AMSI VACATION RESEARCH SCHOLARSHIPS 2019–20

EXPLORE THE MATHEMATICAL SCIENCES THIS SUMMER

Advanced Bayesian Statistical Inference Methods for Simulation-Based models in Cell Biology

Michael Carr

Supervised by Assoc. Prof. Chris Drovandi and Prof. Mathew Simpson

Queensland University of Technology

Vacation Research Scholarships are funded jointly by the Department of Education and the Australian Mathematical Sciences Institute.

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government Department of Education

Contents

Abstract

Understanding the parameters which govern the cell cycle and spread of cells has wide impacts in many areas. One being Cancer research, where drug treatments can be more effective during different phases of the cell cycle. This project develops a parameter estimation method for a stochastic model of melanoma skin cells utilising fluorescent ubiquitination-based cell cycle indicator (FUCCI) technology to visualise the cell cycle. Investigations are based on simulated data. Using the Gillespie algorithm, the simulation model simulates a 2D hexagonal lattice random walk which is dependent on the cell proliferation rates and motility rates. Using Approximate Bayesian Computation (ABC) methods, we estimate the posterior distribution for the proliferation rates and investigate the effectiveness of the summary statistics used.

Introduction

This project explores computational statistic applications in cell biology. More specifically, we investigate algorithms to estimate parameters for cell proliferation utilising a cell invasion model to simulate and draw inference from. The cell invasion model explored involve moving cell fronts driven by cell migration and cell proliferation on a 2D hexagonal lattice (see Figure 1 (b)). Commonly, cell proliferation is conceptualised as a sequence of 4 phases: gap 1 (G1), synthesis (S), gap 2 (G2), mitotic (M) (Haass, 2014).

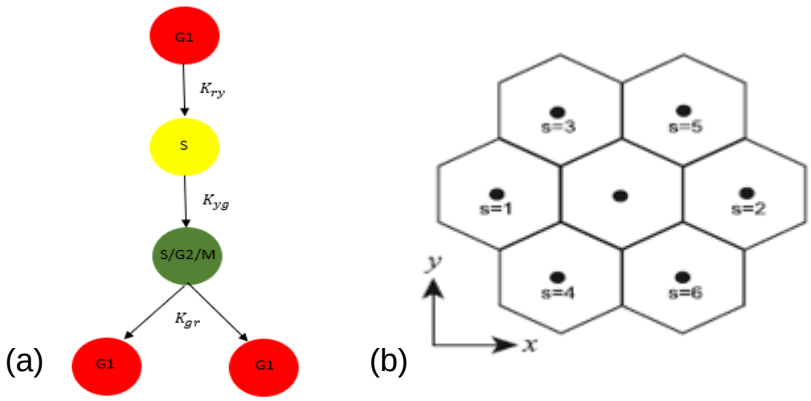

Figure 1: (a) Schematic showing the progression of cell cycle for FUCCI. (b) Hexagonal Lattice (Simpson et al, 2018)

Recent improvements in technology have enabled us to visualise the effects of cell proliferation and migration through different phases in the cell cycle using fluorescent ubiquitination-based cell cycle indicator (FUCCI) technology. FUCCI involves two fluorescent probes in the cell which emit a red colour when the cell is in the G1 phase and green when in the S/G2/M phase; cells in the early S phase will appear yellow (see Figure

1 (a)). FUCCI technology is becoming more important in cancer research because many drug treatments can be more effective during different phases of the cell cycle (Haass & Gabrielli, 2017). This involves understanding the underlying variables which influence the cell cycle.

However, often cell invasion models have intractable likelihood functions; this project develops a parameter estimation method for a stochastic model of melanoma skin cells using likelihood-free Bayesian computation methods. The model depends on 6 parameters (red, yellow and green proliferation and motility rates respectively); proliferation rates: Kry, Kyg, Kgr and motility rates: Pr, Py, Pg. We develop algorithms and explore summary statistics for attaining the marginal posterior distribution for the proliferation parameters and leave the motility parameters for future work.

Statement of Authorship

The workload was divided as follows:

- Michael Carr implemented the SMC ABC algorithm (and other ABC algorithms); altered the simulation model to include left and right, and random initial seeds; interpreted and reported results; and wrote the report.
- Chris Drovandi developed the SMC ABC algorithm used, supervised the work, proofread this report
- Mathew Simpson developed the simulation model used, supervised the work, proofread this report

Methods

Simulation model

The simulation model, which was developed by Simpson et al. is a discrete exclusion process based random walk model on a hexagonal lattice which models cell proliferation and migration on a 2D lattice. Each lattice site is assigned a unique value k ϵ [1 K] and a cartesian coordinate, (x, y). Each lattice site is associated with a set of six neighbouring lattice sites indexed s = 1,2,..,6 (see Figure 1 (b))). To mimic the FUCCI technology the simulated population is composed of three subpopulations: red, yellow and green, which relate to G1, early S, and S/G2/M phases respectively. Each subpopulation undergoes a random walk with exclusion by utilising the Gillespie algorithm.

To simulate cell migration, each cell subpopulation red, yellow and green undergoes a nearest neighbour unbiased random walk with motility rates Pr, Py and

Pg respectively. However, if the chosen neighbouring site is occupied then the migration event will not take place.

To simulate cell proliferation, it was first assumed that transition events were unaffected by crowding. Otherwise, red cells are allowed to transition into yellow cells at a rate Kry, yellow to green with rate Kyg, and green to two red daughter cells with rate Kgr. However, for the green to red transition to take place, the occupancy status of a randomly chosen nearest neighbour must be vacant. While there are multiple ways to handle the event of a non-vacant cell, Simpson et al. abort the event - leaving the green cell unchanged.

One useful question that can be answered from this project is: does the effectiveness of the summary statistic change if you conduct the experiment differently? This project explores the effect of experimental design on the effectiveness of the summary statistics. To explore the effect of experimental design on the effectiveness of the summary statistics, the simulation model was adapted to allow for the initial seeds: middle, left, left and right, and random (10% density) (see Figure 2). Besides the initial seeds, the model will behave as described previously.

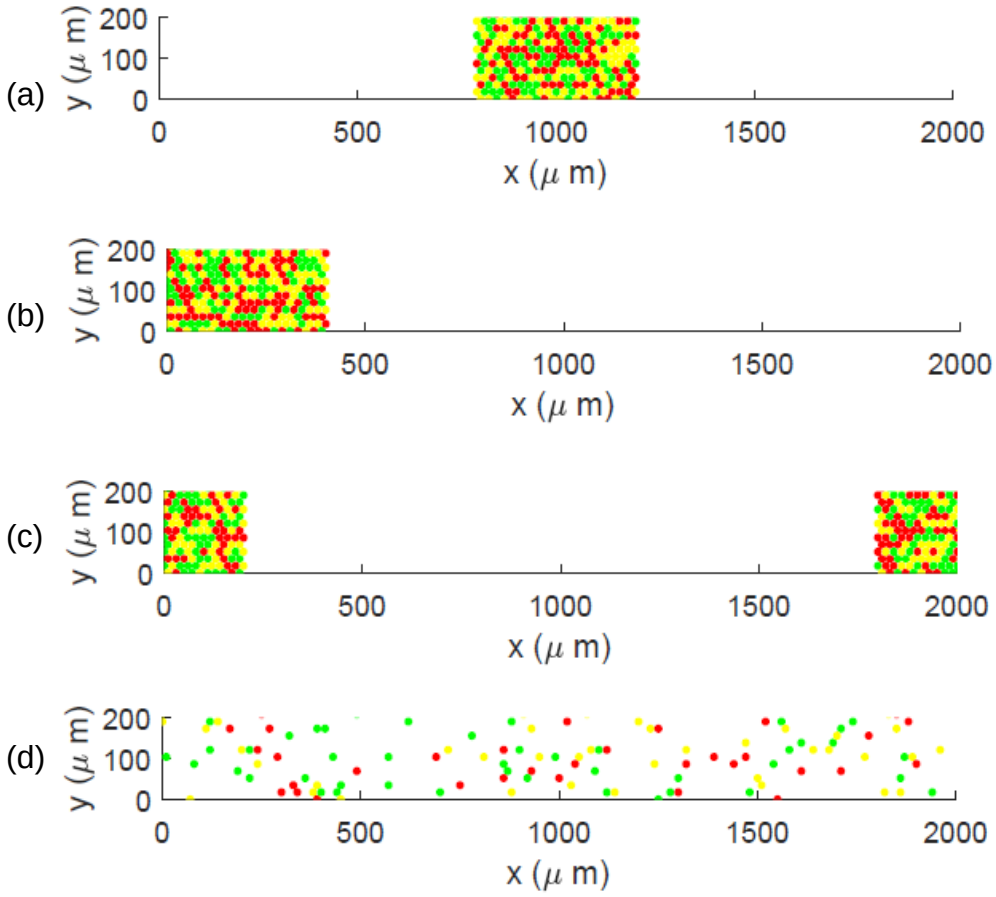

Figure 2:Initial seeds: (a) middle, (b) left, (c) left and right, (d) random (10% density)

In place of using observed data, we simulated a sample of data derived from biologically plausible parameter values which we assume to be true and then take the average of the samples as our “observed” data. This way we know the true values which the marginal posterior distributions should be centred around which allows us to compare the effectiveness of the algorithms and summary statistics

For the purposes of statistical analysis, one difficulty identified in the simulation model was the proposal of parameters which lead to cell type extinction events. A trivial example of this would be seen by setting the transition rate of red cells to 0 (or close to) while the others are relatively higher. This would result in the yellow and green cells going extinct with a large build-up of red cells. This does not happen with the observed data in the laboratory, so we decided to reject proposals leading to an extinction event as these will never match the observed data and will significantly reduce the computation time in the algorithms.

Statistical inference algorithms

Stochastic random walk models often have intractable likelihood functions which makes conventional Bayesian inference methods infeasible. In order to be able to estimate the posterior distributions, we used Approximate Bayesian Computation (ABC) methods. ABC involves summarising observed data and simulated data and comparing their difference by some appropriate discrepancy function (we use the Euclidean distance); where proposed parameter values through the simulation model which have a small discrepancy function are accepted while the others are rejected. The algorithms ability to provide an accurate posterior distribution relies heavily on the quality of the summary statistics used. The summary statistic we explore are the number of agents of each cell subpopulation (red, yellow, green) at time t ϵ $[0\ 48]$ hours. We believe that the count of each cell type to be a good choice because the proliferation rates will only influence the number of cells.

There are a few algorithm choices explored to attain the posterior distribution of the proliferation rate parameters. The Sequential Monte Carlo Approximate Bayesian Computation (SMC ABC) replenishment algorithm (Drovandi & Pettitt, 2011) was preferable due to its greater efficiency over other algorithms such as Monte Carlo Markov Chain ABC (MCMC ABC) (Marjoram, 2003; Sisson & Fan, 2011). The SMC ABC algorithm works by proposing parameter values using an MCMC random walk but will iteratively update the tuning parameter of the proposal distribution and the acceptance tolerance threshold until either a target tolerance or acceptance rate is met. For this problem we used a multivariate normal proposal distribution which takes the previously accepted parameters as the mean and the covariance matrix of the accepted proposals as the variance to undergo the MCMC random walk. Using a multivariate normal proposal distribution may propose values outside our prior knowledge of the proliferation rates (Uniform(0,1)/hour), so we also reject these parameters.

As previously stated, we opt to simulate data as our “observed” data in place of results from the laboratory. Choosing the true proliferation rates (K_{ry}, K_{yg}, K_{gr}) to be (0.04, 0.17, 0.08) and the motility rates to all be equal to 4 μm/hour produced summary statistics for the “observed” data (with initial seed set to be random) recorded at time 24 and 48 hours in Table 1.

Table 1: Summary Statistics produced by ($K_{ry}, K_{yg}, K_{gr}, P_r, P_y, P_g$) = (0.04, 0.17, 0.08, 4, 4, 4) and initial seed set as random

Summary Statistic	Variable name	Value
Number of red cells at t = 24	Nred_t24	128
Number of red cells at t = 48	Nred_t48	187
Number of yellow cells at t = 24	Nyellow_t24	28
Number of yellow cells at t = 48	Nyellow_t48	42
Number of green cells at t = 24	Ngreen_t24	63
Number of green cells at t = 48	Ngreen_t48	97

Results

Using the SMC ABC algorithm and taking the number of each cell subpopulation at time 48 hours as the summary statistics we can produce the marginal posterior distribution for the proliferation rates in Figure 3. Recalling that our prior knowledge of the proliferation rates was the interval [0,1], the SMC ABC is able to produce relatively precise distributions for the proliferation rate parameters which all contain the true value (shown as a red asterisk).

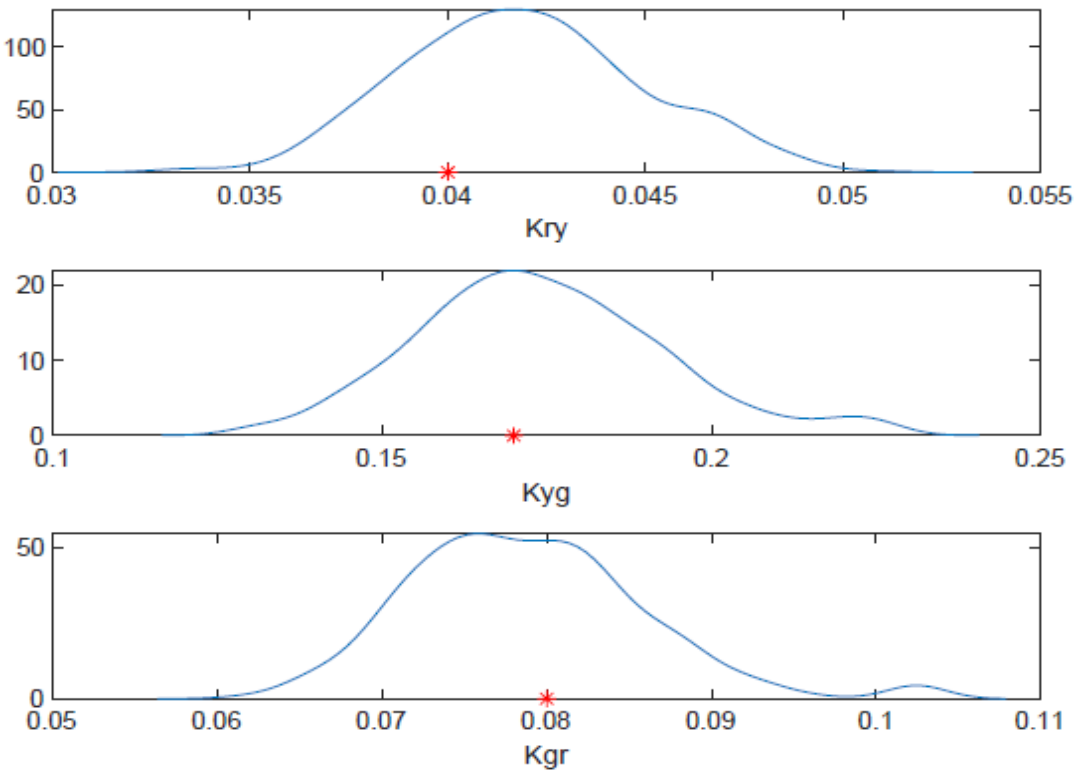

Figure 3: Marginal Posterior Distributions with t =48hrs summary statistic

However, the marginal posterior distributions in Figure 3 do appear to be slightly right skewed. This could be a result of the summary statistics not containing enough information to be certain. One possible solution is to include a second time point. In Figure 4 below, we compare the marginal posterior distributions produced with the summary statistics recorded at 48hrs and summary statistics recorded at 24 and 48 hours. It is clear that there is some benefit to adding the additional time point as it increases the precision of the distribution while still containing the true proliferation rates (K_{ry}, K_{yg}, K_{gr}) to be (0.04, 0.17, 0.08).

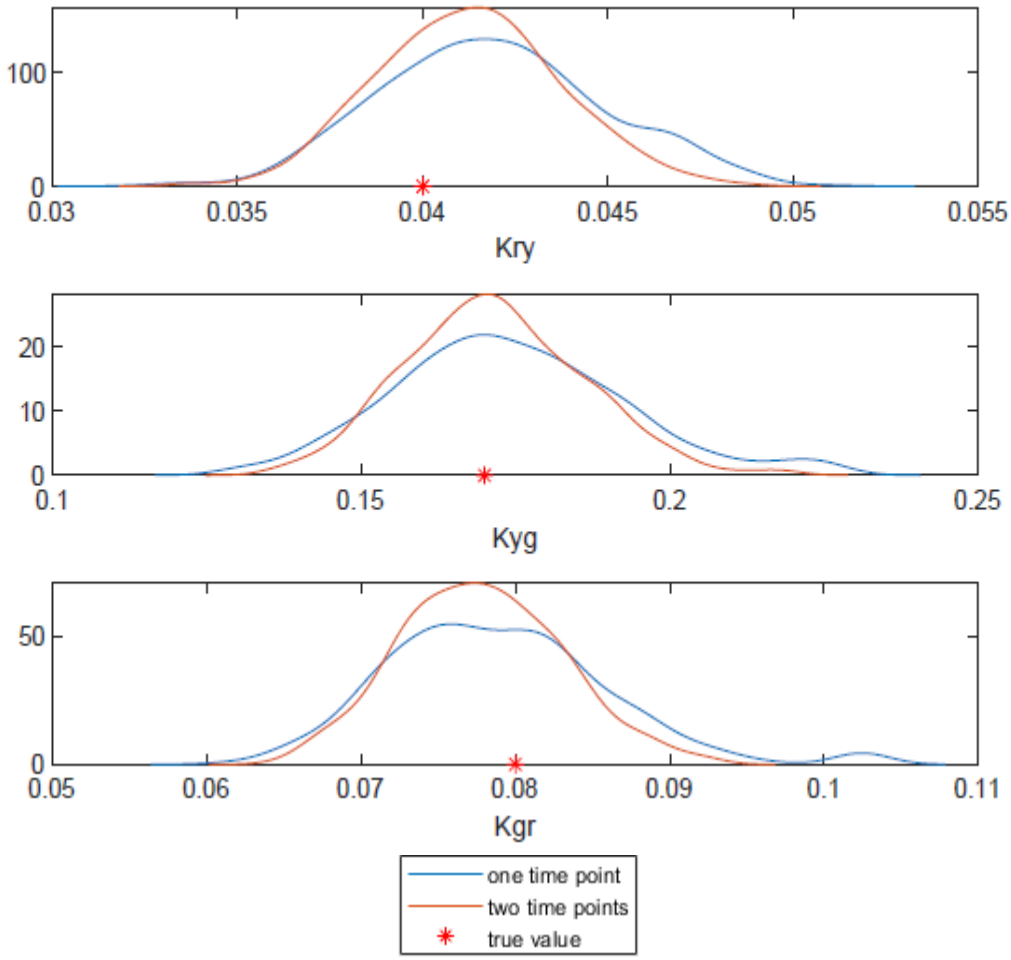

Figure 4: Marginal Posterior Distributions with t =24hrs and/or t =48hrs summary statistic comparison

Finally, we explore the reliability of using the number of cell subpopulations as summary statistics when the way the experiment is conducted changes. For instance, setting the initial seed to be any one of the four options in Figure 2 we would hope that the marginal posterior distributions are still precise and contain the true parameter values. Otherwise, we would require different summary statistics when the experimental design changes.

Using the initial seeds (middle, left, left and right, and random (10% density)) we repeat the SMC ABC algorithm with the true proliferation rates (K_{ry}, K_{yg}, K_{gr}) = (0.04, 0.17, 0.08) / hour, the motility rates equal to 4 μm/hour, and the summary statistics recorded at 24 and 48 hrs. In Figure 5, we can see that each marginal posterior distribution produced from changing the initial seeds are precise, contain the true value, and appear to be approximately equal.

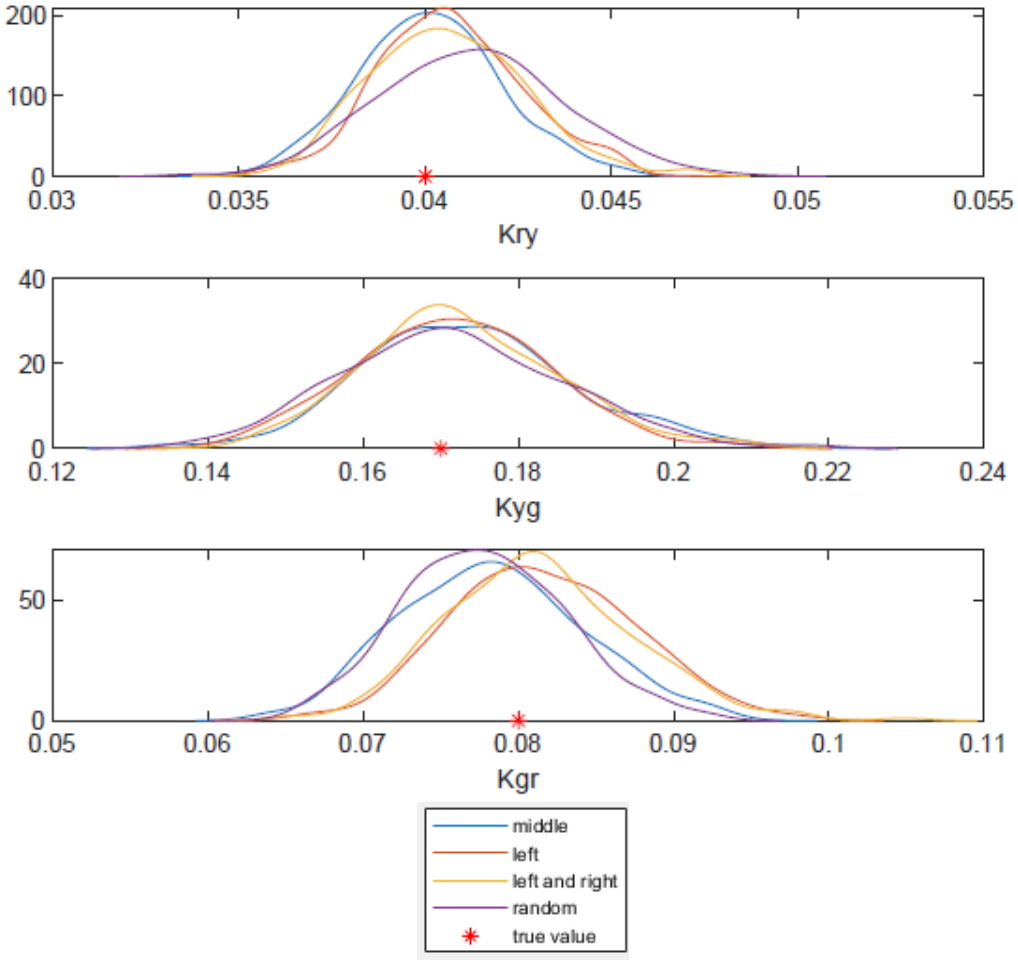

Figure 5: Marginal Posterior distributions produced by changing initial seed

Discussion

During this project we have developed a parameter estimation method for a stochastic model of melanoma skin cells and demonstrated its ability to estimate the proliferation rate parameters. By demonstrating its effectiveness at attaining the true proliferation parameters in a simulation environment, we can now apply the methods to real data and begin conducting inference on the cell cycle to find useful results or more importantly, estimate the actual transition rates between cell phases of melanoma skin cells. However,

further work is also required to estimate the motility rates which we have assumed to be known throughout this project.

Our investigation into the additional benefit added by incorporating a second set of summary statistics at 24 hours in addition to 48 hours was a useful result, especially for future work which will rely on this knowledge. Another interesting result which we have been able to produce from this project was identifying the robustness of using the number of each cell subpopulations across different initial seeds. This is an important finding because the way we conduct experiments can be more flexible without compromising inference methods.

Acknowledgements

- Supervisors: Associate Professor Chris Drovandi and Professor Mathew Simpson
- Queensland University of Technology (QUT)
- Queensland University of Technology High Performance Computing (QUT HPC)
- Australian Mathematical Sciences Institute (AMSI)

References

[1] Drovandi, C. & Pettitt, A. 2011, 'Estimation of Parameters for Macroparasite Population Evolution Using Approximate Bayesian Computation', *Biometrics*, vol. 67, no. 1, pp. 225-233

[2] Gillespie, D. 1977, 'Exact stochastic simulation of coupled chemical reaction, *The Journal of Physical Chemistry',* vol. 81, no. 25, pp. 2340-2361

[3] Haass, N. & Gabrielli, B. 2017, 'Cell cycle- tailored targeting of metastatic melanoma: Challenges and opportunities', *Experimental Dermatology*, vol. 365, pp. 189-195

[4] Haass, N et al. 2014, 'Real-time cell cycle imaging during melanoma growth, invasion, and drug response', *Pigment Cell & Melanoma Research*, vol. 27, no. 1, pp. 764–776.

[5] Marjoram, P et al. 2003, 'Markov chain Monte Carlo without likelihoods', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 26, pp. 15324– 15328

[6] Simpson, M et al. 2018, 'Stochastic models of cell invasion with fluorescent cell cycle Indicators', *Physica A,* vol. 210, pp. 375-386

[7] Sisson, S. A. & Fan, Y. 2011, 'Likelihood-free MCMC', in Handbook of Markov Chain Monte Carlo. CRC Press. pp. 313–336