (d-e) Experimental data showing the time-series fluorescence responses of the networks to the addition/removal of aTc. When aTc is added, the fluorescence of cascade 1 and 3 increases, while cascade 2's output decreases. This pattern is reversed when aTc is removed from the growth medium. Cascade 1 has the quickest response, while cascade 3 exhibits the longest latency.

alleviated by the addition of anhydrotetracycline (aTc). In cascade 2, the *lac* repressor (LacI) replaces EYFP, which is now regulated by P_{lac} (Figure 1b). Hence, the output of cascade 2 is the inverse of cascade 1. In cascade 3, the lambda repressor (CI) replaces EYFP, whose expression is now regulated by the $\lambda P_{(R-O12)}$ promoter. Cascade 3's output follows the same pattern as cascade 1: low EYFP expression with no aTc induction and high EYFP expression with induction. However, the dynamic response of cascade 3 is delayed due to the latency incurred by the additional components in the system, as shown in the experimental results in Figure 1d and 1e. Parameter fitting was performed on cascade 3, and the EYFP output values of cascade 1 and 2 were used to approximate LacI and CI levels respectively.

Cascade 3 is modeled using the following Hill functions that represent regulated gene expression and protein decay:

$$\frac{dT}{dt} = \alpha_T - \gamma_T \cdot T \quad (1)$$

$$\frac{dL}{dt} = \alpha_{0L} + \frac{\alpha_L}{1 + \left(\frac{T}{\beta_T \cdot (1 + (A/\beta_A)^{\eta_A})} \right)^{\eta_T}} - \gamma_L \cdot L \quad (2)$$

$$\frac{dC}{dt} = \alpha_{0C} + \frac{\alpha_C}{1 + \left(\frac{L}{\beta_L} \right)^{\eta_L}} - \gamma_C \cdot C \quad (3)$$

$$\frac{dY}{dt} = \alpha_{0Y} + \frac{\alpha_Y}{1 + \left(\frac{C}{\beta_C} \right)^{\eta_C}} - \gamma_Y \cdot Y \quad (4)$$

The model consists of basal expression (α_{0L} , α_{0C} , α_{0Y}), protein

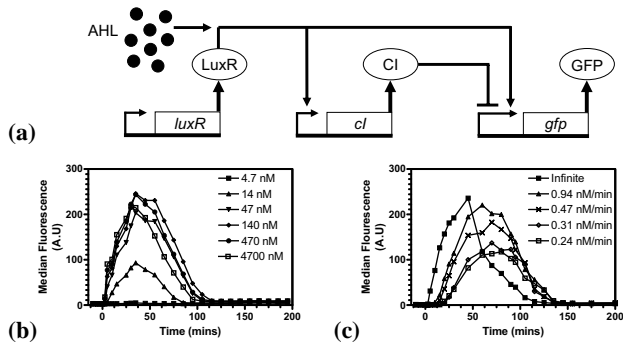


Fig. 2. (a) Schematic diagram for the pulse generating network. (b) Experimental results showing the time-series response to different final concentrations of AHL. (c) Experimental results showing the time-series response to different rates of AHL increase, all reaching the same final concentration of 47 nM.

synthesis ($\alpha_T, \alpha_L, \alpha_C, \alpha_Y$), repressor binding ($\beta_A, \beta_T, \beta_L, \beta_C$), protein decay ($\gamma_T, \gamma_L, \gamma_C, \gamma_Y$) and repression cooperativity ($\eta_A, \eta_T, \eta_L, \eta_C$) for TetR (T), aTc (A), LacI (L), CI (C), and EYFP (Y). TetR (equation 1) is constitutively expressed, and its repression activity is inhibited by the addition of aTc (equation 2).

3.2. Pulse Generating Network

The pulse generating network [5] uses quorum sensing elements from *Vibrio fischeri* [10] to activate the expression of both CI and the green fluorescent protein (GFP) (Figure 2a). An acyl-homoserine lactone (AHL) signal diffuses through the cell membrane and forms a complex with the constitutively expressed LuxR protein. The LuxR-AHL complex then activates the $luxP_R$ promoter of CI and the $luxP_{RCI-OR1}$ hybrid promoter of GFP. The gfp gene is engineered with a much stronger ribosome binding site than the cl gene, and thus GFP expression increases more quickly following induction. Eventually, enough CI accumulates to bind the $luxP_{RCI-OR}$ promoter effectively and repress GFP expression. Thus, after transient GFP expression, its cytoplasmic concentration decreases to a basal level due to protein decay. Experimental results demonstrate that the GFP pulse amplitude, duration and delay depend on the AHL concentration (Figure 2b) and the rate at which AHL is added to the growth medium (Figure 2c).

The pulse generating network was modeled using the following Hill functions that represent gene activation, repression, and protein decay:

$$\frac{dL}{dt} = \alpha_L - \gamma_L \cdot L \quad (5)$$

$$\frac{dC}{dt} = \alpha_{0C} + \frac{\alpha_C \cdot A^{\eta_A}}{(\theta_A)^{\eta_A} + A^{\eta_A}} - \gamma_C \cdot C \quad (6)$$

$$\frac{dG}{dt} = \alpha_{0G} + \left(\frac{\alpha_G \cdot A^{\eta_A}}{(\theta_A)^{\eta_A} + A^{\eta_A}} \right) \frac{1}{1 + (C/\beta_C)^{\eta_C}} - \gamma_G \cdot G \quad (7)$$

where

$$A = \frac{L \cdot H^{\eta_H}}{(\theta_H)^{\eta_H} + H^{\eta_H}}$$

The model includes LuxR (L), LuxR-AHL complex (A) that is based on the AHL input (H), CI (C) and GFP (G). It consists

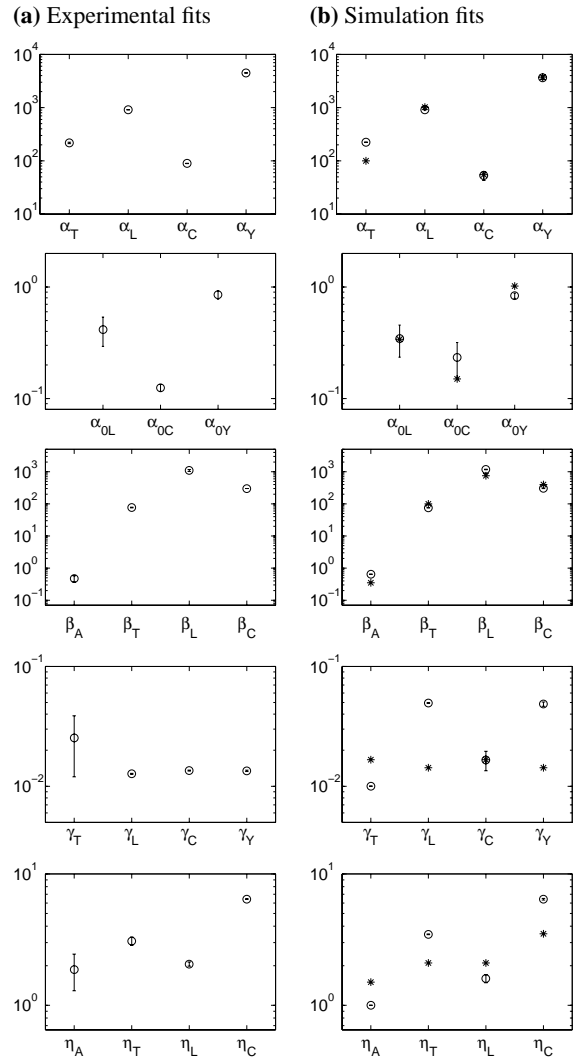


Fig. 3. Parameter fitting results for the transcriptional cascades.

of protein synthesis ($\alpha_G, \alpha_C, \alpha_L$), basal expression (α_{0G}, α_{0C}), protein decay ($\gamma_G, \gamma_C, \gamma_L$), activation and binding of LuxR to AHL (θ_H) and of the LuxR-AHL complex to the promoters (θ_A), repression (β_C), and regulatory cooperativity (η_C, η_A, η_H). GFP expression (equation 7) depends both on LuxR-AHL activation (first term) and on CI repression (second term).

4. PARAMETER FITTING RESULTS

After the experimental data was preprocessed as described in Section 2.1, the smoothed data were used to perform the parameter fit. The parameter estimation results are shown in Figure 3a (transcriptional cascades) and Figure 4a (pulse generator). The optimization was repeated 20 times for each network in order to generate distributions for the parameters. The results are reported as the mean value obtained \pm standard deviation (circle and error bars). Most of the estimations converged to very tight parameter value ranges. The exceptions include α_{0L} , β_A , γ_T and η_A for the cascades as well as α_{0C} and γ_L for the pulse generator.

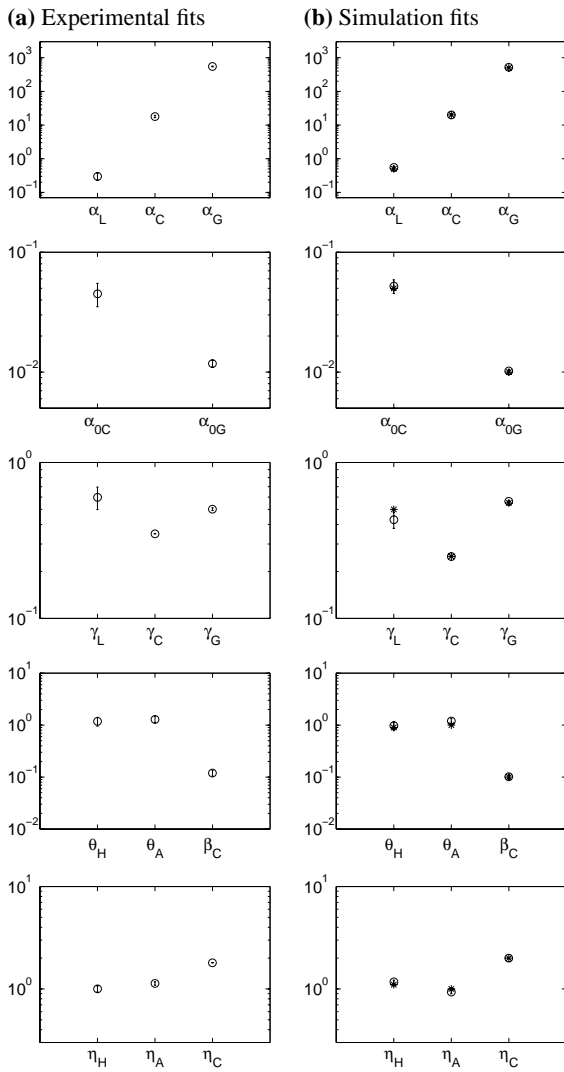


Fig. 4. Parameter fitting results for the pulse generating network.

In order to test the accuracy of the fitting algorithm, we also performed parameter estimation to simulated time-series data. These simulations were performed using the Hill functions described above (equations 1-4 for the cascades and 5-7 for the pulse generator) with reference kinetic parameter values as indicated by the asterisks in Figures 3b and Figures 4b. The parameter fitting for the pulse generator (Figure 4b) provided more accurate estimations than that of the transcriptional cascades (Figure 3b). A possible cause is the lack of sufficient constraints for the transcriptional cascades: 6 time-series curves were used to estimate 19 cascade parameters as compared to 11 time-series curves for 14 pulse generator parameters. Also, the pulse generator network exhibits a greater phenotypical complexity in response to a long-lasting change in inducer concentration, possibly providing it with tighter constraints. Interestingly, despite the fact that many cascade parameter values were not recovered accurately, the distributions of

most of the estimations were quite tight. This observation warrants further investigation.

5. CONCLUSION

The parameter estimation approach described here appears able to recover kinetic parameter values reasonably well for highly constrained gene networks. This method's effectiveness is hard to measure directly, as the actual *in vivo* kinetic parameters are not well known. Given the results of the parameter estimation with simulated data, it is likely that the parameters obtained for the pulse generating network more closely resemble the actual biological parameters than the ones obtained for the transcriptional cascade. The accuracy of the parameter estimation obtained in this study will be verified in future work by constructing networks variants using elements from both systems, and using the estimations computed here to predict the new systems' behaviors. It is also interesting to note that the magnitude of the standard deviations varied significantly from parameter to parameter. It is likely that this value provides some indication of the system sensitivity to the given parameter, and this will be investigated in the future as well.

6. REFERENCES

- [1] V. V. Gusky, J. Jaeger, K. N. Kozlov, J. Reintz, and A. M. Samsonov, "Pattern formation and nuclear divisions are uncoupled in drosophila segmentation: Comparison of spatially discrete and continuous models," *Physica D*, vol. 197, pp. 286–302, 2004.
- [2] M. Ronen, R. Rosenberg, B. Shraiman, and U. Alon, "Assigning numbers to arrows: Parameterizing a gene regulation network by accurate expression kinetics," *PNAS*, vol. 99, pp. 10555–10560, 2002.
- [3] A. L. Ingber, "Simulated annealing: Practice versus theory," *Mathematical Computer Modeling*, vol. 18, no. 11, pp. 29–57, 1993.
- [4] S. Hooshangi, S. Thiberge, and R. Weiss, , "submitted, 2004.
- [5] S. Basu, R. Mehreja, S. Thiberge, M. T. Chen, and R. Weiss, "Spatiotemporal control of gene expression with pulse-generating networks," *PNAS*, vol. 101, pp. 6355–6360, 2004.
- [6] F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM J. Numerical Analysis*, vol. 17, pp. 238–246, 1980.
- [7] U. Alon, L. Camarena, M. G. Surette, B. Aguera y Arcas, Y. Liu, S. Leibler, and J. B. Stock, "Response regulator output in bacterial chemotaxis," *EMBO J.*, vol. 17, pp. 4238–4248, 1998.
- [8] D. B. Hattis and D. E. Burmaster, "Assessment of variability and uncertainty distributions for practical risk assessments," *Risk Analysis*, vol. 14, no. 5, pp. 713–730, 1994.
- [9] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-PLUS, 3rd edition*, Springer, 1999.
- [10] M. B. Miller and B. L. Bassler, "Quorum sensing in bacteria," *Annu. Rev. Microbiol.*, vol. 55, pp. 169–99, 2001.