Contents

1 Background

1.1 Introduction to synthetic biology

Synthetic biology aims at the rational design and construction of biological parts, devices, and systems in order to engineer organisms to perform new tasks (Lu, Khalil, & Collins 2009; Andrianantoandro et al. 2014). A part is a basic unit, like a promoter or a ribosome binding site that when combined with other parts will make a functional unit, a device (Heinemann & Panke 2006). A device processes inputs, performs functions and produces outputs (Andrianantoandro et al. 2014). A system comprises of a collection of devices.

Emphasis is put on the use of engineering principles such as modularity, standardisation, use of predictive models and the separation of design and construction (Agapakis & Silver 2009; Heinemann & Panke 2006). A hierarchy similar to computer science is used, with cells, pathways and biochemical reactions acting as computers, modules and gates respectively (Andrianantoandro et al. 2014).

Numerous applications of synthetic biology have emerged, from altering existing metabolisms to producing synthetic drugs (Holtz & Keasling 2010) or creating new synthetic life forms (Agapakis & Silver 2009). Despite the successes there is still a lack of predictive power due to the stochasticity and lack of complete knowledge of the cellular environment (Andrianantoandro et al. 2014).

Synthetic biology is now entering an age where simple synthetic circuits have been built, such as toggle switches (**Isaacs:2003ht**; **Deans:2007cya**; Gardner, Cantor, & Collins 2000; Kramer et al. 2004; Ham et al. 2008; Friedland et al. 2009), oscillators (Stricker et al. 2008; Fung et al. 2005; Tigges et al. 2009) and pulse generators (Basu et al. 2004), but larger circuits have proven more difficult (XXX). The leap from building low-level circuits to assembling them into complex networks has yet to be made successfully (Lu, Khalil, & Collins 2009), and predictable circuit behaviour remains challenging (XXX). Efforts to do so are plagued by intra-circuit

crosstalk and incompatibility, as well as cellular noise, which can render synthetic networks non-functional *in vivo* (XXX).

1.2 System design in synthetic biology

Creating synthetic devices that are robust to changing cellular contexts will be key to the success of synthetic biology. Unknown initial conditions and parameter values as well as the variability of the cellular environment, extracellular noise and crosstalk makes the majority of synthetic genetic devices non-functional (Chen, Chang, & Lee 2009). Designing devices robust to this environment will lead to reliable behaviour of the systems. When faced with a set of competing designs for a given genetic circuit, one is likely to choose the simplest possible model that can achieve the desired behaviour. However, simple systems are often the least robust. Feedback loops are well known key regulatory motifs (Brandman et al. 2005). Negative feedback loops are essential for homeostasis and buffering (Thomas, Thieffry, & Kaufman 1995) thus increasing robustness to extrinsic noise sources and positive feedback loops can generate multistationarity in a system (Thomas, Thieffry, & Kaufman 1995). Incorporating this kind of additional feedback interactions can make a design more robust and reliable. Maximising production is an important goal for a metabolic engineering project if it is to produce an economically viable substance (Holtz & Keasling 2010). Network topologies and parameter values of different toggle switch designs are explored here in order to identify the design that maximises robustness and distance between steady states. This ensures the reliable production of the product with the greatest distance between the on and off states of the switch. In the future, by selecting the system components accordingly, the parameter values can be adjusted *in vivo*. For example, the parameter value corresponding to the translation initiation rate can be chosen by selecting the appropriate RBS sequence which given a nucleotide sequence will produce the desired rate (Holtz & Keasling 2010), a method developed by Salis, Mirsky, & Voigt (2009). Another method to tweak the parameter values *in vivo* is to select the promoter to have the strength corresponding to the levels of gene expression and repression desired. Activity of each promoter can be measured and standardised (Kelly et al. 2009) making this process possible. For a system requiring more than one promoter, these can be efficiently selected from a promoter library using a genetic algorithm created by Wu, Lee, & Chen (2011). These standardised interchangeable components with known sequence and activity are what synthetic biology classes

as BioBricks (Kelly et al. 2009; Canton, Labno, & Endy 2008). These can be selected and used to construct a desired system and replicate the parameter values found in the scan presented here.

The first computational approach for the tuning of robust synthetic networks was that of Batt et al. (2007) where they examined the problem of finding a subset of the parameter set for which a given property was satisfied for all the parameters. Chen, Chang, & Lee (2009) used the fuzzy dynamic game method to solve the minimax regulation design problem of synthetic genetic networks. In that method the worst case effect of all disturbances is minimised for a given network. An evolutionary algorithm has also been used to solve the robust design problem by evolving the parameters of the system in order to make it more robust to cellular disturbances by **Chen:2011hj** The added value of the methodology presented here is that the network structure in addition to the network parameters are adjusted to select a network that can robustly create the desired behaviour.

1.3 Introduction to Biochemical Modelling

1.3.1 Graphical representation of biochemical systems

It is common to represent coupled biochemical reactions graphically. In a graph, as shown in Figure **??**, 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:2007b**). 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:2007b**).

1.3.2 Deterministic and Stochastic 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**). When modelling a biochemical system it is generally assumed that the rates of a reaction are directly proportional to the concentration of the reactants, raised to the power of their stoichiometry (**wilkinson:2006**). This is known as mass-action kinetics and is used in this work to model the various systems. There are two main ways

of modelling a system, deterministically and stochastically. Deterministic modelling utilises Ordinary differential equation (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). When modelling deterministically the model is viewed as a system which, with sufficient knowledge of the system, its behaviour is entirely predictable. Nevertheless we are still a long way away from having complete knowledge of a system of interesting size (**wilkinson:2006**). Deterministic modelling also assumes a homogenous mixture where species concentrations vary continuously and deterministically, assumptions that often are not met *in vivo*. A cell is spatially and temporally separated, due to small molecule numbers and fluctuations in the timing of processes (de Jong 2002).

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). It takes uncertainty into account and does not assume a homogenous mix. It is thus often more appropriate for modelling cellular systems, although more computationally intensive. 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. Number of molecules in the system initialised
2. Two random numbers generated, one to determine which reaction will occur next and one to determine the time step
3. Time step increased and molecule counts updated according to Step 2
4. Repeat from Step 2 until total simulation time reached

1.3.3 Steady state and stability

In a steady state, the state of a system remains fixed. In non-linear systems, like the ones systems biology deals with, there is generally not an analytical solution thus the system has to be solved numerically. A stable steady state is defined as a fixed point whose nearby points approach the fixed point (**kaplan:1959**). 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**).

1.4 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 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 Importance 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:2007cya**; Kramer et al. 2004). Other synthetic applications of the toggle switch include the construction of a synthetic genetic clock (**Atkinson:2003tu**), 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 (**Wilkinson:2006**; de Jong 2002). 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 1.1. 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 1.1 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)

1.5 Introduction to Bayesian statistics

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

because

$$p(x)p(\theta|x) = p(\theta)p(x|\theta)$$

where $p(x|\theta)$ is the likelihood, $p(\theta)$ is the prior, and $\int p(x|\theta)p(\theta)d\theta$ is the evidence. This is the normalisation.

Bayes factor:

$$B_{12} = \frac{\int p(x|\theta, M_1)p(\theta, M_1)d\theta}{\int p(x|\theta, M_2)p(\theta, M_2)d\theta}$$

In our case, O is the objective, and D is the design. Therefore:

$$p(O|D_1) = \int p(O|\theta, D_1)p(\theta|D_1)d\theta,$$

This is the robustness, or evidence or marginal likelihood

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

where $\underline{\Theta} = \{\theta_1, \theta_2, \theta_3\}$

Assuming the prior is uniform, and $a = 0$:

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

Assuming uniform likelihood:

$$p(O|D_1) = \frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}\iiint_{\underline{\Theta}_F} 1d\theta_1\theta_2\theta_3 + \frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}\iiint_{\underline{\Theta}_F} Od\underline{\Theta}$$

1.5.1 Bayes' theorem

1.5.2 Bayesian inference

1.5.3 Model checking

1.5.4 Prior selection

A circuit must be robust to a fluctuating cellular environment and its response and sensitivity must be able to be fine tuned in order to orchestrate a network of circuits that function together. A robust circuit can tolerate the compound stochasticity that a chain of circuits brings, and fine tuning of its response and sensitivity enables the researcher to make it sensitive to an upstream signal as well as influence a downstream subsystem. Parts can be fine tuned by developing component libraries (Lu, Khalil, & Collins 2009), but this will be of little use if the required parameter ranges for parts to make a functional complex network are unknown, and will only perpetuate the cycles of trial-and-error. A computational method to find the range of parameter values that will produce the behaviour of choice is crucial to the design process by enabling the informed selection of appropriate parts from the libraries. For example, if it is known that gene expression must be low for a given stability, one can select a weak promoter or a low copy plasmid for the desired construct.

Both analytical and computational approaches have been deployed for the study of the toggle switch. Analytical approaches are limited to simpler models and thus require a number of assumptions to be made. The system under consideration has to be reduced to very few equations and parameters in order to make the system solvable. This requires assumptions to be made about the system that cannot always be justified, such as the quasi-steady state approximation (QSSA). The QSSA assumes that the binding/unbinding processes are much faster than any other process (**Loinger:2007vma**), thus the bound intermediate is assumed to always be in steady state. The QSSA assumption is met *in vitro* but often does not hold *in vivo* and its misuse can lead to large errors and incorrectly estimated parameters (Pedersen, Bersani, & Bersani 2007). Moreover, it is generally not possible to solve even simple stochastic models analytically, and these methods are restricted to deterministic models. The computational and graph-theoretic approaches developed for the study of multistationarity generally focus on deciding on whether a given system is incapable of producing multiple steady states (Conradi et al. 2007; Banaji & Craciun 2010; Feliu & Wiuf 2013). For example, Feliu & Wiuf (2013) developed an approach using chemical reaction theory and generalised mass action modelling

(Feliu & Wiuf 2013). No approach exists that can handle both deterministic and stochastic systems in an integrated manner.

For this purpose, I developed a computational framework based on sequential Monte Carlo that takes a model and determines whether it is capable of producing a given number of (stable) steady states and the parameter space that gives rise to the behaviour. Uniquely, this can be done for both deterministic and stochastic models, and also complex models with many parameters, thus removing the need for simplifying assumptions. This framework can be used for comparing the conclusions drawn by various modelling approaches and thus provides a way to investigate appropriate abstractions. I have made this framework into a python package, called Stability Finder.

I use this methodology to investigate genetic toggle switches and uncover the design principles behind making a bistable switch, as well as those necessary to make a tristable and a quadristable switch (4 steady states). I also demonstrate the ability of Stability Finder to examine more complex systems and examine the design principles of a three gene switch. The examples I used demonstrate that Stability Finder will be a valuable tool in the future design and construction of novel gene networks.

1.5.5 Model parametric Robustness

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:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int p(x|\theta)p(\theta)d\theta} \tag{1.1}$$

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.

$$p(O|D_1) = \int p(O|\theta, D_1)p(\theta|D_1)d\theta, \tag{1.2}$$

For a design with three parameters this becomes:

$$p(O|D_1) = \iiint_{\underline{\Theta}} p(O|\underline{\Theta})p(\underline{\Theta}|D_1)d\underline{\Theta}, \tag{1.3}$$

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:

$$p(O|D_1) = \iiint_{\underline{\Theta}} p(O|\underline{\Theta})\frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}d\underline{\Theta} \text{ , and} \tag{1.4}$$

$$p(O|D_1) = \frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}\iiint_{\underline{\Theta}} p(O|\underline{\Theta})d\underline{\Theta} \tag{1.5}$$

since $\frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}$ is a constant. Then assuming that the likelihood is uniform Equation1.5 becomes:

$$p(O|D_1) = \frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}\Big[\iiint_{\underline{\Theta}_F} 1d\underline{\Theta} + \cancelto{0}{\iiint_{\underline{\Theta}_{\not F}} 0d\underline{\Theta}}\Big] \tag{1.6}$$

$$\tag{1.7}$$

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

$$p(O|D_1) = \overbrace{\frac{1}{b_1}\frac{1}{b_2}\frac{1}{b_3}}^{|P|}\underbrace{\iiint_{\underline{\Theta}_F} 1d\underline{\Theta}}_{|F|}, \tag{1.8}$$

$$\therefore p(O|D_1) = \frac{|F|}{|P|}, \tag{1.9}$$

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, Equation1.5 becomes:

$$p(O|D_1) = \frac{1}{|P|} \iiint_{\underline{\Theta}} f(\underline{\Theta}; \mu, \Sigma) d\underline{\Theta} \tag{1.10}$$

$$\therefore p(O|D_1) = \frac{1}{|P|} \underbrace{\times (2\pi)^{\frac{k}{2}} \times |\Sigma|^{\frac{1}{2}}}_{|F|} \tag{1.11}$$

$$\therefore p(O|D_1) = \frac{|F|}{|P|}, \tag{1.12}$$

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

$$B_{ab} = \frac{\int p(x|\theta, D_a) p(\theta, D_a) d\theta}{\int p(x|\theta, D_b) p(\theta, D_b) d\theta} \tag{1.13}$$

$$\therefore B_{ab} = \frac{|Fa|}{|Pa|} / \frac{|Fb|}{|Pb|} \tag{1.14}$$

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.

1.6 Approximate Bayesian Computation (ABC)

1.6.1 ABC algorithms

Stability Finder is based on a statistical inference method which combines ABC with Sequential Monte Carlo (SMC) (Toni et al. 2009). This simulation-based method uses an iterative process to arrive at a distribution of parameter values that can give rise to observed data or a desired system behaviour (Barnes et al. 2011).

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: $$w_t^{(i)} = \begin{cases} 1, & \text{if } n = 0 \\ \frac{P(\theta_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta_t^{(i)})}, & \text{if } n>0. \end{cases}$$

1.6.2 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.

1.6.3 Perturbation

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

$$K_p(\theta|\theta*) = \theta * + U(+s_p, -s_p), \text{ where:} \tag{1.15}$$

$$s_p = \frac{1}{2}\big(max(\theta_{p-1}) - min(\theta_{p-1})\big) \tag{1.16}$$

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.

1.6.4 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.

1.6.5 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.

$$w_t^{(i)} = \frac{P(\theta_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta_t^{(i)})} \text{ for n} > 0 \tag{1.17}$$

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.

$$\frac{d[B]}{dt} = p[A] \tag{1.18}$$

$$\frac{d[A]}{dt} = -p[A] \tag{1.19}$$

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 1.1.

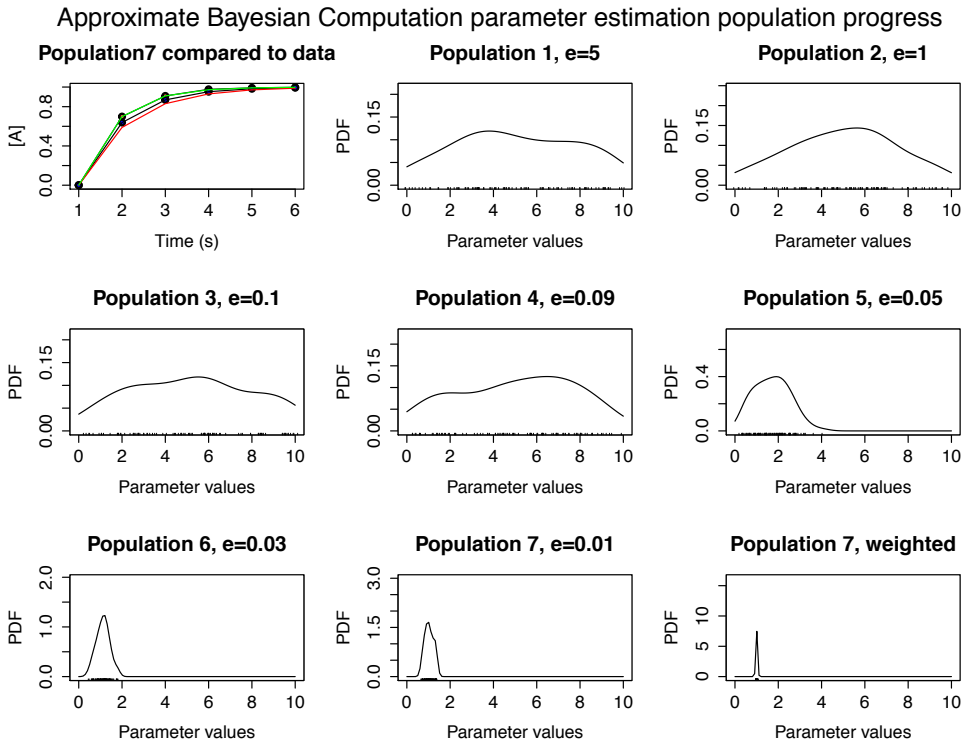

Figure 1.1 ABC SMC parameter inference. The posterior parameter is equal to 1 and its time course shown in red in the top left panel. The blue time course is that of the final population, green is the upper quartile and red is the lower quartile range of values. The progress of the selection process can be seen the εschedule 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 1.1 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.

1.7 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 1.2. 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 used in synthetic biology for BioBrick characterisation (Kelly et al. 2009), enzyme screening (Choi et al. 2014) and industrial bioprocesses (Díaz et al. 2010) among others. 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).

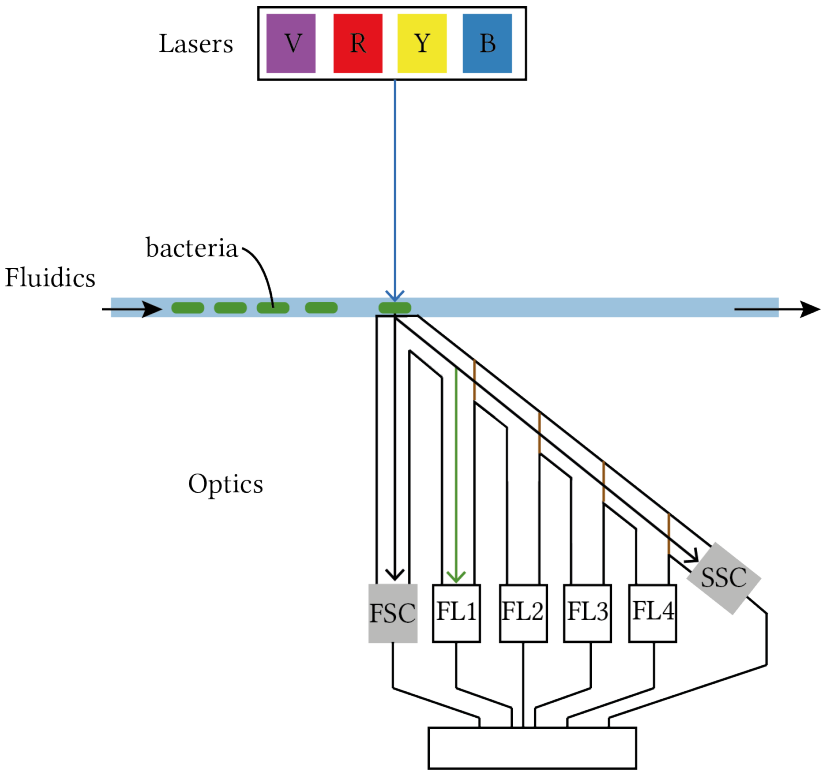

Figure 1.2 : Flow cytometry. A laser excites the fluorescent proteins present in each cell. The cytometer has up to 4 lasers, violet (V), red (R), yellow (Y) and blue (B). The detectors in the optics, FL1-4 pick up the signals. The cytometer also picks up size and granularity information via the forward scatter (FSC) and side scatter (SSC) detectors.

Bibliography

Agapakis, C. M. & Silver, P. A. (2009). 'Synthetic biology: exploring and exploiting genetic modularity through the design of novel biological networks'. *Molecular BioSystems* **5**(7), 704.

Andrianantoandro, E., Basu, S., Karig, D. K., & Weiss, R. (2014). 'Synthetic biology: new engineering rules for an emerging discipline.' *Molecular systems biology* **2**(1), 2006.0028.

Attune NxT Acoustic Focusing Cytometer (2015). CO016625.

Banaji, M. & Craciun, G. (2010). 'Graph-theoretic criteria for injectivity and unique equilibria in general chemical reaction systems'. *Advances in Applied Mathematics* **44**(2), 168–184.

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

Barnes, C. P., Silk, D., Sheng, X., & Stumpf, M. P. H. (2011). 'Bayesian design of synthetic biological systems.' *Proceedings of the National Academy of Sciences of the United States of America* **108**(37), 15190–15195.

Basu, S., Mehreja, R., Thiberge, S., Chen, M.-T., & Weiss, R. (2004). 'Spatiotemporal control of gene expression with pulse-generating networks.' *Proceedings of the National Academy of Sciences of the United States of America* **101**(17), 6355–6360.

Batt, G., Yordanov, B., Weiss, R., & Belta, C. (2007). 'Robustness analysis and tuning of synthetic gene networks.' *Bioinformatics (Oxford, England)* **23**(18), 2415–2422.

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*.

Brandman, O., Ferrell, J. E., Li, R., & Meyer, T. (2005). 'Interlinked fast and slow positive feedback loops drive reliable cell decisions.' *Science* **310**(5747), 496–498.

Canton, B., Labno, A., & Endy, D. (2008). 'Refinement and standardization of synthetic biological parts and devices.' *Nature Biotechnology* **26**(7), 787–793.

Chen, B.-S., Chang, C.-H., & Lee, H.-C. (2009). 'Robust synthetic biology design: stochastic game theory approach.' *Bioinformatics (Oxford, England)* **25**(14), 1822–1830.

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.

Choi, S.-L., Rha, E., Lee, S. J., Kim, H., Kwon, K., Jeong, Y.-S., Rhee, Y. H., Song, J. J., Kim, H.-S., & Lee, S.-G. (2014). 'Toward a generalized and high-throughput enzyme screening system based on artificial genetic circuits.' *ACS Synthetic Biology* **3**(3), 163–171.

Conradi, C., Flockerzi, D., Raisch, J., & Stelling, J. (2007). 'Subnetwork analysis reveals dynamic features of complex (bio)chemical networks'. *PNAS* **104**(49), 19175–19180.

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

Díaz, M., Herrero, M., García, L. A., & Quirós, C. (2010). 'Application of flow cytometry to industrial microbial bioprocesses'. *Biochemical Engineering Journal* **48**(3), 385–407.

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

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.

Feliu, E. & Wiuf, C. (2013). 'A computational method to preclude multistationarity in networks of interacting species.' *Bioinformatics (Oxford, England)* **29**(18), 2327–2334.

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.

Friedland, A. E., Lu, T. K., Wang, X., Shi, D., Church, G., & Collins, J. J. (2009). 'Synthetic gene networks that count.' *Science* **324**(5931), 1199–1202.

Fung, E., Wong, W. W., Suen, J. K., Bulter, T., Lee, S. G., & Liao, J. C. (2005). 'A synthetic gene–metabolic oscillator '. *Nature* **435**(7038), 118–122.

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.

Ham, T. S., Lee, S. K., Keasling, J. D., & Arkin, A. P. (2008). 'Design and Construction of a Double Inversion Recombination Switch for Heritable Sequential Genetic Memory'. *PLoS ONE* **3**(7), e2815.

Heinemann, M. & Panke, S. (2006). 'Synthetic biology–putting engineering into biology'. *Bioinformatics (Oxford, England)* **22**(22), 2790–2799.

Holtz, W. J. & Keasling, J. D. (2010). 'Engineering Static and Dynamic Control of Synthetic Pathways'. *Cell* **140**(1), 19–23.

Kelly, J. R., Rubin, A. J., Davis, J. H., Ajo-Franklin, C. M., Cumbers, J., Czar, M. J., de Mora, K., Glieberman, A. L., Monie, D. D., & Endy, D. (2009). 'Measuring the activity of BioBrick promoters using an in vivo reference standard.' *Journal of Biological Engineering* **3**(1), 4–4.

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

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.

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.

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.

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.

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.

Pedersen, M. G., Bersani, A. M., & Bersani, E. (2007). 'Quasi steady-state approximations in complex intracellular signal transduction networks – a word of caution'. *Journal of Mathematical Chemistry* **43**(4), 1318–1344.

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.

Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). 'Automated design of synthetic ribosome binding sites to control protein expression.' *Nature Biotechnology* **27**(10), 946–950.

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.

Stricker, J., Cookson, S., Bennett, M. R., Mather, W. H., Tsimring, L. S., & Hasty, J. (2008). 'A fast, robust and tunable synthetic gene oscillator'. *Nature* **456**(7221), 516–519.

Thomas, R., Thieffry, D., & Kaufman, M. (1995). 'Dynamical behaviour of biological regulatory networks–I. Biological role of feedback loops and practical use of the

concept of the loop-characteristic state.' *Bulletin of mathematical biology* **57**(2), 247–276.

Tigges, M., Marquez-Lago, T. T., Stelling, J., & Fussenegger, M. (2009). 'A tunable synthetic mammalian oscillator.' *Nature* **457**(7227), 309–312.

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.

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.

Wu, C.-H., Lee, H.-C., & Chen, B.-S. (2011). 'Robust synthetic gene network design via library-based search method'. *Bioinformatics (Oxford, England)* **27**(19), 2700–2706.

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.