Algorithm for estimating the parameters in the multi-stage model from cell proliferation data utilising FUCCI-expressing cell lines

Sean T. Vittadello

sean.vittadello@qut.edu.au

School of Mathematical Sciences

Queensland University of Technology

Last Update: 3 June 2019

1 Introduction

The algorithm herein details the implementation of the MATLAB script **MultiStageModel_fit.m** with associated function **MultiStageModel_fit_function.m**. This script and function are used to optimise the fit of the multi-stage model to cell proliferation data using FUCCI-expressing cell lines, as employed in our manuscript (Sean T. Vittadello, Scott W. McCue, Gency Gunasingh, Nikolas K. Haass and Matthew J. Simpson. 2019. *Mathematical models incorporating a multi-stage cell cycle replicate normally-hidden innate synchronisation in cell proliferation.* bioRxiv: 557702.)

The multi-stage model requires specification of the number of stages, the transition rates from each stage to the successive stage, and the initial population in each stage. We aim to achieve the best fit of the model to our data while keeping the number of model parameters with distinct values to a minimum.

MATLAB optimisation is used to estimate most of the parameters, however we choose certain parameter values beforehand, namely S_{r}, S_{y}, S_{g}, I_{r}, I_{y} and I_{g}. The numbers of stages in each phase, S_{r}, S_{y} and S_{g}, determine the sustainment of the oscillation amplitude of the ratio $Q(t) = R(t)/\big(Y(t) + G(t)\big)$, where $R(t)$, $Y(t)$ and $G(t)$ are the numbers of cells in G1, eS and S/G2/M over time. We always set $S_{\mathrm{r}} = S_{\mathrm{y}} = S_{\mathrm{g}}$. The appropriate number of stages is determined by optimising the fit of $Q(t)$ for the multi-stage model to $Q(t)$ for the data, which can be assisted by visually comparing the data and model solution. The parameters I_{r}, I_{y} and I_{g} are the numbers of consecutive stages in each cell cycle phase highlighted by FUCCI for which the initial cell numbers are equal. We always set $I_{\mathrm{r}} = I_{\mathrm{y}} = I_{\mathrm{g}}$. Once the number of stages is determined, the possible values for I_{r}, I_{y} and I_{g} are among the divisors of S_{r}, S_{y} and S_{g}, respectively. Optimising the fit of $Q(t)$ for the multi-stage model to $Q(t)$ for the data, with the possible values for I_{r}, I_{y} and I_{g}, will determine the appropriate values. Further, we set the transition rates between successive stages to be equal within each cell cycle phase highlighted by FUCCI.

We find that the objective function $\mathbf{F}$ typically has multiple local minima for which the 2-norm of the residual has similar small values. It may therefore be necessary to choose the parameters which correspond to the best visual fit to the data, for which we find there is usually a unique choice.

The starting point for parameterising the multi-stage model is to determine the appropriate number

of stages for the model. We initially set all of the weights $w_2, \ldots, w_7$ to zero, so that the objective function fits the model only to the ratio $Q(t)$. In choosing the number of stages, typically there is a maximum number of stages above which more stages do not improve the fit to $Q(t)$, and there is a minimum of stages below which the oscillations in $Q(t)$ decay too rapidly. The appropriate number of stages is based on the norm of the residual from the optimised fit, and can be assisted by comparing the data and model solution visually. The result is at most a few similar values for the number of stages.

We then additionally fit the model to the data for the three subpopulations of cells in G1 (red), eS (yellow) and S/G2/M (green). For simplicity, we set the weights $w_2, \ldots, w_4$, corresponding to the red, yellow and green subpopulations, to the same value. Choosing different weights for the subpopulations often results in a poor fit to at least one of the subpopulations. We begin by setting $w_2, \ldots, w_4$ to 10^{-8}, which has no effect on the optimisation, and then progressively increase the weights by an order of magnitude until the model fits the subpopulation data well. If required, this process can be repeated for any different options for the number of stages, which tends to narrow down these options to a best number of stages.

When the ratio $Q(t)$ from the data consists of only a single oscillation, which in our experiments is the case for cell lines with relatively long cell cycle times, such as WM983C and 1205Lu, the optimisation may result in a much larger estimated cell cycle duration, reflected in the estimates for the phase durations L_{r}, L_{y} and L_{g} of G1, eS and S/G2/M, than is observed experimentally. In this case it becomes necessary to constrain the estimated phase duration parameters L_{r}, L_{y} and L_{g} to sum to the expected cell cycle time. (Haass NK, Beaumont KA, Hill DS, Anfosso A, Mrass P, Munoz MA, et al. *Real-time cell cycle imaging during melanoma growth, invasion, and drug response.* Pigment Cell Melanoma Res. 2014;27:764–776.) This is achieved by setting the weight w_6 to 10^{-8}, which has no effect on the optimisation, and then progressively increasing w_6 by an order of magnitude until the phase durations assume more realistic values.

The weight w_7 is used to fit the model to the total population data. Most of the time we can set $w_7 = 0$, as fitting the model to the three subpopulations automatically yields a good fit to the total

population. Sometimes, however, we find it is necessary to constrain the model to the total population data. This is achieved by setting the weight w_7 to 10^{-8}, which has no effect on the optimisation, and then progressively increasing w_7 by an order of magnitude until a good fit results.

Finally, the weight w_5 is used to constrain the final data point of the green subpopulation. This is sometimes required as the model may not fit the green subpopulation well toward the end of the experiment. This is likely due to a decreased pH of the growth medium due to the increasing concentration of acidic cell-metabolites over the course of the experiment, causing a small degree of arrest in G1 near the end of the experiment, and therefore slightly shifting the normal proportions of red, yellow and green cells. We have been able to minimise this effect by increasing the volume of growth medium in the wells from the standard 1 ml to 2.5 ml, however it is still evident in the final hours of the experiment. Often we can let $w_5 = 0$, otherwise we set the weight w_5 to 10^{-8}, which has no effect on the optimisation, and then progressively increase w_5 by an order of magnitude until a reasonable fit results.

Algorithm 1 Parameterisation of the multi-stage model - part 1

/* The MATLAB script is **MultiStageModel_fit.m** and the associated */
/* function is **MultiStageModel_fit_function.m** */

1: Input the mean cell-cycle duration **Cdata** for the cell line, based on experimental data. For example:

- C8161 cell line - 18 hours;
- WM983C cell line - 27 hours;
- 1205Lu cell line - 36 hours.

/* Haass NK, Beaumont KA, Hill DS, Anfosso A, Mrass P, Munoz MA, et al. *Real-time cell cycle imaging during melanoma growth, invasion, and drug response.* Pigment Cell Melanoma Res. 2014;27:764–776. */

2: Input the number of stages **Sr** in G1, the number of stages **Sy** in eS, and the number of stages **Sg** in S/G2/M.

3: Input the number of consecutive stages **Ir** in G1 for which the initial cell numbers are equal. Note that **Ir** must be a divisor of **Sr**.

4: Input the number of consecutive stages **Iy** in eS for which the initial cell numbers are equal. Note that **Iy** must be a divisor of **Sy**.

5: Input the number of consecutive stages **Ig** in S/G2/M for which the initial cell numbers are equal. Note that **Ig** must be a divisor of **Sg**.

6: Input the vector of starting parameters **param** for the vector objective function **F**. The vector **param** has the form

$$[\,\mathcal{R}_1 \ldots \mathcal{R}_{Sr/Ir} \quad \mathcal{Y}_1 \ldots \mathcal{Y}_{Sy/Iy} \quad \mathcal{G}_{Sg/Ig} \ldots \mathcal{G}_{Sg/Ig} \quad \mathrm{Lr} \quad \mathrm{Ly} \quad \mathrm{Lg}\,]$$

where $\mathcal{R}_i$ is the number of cells in each of the stages in the ith block of **Ir** stages in G1, and similarly for $\mathcal{Y}_i$ and $\mathcal{G}_i$.

Algorithm 1 Parameterisation of the multi-stage model - part 2

7: Input the weights **w2**, ..., **w7** for the vector objective function:

- **w2** weights the fit of the model to the G1 subpopulation data.
- **w3** weights the fit of the model to the eS subpopulation data.
- **w4** weights the fit of the model to the S/G2/M subpopulation data.
- **w5** weights the fit of the model to the S/G2/M subpopulation data at the final time point, and is only required if the cells are starting to arrest in G1 near the end of the experiment due to the decreased pH of the growth medium.
- **w6** weights the constraint of the estimated phase durations to sum to the expected cell cycle time, and is generally required only when there are an insufficient number of oscillations in Qdata(t) to bound the estimated cell-phase durations to physically realistic values.
- **w7** weights the fit of the model to the total population data, and is often not required as a good fit usually follows from fitting to the three subpopulations.

8: Input the path of the data file for **path**.