A stochastic model of gene switches

Lucia Santoso* Hilary S. Booth* Conrad J. Burden†
Markus Hegland‡

(Received 19 November 2004; revised 30 March 2005)

Abstract

We present a stochastic model of genetic regulation where the expression of genes are controlled by protein levels. In particular, we examine a genetic toggle switch with two competing proteins where one protein switches off the other gene. We model this switching behaviour in the framework of the Stochastic Master Equation (SME), which is a continuous time variant of a Markov model used in chemical systems. Thus far, the SME is mainly solved by stochastic simulation

*Centre for Bioinformation Science, Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, AUSTRALIA
mailto:lucia.santoso@anu.edu.au

†Centre for Bioinformation Science, John Curtin School of Medical Research & Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, AUSTRALIA

‡Centre for Mathematics and its Application, Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, AUSTRALIA

See http://anziamj.austms.org.au/V46/CTAC2004/Sant for this article, Published June 29, 2005. ISSN 1446-8735

due to the perceived high computational demands. We explore approximation techniques which allow the numerical solution of the SME to be tractable.

Contents

1 Introduction

The expression of genetic information is regulated by mechanisms activating or repressing the transcription of genes. This process has been studied extensively in prokaryotes (organisms whose cells have no nucleus). Highly efficient mechanisms evolved that turn genes on and off, depending on the cell's needs in particular environments.

Gene regulatory mechanisms typically form large, complex networks, with proteins as regulatory agents, which exist in quantities ranging from 5–10 molecules to 100,000 molecules per cell, depending on their function in the

cell [4]. The actual regulatory action is protein-protein binding, or protein-DNA binding, which yields either the expression or repression of genes.

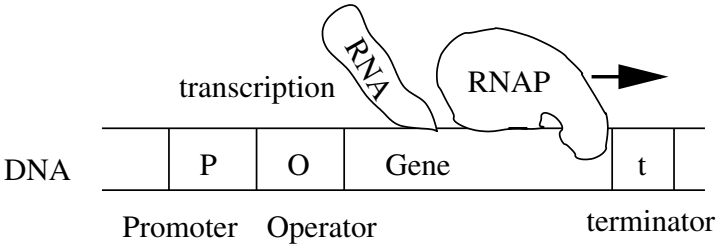
We investigate modelling gene regulatory networks using a continuous time Markovian framework, known as the Stochastic Master Equation (SME). We aim for a model which is sufficiently generic and transferable between systems. Thus we develop this probabilistic framework and a modular approach to the model. Here we focus on gene switches, which are a common module in regulatory networks.

First, we give a background on gene regulation, followed by a description of the SME. Second, we give approximations of the SME for computability purposes, concluded by discussion and future directions.

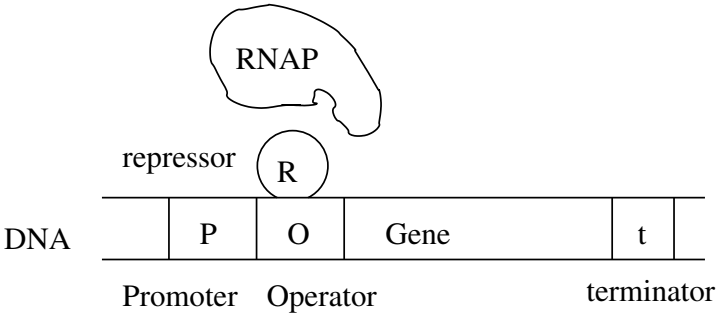
2 Regulation of gene expression

Simplistically, there are two main components in the mechanism of gene regulation: gene switches (ON and OFF) and dynamics of gene products (synthesis and degradation), such as protein. Genes are parts of the DNA which serve as recipes for proteins. This machinery functions as follows (Figure 1(a)): a molecule RNA polymerase (RNAP) binds to a *promoter* site, juxtaposed upstream to the start of the gene; moves along the gene producing RNA; and stops at the *terminator* site. This process is called *transcription*. A subset of RNAs called *messenger* RNA*s* (mRNA) is then *translated* to proteins. RNAP can be prevented from binding to a promoter by proteins called *repressors*, which bind to *operator* sites adjacent to the promoter (Figure 1(b)) [4]. When RNAP is able to bind to a promoter site and transcribe the gene, we refer to this gene as being in the ON state, otherwise it is OFF.

Some gene products control (facilitate or repress) the expression of genes (its own as well as others), as illustrated in Figure 2.



(a) Gene transcription



(b) Gene repression

FIGURE 1: Control of transcription

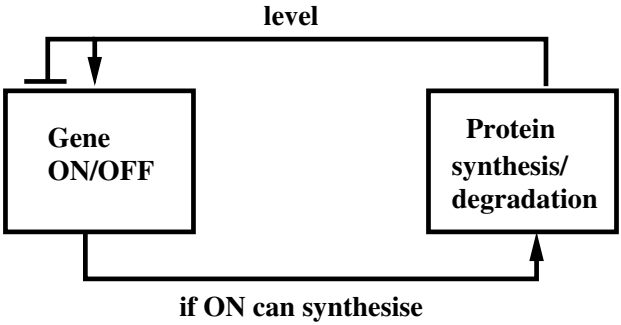


FIGURE 2: Gene-protein feedback system

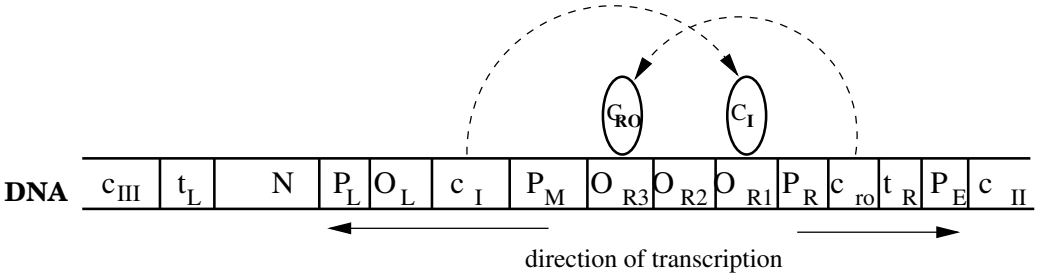


FIGURE 3: The main agents in the gene regulation in the λ-phage system

An example of a feedback switching mechanism is the switch which occurs in the life cycle of a bacteria infecting virus called the λ-phage. Figure 3 shows two competing proteins, C_I and C_{RO}, which repress each other's genes by binding to the juxtaposed operators.

3 The Stochastic Master Equation

The state space of the simplest genetic switch model comprises four dimensions, two each for genes $\in [0, 1]$ and two for the proteins $\in N = 0, 1, 2, \ldots$. The presence of a protein only affects the likelihood that the related event

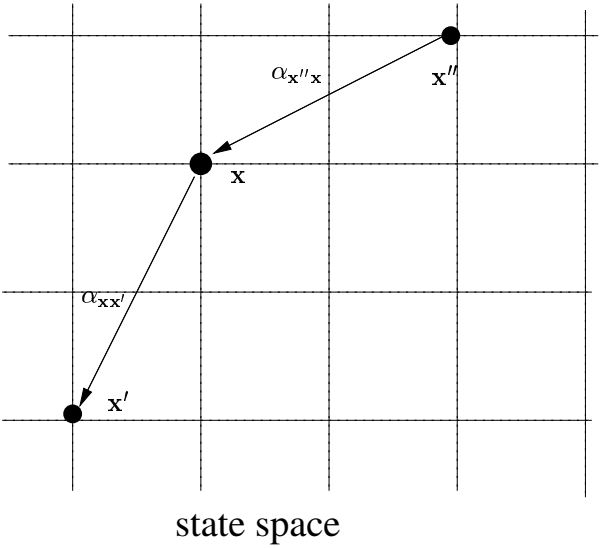


FIGURE 4: Representation of chemical reactions in state space

will occur, thus at any time t, our system has a probability $p(\mathbf{x}, t)$ of being in a state $\mathbf{x}$, and has a propensity $\alpha_{\mathbf{x}\mathbf{x}'}$ of moving to another state $\mathbf{x}'$ (Figure 4). Assuming that this system is Markovian, we write this as

$$\frac{\partial p}{\partial t}(\mathbf{x}, t) = \sum_{\mathbf{x}''} \alpha_{\mathbf{x}''\mathbf{x}} p(\mathbf{x}'', t) - \sum_{\mathbf{x}'} \alpha_{\mathbf{x}\mathbf{x}'} p(\mathbf{x}, t)\,, \tag{1}$$

where the two terms indicate the transitions from and to other states respectively. We write Equation 1 in matrix form:

$$\frac{\partial p}{\partial t}(\mathbf{x}, t) = -A p(\mathbf{x}, t)\,, \tag{2}$$

where the matrix A contains the propensities α. The above equation is known as the Stochastic Master Equation [3, 7].

This framework expresses a stochastic systems in terms of a system of linear ordinary differential equations. It can be solved by numerical integra-

tion, but many have thus far shied away from this approach due to perceived high computational costs. Instead, solution methods such as stochastic simulations are popular [1, 6]. In the following section we describe approximation methods, applied to the λ-phage system, to allow numerical solution for the SME to be feasible.

4 Two stochastic models and their approximation

We look at two basic components of the gene regulatory networks: protein production (synthesis and degradation) and the gene switching mechanism. For simplicity we do not impose constraints on the numbers of proteins.

4.1 Protein production

The state is described by the number of proteins $n \in N$. In this case a probability distribution is a nonnegative sequence $p = (p_0, p_1, \ldots)$ which sums to one. Any bounded sequence $x = (x_0, x_1, \ldots)$ defines a functional[1] acting on the probability distribution as

$$\langle x, p \rangle = \sum_{i=0}^{\infty} x_i p_i \,. \tag{3}$$

In a prokaryote, such as bacteria, protein production simply follows translation. In this simplified production process (Figure 5) protein is synthesised at a constant rate α and degrades at a rate βn proportional to the the amount of protein available. The propensity matrix (or more precisely, operator) is

[1] A function that takes functions as its argument.

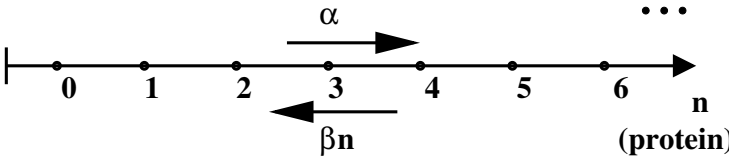


FIGURE 5: State space representation of protein synthesis and degradation.

$$A = \begin{bmatrix} \alpha & -\beta & & \\ -\alpha & \alpha+\beta & -2\beta & \\ & -\alpha & \alpha+2\beta & \\ & & -\alpha & \ddots \end{bmatrix} . \tag{4}$$

For modelling protein production one requires three operators defined on probability distributions: the shift S, its adjoint S^*, and a diagonal operator D. Their effects on a probability distribution p are: $(Sp)_0 = 0$ and $(Sp)_{i+1} = p_i$; $(S^*p)_i = p_{i+1}$; and $(Dp)_i = ip_i$; respectively, where $i = 0, 1, \ldots$, and S^* is the adjoint of S with respect to $\langle \cdot, \cdot \rangle$.

One can easily verify that S and D satisfy the commutator relation

$$[D, S] := DS - SD = S\,, \tag{5}$$

and consequently that S^* and D satisfy

$$[D, S^*] := DS^* - S^*D = -S^*\,. \tag{6}$$

Furthermore, note that $S^*e = e$ and $S^*S = I$ where I is the identity and

$$SS^*D = DSS^* = D\,. \tag{7}$$

The probability distribution of protein dynamics depends on (continuous) time, is denoted by $p(t)$ and satisfies the SME 2, where the operator A consists of a production term and a decay term:

$$A = \alpha(I - S) + \beta(I - S^*)D\,. \tag{8}$$

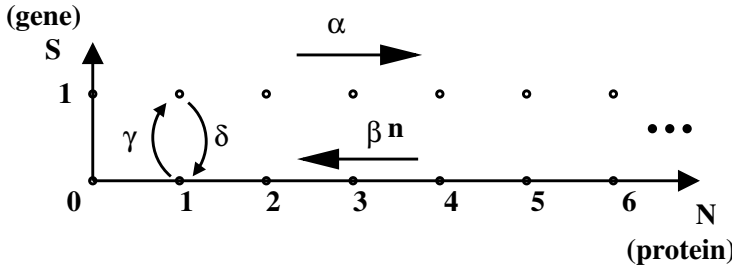


FIGURE 6: State space representation of one gene – one protein system

Let the mean (or first moment) be $\mu_1 = \langle e, Dp \rangle$, where $e = (1, 1, 1, \ldots)$. We determine the differential equation for the first moment to be

$$\frac{d\mu_1}{dt} = \alpha - \beta\mu_1 \,, \tag{9}$$

which is an ordinary reaction kinetic equation with a stable stationary point at $\mu_1 = \alpha/\beta$.

The state space is now augmented by the state of the gene switch which can be either ON or OFF. Thus one gets a probability distribution $p_{s,n}$ where $s \in [0, 1]$ and $n \in N$. Protein is only produced when the switch is ON, with propensities α, and decays in both gene switch states with propensity β, and the switch turns ON and OFF with propensities γ and δ respectively. We write the SME as

$$\frac{dp}{dt} = -Bp \,, \tag{10}$$

$$B = \alpha \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes (I - S) + \beta I_2 \otimes ((I - S^*)D) + \begin{bmatrix} \gamma & -\delta \\ -\gamma & -\delta \end{bmatrix} \otimes I \,.$$

If the first component in the tensor product corresponds to the lowest bit in the index then the matrix B consists of two by two blocks. The genetic switches are *much faster* than the protein production. As a first approximation we assume that the genetic switch instantly settles into a stationary

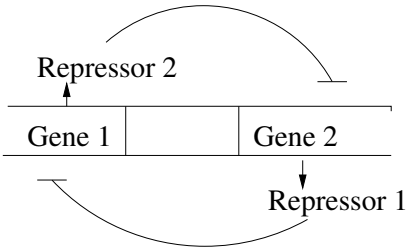


FIGURE 7: Gene switch

probability distribution which is $\pi = (\delta, \gamma)^T/(\gamma + \delta)$. The overall probability distribution is factorised as

$$p_{s,n} = q_n \pi_{s|n}\,, \tag{11}$$

and the conditional probability $\pi_{s|n}$ approximated by the stationary distribution above. The system matrix for the probability q_n is found as

$$(e^T \otimes I)B(\pi \otimes I) = \frac{\gamma}{\gamma + \delta}\alpha(I - S) + \beta(I - S^*)D\,, \tag{12}$$

where $e^T = (1, 1)$. Note that the decay is the same as in the previous example but now the production is reduced due to the fact that it occurs only when the switch is ON.

4.2 Switching feedback mechanism

We consider the feedback mechanism behind the gene switch, that is, when the gene product represses another gene. The presence of protein 1 reduces the propensity that gene 2 will be expressed. The state is now described by the state of two switches and two protein counts and is thus $(s, n) = (s_1, s_2, n_1, n_2)$. The master equation is

$$\frac{dp}{dt} = -Bp\,, \tag{13}$$

where $B = B_1 + B_2 + B_3$. Production of the two proteins is described by the matrix

$$B_1 = \alpha_1 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes I_2 \otimes (I - S) \otimes I + \alpha_2 I_2 \otimes \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes I \otimes (I - S), \quad (14)$$

and protein decay modelled by

$$B_2 = \beta_1 I_2 \otimes I_2 \otimes ((I - S^*)D) \otimes I + \beta_2 I_2 \otimes I_2 \otimes ((I - S^*)D) \otimes I. \quad (15)$$

The matrix B_3 describes the transitions of the gene switches:

$$B_3 = \left(I_2 \otimes \bigoplus_{n_1=0}^{\infty} \begin{bmatrix} \gamma & -n_1\delta \\ -\gamma & n_1\delta \end{bmatrix} \otimes I \right) + P \left(I_2 \otimes \bigoplus_{n_2=0}^{\infty} \begin{bmatrix} \gamma & -n_2\delta \\ -\gamma & n_2\delta \end{bmatrix} \otimes I \right) P^T, \quad (16)$$

and where P is a permutation matrix which swaps indices s_1, n_2 with s_2, n_1. The system matrix B is thus a block matrix with four by four blocks.

From these equations we now derive approximate ordinary differential equations for the protein concentrations using a variant of the aggregation method [5]. In this approximation it is assumed that the equations for the switching can be decoupled from the protein production equations due to the large difference in the time scale. For a given fixed number of proteins n_1 and n_2 one gets a stationary "switch" distribution as

$$\pi_{s|n} = \frac{1}{(\delta_1 n_2 + \gamma_1)(\delta_2 n_1 + \gamma_2)} \begin{bmatrix} \delta_1 n_2 \\ \gamma_1 \end{bmatrix} \otimes \begin{bmatrix} \delta_2 n_1 \\ \gamma_2 \end{bmatrix}. \quad (17)$$

Using this approximation, one gets the following agglomerated approximation of B as

$$\begin{aligned} \tilde{A} &= (e_2^T \otimes e_2^T \otimes I \otimes I) B \bigoplus_n \left[\pi_{s|n}\right] \otimes I \otimes I \\ &= \alpha_1\gamma_1 (I - S) \otimes (\gamma_1 I + \delta_1 D)^{-1} + \alpha_2\gamma_2 (\gamma_2 I + \delta_2 D)^{-1} \otimes (I - S) \\ &\quad + \beta_1 ((I - S^*)D) \otimes I + \beta_2 I \otimes ((I - S^*)D). \end{aligned} \quad (18)$$

The expected number of proteins

$$\mu_{11} = \langle (De) \otimes e, q \rangle \,, \tag{19}$$

and one gets the coupled differential equations

$$\frac{d\mu_{11}}{dt} = \frac{\alpha_1 \gamma_1}{\gamma_1 + \delta_1 \mu_{12}} - \beta_1 \mu_{11} \,, \tag{20}$$

$$\frac{d\mu_{12}}{dt} = \frac{\alpha_2 \gamma_2}{\gamma_2 + \delta_2 \mu_{11}} - \beta_2 \mu_{12} \,. \tag{21}$$

This is simply the pair of equations describing the dynamics of two competing proteins, such as in an engineered genetic toggle [2].

It can be shown that these equations have at most one meaningful stationary solution. Thus whereas this approximation may lead to some initially useful results, they do not provide an explanation for the bistability. For this one requires a more refined approximation which will be discussed in a future paper. It is known that the "exact" agglomeration technique does preserve the stationary solution of the master equation [5]. In our case it will have a bimodal distribution which corresponds to bistability.

5 Discussion and conclusion

We have presented here mathematical tools to enable solving numerically the Stochastic Master Equation (with infinite states), namely the representation of propensity matrices in terms of operators. Using this method we are able to compute the dynamics of the probability distribution of states (gene expression and proteins) for a coupled genetic switch system incorporating feedback and competing proteins.

This system is a main building block of a simple regulatory network, in the bacteriophage-λ. Most genetic regulatory networks are large and complex.

Therefore modelling frameworks for such systems need to have a generic form and be computationally viable. The stochastic master equation construct allows for fuller representation of the system, including both genes and products, and the stochasticity suits the nature of gene interactions. Furthermore, its simple form reduces the computational issue to merely solving a set (albeit large and often infinite) of ordinary differential equations. This has traditionally been considered intractable, hence the popular solution approach of stochastic simulation. But we have shown here, on a simple system, that using some "tricks", the numerical solution of the full SME is feasible. We showed how three simple scenarios, of protein dynamics alone, one gene–one protein and a competing two gene–two protein systems, can be represented in the SME framework in terms of three operators, allowing numerical solution methods.

There remain challenges, of course, namely the representation of (more) complicated networks and calculation of propensity constants. The latter is particularly problematic due to the lack of biological data containing measurements of absolute intracellular substance concentrations and free energy levels. Thus far, using relative values, we manage to simulate protein dynamics which display qualitatively the correct behaviour.

We have investigated model reduction techniques which preserve the Markov property, to enable modelling of (typically large and complicated) genetic regulatory networks. These techniques reduce the system to its simpler models which display an approximate behaviour of the original system and yet are still biologically meaningful.

References

[1] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *escherichia coli*. *Genetics*, 149:1633–1648, 1998. C536

[2] T. S. Gardner, C.R. Cantor, and J.J. Collins. Construction of a genetic toggle switch in *escherichia coli*. *Nature*, 403:339–342, 2000. C541

[3] D. T. Gillespie. *Markov Processes: an introduction for physical scientists*. Academic Press, San Diego, USA, 1992. C535

[4] B. Lewin. *Genes V*. Oxford University Press, 1994. C532

[5] W. Stewart. *Introduction to the numerical solution of Markov Chains*. Princeton University Press, 1994. C540, C541

[6] T. Tian and K. Burrage. Bistability and switching in the lysis/lysogeny genetic regulatory network of bacteriophage - λ. *Journal of Theoritical Biology*, 227:229–237, 2004. C536

[7] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North Holland, Amsterdam, the Netherlands, 1981. C535