Analysis of CFSE time-series data using division-, age- and label-structured population models

Sabrina Hross and Jan Hasenauer*

Institute of Computational Biology, Helmholtz Zentrum München,
Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

Chair of Mathematical Modeling of Biological Systems, Center for Mathematics,
Technische Universität München, Boltzmannstraße 3, 85748 Garching, Germany.

*corresponding author (jan.hasenauer@helmholtz-muenchen.de)

1 Abstract

Motivation: *In vitro* and *in vivo* cell proliferation is often studied using the dye Carboxyfluorescein succinimidyl ester (CFSE). The CFSE time-series data provides information about the proliferation history of populations of cells. While the experimental procedures are well established and widely used, the analysis of CFSE time-series data is still challenging. Many available analysis tools do not account for cell age and employ optimisation methods which are inefficient (or even unreliable).

Results: We present a new model-based analysis method for CFSE time-series data. This method uses a flexible description of proliferating cell populations, namely, a division-, age- and label-structured population model. Efficient maximum likelihood and Bayesian estimation algorithms are introduced to infer the model parameters and their uncertainties. These methods exploit the forward sensitivity equations of the underlying partial differential equation model for efficient and accurate gradient calculation, thereby improving computational efficiency and reliability compared to alternative approaches and accelerating uncertainty analysis. The performance of the method is assessed by studying a dataset for immune cell proliferation. This revealed the importance of different factors on the proliferation rates of individual cells. Among others, the predominate effect of the cell-age on the division rate is found which was not revealed by available computational methods.

Availability: The MATLAB source code implementing the models and algorithms is available from http://janhasenauer.github.io/ShAPE-DALSP/.

Contact: jan.hasenauer@helmholtz-muenchen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

2 Introduction

Proliferation is essential in many biological processes, ranging from development to wound healing, immune response and stem cell renewal. Due to this eminent role, a key challenge in cell biology is to quantify proliferation dynamics. Nowadays there are different experimental methods available to assess cell proliferation, e.g., time-lapse microscopy (Schroeder, 2011) and proliferation assays based upon Carboxyfluorescein succinimidyl ester (CFSE) labeling (Lyons and Parish, 1994). While time-lapse microscopy is highly informative, it is difficult to apply in vivo. CFSE-based proliferation assays can be employed *in vivo* and *in vitro* but the data analysis is more intricate and no single-cell time courses are available.

To study proliferation using CFSE, cells are incubated with carboxyfluorescein diacetate succinimidyl ester (CFDA-SE). CFDA-SE can diffuse across the cell membrane and once in the cytoplasm is converted to CFSE, which binds covalently to intracellular proteins. CFSE is a fluorescent dye and its concentration is reduced by protein degradation as well as cell division. During cell division, CFSE is distributed approximately equally among daughter cells (Lyons and Parish, 1994), hence, proliferation results in a progressive dilution of the dye (Figure 1A) which can be recorded using flow cytometry. As the individual cells do not, in general, divide with the same rates, an initially uni-modal distribution (Figure 1B) becomes multi-modal as time progresses (Figure 1C). The modes are related to different numbers of cell divisions (Hawkins *et al.*, 2007a); however, often the cells with different numbers of divisions are not strictly separated. For this reason, the calculation of the number of cells with a certain division number using peak detection and deconvolution methods (Hawkins *et al.*, 2007a; Luzyanina *et al.*, 2007a) is error-prone. Furthermore, even knowledge about the number of cells with a particular division number does not enable the comparison of different hypotheses regarding, e.g., the proliferation rates.

The need for quantitative data analysis and hypothesis testing inspired the development of a multitude of model-based approaches. Nowadays, age-structured population (ASP) models (Bernard *et al.*, 2003; Hawkins *et al.*, 2007b), division-structured population (DSP) models (De Boer *et al.*, 2006), label-structured population (LSP) models (Banks *et al.*, 2010; Luzyanina *et al.*, 2007b), age- and division-structured population (ADSP) models (Hawkins *et al.*, 2007b), and division- and label-structured population (DLSP) models (Hasenauer *et al.*, 2012; Schittler *et al.*, 2011) are used to study CFSE data. These classes of population balance models account for up to two properties of individual cells:

- number of divisions i ($\rightarrow$ DSP and DLSP model)
- CFSE concentration x ($\rightarrow$ LSP and DLSP model)
- cell's age a ($\rightarrow$ ASP and ADSP model)

and are mostly written as systems of ordinary differential equations (ODEs) or partial differential equations (PDEs). Population balance models which account for the cell age – the time passed since cell division –, i.e., ASP and ADSP, can also be formulated as sets of nested integrals (De Boer and Perelson, 2013). This alternative formulation is employed for the Smith-Martin model (Smith and Martin, 1973) and the cyton model (Hawkins *et al.*, 2007b). For the Smith-Martin model and the cyton model (with progression fraction equal to zero), equivalent PDE formulations are available (Bernard *et al.*, 2003; De Boer and Perelson, 2013).

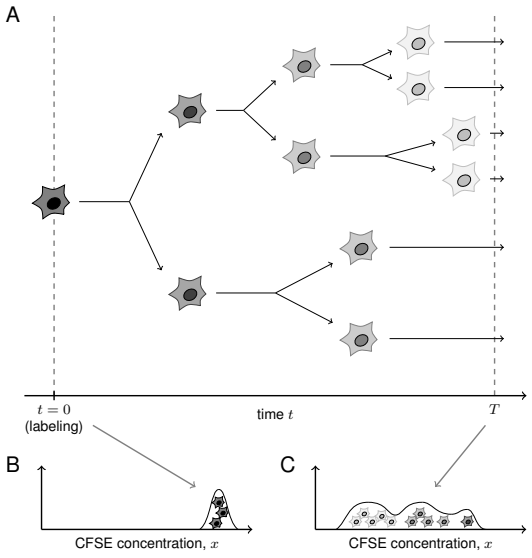

Figure 1: Illustration of proliferation assay showing (**A**) one labeled cell and its descendants, (**B**) the frequency of different CFSE concentration in the cell population at $t = 0$, and (**C**) the frequency of different CFSE concentration in the cell population at $t = T$. Shades from dark to light gray correspond to high and low CFSE concentration, respectively.

For a comprehensive introduction to structured population models, we refer to the review by De Boer and Perelson (2013).

As the population models available in the literature did not capture the complexity of the process, we recently introduced the division-, age- and label-structured population (DALSP) model (Metzger *et al.*, 2012). The DALSP model, a system of coupled partial differential equations (PDEs), provides a flexible description of proliferation dynamics (Metzger *et al.*, 2012), but it has never been used to analyse CFSE data. Accordingly, no parameter estimation, model selection and uncertainty analysis methods have been available for this type of model until now.

In this manuscript, we introduce methods to infer the parameters of DALSP models from CFSE distribution time-series data and to assess the dependency of the proliferation rates on factors such as the cell age. For this we introduce a statistical model linking predictions of the DALSP model to measured CFSE data and formulate the corresponding inverse problem. As the optimisation problem is nonlinear, we compare different optimisation procedures. Furthermore, we implement the first identifiability and uncertainty analysis methods for DALSP models. Combined with model selection methods, these methods facilitate an in-depth analysis of CFSE time-series data, which we illustrate for a published dataset.

3 Methods

To analyse CFSE data, we combine mechanistic and statistical models of the biological process and the measurement process. In the following we present the different ingredients as well as the inference methods. A more detailed description is provided in the *Supplementary*

Information.

Notation: We denote the set of non-negative real numbers by $\mathbb{R}_+ := [0, \infty)$ and the set of natural numbers with zero by $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$. The units used in the following equations are number of cells (cells), unit of concentration (UC), unit of fluorescence intensity (UFI) and unit of time (UT). For simplicity, we assume that age and time are measured in the same units, generalizations are straight forward.

3.1 DALSP model

The state variable of the DALSP model is the *joint number density of age, label concentration and cell division number*, $n(a, x, i|t)$ (in cells/UC/UT). Its dynamics are governed by a system of coupled 2-dimensional PDEs,

$$\begin{aligned} \frac{\partial n(a,x,i|t)}{\partial t} + \frac{\partial n(a,x,i|t)}{\partial a} + \frac{\partial (\nu(t,x) n(a,x,i|t))}{\partial x} \\ = -\left(\alpha_i(t,a) + \beta_i(t,a)\right) n(a,x,i|t) \end{aligned} \tag{1}$$

with initial conditions (ICs)

$$\begin{aligned} i = 0: &\quad n(a,x,0|0) \equiv n_0(a) p_0(x), \\ \forall i \geq 1: &\quad n(a,x,i|0) \equiv 0, \end{aligned} \tag{2}$$

and boundary conditions (BCs)

$$\begin{aligned} &i = 0: n(0,x,0|t) \equiv 0, \\ &i \geq 1: n(0,x,i|t) \equiv 4 \int_{\mathbb{R}_+} \alpha_{i-1}(t,a) n(a,2x,i-1|t) da. \end{aligned} \tag{3}$$

The i-th PDE describes the dynamics of cells with i divisions. The factorization of the initial condition, $n(a, x, 0|0) \equiv n_0(a)p_0(x)$, in initial age distribution $n_0(a)$ (in cells/UT) and initial label density $p_0(x)$ (in 1/UC), is biologically plausible as the labeling efficiency should not depend on the cell age. This factorization will allow for an efficient numerical solution algorithm.

In (1)-(3), $\alpha_i(t, a)$ and $\beta_i(t, a)$ denote the rates (in 1/UT) of cell division and cell death for cells with i divisions. The rate of cellular label degradation is denoted by $\nu(t, x) = -k(t)x$ (in 1/UT), with rate constant $k(t)$ (in 1/UT/UC). Accordingly, the following terms contribute to the temporal change of the density $n(a, x, i|t)$:

- $\partial(\nu(t, x)n(a, x, i|t))/\partial x$, change of label x with rate $\nu(t, x)$,
- $\partial(n(a, x, i|t))/\partial a$, increase of cell age a, and
- $-\left(\alpha_i(t, a) + \beta_i(t, a)\right) n(a, x, i|t)$, loss of cells from the i-th subpopulation due to cell division and due to cell death,

and possess units cells/UC/UT^2. The loss of cells from the $(i-1)$-th subpopulation due to cell division results in the birth of cells in the i-th subpopulation with age $a = 0$, defining the BCs (3). The BCs are obtained by integrating $\alpha_{i-1}(t, a)n(a, 2x, i-1|t)$ over the age. As the

cells double (factor 2) and the label distribution is rescaled due to the halving of the label concentration (factor 2), this integral is multiplied by a factor 4.

The rates as well as the initial conditions are usually unknown and have to be estimated from experimental data. Therefore, rates and initial conditions are parameterized in terms of a parameter vector $\theta \in \mathbb{R}^{n_\theta}$ (see *Results* section). For a more detailed statement of the model we refer to the *Supplementary Information, Section 1.1.1.*

3.2 Modeled and measured quantities

The number density $n(a, x, i|t)$ encodes the properties of the proliferating cell population. By marginalizing $n(a, x, i|t)$ over all cell ages and label concentrations, the number of cells which underwent i divisions,

$$N(i|t) = \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} n(a, x, i|t) dx da,$$

is obtained. The subsequent summation over the division number i yields the overall number of cells, $N(t) = \sum_{i \in \mathbb{N}_0} N(i|t)$. Marginalization over label concentrations and division number and subsequent normalization with the number of cells yields the age distribution,

$$p(a|t) = \frac{1}{N(t)} \sum_{i \in \mathbb{N}_0} \int_{\mathbb{R}_+} n(a, x, i|t) dx.$$

The label distribution, $p(x|t)$, is obtained by marginalizing over cell age and division number and normalizing with the overall cell number,

$$p(x|t) = \frac{1}{N(t)} \sum_{i \in \mathbb{N}_0} \int_{\mathbb{R}_+} n(a, x, i|t) da.$$

Proliferation assays provide information about the overall number of cells as well as the sum of label induced fluorescence (Figure 1B and C) and cellular background fluorescence. The label induced fluorescence, y (in UFI), is proportional to the label concentration, $y = cx$ with $x \sim p(x|t)$ and proportionality constant $c > 0$ (UFI/UC). The background fluorescence y_b is a random variable, $y_b \sim p_b(y_b)$ (in UFI), whose distribution (p_b) depends potentially on the biological system, the measurement procedure and technical factors. The distribution of the total measured fluorescence, $y_m = y + y_b$ (in UFI), obeys the convolution integral,

$$p(y_m|t) = \int_0^{y_m} p(y|t) p_b(y_m - y) dy,$$
$$\text{with} \qquad p(y|t) = \frac{1}{c} p\left(x = \frac{y}{c} \middle| t\right)$$

(see *Supplementary Information, Section 1.1.2*). In the presence of outliers, $p(y_m|t)$ can be mixed with an outlier distribution, $p_{\text{outliers}}(y_o)$ (see *Supplementary Information, Section 2.1*). The measurement of the fluorescence distribution $p(y_m|t)$ does not provide information about the absolute values of the concentration x. Any changes in c can be compensated by changes in the initial label distribution $p_0(x)$ and $k(t)$ (Hasenauer, 2013), rendering $c = 1$ (UFI/UC) a valid parameterization.

Commonly used measurement devices possess a finite resolution and collect interval censored samples from $p(y_m|t)$. The resulting binned snapshot data provides the number of cells

$\bar{H}_k^j$ observed in a bin j with intensity range I_j at time t_k. These counts $\{\bar{H}_k^j\}_{j=1}^J$ along with intervals $\{I_j\}_{j=1}^J$ provide a histogram. The probability $p(y_m \in I_j|t_k)$ of observing an individual cell at time point t_k in bin j is the integral of $p(y_m|t_k)$ over I_j. The overall cell count measured at time point t_k is denoted by $\bar{N}_k = \sum_{j=1}^J \bar{H}_k^j$.

3.3 Numerical simulation

To compute the number density $n(a, x, i|t)$ and further model properties, the DALSP model (1)-(3) is solved numerically. For this we exploit that the solution of the system of coupled two-dimensional PDEs (1)-(3) is factorable (Metzger *et al.*, 2012). The first factor is the solution to a system of coupled one-dimensional PDEs describing an age- and division-structured population. This solution is computed using an efficient iterative numerical scheme. The second factor is the solution to a set of decoupled one-dimensional PDEs describing the label distribution in cellular subpopulations with similar division number. This solution can be determined analytically. The factorization accelerates the numerical evaluation by several orders of magnitude compared to naive numerical methods. A similar decomposition approach is used to compute the sensitivities of $n(a, x, i|t)$ with respect to the parameters θ_i, $\partial n(a, x, i|t)/\partial\theta_i$. Given $n(a, x, i|t)$ and its sensitivities, the model properties and their derivatives are determined via numerical integration. The convolution integral defining the measured fluorescence intensity is evaluated using Fenton's approximation (Fenton, 1960). For further details we refer to the *Supplementary Information, Section 1.2 to 1.5.*

3.4 Parameter estimation and uncertainty analysis

We employed maximum likelihood and Bayesian parameter estimation to determine the unknown model parameters θ, with $\theta \in \Theta \subseteq \mathbb{R}^{n_\theta}$, from the collection of binned snapshot data $\mathcal{D} = \{\{\bar{H}_k^j\}_{j=1}^J, \bar{N}_k\}_{k=1}^K$. To account for measurement noise in the measured overall cell number and the histograms, the likelihood function

$$\mathbb{P}(\mathcal{D}|\theta) = \prod_{k=1}^{K} \mathbb{P}(\{\bar{H}_{t_k}^j\}_{j=1}^J|\theta)\mathbb{P}(\bar{N}_k|\theta)$$

is used (Hasenauer, 2013). The likelihood $\mathbb{P}(\{\bar{H}_{t_k}^j\}_{j=1}^J|\theta)$ of observing the histogram $\{\bar{H}_{t_k}^j\}_{j=1}^J$ follows a multinomial distribution with category probabilities $\{p(x_m \in I_j|t_k)\}_{j=1}^J$. The likelihood $\mathbb{P}(\bar{N}_k|\theta)$ of observing the overall population size $\bar{N}_k$ assumes a log-normally distributed measurement error. A detailed statement of the likelihood functions is provided in the *Supplementary Information, Section 2.1.* Prior and posterior distribution for Bayesian parameter estimation are denoted by $\pi(\theta)$ and $\pi(\theta|\mathcal{D})$, with $\pi(\theta|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)$. A log-uniform prior $\pi(\theta)$ is employed.

The maximum likelihood estimates θ^{ML} and maximum a posteriori parameter estimates θ^{MAP} are obtained by maximizing the respective objective functions. To improve the numerical evaluation and the optimiser convergence, we maximize the logarithms of the objective functions, e.g.

$$\theta^{\mathrm{ML}} = \arg\max_{\theta\in\Theta} \log \mathbb{P}(\mathcal{D}|\theta).$$

These nonlinear optimisation problems are solved using stochastic global optimisation (Weise, 2009) and multi-start local optimisation (Raue *et al.*, 2013). For stochastic global optimisation

the particle-swarm pattern search method PSwarm (Vaz and Vicente, 2007) is employed. For multi-start local optimisation, parameters are drawn from the parameter domain Θ and used as starting points for the local optimiser. For local optimisation the MATLAB routine `fmincon.m` is used with and without user-supplied gradients. The user supplied gradients are computed using the forward sensitivity equations of the DALSP model (see *Supplementary Information, Section 1.2 and 2.1.*)

The estimated parameter values reveal properties of the proliferation dynamics. To assess the parameter identifiabilities and uncertainties we calculate profile likelihoods (Raue *et al.*, 2009) using our in-house Parameter Estimation Toolbox (PESTO) and sample the posterior distribution using the Delayed Rejection Adaptive Metropolis (DRAM) sampler developed by Haario *et al.* (2006). The posterior sample is employed to study parameter correlations and prediction uncertainties (Hug *et al.*, 2013). For details on the optimiser and sampler settings we refer to the *Supplementary Information, Section 2.2-2.4.*

3.5 Hypothesis testing

Competing hypotheses regarding the mode of proliferation can be encoded in the rates $\alpha(t,a)$ and $\beta(t,a)$. Each of the resulting models is assessed using the Akaike information criterion (AIC),

$$\mathrm{AIC} = -2\log\mathbb{P}(\mathcal{D}|\theta^{\mathrm{ML}}) + 2n_\theta,$$

and the Bayesian information criterion (BIC)

$$\mathrm{BIC} = -2\log\mathbb{P}(\mathcal{D}|\theta^{\mathrm{ML}}) + \log(n_\mathcal{D})n_\theta,$$

in which the maximum likelihood estimate and the number of parameters for the model of interest are denoted by θ^{ML} and n_θ. The number of independent data points is denoted by $n_\mathcal{D}$. AIC and BIC account for the likelihood of the data and penalize model complexity. Low AIC and BIC values are favorable. We consider a difference > 10 between AIC/BIC values of different models as substantial (Burnham and Anderson, 2002).

3.6 Implementation

All implementations are available as MATLAB toolbox from GitHub (http://janhasenauer.github.io/ShAPE-DALSP/). As the MATLAB Symbolic Math Toolbox is used to construct the models and sensitivity equations, a variety of alternative rates $\alpha_i(t,a)$, $\beta_i(t,a)$ and $\nu(t,x) = -k(t)x$ can be analysed easily.

4 Application

We illustrate the proposed model-based quantification method by studying T lymphocyte proliferation. T lymphocytes are part of the adaptive immune system and their pool expands upon pathogen recognition. This expansion is frequently monitored using CFSE labeling (Hawkins *et al.*, 2007a)

4.1 Experimental data and mechanistic model

We considered T lymphocyte proliferation upon treatment with antibodies against CD3 and CD28 receptors. Experimental data for this setting has been collected by Luzyanina *et al.* (2007b) and analysed in a series of studies (Banks *et al.*, 2010, 2011; Hasenauer, 2013; Luzyanina *et al.*, 2007b; Thompson, 2012). This multitude of preceding studies underlines the importance of the dataset and renders it ideal for the evaluation of our approach.

In this study, an ensemble of DALSP models is used to study the CFSE data collected on days 1-5 post treatment. To capture the observed heterogeneity, the initial CFSE distribution (on day 1), $p_0(x)$, is modeled by a weighted sum of two log-normal distributions and the distribution of the autofluorescence, $p_a(x_a)$ is modeled by a single log-normal distribution. After considering different alternatives, we set initial cell age to zero, such that the initial condition is given by $n_0(a) = N_0\delta(a)$. N_0 is the initial number of cells and δ denotes the Dirac delta distribution. We considered rates of cell division which are constant, age- and/or division number-dependent, i.e.,

$$\alpha_i(a) = \begin{cases} k_\alpha & \rightarrow \text{constant} \\ k_{\alpha,i} & \rightarrow \text{division number dependent} \\ \frac{k_\alpha a^{n_\alpha}}{K_\alpha^{n_\alpha} + a^{n_\alpha}} & \rightarrow \text{age dependent} \\ \frac{k_{\alpha,i} a^{n_\alpha}}{K_\alpha^{n_\alpha} + a^{n_\alpha}} & \rightarrow \text{division number and age dependent,} \end{cases}$$

and the same holds for the rate of cell death, $\beta_i(a)$. The age-dependence is modeled by Hill-type functions with maximal rates $k_{\alpha,i}$ and $k_{\beta,i}$ which can depend on the division number i. Similar to previous publications the intracellular CFSE degradation is described using Gompertz decay, $k(t) = -k_{\text{deg}} \exp(-c_{\text{deg}} t)$ (Banks *et al.*, 2013b). The constants used to model the rates, e.g., $k_{\alpha,i}$, $k_{\beta,i}$, k_{deg} and c_{deg}, and the parameters of the initial conditions are part of the parameter vector θ and estimated from the data. Lower and upper bounds for these parameters are reported and discussed in the *Supplementary Information, Section 3.1.*

The combinations of the different hypotheses regarding $\alpha_i(a)$ and $\beta_i(a)$ give rise to 16 model alternatives with 12 to 29 parameters. We denote these models with $\mathcal{M}_1$ through $\mathcal{M}_{16}$ and their properties are summarized in Table 1. In the following sections, these model alternatives are compared and used to interpret the CFSE time-series data.

Table 1: Number of parameters, negative log-likelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC) for the 16 model alternatives. Rows two to five indicate the dependencies considered in the different models. Dependencies are indicated using fnc, e.g. $\alpha = \text{fnc}(a)$ indicates that α depends on a. The values for AIC and BIC are for the maximum likelihood estimate found after 250 runs of a deterministic local optimiser exploiting forward sensitivity equations.

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$	$\mathcal{M}_8$	$\mathcal{M}_9$	$\mathcal{M}_{10}$	$\mathcal{M}_{11}$	$\mathcal{M}_{12}$	$\mathcal{M}_{13}$	$\mathcal{M}_{14}$	$\mathcal{M}_{15}$	$\mathcal{M}_{16}$
$\alpha = \text{fnc}(a)$		×		×		×		×		×		×		×		×
$\alpha = \text{fnc}(i)$			×	×			×	×			×	×			×	×
$\beta = \text{fnc}(a)$					×	×	×	×					×	×	×	×
$\beta = \text{fnc}(i)$									×	×	×	×	×	×	×	×
n_θ	12	14	18	20	14	16	20	22	19	21	25	27	21	23	27	29
$-\log \mathbb{P}$ ($\times 10^4$)	2.242	1.521	2.185	1.476	2.089	1.521	2.016	1.476	2.176	1.508	2.138	1.350	2.033	1.497	1.987	1.345
AIC ($\times 10^4$)	4.486	3.044	4.374	2.956	4.180	3.045	4.036	2.956	4.356	3.021	4.281	2.705	4.070	2.998	3.980	2.696
BIC ($\times 10^4$)	4.497	3.058	4.391	2.974	4.194	3.060	4.056	2.977	4.374	3.041	4.305	2.730	4.090	3.020	4.006	2.723

4.2 Comparison of optimisation methods

We estimated the parameters of all 16 model alternatives using the particle-swarm optimisation algorithm implemented in PSwarm and the multi-start local optimisation method implemented in PESTO. The latter performs individual deterministic optimisations using gradients obtained by finite differences or sensitivity equations. For all optimisation methods we assessed the percentage of 'converged starts' using the likelihood ratio test with a significance level of 0.05 (see *Supplementary Information, Section 2.4*). This is a weak, statistically motivated measure of convergence and can be checked easily also for complex problems. Beyond converged starts, we considered the distribution of computation times for individual starts as well as the average computation time per converged start. The latter is computed by dividing the overall computation time by the number of converged starts (see *Supplementary Information, Section 2.4*). This yields a measure for the overall optimiser efficiency.

The results for 250 runs of PSwarm, 250 runs of the deterministic optimisers used in multi-start local optimisation and 250 random parameter samples for different model alternatives are depicted in Figure 2 and Figure S2. For $\mathcal{M}_2$ all optimisers achieve a significant improvement compared to random samples. For $\mathcal{M}_2$ the percentages of converged starts were for all optimisers comparable (Figure 2A), but deterministic optimisers using sensitivity equations were two orders of magnitude faster than PSwarm (Figure 2B and C) and the only algorithms able to reproduce the global optimum. For model alternatives with more parameters, the difference in the performance of multi-start methods and PSwarm becomes even more apparent (Figure S2). Hence, although PSwarm is known to outperform most available global optimisation algorithms (Vaz and Vicente, 2007), we found that multi-start local optimisation yields better results for DALSP models.

4.3 Evaluation of age-dependent proliferation rates

Model $\mathcal{M}_1$ to $\mathcal{M}_{16}$ describe different dependencies of the proliferation rates on cell age and division number. A comparison of the measured distributions and cell counts (Figure 3A) with the best fits for these model hypotheses is provided in Figure 3, S4 and S5. We found that already model $\mathcal{M}_1$, which assumes constant rates of cell division and cell death, fits a large fraction of the data well (Figure 3B). Model $\mathcal{M}_2$, which accounts for age-dependence of the division rate, already provides a good visual agreement of model and data (Figure 3C). The fit tends to improve further as additional degrees of freedom (i.e., model parameters) are introduced (Figure 3D). To our surprise, we found only a weak correlation between the log-likelihood and the number of parameters of a model (Figure 4A). In contrast, age-dependence of the division rate separates the models in those with high log-likelihood values ($\rightarrow$ good fits) and those with low log-likelihood values ($\rightarrow$ bad fits). Taken together, this indicates that missing age-dependence of the rates of cell division cannot be compensated by increasing the number of parameters and the model complexity.

To assess the importance of individual dependencies of proliferation rates we computed the AIC and the BIC values (Table 1). Both information criteria provided identical rankings and revealed the model hierarchy visualized in Figure 4B and C. The best eight models were those with age-dependence of the rates of cell division, including $\mathcal{M}_2$. The best four models were those with age- and division-dependent rates of cell division. The best two models were those possessing in addition a dependence of the rates of cell death on the division number.

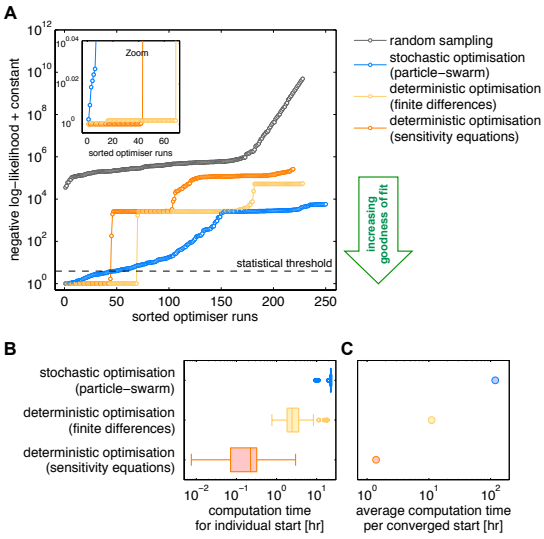

Figure 2: Performance of different optimisation methods for model $\mathcal{M}_2$. (**A**) Negative log-likelihoods for 250 runs of deterministic local and global optimisers and 250 randomly sampled parameter values. Missing points indicate failed objective function evaluations and optimisation runs. The dashed line indicates the significance threshold for converged starts. (**B**) Box plot of the computation time per optimisation run. (**C**) Average computation time per converged start.

In the best model the rates of cell division and cell death depended on cell age and division number.

Previous studies relying on LSP and DLSP models assumed time- and/or CFSE concentration-dependent division rates. The estimates of these division rates were multi-modal, difficult to interpret and possessed many unknowns (see, e.g., (Banks *et al.*, 2010, Figure 8) and (Thompson, 2012, Figure 3.10)). The DALSP models proposed here possess fewer parameters and the age-dependent rates of cell division are interpretable, e.g. in terms of inter-division times (Metzger *et al.*, 2012). Furthermore, age-dependent rates provide a direct link to the cell cycle-dependent gene expression during T lymphocyte proliferation and differentiation (Bird *et al.*, 1998). Accordingly, DALSP models facilitated an accurate description and a meaningful interpretation of CFSE time-series data in terms of model parameters.

4.4 Analysis of parameter and prediction uncertainties

As estimated parameters and predicted model properties are potentially non-identifiable from available CFSE time-series data, we assessed their uncertainties using profile likelihoods and MCMC sampling. We evaluated the optimisation procedures and found that only deterministic optimisation with sensitivity equations yields accurate profile likelihoods. Furthermore, despite the use of state-of-the-art adaptive MCMC sampling procedures, random initialization of the starting point did not yield a sample within 30 CPU days that passes the Geweke test (Brooks and Roberts, 1998), a convergence diagnostic. In contrast, initialization at the MAP estimate resulted after only 2 CPU days in a sample which passes the Geweke test. This demonstrated the importance of reliable and efficient optimisation methods for uncertainty

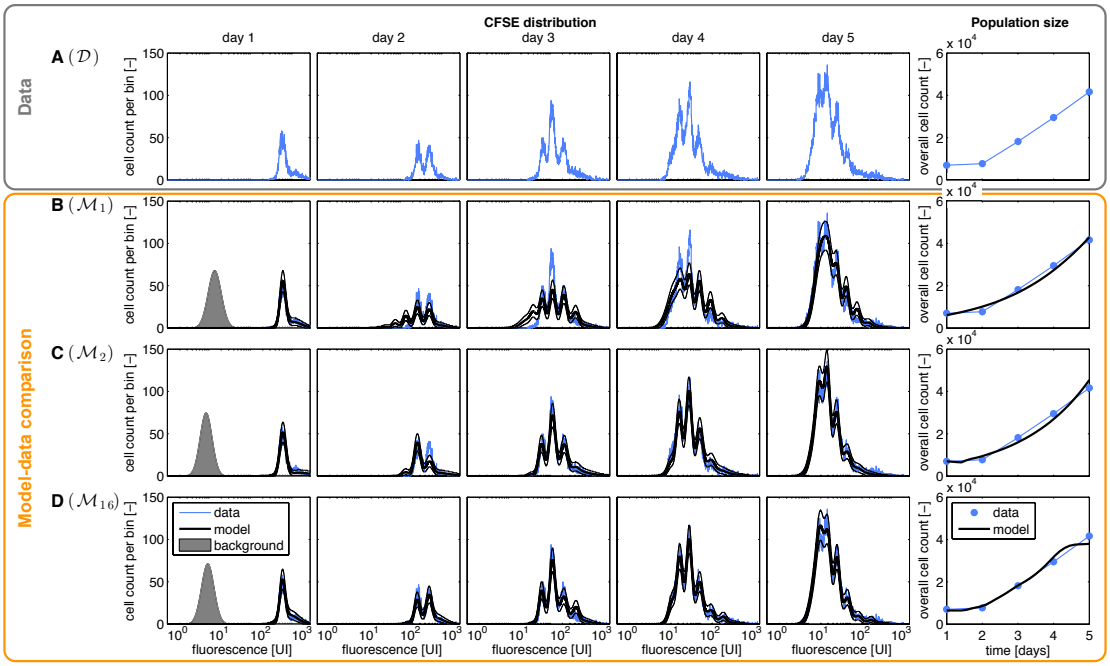

Figure 3: Comparison of measured CFSE distributions (**A**, left) and measured overall cell counts (**A**, right) with the best fits of model $\mathcal{M}_1$ (**B**), $\mathcal{M}_2$ (**C**), and $\mathcal{M}_{16}$ (**D**). The region in between the fine black lines (═) indicates the 90% confidence interval (5-th to 95-th percentile) of the bin counts for the particular number of measured cells. A coverage/overplotting of the experimental data by the confidence interval indicates a good fit.

analysis. The results of the uncertainty analysis for $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_{16}$ are visualized in Figure S7-S12.

For $\mathcal{M}_2$, the simplest model which provides a good visual agreement of model and data, we found that most parameters are practically identifiable (Figure S8). The maximum rate of cell division, k_α, as well as the age at which the half maximum value of the division rates is reached, K_α, were tightly constrained (Figure 5A). For the Hill coefficient, n_α, we considered the range of 10^{-6} to 10^2 and found a lower bound of 45.4. Accordingly, the age-dependent division rates, $\alpha_i(a)$, is close to the scaled Heaviside step function $k_\alpha \mathrm{h}(a - K_\alpha)$ which would be reached for $n_\alpha \to \infty$ (Figure 5B). This indicates that the upper bound for n_α chosen in the study, $n_\alpha \leq 10^2$, does not influence the estimation results significantly. For the estimated division rate, the inter-division time (in the absence of cell death) is similar to a shifted exponential distribution (Figure 5C). This is the distribution assumed in the Smith-Martin model (Smith and Martin, 1973). The rate of cell death, $\beta(a) = k_\beta$, was significantly smaller than the maximal rate of cell division, k_α (Figure 5D). As the rate of cell death is constant, the time to cell death (in the absence of cell division) is exponentially distributed (Figure 5D). The estimated parameters of the rates of cell division and cell death are correlated (Figure S11). The good agreement of profile likelihoods and histograms obtained using Markov chain Monte Carlo sampling underpin the functioning of the methods. Confidence intervals for all parameters of model $\mathcal{M}_2$ are provided in Table 2.

Beyond the quantification of parameter uncertainties, we assessed model predictions. As

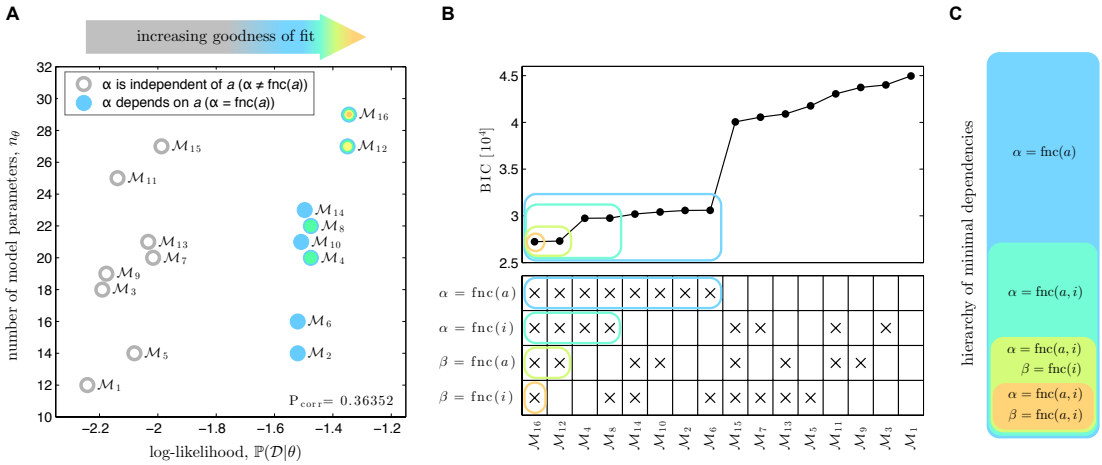

Figure 4: Model selection results for proliferation models $\mathcal{M}_1$ - $\mathcal{M}_{16}$. (**A**) Optimal log-likelihood value vs. number of parameter for the individual models. (**B**) Ordering of models implied by BIC values, starting with the best model. Crosses in the table indicate dependencies of the rate of cell division, α and cell death, β. (**C**) Visualization of the hierarchy of models according to the minimal set of dependencies, which are color-coded consistently across all subplots. (**A**,**B**,**C**) Visualization of AIC instead of BIC yields the same results as AIC and BIC values differ merely in the third significant digit.

CFSE time-series data are most frequently used to determine the number of cells with a certain number of divisions, $N(i|t)$, we used samples from the posterior distribution $\pi(\theta|\mathcal{D})$ obtained by MCMC to quantify this property and its uncertainty. The results for $\mathcal{M}_2$ (Figure 5C) revealed that the uncertainties are relatively small. A second property providing insights into the proliferation dynamics is the age distribution. While this property could not be assessed with LSP and DLSP models, the DALSP model revealed its structure. Due to the continuous renewal, cells tended to be young with a maximum at $a = 0\,\mathrm{d}$. Between $a = 0\,\mathrm{d}$ and $a = 0.41\,\mathrm{d}$ the density slowly decreases, mostly due to cell death. At $a = 0.41\,\mathrm{d} \approx K_\alpha$ cell division sets in and results in an accelerated decline. Although CFSE time-series data do not directly provide information about the cell age distribution, DALSP models could be used to quantify it accurately, which is indicated by narrow credible intervals. This is also true for other properties such as intracellular CFSE degradation.

5 Conclusion

In vivo and in vitro proliferation assays using CFSE require accurate tools for quantifying biologically meaningful parameters (Hawkins *et al.*, 2007a). To meet this requirement, we propose to use DALSP models which account for the age of individual cells. In contrast to the frequently used LSP and DLSP models, DALSP models allow for non-exponential inter-division time. To exploit DALSP models for the analysis of CFSE time-series data, we derived sensitivity equations for gradient calculation and we developed a tailored numerical scheme exploiting the hierarchical structure of DALSP models.

As an accurate model-based analysis of experimental data requires reliable inference, we

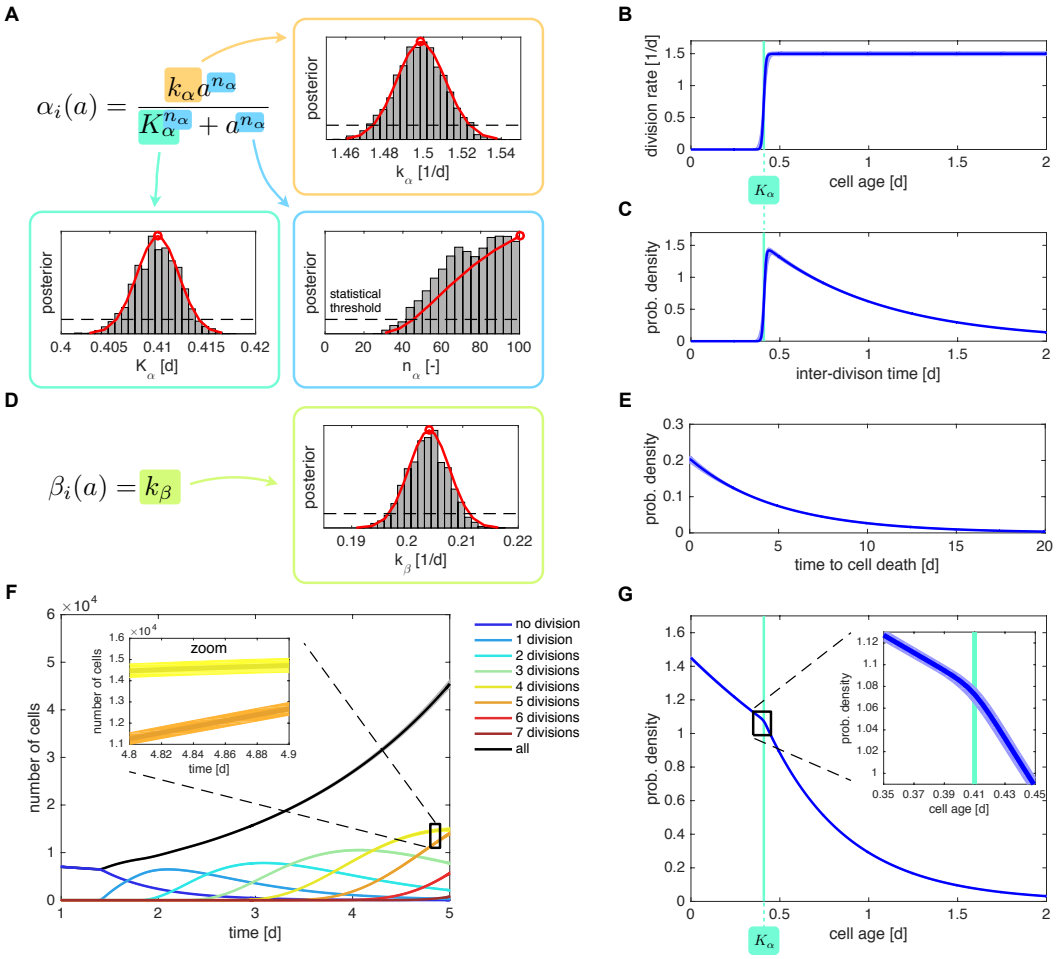

Figure 5: Parameter and prediction uncertainties for model $\mathcal{M}_2$. Parameterization of (**A**) age-dependent division rate and (**D**) death rates along with the uncertainty of the corresponding model parameters. Histograms of the MCMC samples (gray boxes) and profile likelihoods the individual parameters (red line) are depicted along with the statistical threshold for the 95% confidence intervals (dashed black line). Median (lines) and 99% credible interval (semi-transparent areas) for (**B**) the age-dependent rate of cell division, (**C**) the distribution of the inter-division time (in the absence of cell death), (**E**) the distribution of the time to cell death (in the absence of cell division), (**F**) the size of the subpopulations and the overall population and (**G**) the age distribution on day 5. Uncertainty assessment reveals that parameter and prediction uncertainties are small. (**F**,**G**) Zooms are provided to visualize the uncertainties.

compared different optimisation algorithms with respect to convergence and computation time. Our comparison, which is to the best of our knowledge the first comparison of this kind for structured population models, revealed that multi-start local optimisation outperforms stochastic global optimisation for this class of PDE models. This confirms results by Raue *et al.* (2013) for ODE models. Indeed, for some DALSP models the evaluated stochastic global optimisers did not even converge if a large number of function evaluations ($> 10^5$)

Table 2: Parameter estimates and confidence intervals for model $\mathcal{M}_2$.

Name	ML estimate	95% confidence interval	Unit
k_α	1.499	$(1.474, 1.524)$	d^{-1}
n_α	100.0	$(45.39, \geq 100)$	-
K_α	0.410	$(0.406, 0.414)$	d
k_β	0.204	$(0.197, 0.211)$	d^{-1}
k_{deg}	0.190	$(0.184, 0.196)$	d^{-1}
c_{deg}	0.076	$(0.057, 0.095)$	d^{-1}
μ_{noise}	2.150	$(2.132, 2.169)$	-
σ_{noise}	0.333	$(0.321, 0.344)$	-
N_0	7012	$(6902, 7123)$	cells
$r_{\mathrm{x},1}$	0.746	$(0.735, 0.757)$	-
$\mu_{\mathrm{x},1}$	6.402	$(6.396, 6.407)$	-
$\mu_{\mathrm{x},2}$	7.302	$(7.270, 7.335)$	-
$\sigma_{\mathrm{x},1}$	0.177	$(0.174, 0.179)$	-
$\sigma_{\mathrm{x},2}$	0.741	$(0.724, 0.758)$	-

was allowed. For the same models, multi-start local optimisation with accurate gradients provided reproducible results. This underpined the importance of using sensitivity equations, which have so far not been described for LSP and DLSP models. The latter raising questions regarding the reliability of previous estimation results.

The efficient deterministic optimisers were used to maximize a likelihood function which accounts for the stochasticity of the acquisition process. This likelihood function is statistically more meaningful than least-squares type objective functions (Banks *et al.*, 2010, 2011). We employed the proposed likelihood function for model selection and found evidence for a dependence of the division rate on the cell age. More precisely, the Hill-type division rate with a high degree, $n_\alpha \gg 1$, indicated a lower bound for the inter-division time, which could be interpreted as a minimal length of the cell cycle. This seems biologically more plausible than the time- and CFSE concentration-dependent division rates proposed in previous publications (Banks *et al.*, 2010, 2011; Luzyanina *et al.*, 2007b). Furthermore, age-dependent rates of cell division have been reported for similar biological systems using time-lapse microscopy data (Duffy *et al.*, 2012; Shokhirev *et al.*, 2015).

Beyond model selection, we provided the first detailed Bayesian uncertainty analysis for structured population models. This was challenging due to the computational complexity of PDE models and the number of unknown parameters. We established computational feasibility by exploiting tailored numerical methods and an initialization using optimisation results. The posterior samples obtained using MCMC methods revealed a low level of uncertainties for latent properties, such as the age distribution and the parameters of the cell division rate. In addition, the marginal distribution of the parameters are consistent with the profile likelihoods, which substantiated the results further.

All the modeling, simulation, parameter estimation and model selection methods used in this publication are implemented in the open-source MATLAB toolbox ShAPE-DALSP. The availability of the code will facilitate the application of the method and simplify the

development of extensions, e.g., towards multiple cell-types (Schittler *et al.*, 2012), asymmetric cell division (Banks *et al.*, 2015; Bocharov *et al.*, 2013; Kapraun, 2014; Luzyanina *et al.*, 2014) and alternative DALSP models (Banks *et al.*, 2013a, 2014, 2015; Kapraun, 2014; Luzyanina *et al.*, 2014). In particular the implementation of a cyton-based model (Banks *et al.*, 2013a, 2014, 2015; Kapraun, 2014) would be interesting as these models allow for a fraction of non-dividing cells. In addition to extensions, alternative parameterizations can be included, e.g., a parameterization of the probability density functions for a cell to divide and die at age a (as used in the cyton models) in contrast to a parameterization of the age-dependent rate of cell division and cell death. The incorporation of the features in ShAPE-DALSP would allow for an even broader spectrum of applications and improve user-friendliness.

In summary, this study presents a novel model-based analysis method for CFSE time-series data. Besides a statistical model, we present findings regarding optimiser performance and uncertainty analysis. These findings and the methods we developed can be easily transferred to other structured population models and might also be applicable to other types of population balance models. Accordingly, this study will help to make the most of CFSE time-series data and other data requiring cell-cycle corrections.

Funding

This work has been supported by the Postdoctoral Fellowship Program (PFP) of the Helmholtz Zentrum München.

References

Banks, H. T., Sutton, K. L., Thompson, W. C., Bocharov, G., Roose, D., Schenkel, T., and Meyerhans, A. (2010). Estimation of cell proliferation dynamics using CFSE data. *Bull. Math. Biol.*, **73**(1), 116–150.

Banks, H. T., Sutton, K. L., Thompson, W. C., Bocharov, G., Doumic, M., Schenkel, T., Argilaguet, J., Giest, S., Peligero, C., and Meyerhans, A. (2011). A new model for the estimation of cell proliferation dynamics using CFSE data. *J. Immunological Methods*, **373**(1–2), 143–160.

Banks, H. T., Kapraun, D. F., Thompson, W. C., Peligero, C., Argilaguet, J., and Meyerhans, A. (2013a). A novel statistical analysis and interpretation of flow cytometry data. *J. Biol. Dyn.*, **7**(1), 96–132.

Banks, H. T., Choi, A., Huffman, T., Nardini, J., Poag, L., and Thompson, W. C. (2013b). Quantifying CFSE label decay in flow cytometry data. *Appl. Math. Lett.*, **26**(5), 571–577.

Banks, H. T., Kapraun, D. F., Link, K. G., Thompson, W. C., Peligero, C., Argilaguet, J., and Meyerhans, A. (2014). Analysis of variability in estimates of cell proliferation parameters for cyton-based models using cfse-based flow cytometry data. *J. Inverse Ill-Posed Probl.*, **23**(2), 135–171.

Banks, H. T., Kapraun, D. F., Peligero, C., Argilaguet, J., and Meyerhans, A. (2015). Evaluating the importance of mitotic asymmetry in cyton-based models for cfse-based flow cytometry data. *Int. J. Pure Appl. Math.*, **100**(1), 131–156.

Bernard, S., Pujo-Menjouet, L., and Mackey, M. C. (2003). Analysis of cell kinetics using a cell division marker: mathematical modeling of experimental data. *Biophys. J.*, **84**(5), 3414–3424.

Bird, J. J., Brown, D. R., Mullen, A. C., Moskowitz, N. H., Mahowald, M. A., Sider, J. R., Gajewski, T. F., Wang, C.-R., and Reiner, S. L. (1998). Helper T cell differentiation is controlled by the cell cycle. *Immunity*, **9**(2), 229–237.

Bocharov, G., Luzyanina, T., Cupovic, J., and Ludewig, B. (2013). Asymmetry of cell division in CFSE-based lymphocyte proliferation analysis. *Front. Immunol.*, **4**(264).

Brooks, S. P. and Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Stat. Comp.*, **8**(4), 319–335.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer, New York, NY, 2nd edition.

De Boer, R. J. and Perelson, A. S. (2013). Quantifying t lymphocyte turnover. *J. Theor. Biol.*, **327**, 45–87.

De Boer, R. J., Ganusov, V. V., Milutinoviò, D., Hodgkin, P. D., and Perelson, A. S. (2006). Estimating lymphocyte division and death rates from CFSE data. *Bull. Math. Biol.*, **68**(5), 1011–1031.

Duffy, K. R., Wellard, C. J., Markham, J. F., Zhou, J. H. S., Holmberg, R., Hawkins, E. D., Hasbold, J., Dowling, M. R., and Hodgkin, P. D. (2012). Activation-induced B cell fates are selected by intracellular stochastic competition. *Science*, **335**(6066), 338–341.

Fenton, L. F. (1960). The sum of lognormal probability distributions in scatter transmission systems. *IRE Trans. Commun. Syst.*, **8**(1), 57—67.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Stat. Comp.*, **16**(4), 339–354.

Hasenauer, J. (2013). *Modeling and parameter estimation for heterogeneous cell populations*. Ph.D. thesis, University of Stuttgart.

Hasenauer, J., Schittler, D., and Allgöwer, F. (2012). Analysis and simulation of division- and label-structured population models: A new tool to analyze proliferation assays. *Bull. Math. Biol.*, **74**(11), 2692–2732.

Hawkins, E. D., Hommel, M., Turner, M. L., Battye, F. L., Markham, J. F., and Hodgkin, P. D. (2007a). Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. *Nat. Protoc.*, **2**(9), 2057–2067.

Hawkins, E. D., Turner, M. L., Dowling, M. R., van Gend, C., and Hodgkin, P. D. (2007b). A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proc. Natl. Acad. Sci. U S A*, **104**(12), 5032–5037.

Hug, S., Raue, A., Hasenauer, J., Bachmann, J., Klingmüller, U., Timmer, J., and Theis, F. J. (2013). High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. *Math. Biosci.*, **246**(2), 293–304.

Kapraun, D. F. (2014). *Cell proliferation models, CFSE-based flow cytometry data, and quantification of uncertainty*. Phd thesis, North Carolina State University, Raleigh, North Carolina, USA.

Luzyanina, T., Mrusek, S., Edwards, J., Roose, D., Ehl, S., and Bocharov, G. (2007a). Computational analysis of CFSE proliferation assay. *J. Math. Biol.*, **54**(1), 57–89.

Luzyanina, T., Roose, D., Schenkel, T., Sester, M., Ehl, S., Meyerhans, A., and Bocharov, G. (2007b). Numerical modelling of label-structured cell population growth using CFSE distribution data. *Theor. Biol. Med. Model.*, **4**, 26.

Luzyanina, T., Cupovic, J., Ludewig, B., and Bocharov, G. (2014). Mathematical models for CFSE labelled lymphocyte dynamics: asymmetry and time-lag in division. *J. Math. Biol.*, **69**(6–7), 1547–1583.

Lyons, A. and Parish, C. (1994). Determination of lymphocyte division by flow cytometry. *J. Immunol. Methods.*, **171**(1), 131–137.

Metzger, P., Hasenauer, J., and Allgöwer, F. (2012). Modeling and analysis of division-, age-, and label-structured cell populations. In *Proceedings of 9th International Workshop on Computational Systems Biology*, pages 55–58, Ulm, Germany.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinf.*, **25**(25), 1923–1929.

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U., and Timmer, J. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, **8**(9), e74335.

Schittler, D., Hasenauer, J., and Allgöwer, F. (2011). A generalized population model for cell proliferation: Integrating division numbers and label dynamics. In *Proceedings of 8th International Workshop on Computational Systems Biology*, pages 165–168, Zürich, Switzerland.

Schittler, D., Hasenauer, J., and Allgöwer, F. (2012). A model for proliferating cell populations that accounts for cell types. In *Proceedings of 9th International Workshop on Computational Systems Biology*, pages 79–82, Ulm, Germany.

Schroeder, T. (2011). Long-term single-cell imaging of mammalian stem cells. *Nat. Methods*, **8**(4), 30–35.

Shokhirev, M. N., Almaden, J., Davis-Turak, J., Birnbaum, H. A., Russell, T. M., Vargas, J. A. D., and Hoffmann, A. (2015). A multi-scale approach reveals that NF-κB cRel enforces a B-cell decision to divide. *Mol. Syst. Biol.*, **11**(783).

Smith, J. A. and Martin, L. (1973). Do cells cycle? *Proc. Natl. Acad. Sci. U S A*, **70**(4), 1263–1267.

Thompson, W. C. (2012). *Partial differential equation modeling of flow cytometry data from CFSE-based proliferation assays*. Ph.d. thesis, North Carolina State University.

Vaz, A. and Vicente, L. (2007). A particle swarm pattern search method for bound constrained global optimization. *J. Global Optim.*, **39**(2), 197–219.

Weise, T. (2009). Global optimization algorithms: Theory and application. ebook, Nature Inspired Computation and Applications Laboratory (NICAL), University of Science and Technology, China.