Estimating cell cycle model parameters using systems identification

Edwin Juarez, Ahmadreza Ghaffarizadeh, Samuel H. Friedman, Edmond Jonckheere, and Paul Macklin

***Abstract* — A current challenge in data-driven mathematical modeling of cancer is identifying biologically-relevant parameters of mathematical models from sparse and often noisy experimental data of mixed types. We describe a cell cycle model and outline how to use the Optimization Toolbox in Matlab to estimate its timescale parameters, given flow cytometry and cell viability (synthetic) data, and illustrate the technique with simulated data. This technique can be similarly applied to a variety of cell cycle models, particularly as more laboratories begin to use high-content, quantitative cell screening and imaging platforms. An advanced version of this work (CellPD: cell line phenotype digitizer) will be released as open source in early 2016 at MultiCellDS.org.**

I. Introduction

Cell cycle time scales are parameters often needed when developing a mathematical model to describe a (cancer) cell population. Measuring these time scales experimentally can be very challenging, and it often requires technologies that are not accessible in most labs. However, using cell flow cytometry [1] we can measure the fraction of cells at three different stages of the cell cycle (G_0/G_1, S, G_2/M). Additionally, propidium iodide (PI) can be used to measure the unviable/apoptotic fraction of the same cell population [2]. Combining these fractions with an automated cell counter (such as the Bio-Rad TC20), we can measure the number of cells at each stage of the cell cycle, and the number of cells undergoing apoptosis.

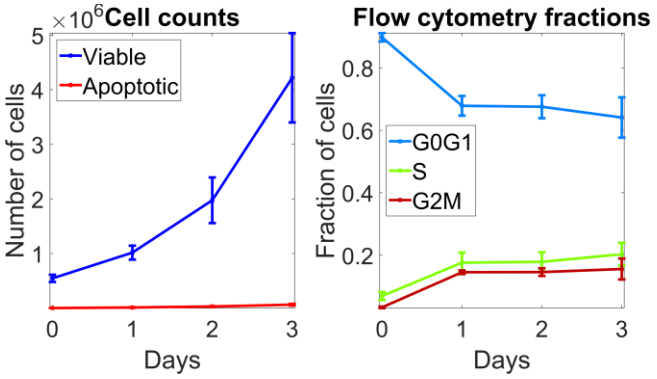

Figure 1 - Synthetic data representative of experimental data. Sample mean ± standard deviation are plotted.

Research supported by The Breast Cancer Research Foundation, USC James H. Zumberge Research & Innovation Fund, and USC Provost's PhD Fellowship.

E. Juarez, A. Ghaffarizadeh, S. H. Friedman, and P. Macklin are with the Center for Applied Molecular Medicine, Department of Medicine, University of Southern California, 2250 Alcazar St., HSC-CSC 240, Los Angeles, CA 90033-9075, USA. E-mail: [juarezro, aghaffar, samuelf, paul.macklin]@usc.edu.

P. Macklin is the corresponding author (Tel: (310) 701-5785).

To study the cell cycle of a cell line, in the lab we would typically record these measurements at a few time points per microenvironmental condition. In this example, we have synthetized data by running the model with a given set of parameters (referred to as the "true parameters") and multiplying each sample point by a "noise" factor f with Gaussian distribution (with a mean of 1 and a standard deviation of 0.2). This process was repeated 3 times (the mean and standard deviation of those 3 samples was computed) to mimic the low number of replicates in many experiments. Fig. 1 summarizes these synthetic data. The methods described in this example can be directly applied to lab data under each specified microenvironmental condition to understand the impact of the microenvironment on the parameters of the model (e.g. cell cycle timescales).

Authors' note: The methods described in this work will be released in a more advanced form through the CellPD (cell line phenotype digitizer) tool. Please check MathCancer.org and MultiCellDS.org/Tools.php for the most up-to-date project information and downloads.

II. Model Description and Parameter Estimation

We can use a system of ordinary differential equations (ODEs) to describe the cell cycle dynamics:

$$\frac{dG_1}{dt} = \frac{2}{\tau_M}M - \frac{1}{\tau_{G1}}G_1 - r_A G_1 \quad (1)$$

$$\frac{dS}{dt} = \frac{1}{\tau_{G1}}G_1 - \frac{1}{\tau_S}S - r_A S \quad (2)$$

$$\frac{dG_2}{dt} = \frac{1}{\tau_S}S - \frac{1}{\tau_{G2}}G_2 - r_A G_2 \quad (3)$$

$$\frac{dM}{dt} = \frac{1}{\tau_{G2}}G_2 - \frac{1}{\tau_M}M - r_A M \quad (4)$$

$$\frac{dA}{dt} = r_A(G_1 + S + G_2 + M) - \frac{1}{\tau_A}A \quad (5)$$

Each equation represents the net change on the number of cells in each modeled stage of the cell cycle (G_0/G_1, S, G_2, and M) and nonviable (apoptotic) cells (A). The parameters (τ_{G1},τ_S,τ_{G2},τ_M,τ_A,r_A) represent the mean values of an unknown distribution and are, in general, microenvironment- and time-dependent. For this demonstration, we assume them to be constant in time, and we also assume that the microenvironment does not change significantly enough to modify the value of those parameters during the experiment. Our goal will be to find the set of parameters that best describes the data.

In order to find this optimal parameter set, we select an initial guess for each of the parameters based on literature [3-5] and observations from the lab. We define an error vector, $\vec{E}$, to quantify the discrepancy between the experimental measurements and the model predictions. For any data observation time t, we define the raw error vector $\vec{E}(t)$ to be the differ-

ence between the simulated and experimental values of each cell –sub-population. In order to account for measurement errors, we divide each vector component by its corresponding coefficient of variation (CV); this gives the greatest weighting to data with the least uncertainty. The overall error vector $\vec{E}$ concatenates each sample time's error vector. In the example presented here, after including cell viability, flow cytometry and cell count data we have 5 "Target" populations: Viable, Apoptosing, G_0/G_1, S, and G_2/M cells. Each of these target populations have 4 time points, so $\vec{E}$ has 20 elements. The sum of the squares of the elements on $\vec{E}$ can be computed to find the Sum of Squared Errors (SSE). We can now define an optimization problem by attempting to minimize the SSE while maintaining the timescales within predefined constraints. This minimization is performed by using Matlab's lsqnonlin function.

Sample code to generate the synthetic data and estimate the timescales of the cell cycle model used in this example will be made available for download from:

http://MathCancer.org/NCI_handbook_parameters, and

http://MultiCellDS.org/Tools.php.

III. Results of Method

After applying the optimization described above, to the optimal parameters (as shown in Table 1, column "Estimated parameter, dataset 1") lead to a good fit of the data as shown in Fig. 2. But it is worth noting that due to the low number and sparse frequency of the samples, the true parameters could not be recovered exactly. But if we repeat this method and we sample the data more often, more times, and with smaller noise, we can recover the true parameters even if we start from the same initial guess as before. Table 1, column "Estimated parameter, dataset 2" shows the results of the optimization after sampling the data twice per day, and take 30 samples each time with a Gaussian noise with mean 1 and standard deviation of 0.001. Synthetic dataset 2 is virtually impossible to replicate in the lab given the large number of samples and the low level of measurement error required, but it shows that it is possible to extract the true parameters given a sufficiently clean set of data. Furthermore, even with the noisier dataset, we can obtain parameters that not only lead to a good simulation fit but are also close to the true parameters.

Table 1: Optimization Results

Parameter name	Estimated parameter, dataset 1	Estimated parameter, dataset 2	True parameter	Units
τ_{G1}	16.3	15.998	16	Hours
τ_S	5.2	5.999	6	Hours
τ_{G2}	3.1	3.016	3	Hours
τ_M	1.4	1.984	2	Hours
τ_A	42.9	23.963	24	Hours
r_A	0.001	0.001	0.001	1/Hour

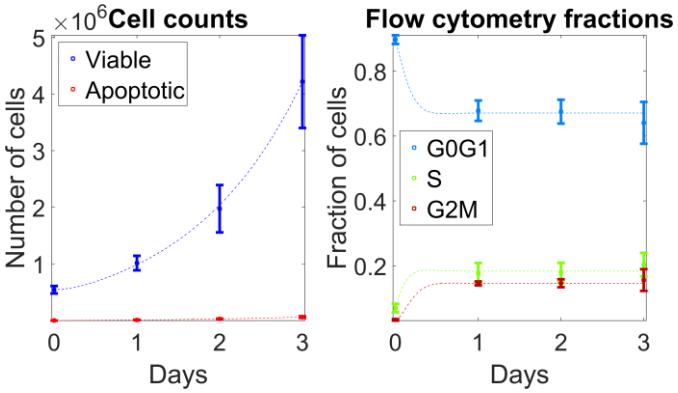

Figure 2: Simulation results using estimated parameters from dataset 1 (3 samples every 24h with a Gaussian noise of mean 1 and standard deviation of 0.2). Sample mean ± standard deviation are plotted, dotted lines represent simulation output.

IV. Type of Settings in which these Methods are Useful

This method can be applied in virtually the same way to other cell cycle models, where the errors and model are chosen to suit the available data. The same methodology can be used to calibrate agent-based models by defining an error metric and tuning simulation parameters to minimize that error function. Lastly, we note that the optimization technique can be iterated to help quantify uncertainty in the parameters.

REFERENCES

[1] P. Pozarowski and Z. Darzynkiewicz, "Analysis of cell cycle by flow cytometry," *Methods Mol Biol*, vol. 281, pp. 301-11, 2004.

[2] I. Nicoletti, G. Migliorati, M. C. Pagliacci, F. Grignani, and C. Riccardi, "A rapid and simple method for measuring thymocyte apoptosis by propidium iodide staining and flow cytometry," *J Immunol Methods*, vol. 139, pp. 271-9, Jun 3 1991.

[3] K. Simms, N. Bean, and A. Koerber, "A mathematical model of cell cycle progression applied to the MCF-7 breast cancer cell line," *Bull Math Biol*, vol. 74, pp. 736-67, Mar 2012.

[4] M. R. Dowling, A. Kan, S. Heinzel, J. H. Zhou, J. M. Marchingo, C. J. Wellard, J. F. Markham, and P. D. Hodgkin, "Stretched cell cycle model for proliferating lymphocytes," *Proc Natl Acad Sci U S A*, vol. 111, pp. 6377-82, Apr 29 2014.

[5] R. L. Sutherland, R. E. Hall, and I. W. Taylor, "Cell proliferation kinetics of MCF-7 human mammary carcinoma cells in culture and effects of tamoxifen on exponentially growing and plateau-phase cells," *Cancer Res*, vol. 43, pp. 3998-4006, Sep 1983.