Contents

1 Background

1.1 Methods in biochemical modelling

Modelling attempts to describe the elements and dynamics of the biochemical system of interest. It is a tool used for integrating knowledge and experimental data as well as for making predictions about the behaviour of the system (Wilkinson 2006).

1.1.1 Representation of transcription networks

A transcription network can be represented in a number of ways. A network can be described by using a diagram with accompanying verbal explanations or a set of differential equations. A diagram with a lengthy verbal explanation risks the not providing sufficient clarity whereas a set of differential equations cannot easily be separated from the underlying assumptions made on the kinetics of the network. A convenient way of describing sufficient information about a system while avoiding the addition of the particular interpretation of the underlying kinetics is the use of coupled chemical reactions (Wilkinson 2006).

1.1.1.1 Coupled chemical reactions and the law of mass action

Coupled chemical reactions are often used to describe transcription networks in systems biology. They have the advantage of describing a system concisely while they can be used subsequently used for a variety of different simulation methods, each with their associated interpretation of chemical kinetics (Wilkinson 2006). Coupled chemical reactions take the form

$$R_a + R_b \xrightarrow{k} P, \tag{1.1}$$

where R represents a reactant and P a product. Each reaction has an associated rate k. A biological transcription network can be represented using the above notations.

Table 1.1 Examples of common genetic coupled chemical reactions. p stands for promoter, and A represents a protein. A_2 is the dimer of protein A.

Event	Coupled chemical reaction
Transcription	$p \xrightarrow{k_1} p + RNA$
Dimerization	$2A \underset{k_3}{\overset{k_2}{\rightleftharpoons}} A_2$
Promoter repression	$A_2 + p \underset{k_5}{\overset{k_4}{\rightleftharpoons}} p \bullet A_2$
Activation	$A_2 + p \xrightarrow{k_6} p \bullet A_2 + RNA$
Degradation	$A \xrightarrow{k_7} \emptyset$

Some common examples of coupled chemical reactions used in a biological network are given in Table 1.1. A double headed arrow represents a reversible reaction.

The law of mass action allows us to derive these reaction rates k_1-k_7 from the coupled chemical reactions. The assumption made in the law of mass action is that the system exists in a well mixed solution. The law of mass action states that the reaction rate is proportional to the concentration of the reactants. So for a given chemical equation as the one shown in Equation 1.1, the rate of the reaction k is defined by:

$$k = [R_a][R_b]$$

1.1.1.2 Graphical representation of biochemical systems

It is common to represent coupled biochemical reactions graphically. In a graph, as shown in Figure 1.1, nodes represent the species and the edges represent an interaction between the species it connects, in which a transcription factor directly affects the transcription of a gene (Alon 2007). An arrow at the end of an arc represents activation, i.e. that when the transcription factor binds to the promoter the rate of transcription of the gene increases. A flat line perpendicular to the arc at the end of an arc represents repression, i.e. that when the transcription factor binds to the promoter the rate of transcription of the gene decreases (Alon 2007).

1.1.1.3 Systems Biology Markup Language (SBML)

The Systems Biology Markup Language (SBML) was developed by Hucka et al. (2004) in order to allow for the exchange of biochemical models between software. It is an extension of the XML encoding (DuCharme 1999) with additional fields specific

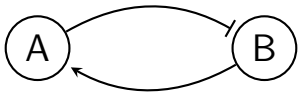

Figure 1.1 A graphical representation of a biochemical system. The two nodes, A and B represent species and the edges (arrows) a reaction between the two. An arrow represents activation and a flat line represents repression.

to biochemical models. Software like Copasi (Hoops et al. 2006) can be used to convert a set of coupled chemical reactions to an SBML model. SBML models have been key resource for model sharing within the systems biology community (Wilkinson 2006) in databases like the BioModels database (Le Novère et al. 2006).

1.1.2 Modelling promoter regulation

The processes of transcription regulation in prokaryotes is complex and there have been a number of mathematical descriptions developed to approximate the dynamics observed. These include the Hill equation and the Shea-Ackers formalism.

1.1.2.1 Hill formalism

The Hill formalism is often used to describe a biochemical system where an activator or repressor is present (Alon 2007). The Hill function is often represented as

$$\frac{dP}{dt} = V_{max}\frac{S^n}{K^n + S^n},$$

if activation is being modelled. Parameter n is the Hill coefficient and K the Hill constant. $Vmax$ is the maximum amount of product and S is the substrate concentration. The Hill constant represent the substrate concentration that results in half of the response and the Hill coefficient affects the steepness of the function and represents the cooperativity of the binding to the promoter (Alon 2007). If repression is being modelled, the Hill function is represented as

$$\frac{dP}{dt} = V_{max}\frac{K}{K + S^n}$$

An example of the effect that the value of n has on the shape of the Hill function in both activation and repression is given in Figure 1.2.

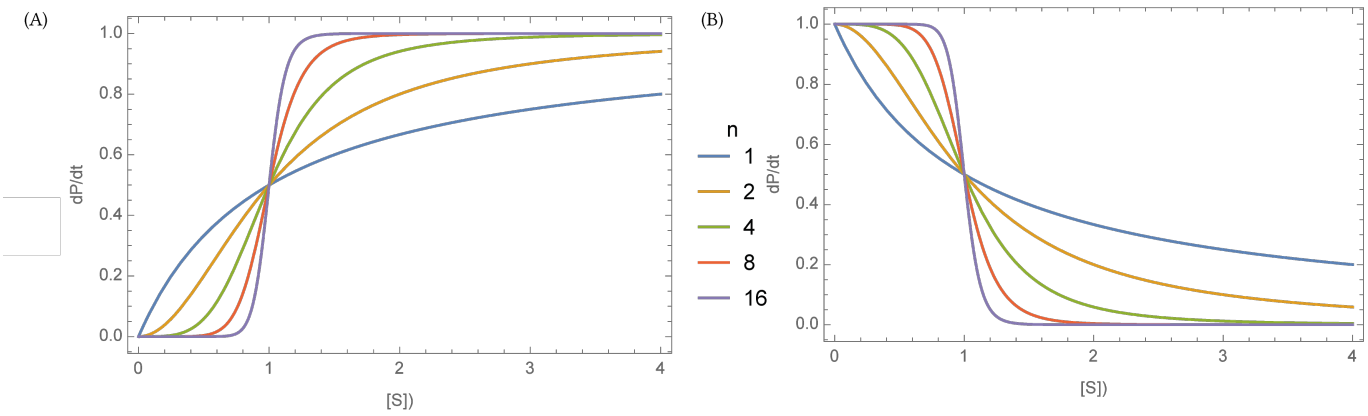

Figure 1.2 The effect of different values of n on the Hill function when K is kept constant in the case of (A) activation and (B) repression.

1.1.2.2 Shea-Ackers formalism

The Shea-Ackers formalism developed by Ackers, Johnson, & Shea (1982) uses a statistical thermodynamic model to represent the bindinf of transcription factors to their promoters. A system is described by the various states the promoter can have. An exampled of possible states is given in Figure 1.3. Each state has an associated term, or weight, and the probability of transcription is given by the ration of the producing states over all possible states. This is referred to as the partition function.

The partition function is thus given by

$$P_T = \alpha \frac{k_1 + k_3 A^2}{1 + k_1 + k_2 R^2 + k_3 A^2}$$

Here I assume that repression and activation is cooperative, thus two transcription factors must bind to the promoter to repress or activate it.

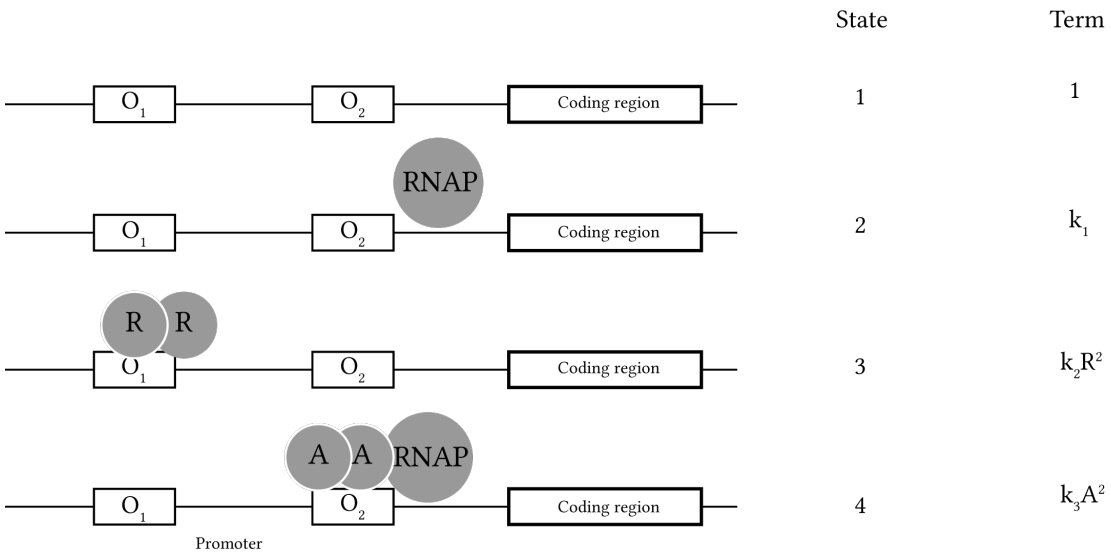

Figure 1.3 An example of a promoter regulated by a repressor (R) and an activator (A) modelled using the Shea-Ackers formalism. Figure adapted from Woods et al. (2016)

1.1.3 Simulation of dynamical systems

There are two main ways of simulating a system, deterministically and stochastically: Deterministic modelling utilises ODE and models the concentrations of the species (proteins or other molecules) by time-dependent variables (de Jong 2002). Rate equations are used to model gene regulation where the rate of production of a species is a function of the concentrations of the other species (de Jong 2002).

1.1.3.1 Deterministic mass action kinetics

ODEs are used to represent the quantitative dynamics of a biochemical network. The ODEs describing a system can be derived from the coupled chemical reactions descibing the system as well as their associated rates. This will be illustrated using a simple example, the Lotka-Voltera predator-prey model (Lotka 1925). This system describes the dynamics between two interacting species, a predator and a prey. The chemical reactions describing the system are given in Table 1.2. The rates of the system are organised to vector form:

$$h = \begin{pmatrix} k_1 x \\ k_2 xy \\ k_3 y \end{pmatrix}$$

Table 1.2 Predator-prey chemical reactions

Name	Reaction	Rate
prey birth	$x \xrightarrow{k_1} 2x$	$k_1 x$
predation	$x + y \xrightarrow{k_2} 2y$	$k_2 xy$
predator death	$y \xrightarrow{k_3} \emptyset$	$k_3 y$

The stoichiometry matrix of the system is $m \times n$ matrix, where m is the number of species and n the number of reactions and it summarises the stoichiometries of the system.

$$S = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \tag{1.2}$$

The ODEs can then be constructed by multiplying the stoichiometry matrix S by the matrix containing the rates h. Therefore

$$s(t) = \frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}\begin{pmatrix} k_1[x] \\ k_2[x][y] \\ k_3[y] \end{pmatrix} \tag{1.3}$$

and thus we get the two ODEs describing the system as

$$\frac{dx}{dt} = k_1 x - k_2 xy \tag{1.4}$$

$$\frac{dy}{dt} = k_2 xy - k_3 y \tag{1.5}$$

These differential equations can be simulated numerically over time using software packages like Mathematica (Mathematica 2016) and Python.

1.1.3.2 Assumptions of deterministic modelling

Two key assumptions are made when modelling a biochemical system using ODEs. Firstly, the species present in the system are measured continuously rather than discreetly. This means that the species are measured in concentration over time and not number of molecules over time. This assumption requires at least 1000 molecules to be present in order to be met (Ingalls & Iglesias 2010). The second assumption made is that the reactants are in a well mixed solution. This means that the species in the system can interact each other constantly and freely.

1.1.4 Phase plane analysis

An alternative to studying the trajectory of a dynamical system over time is to study its behaviour in the phase plane. During a phase plane analysis the dependent variables x and y are plotted against eachother, as shown in Figure 1.4.

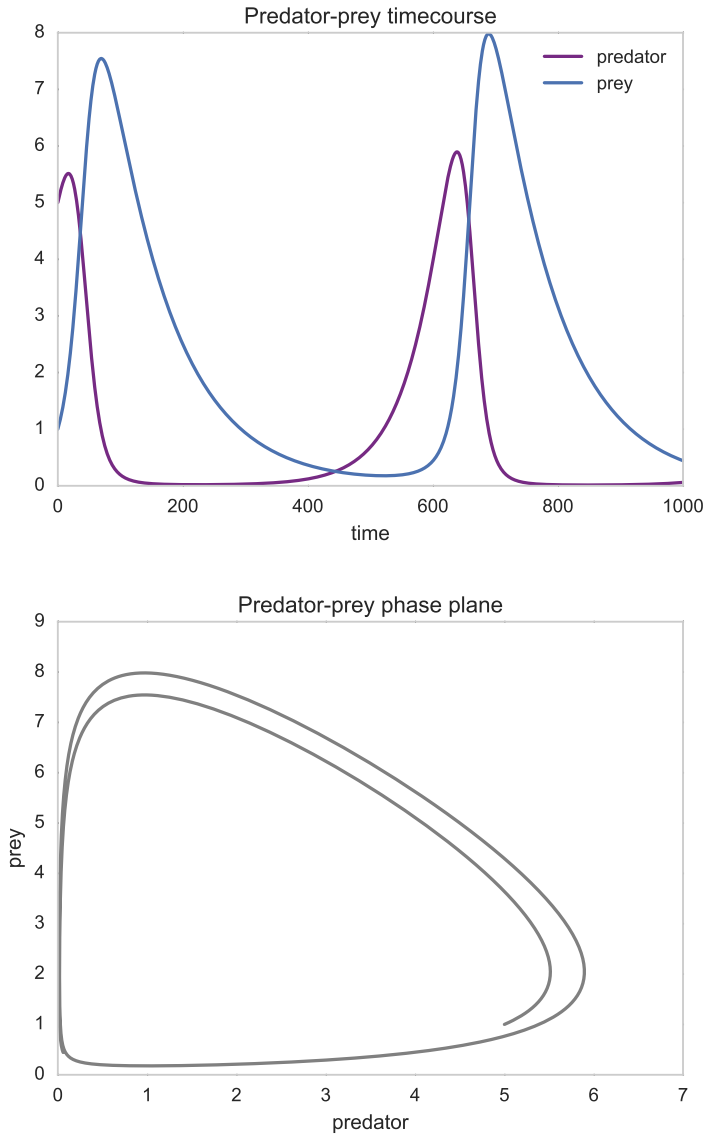

Figure 1.4 (A) The predator-prey system of equations as defined in Section 1.1.3.1 trajectory over time (B) Phase plane plot of the predator-prey system of equations. The parameters used here are $k_1 = 2$, $k_2 = 1$ and $k_3 = 1$.

1.1.5 Steady states

For a system s, any system satisfying $\frac{d}{dt}s(t) = 0$ is considered a fixed point, or steady state. At that point the dynamics of the system are considered in equilibrium and will not change with increasing time. Using the example of the predator-prey system, a steady state exists when the system of Equations 1.5 are equal to 0:

$$\frac{dx}{dt} = f_x(x, y) = k_1 x - k_2 xy = 0 \tag{1.6}$$

$$\frac{dy}{dt} = f_y(x, y) = k_2 xy - k_3 y = 0 \tag{1.7}$$

By solving this system of equations, we get two steady states. One when $x = y = 0$ and one when $x = \frac{k_3}{k_2}$ and $y = \frac{k_1}{k_2}$. The stability of each steady state can then be determined.

1.1.6 Steady state stability

A stable steady state is defined as a fixed point whose nearby points approach the fixed point (Kaplan & Glass 1995). This means that after a small perturbation the system will quickly return to the steady state. An unstable steady state is one which if the system is perturbed slightly then it moves away from the steady state (Konopka 2007). The stability of the fixed points can be determined by the sign of the eigenvalues of the Jacobian matrix at each point. The Jacobian matrix is given by

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f_x}{\partial x} & \frac{\partial f_x}{\partial y} \\ \frac{\partial f_y}{\partial x} & \frac{\partial f_y}{\partial y} \end{pmatrix} \tag{1.8}$$

Using the predator-prey system as an example, the Jacobian matrix is given by

$$\mathbf{J} = \begin{pmatrix} k_1 - k_2 y & -k_2 x \\ k_2 y & k_2 x - k_3 \end{pmatrix} \tag{1.9}$$

The eigenvalues λ are given by

$$\det(\mathbf{J} - \lambda \mathbf{I}) \tag{1.10}$$

where det is the determinant and $\mathbf{I}$ the identity matrix. If both eigenvalues are real and negative or imaginary with a negative real part then the steady state is stable.

If both eigenvalues have a positive real part then the steady state is unstable and if one has a positive and one has a negative fixed part the steady state is an unstable saddle node. If both eigenvalues are purely imaginary, the system oscillates around the fixed point. Solving the above for the fixed points in the predator-prey system, we find one stable steady state and one oscillatory fixed point.

1.1.7 Stochastic modelling of dynamical systems

The assumptions that have to be made to model a system deterministically cannot always be met. This can occur when the molecule numbers in the system are low or the solution cannot be assumed to be well mixed. When this is the case, stochastic dynamics are more appropriate to model dynamical system. In stochastic modelling species are measured in discrete amounts rather than concentrations and a joint probability distribution is used to express the probability that at time t the cell contains a number of molecules of each species (de Jong 2002; Khammash 2010). It takes uncertainty into account and does not assume a well mixed solution.

Biological processes are well known to include randomness. The source of this randomness originates from the random collisions between molecules that govern biological reactions (Khammash 2010). This randomness affects downstream events and affect the phenotypic behaviour of cells. This is known as cellular noise and it is known to be key for various cellular processes (Eldar & Elowitz 2010). Cellular noise can be classified into two categories, intrinsic noise and extrinsic noise. Intrinsic noise originates from the inherently random collisions between the species of the system under consideration. Extrinsic noise originates from fluctuations in the environment within which the system of interest resides, like the number of available RNA polymerases or other protein numbers (Khammash 2010). The noisiness of biological processes often make stochastic dynamics more appropriate for modelling cellular systems.

1.1.7.1 Simulating stochastic models

Stochastic models are often analytically intractable but can be studied using numerical simulation. A well known algorithm for the simulation of such models is the Direct method proposed by Gillespie (1977).

1.1.7.2 The Gillespie algorithm

In stochastic systems the Gillespie algorithm is widely used to simulate the time-evolution of the state of the system (Wilkinson 2006). The algorithm, developed by Gillespie (1977) can be summarised in four steps:

1. Initialise time t and number of species s and state of system x
2. Draw a sample timestep τ from the distribution of time T
3. Draw a sample reaction from all reactions R
4. Update time by $t = t + \tau$ and state of system by $x = x + s_\mu$
5. Repeat from Step 2 until total simulation time reached

This algorithm results in one trajectory of the system. It has to be repeated a number of times to obtain enough realisations of the trajectory to compute appropriate summary statistics.

1.1.7.3 Stochastic mass action kinetics

Here I will consider the predator-prey system introduced in Section 1.1.3.1. A set of reactions is defined, as shown in Table 1.2, and each one has an associated stochastic rate constant c_i. The rate constant, or hazard function, of each reaction i is defined as $h_i(x, c_i)$, where x is the current state of each species in the system.The form of each hazard function is defined by the order of the given reaction (Wilkinson 2006), as shown in Table 1.3.

Table 1.3 Defining reaction hazards

Order	Reaction	Hazard
Zeroth	$\emptyset \xrightarrow{c_i} X$	$h_i(x, c_i) = c_i$
First	$X_j \xrightarrow{c_i} ?$	$h_i(x, c_i) = c_i x_j$
Second	$X_j + X_k \xrightarrow{c_i} ?$	$h_i(x, c_i) = c_i x_j x_k$
Dimerization	$X_j + X_j \xrightarrow{c_i} X_{2j}$	$h_i(x, c_i) = c_i \frac{x_j(x_j-1)}{2}$

Therefore, when simulating a stochastic system using the Gillespie algorithm, the state of the system x is defined as the sum of all the reaction hazards, namely $h_0(x, c) = \sum_{i=1}^{n} h_i(x, c_i)$ (Wilkinson 2006).

1.2 Parameter inference

For a model to be used in a systems biology setting the parameters must first be defined. The parameters of a model represent the biochemical rates that are involved in the system under study, like degradation rates, transcription rates and polymerization rates. These rates are cannot always be measured experimentally and taking generalised estimates from existing literature can be inaccurate. Iterative statistical methods have been developed for the inference of the model parameters.

The methods available can largely be separated into two categories: frequentist methods and bayesian methods.

All methods are comprised largely of a cost function that compares the model data to the experimental data and an optimization function that aims to optimize the cost function. There is a wide range of such optimization algorithms used like gradient decent(XXX), simulated annealing(XXX) and eveolutionary algorithms(XXX).

Bayesian approaches to parameter inference have the advantage of offering a range of values that give rise to the data, rather than point estimates. Bayesian approaches have been successfully developed for the estimation of parameters in biological models(XXX)

1.2.1 Approximate Bayesian Computation (ABC)

ABC methods are used for inferring the posterior distribution in cases where it is too computationally expensive to evaluate the likelihood function. Instead of calculating the likelihood, ABC methods simulate the data and then compare the simulated and observed data through a distance function (Toni et al. 2009). Given the prior distribution $\pi(\theta)$ we can approximate the posterior distribution, $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$, where $f(x \mid \theta)$ is the likelihood of a parameter, θ, given the data, x. There are a number of different variations of the ABC algorithm depending on how the the approximate posterior distribution is sampled.

The simplest ABC algorithm is the ABC rejection sampler (Pritchard et al. 1999). In this method, parameters are sampled from the prior and data simulated through the data generating model. For each simulated data set, a distance from that of the desired behaviour is calculated, and if greater than a threshold, ε, the sample is rejected, otherwise it is accepted.

Algorithm 1 ABC rejection algorithm

1: Sample a parameter vector θ from prior $\pi(\theta)$
2: Simulate the model given θ
3: Compare the simulated data with the desired data, using a distance function d and tolerance ε. if $d \leq \varepsilon$, accept θ

The main disadvantage of this method is that if the prior distribution is very different from the posterior, the acceptance rate is very low (Toni et al. 2009). An alternative method is the ABC Markov Chain Monte Carlo (MCMC) developed by Marjoram et al. (2003). The disadvantage of this method is that if it gets stuck in an area of low probability it can be very slow to converge (**Sisson:wf**).

The method used here is based on Sequential Monte Carlo, which avoids both issues faced by the rejection and MCMC methods. It propagates the prior through a series of intermediate distributions in order to arrive at an approximation of the posterior. The tolerance, ε, for the distance of the simulated data to the desired data is made smaller at each iteration. When ε is sufficiently small, the result will approximate the posterior distribution (Toni et al. 2009).

ABC SMC can identify the parameter values within a predefined range of values that can achieve the desired behaviour. It works by first sampling at random from the initial range set by the user, i.e. form the prior distribution of values. Each sample from the priors is called a particle. It then simulates the model given those values and compares that to the target behaviour. If the distance between the simulation and the target behaviour is greater than a predefined threshold distance ε, then the parameter values that produced that simulation are rejected. This is repeated for a predefined number of samples which are collectively referred to as a population. Each particle in a population has a weight associated with it, which represents the probability of it producing the desired behaviour. At subsequent iterations the new samples are obtained from the previous populations and the ε is set to smaller value, thus eventually reaching the desired behaviour. The algorithm proceeds as follows:

Algorithm 2 ABC SMC algorithm

1: Select ε and set population t = 0
2: Sample particles (θ). If t = 0, sample from prior distributions (P). If t > 0, sample particles from previous population.
3: If t > 0: Perturb each particle by ± half the range of the previous population (j) to obtain new perturbed population (i).
4: Simulate each particle to obtain time course.
5: Reject particles if d > ε.
6: Calculate the weight for each accepted particle. At the first population assign a weight equal to 1 for all particles. In subsequent populations the weight of a particle is equal to the probability of observing that particle divided by the sum of the probabilities of the particle arising from each of the particles in the previous population:
7: $w_t^{(i)} = \begin{cases} 1, & \text{if } n = 0 \\ \frac{P(\theta_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta_t^{(i)})}, & \text{if } n>0. \end{cases}$

1.2.1.1 Particle sampling

For the first population, particles are sampled from the priors. Random samples are taken from the distribution specified by the user for each parameter.

For subsequent populations particles are sampled from the previous population. The weight of each particle in the previous population dictates the probability of it being sampled. The number of samples to be drawn is specified by the user in the input file.

1.2.1.2 Perturbation

Each sampled particle is perturbed by a kernel defined by the distribution of the previous population, as developed by Toni et al. (2009).

$$K_p(\theta|\theta*) = \theta * + U(+s_p, -s_p), \text{ where:} \tag{1.11}$$

$$s_p = \frac{1}{2}\big(max(\theta_{p-1}) - min(\theta_{p-1})\big) \tag{1.12}$$

If the θ* falls out of the limits of the priors then the perturbation is rejected and repeated until an acceptable θ* is obtained. This method is successful in perturbing the particles by a small amount in order to explore the parameter space, but can be slow to complete.

1.2.1.3 Particle simulation

Each particle is simulated using cuda-sim (Zhou et al. 2011). The model is provided by the user in SBML format and is converted into CUDA® code by cuda-sim. The model in CUDA® code format can then be run on NVIDIA®. CUDA®. GPUs. This allows the user to take advantage of the speed of parallelised simulations without any CUDA® knowledge.

1.2.1.4 Weight calculation

For the first population the weights are all given a value of 1, and then normalised over the number of particles. For subsequent populations the weights of the particles are calculated by considering the weights of the previous population (Toni et al. 2009). The weights are then normalised over the total number of particles.

$$w_t^{(i)} = \frac{P(\theta_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta_t^{(i)})} \text{ for n} > 0 \tag{1.13}$$

1.2.2 ABC algorithm example

This algorithm is implemented on a simple example for illustration. A simple model was used, consisting of one species, A converting to another, B. The model is described by two differential equations, where A is the reactant and B the product, produced at a rate p.

$$\frac{d[B]}{dt} = p[A] \tag{1.14}$$

$$\frac{d[A]}{dt} = -p[A] \tag{1.15}$$

The priors were set to $p \sim U(0, 10)$. Initial conditions for A and B were set to 1 and 0 respectively. The data to which the model was compared to was generated by simulating the same model with the parameter set to 1, as shown in Figure 1.5.

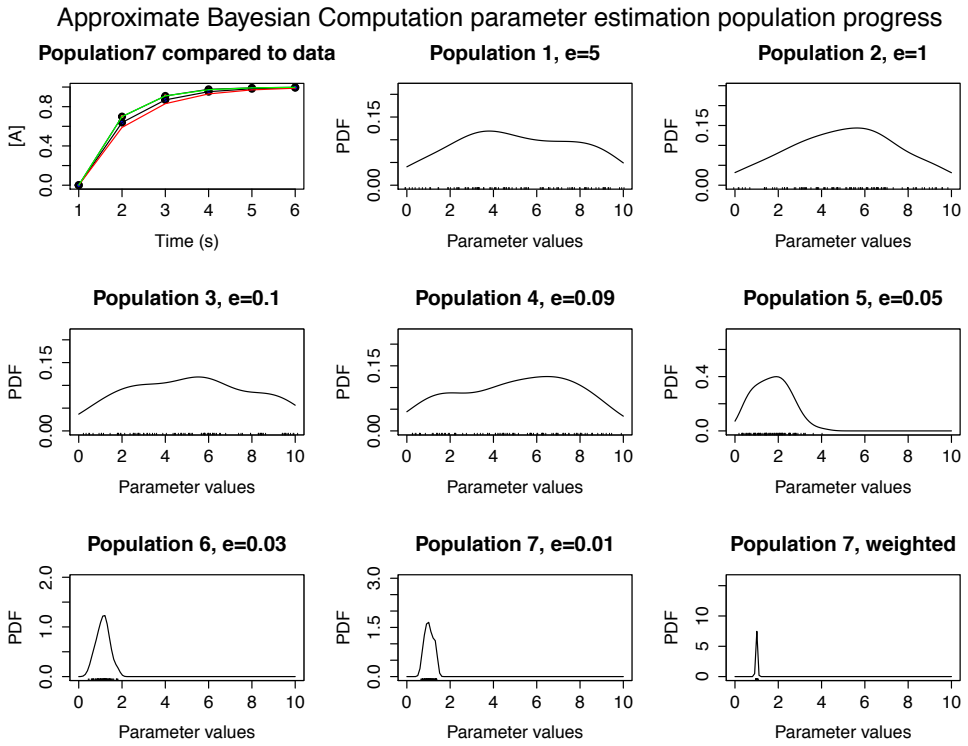

Figure 1.5 ABC SMC parameter inference. The posterior parameter is equal to 1 and its time course shown in red in the top left panel. The blue time course is that of the final population, green is the upper quartile and red is the lower quartile range of values. The progress of the selection process can be seen the ε schedule proceeds from the top left to the bottom right. The bottom far right panel is a density plot of ε= 0.01 with their weights taken into account.

Figure 1.5 demonstrates, using a simple example, that ABC SMC is capable of fitting a model to the data. During the course of 7 populations, the accepted distance ε of the simulated particles to the data is incrementally decreased. This leads to a final population where the distance of the data to the particles is very small, and there is a good agreement between the two. The algorithm concludes with a set of parameter values that produced this behaviour, which approximate the posterior distribution. The posterior distribution found in this model is in good agreement with the parameter value used to generate the data. This example successfully demonstrates the effectiveness of the ABC SMC algorithm in fitting models to data.

1.2.3 Derivation of model parametric robustness defined via Bayesian statistics

During this thesis I define robustness as the ability of a system to retain its function despite parameter perturbations (Stelling et al. 2004). The robustness of biological systems has been studied extensively (Barkai & Leibler 1997; Stelling et al. 2004; Prill, Iglesias, & Levchenko 2005; Kim et al. 2006; Kitano 2007; Hafner et al. 2009; Shinar & Feinberg 2010; Zamora-Sillero et al. 2011; Woods et al. 2016). and it is well known that feedback loops can increase the robustness of a system (Becskei & Serrano 2000; Doyle & Csete 2005).

The robustness of a model can be calculated by dividing the volume of its functional region by the volume of its priors. This is a measure of the volume of the posterior distribution is compared to the priors. It comes from Bayes' rule that:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int p(x|\theta)p(\theta)d\theta} \tag{1.16}$$

where $p(x|\theta)$ is the likelihood, $p(\theta)$ is the prior, and $\int p(x|\theta)p(\theta)d\theta$ is the evidence. The evidence is the normalisation added so that the distribution integrates to 1. For a given model design D and objective O we define the functional region F as the region within the prior where O is satisfied. So within the prior we can assign 1 to any region that falls within F and 0 to any region outside that.

$$p(O|D_1) = \int p(O|\theta, D_1)p(\theta|D_1)d\theta, \tag{1.17}$$

For a design with three parameters this becomes:

$$p(O|D_1) = \iiint_{\underline{\Theta}} p(O|\underline{\Theta}) p(\underline{\Theta}|D_1) d\underline{\Theta}, \tag{1.18}$$

where $\underline{\Theta}$ is a vector containing the three parameters = $\theta_1, \theta_2, \theta_3$. To calculate the robustness, or model evidence, we integrate this with respect to $\underline{\Theta}$. We assume all parameters $\theta_1, \theta_2, \theta_3$ are uniform, $p(\underline{\Theta}|D_1) \sim U(a, b)$. If we assume a = 0 this integral becomes:

$$p(O|D_1) = \iiint_{\underline{\Theta}} p(O|\underline{\Theta}) \frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3} d\underline{\Theta} \text{ , and} \tag{1.19}$$

$$p(O|D_1) = \frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3} \iiint_{\underline{\Theta}} p(O|\underline{\Theta}) d\underline{\Theta} \tag{1.20}$$

since $\frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}$ is a constant. Then assuming that the likelihood is uniform Equation 1.20 becomes:

$$p(O|D_1) = \frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}\left[\iiint_{\underline{\Theta}_F} 1 d\underline{\Theta} + \iiint_{\underline{\Theta}_{\not F}} \overset{0}{\cancel{0}} d\underline{\Theta} \right] \tag{1.21}$$

$$\tag{1.22}$$

since we assign 1 to any region within F and 0 to any region outside it. This becomes:

$$p(O|D_1) = \overbrace{\frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}}^{|P|} \underbrace{\iiint_{\underline{\Theta}_F} 1 d\underline{\Theta}}_{|F|}, \tag{1.23}$$

$$\therefore p(O|D_1) = \frac{|F|}{|P|}, \tag{1.24}$$

where |P| is the volume of the prior P and |F| the volume of the functional region F. Therefore, in the case where both the prior and the likelihood are uniform, the robustness R of the design is the ratio of the volumes of the two. If on the other hand we assume the likelihood is multivariate normal, with priors remaining uniform, Equation1.20 becomes:

$$p(O|D_1) = \frac{1}{|P|} \iiint_{\underline{\Theta}} f(\underline{\Theta}; \mu, \Sigma) d\underline{\Theta} \tag{1.25}$$

$$\therefore p(O|D_1) = \frac{1}{|P|} \underbrace{\frac{2\pi^{\frac{k}{2}}}{k\Gamma(\frac{k}{2})} \left[\chi_k^2(\alpha)\right]^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \tag{1.26}$$

$$\therefore p(O|D_1) = \frac{|F|}{|P|}, \tag{1.27}$$

We can use the Bayes' factor in order to compare the robustness between two model designs. The Bayes' factor is defined as follows:

$$B_{ab} = \frac{\int p(x|\theta, D_a)p(\theta, D_a)d\theta}{\int p(x|\theta, D_b)p(\theta, D_b)d\theta} \tag{1.28}$$

$$\therefore B_{ab} = \frac{|Fa|}{|Pa|} / \frac{|Fb|}{|Pb|} \tag{1.29}$$

Therefore, we can use the ratio of the two robustness measures to calculate the Bayes' factor. If two models have a different number of parameters, the robustness of the system will only increase if |F| increases by more than the proportion by which |P| increased (Woods et al. 2016). A model will be penalised for an additional if it does not increase the volume of the functional region by more than the volume that the added parameter added to the prior. This is true for nested models, where one model is wholly contained in the other.

1.3 Flow Cytometry

Flow cytometry detects the fluorescent intensity levels in individual cells. It can also provide physical information about the size and granularity of a cell via the forward and side scattering respectively. An overview of flow cytometry is shown in Figure 1.6. A laser excites the fluorochrome present in the bacterial cells. The fluorochromes emit a signal that is detected by channels in the optics. The signals are then all collected and analysed. A sample typically consists single cell measurements of 10^4-10^5 cells. Flow cytometry is a powerful tool for synthetic biology as it can measure multiple parameters in single cells, and process up to 35,000 cells sec^{-1} (*Attune NxT Acoustic Focusing Cytometer* 2015).

1.4 Current understanding of the genetic toggle switch

One of the most common devices used in synthetic biology is the genetic toggle switch. A toggle switch consists of a set of transcription factors that mutually

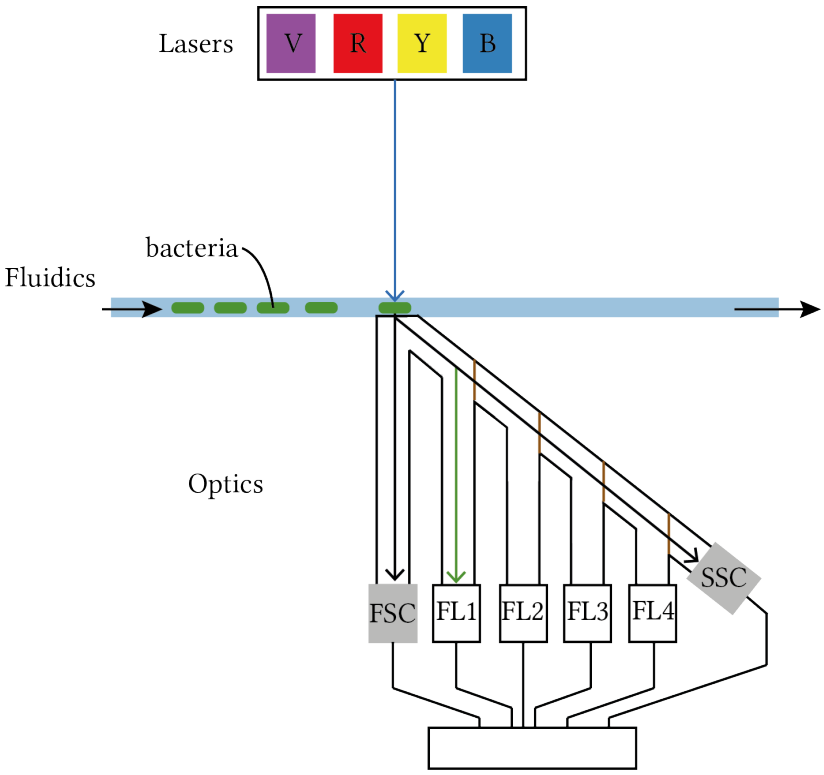

Figure 1.6 Flow cytometry. A laser excites the fluorescent proteins present in each cell. The cytometer has up to 4 lasers, violet (V), red (R), yellow (Y) and blue (B). The detectors in the optics, FL1-4 pick up the signals. The cytometer also picks up size and granularity information via the forward scatter (FSC) and side scatter (SSC) detectors. Diagram adapted from (*What is Flow Cytometry* n.d.)

repress each other (Gardner, Cantor, & Collins 2000). Genetic switches play a major role in binary cell fate decisions like stem cell differentiation, as they are capable of exhibiting bistable behaviour. Bistability of a system is defined by the existence of two distinct phenotypic states but no intermediate state. Bistability is a property that is important in nature and a valuable resource to tap into in synthetic biology. It allows cells to alter their response to environmental cues and increases the overall population fitness by 'hedge-betting' the response of the population (XXX).

1.4.1 The genetic toggle switch in natural systems

In developmental processes, bistability ensures that the differentiating cell will follow one pathway, or the other, with no possible intermediate phenotypes. This is vital for the correct development of a cell in a specific pathway. One example is

the trophectoderm differentiation pathway, in which a mutually inhibitory toggle switch exists between Oct3/4 and Cdx2. This determines whether an Embryonic Stem cell will differentiate into a Trophectoderm cell, if Cdx2 dominates the system, or an Inner Cell Mass cell if Oct3/4 dominates (Niwa et al. 2005). Bistability is critical in this system as a cell must differentiate into either a trophectoderm cell or an inner cell mass cell, thus the signal to do so must be straightforward. In the case of the GATA1 and PU.1 toggle switch, the transcription factor pair controls the fate of the common myeloid progenitors, and the two possible differentiation paths are erythroid and myeloid blood cells (Chickarmane, Enver, & Peterson 2009). The double-negative feedback loop created by the mutually repressive pair of transcription factors sustains the system in balance until an external stimulus causes one of the two transcription factors to increase in concentration. The increased concentration of one transcription factor causes the increased repression of the production of the antagonistic transcription factor, tipping the balance towards the dominance of the first transcription factor. The double negative feedback loop reinforces this dynamic and the system remains in the same state, until an external stimulus disturbs it (Ferrell 2002).

1.4.2 Uses in synthetic biology

Despite their simplicity, toggle switches can be powerful building blocks with which to create complex responses in a synthetic network. They can be used in isolation or in tandem to create complex networks and signalling cascades. The toggle switch has been used for the regulation of mammalian gene expression (Deans, Cantor, & Collins 2007; Kramer et al. 2004). Other synthetic applications of the toggle switch include the construction of a synthetic genetic clock (Atkinson et al. 2003), of a predictable genetic timer (Ellis, Wang, & Collins 2009), and the formation of biofilms in response to engineered stimuli (Kobayashi et al. 2004). These applications are modifications of the classical toggle switch (Gardner, Cantor, & Collins 2000), and to our knowledge no application made of a cascade or collection of the switch has been successful. This would make more complex applications possible and could be used to solve real-life problems. For example, an analog-to-digital converter to translate external stimuli like the concentration of an inducer into an internal digital response, or programmable bacteria to move from point to point up different chemical gradients (Lu, Khalil, & Collins 2009). For a review on current circuits see (Khalil & Collins 2010) and for possible future applications see (Lu, Khalil, & Collins 2009). This leap will be difficult to achieve before first being able to build

robust and well characterised individual switches.

1.4.3 Modelling the genetic toggle switch

The toggle switch motif has been studied extensively and there are numerous studies based on a number of different methods of modelling and analysis of the dynamics, including both deterministic and stochastic approaches. Deterministic modelling utilises ordinary differential equations (ODE) and models the concentrations of the species (proteins or other molecules) by time-dependent variables (de Jong 2002). When modelling deterministically the model is viewed as a system whose behaviour is entirely predictable, given sufficient knowledge. In stochastic modelling, species are measured in discrete amounts rather than concentrations and a joint probability distribution is used to express the probability that at time t the cell contains a number of molecules of each species (de Jong 2002; Wilkinson 2006). It takes uncertainty into account and is thus often more appropriate for modelling cellular systems, although more computationally expensive. In stochastic systems the Gillespie algorithm is widely used to simulate the time-evolution of the state of the system (Warren & ten Wolde 2005).

The conclusions drawn about the stability and robustness of the toggle switch also vary between the different modelling approaches. Numerous studies have concluded that cooperativity is a necessary condition for bistability to arise (Gardner, Cantor, & Collins 2000; Walczak, Onuchic, & Wolynes 2005; Warren & ten Wolde 2004; Warren & ten Wolde 2005; Cherry & Adler 2000). However, Lipshtat et al. (2006) found that stochastic effects can give rise to bistability even without cooperativity in three kinds of switch; the exclusive switch, in which there can only be one repressor bound at any one time, a switch in which there is degradation of bound repressors, and the switch in which free repressor proteins can form a complex, which renders them inactive as transcription factors (Lipshtat et al. 2006). In another study, Ma et al. (2012) found that the stochastic fluctuations in a system involving such a small number of molecules, like the toggle switch, uncovers effects that can not be predicted by the fully deterministic case (Ma et al. 2012). In their system, the toggle switch was found to be tristable, as small number effects render the third unstable steady state stable. Biancalani & Assaf (2015) identified multiplicative noise as the source of bistability in the stochastic case (Biancalani & Assaf 2015). Warren & ten Wolde (2005) concluded that the exclusive switch is always more robust than the general switch, since the free energy barrier is higher (Warren & ten Wolde 2005). A summary of the toggle switch models is shown in Table 1.4.

As is clear from above, there is yet to exist a consensus on the stability a switch is capable of, and the most appropriate method of modelling it. Different methods arrive at different conclusions, creating confusion on which behaviour to be expected by the experimentalist for even a simple system like the toggle switch, consisting of just two genes. The toggle switch cannot be used as a building block of larger, more complex systems until its behaviour can be predicted accurately. Until then, designing systems with predictable behaviour will be near impossible.

Table 1.4 Summary of stability for the CS and DP switches found via different modelling approaches

	CS		DP	
Stability	Deterministic	Stochastic	Deterministic	
Monostable	(Loinger & Biham 2009)	(Loinger & Biham 2009)	(Guantes & Poyatos 2008)	
	(Gardner, Cantor, & Collins 2000)	(Lu, Onuchic, & Ben-Jacob 2014),		
	(Loinger & Biham 2009)	(Lipshtat et al. 2006),		
Bistable		(Biancalani & Assaf 2015),	(Guantes & Poyatos 2008)	
		(Loinger & Biham 2009)		
		(Loinger & Biham 2009),	(Guantes & Poyatos 2008),	
Tristable		(Ma et al. 2012)	(Lu, Onuchic, & Ben-Jacob 2014)	
Quadristable			(Guantes & Poyatos 2008)	

Bibliography

Ackers, G. K., Johnson, A. D., & Shea, M. A. (1982). 'Quantitative model for gene regulation by lambda phage repressor.' *Proceedings of the National Academy of Sciences of the United States of America* **79**(4), 1129–1133.

Alon, U. (2007). *An Introduction To The Systems Biology*. Chapman & Hall/CRC.

Atkinson, M., Savageau, M., Myers, J., & Ninfa, A. (2003). 'Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in *Escherichia coli* '. *Cell*.

Attune NxT Acoustic Focusing Cytometer (2015). CO016625.

Barkai, N. & Leibler, S. (1997). 'Robustness in simple biochemical networks.' *Nature* **387**(6636), 913–917.

Becskei, A. & Serrano, L. (2000). 'Engineering stability in gene networks by autoregulation.' *Nature* **405**(6786), 590–593.

Biancalani, T. & Assaf, M. (2015). 'Genetic Toggle Switch in the Absence of Cooperative Binding: Exact Results'. *Physical review letters*.

Cherry, J. L. & Adler, F. R. (2000). 'How to make a biological switch.' *Journal of Theoretical Biology* **203**(2), 117–133.

Chickarmane, V., Enver, T., & Peterson, C. (2009). 'Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility.' *PLoS Computational Biology* **5**(1), e1000268–e1000268.

De Jong, H. (2002). 'Modeling and simulation of genetic regulatory systems: a literature review.' *Journal of Computational Biology* **9**(1), 67–103.

Deans, T. L., Cantor, C. R., & Collins, J. J. (2007). 'A Tunable Genetic Switch Based on RNAi and Repressor Proteins for Regulating Gene Expression in Mammalian Cells'. *Cell* **130**(2), 363–372.

Doyle, J. & Csete, M. (2005). 'Motifs, Control, and Stability'. *PLoS Biology* **3**(11), e392.

DuCharme, B. (1999). *XML : the annotated specification / Bob DuCharme*. English. Prentice Hall PTR Upper Saddle River, NJ, xix, 339 p. : ISBN: 0130826766.

Eldar, A. & Elowitz, M. B. (2010). 'Functional roles for noise in genetic circuits'. *Nature* **467**(7312), 167–173.

Ellis, T., Wang, X., & Collins, J. J. (2009). 'Diversity-based, model-guided construction of synthetic gene networks with predicted functions.' *Nature Biotechnology* **27**(5), 465–471.

Ferrell Jr, J. E. (2002). 'Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability'. *Current opinion in cell biology* **14**(2), 140–148.

Gardner, T. S., Cantor, C. R., & Collins, J. J. (2000). 'Construction of a genetic toggle switch in Escherichia coli'. *Nature* **403**(6767), 339–342.

Gillespie, D. T. (1977). 'Exact Stochastic Simulation of Coupled Chemical-Reactions'. *Journal of Physical Chemistry* **81**(25), 2340–2361.

Guantes, R. & Poyatos, J. F. (2008). 'Multistable decision switches for flexible control of epigenetic differentiation.' *PLoS Computational Biology* **4**(11), e1000235.

Hafner, M., Koeppl, H., Hasler, M., & Wagner, A. (2009). ''Glocal' Robustness Analysis and Model Discrimination for Circadian Oscillators'. *PLoS Computational Biology* **5**(10), e1000534.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., & Kummer, U. (2006). 'COPASI–a COmplex PAthway SImulator.' *Bioinformatics (Oxford, England)* **22**(24), 3067–3074.

Hucka, M., Finney, A., Bornstein, B. J., Keating, S. M., Shapiro, B. E., Matthews, J., Kovitz, B. L., Schilstra, M. J., Funahashi, A., Doyle, J. C., & Kitano, H. (2004). 'Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project.' *Systems Biology, IEE Proceedings* **1**(1), 41–53.

Ingalls, B. & Iglesias, P. (2010). 'A primer on control engineering'. In: *Control theory and systems biology*. Ed. by B. Ingalls & P. Iglesias. Cambridge, MA: MIT Press. Chap. 1, 1–28.

Kaplan, D. & Glass, L. (1995). *Understanding nonlinear dynamics*. Springer-Verlag.

Khalil, A. S. & Collins, J. J. (2010). 'Synthetic biology: applications come of age'. *Nature Publishing Group* **11**(5), 367–379.

Khammash, M. (2010). 'Modeling and analysis of stochastic biochemical networks'. In: *Control theory and systems biology*. Ed. by B. Ingalls & P. Iglesias. Cambridge, MA: MIT Press. Chap. 2, 29–44.

Kim, J., Bates, D. G., Postlewaite, I., Ma, L., & Iglesias, P. A. (2006). 'Robustness analysis of biochemical network models'. *Systems biology*.

Kitano, H. (2007). 'Towards a theory of biological robustness'. *Molecular systems biology* **3**.

Kobayashi, H., Kaern, M., Araki, M., Chung, K., Gardner, T. S., Cantor, C. R., & Collins, J. J. (2004). 'Programmable cells: interfacing natural and engineered gene networks.' *Proceedings of the National Academy of Sciences of the United States of America* **101**(22), 8414–8419.

Konopka, A. (2007). *Systems Biology, Principles, methods and concepts*. CRC Press.

Kramer, B. P., Viretta, A. U., Daoud-El-Baba, M., Aubel, D., Weber, W., & Fussenegger, M. (2004). 'An engineered epigenetic transgene switch in mammalian cells.' *Nature Biotechnology* **22**(7), 867–870.

Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J. L., & Hucka, M. (2006). 'BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.' *Nucleic Acids Research* **34**(Database issue), D689–D691.

Lipshtat, A., Loinger, A., Balaban, N. Q., & Biham, O. (2006). 'Genetic toggle switch without cooperative binding.' *Physical review letters* **96**(18), 188101.

Loinger, A. & Biham, O. (2009). 'Analysis of genetic toggle switch systems encoded on plasmids.' *Physical review letters* **103**(6), 068104.

Lotka, A. J. (1925). *Elements of Physical Biology*. Williams and Wilkins.

Lu, M., Onuchic, J., & Ben-Jacob, E. (2014). 'Construction of an Effective Landscape for Multistate Genetic Switches'. *Physical review letters* **113**(7), 078102.

Lu, T. K., Khalil, A. S., & Collins, J. J. (2009). 'Next-generation synthetic gene networks'. *Nature Biotechnology* **27**(12), 1139–1150.

Ma, R., Wang, J., Hou, Z., & Liu, H. (2012). 'Small-number effects: a third stable state in a genetic bistable toggle switch'. *Physical review letters* **109**(24), 248107.

Marjoram, P., Molitor, J., Plagnol, V., & Tavare, S. (2003). 'Markov chain Monte Carlo without likelihoods.' *Proceedings of the National Academy of Sciences of the United States of America* **100**(26), 15324–15328.

Mathematica (2016). *version 10.3*. Wolfram Research, Inc.

Niwa, H., Toyooka, Y., Shimosato, D., Strumpf, D., Takahashi, K., Yagi, R., & Rossant, J. (2005). 'Interaction between Oct3/4 and Cdx2 Determines Trophectoderm Differentiation'. *Cell* **123**(5), 13–13.

Prill, R. J., Iglesias, P. A., & Levchenko, A. (2005). 'Dynamic properties of network motifs contribute to biological network organization.' *PLoS Biology* **3**(11), e343–e343.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). 'Population growth of human Y chromosomes: A study of Y chromosome microsatellites'. *Molecular Biology and Evolution* **16**(12), 1791–1798.

Shinar, G. & Feinberg, M. (2010). 'Structural Sources of Robustness in Biochemical Reaction Networks'. *Science* **327**(5971), 1389–1391.

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., & DOYLE, J. (2004). 'Robustness of cellular functions'. *Cell* **118**(6), 675–685.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. H. (2009). 'Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems.' *Journal of the Royal Society, Interface / the Royal Society* **6**(31), 187–202.

Walczak, A. M., Onuchic, J. N., & Wolynes, P. G. (2005). 'Absolute rate theories of epigenetic stability'. *Proceedings of the National Academy of Sciences of the United States of America* **102**(52), 18926–18931.

Warren, P. B. & ten Wolde, P. R. (2004). 'Enhancement of the Stability of Genetic Switches by Overlapping Upstream Regulatory Domains'. *Physical review letters* **92**(12), 128101.

Warren, P. B. & ten Wolde, P. R. (2005). 'Chemical models of genetic toggle switches'. *The Journal of Physical Chemistry B* **109**(14), 6812–6823.

What is Flow Cytometry. Accessed: 2016-08-01.

Wilkinson, D. J. (2006). *Stochastic Modelling for Systems Biology.* CRC Press.

Woods, M. L., Leon, M., Perez-Carrasco, R., & Barnes, C. P. (2016). 'A Statistical Approach Reveals Designs for the Most Robust Stochastic Gene Oscillators.' *ACS Synthetic Biology* **5**(6), 459–470.

Zamora-Sillero, E., Hafner, M., Ibig, A., Stelling, J., & Wagner, A. (2011). 'Efficient characterization of high-dimensional parameter spaces for systems biology.' *BMC systems biology* **5**, 142.

Zhou, Y., Liepe, J., Sheng, X., Stumpf, M. P. H., & Barnes, C. (2011). 'GPU accelerated biochemical network simulation.' *Bioinformatics (Oxford, England)* **27**(6), 874–876.