Markov process modelling of gene regulation

Hilary S. Booth[1,2,3], Conrad J. Burden[1,2,3], Markus Hegland[2,4], Lucia Santoso[1,2,4]

Abstract

This paper discusses the mathematical modelling of gene regulation with emphasis on the bottom-up modelling of genetic componentry rather than the reverse-engineering of networks from gene expression data. Reflecting the stochastic nature of gene regulation, the chemical master equation is used as a tool to study Markovian models of networks of gene states between which probabilistic transitions occur. These states represent the binding/unbinding of protein complexes to DNA, resulting in a gene being expressed/not expressed in the cell, and concentrations of RNA, protein and any other chemical species required in the model. Basic genetic components such as gene repression and promotion and the gene cascade are described. We then describe a more complex system, the switching mechanism of the *Bacteriophage* λ, at it moves stochastically into one or other of its alternate lifestyles.

1   The central dogma of genetics

Within the nucleus of every cell of every human, long coils of DNA (deoxyribonucleic acid) form the chromosomes that contain encoded information necessary for the human being to develop, within a changing external environment, from a foetus to a child and as an adult, to reproduce and eventually to die. More DNA is present in the *mitochondria*, the energy-producing organelles within the cell, and this information is passed exclusively down the maternal line. The *human genome* is the sum of all the DNA (chromosomal and mitochondrial) in the cells of a human. The genome includes genes i.e. those sections of DNA that contribute to a function, which in turn determine physical appearance, certain behavioral characteristics, how well the organism combats specific diseases and other characteristics. The genome also includes mysteriously uninteresting regions of unknown function, often referred to as "junk DNA". The four chemical bases (or nucleotides) – adenine, guanine, thymine, cytosine – are abbreviated as A, G, T and C. The DNA strand that encodes the gene products is accompanied by a second *complementary* strand that is fully determined by the coding strand ($A \rightarrow T; T \rightarrow A; G \rightarrow C; C \rightarrow G$). The two strands form the double helix whose structure was discovered by Watson and Crick in 1965. The human genome consists of approximately 3×10^9 pairs of bases.

Many mathematical models in bioinformatics ultimately aim to model the relationship between *genotype* i.e. the DNA of an individual, and *phenotype* i.e. the physical characteristics of the individual [22]. In Figure 1 we show the "central dogma" of genetics: once a gene has been activated by the gene regulatory network it is *expressed* in the cell i.e. the gene's DNA is *transcribed* into mRNA which is in turn *translated* (via the codon alphabet) into a protein sequence made up of twenty amino acids. The protein folds into a 3D structure. In response to the needs of the cell and the demands of the external environment, the proteins perform functions, resulting in a phenotype. This hierarchical structure can be modelled at any level of detail [11, 9, 35].
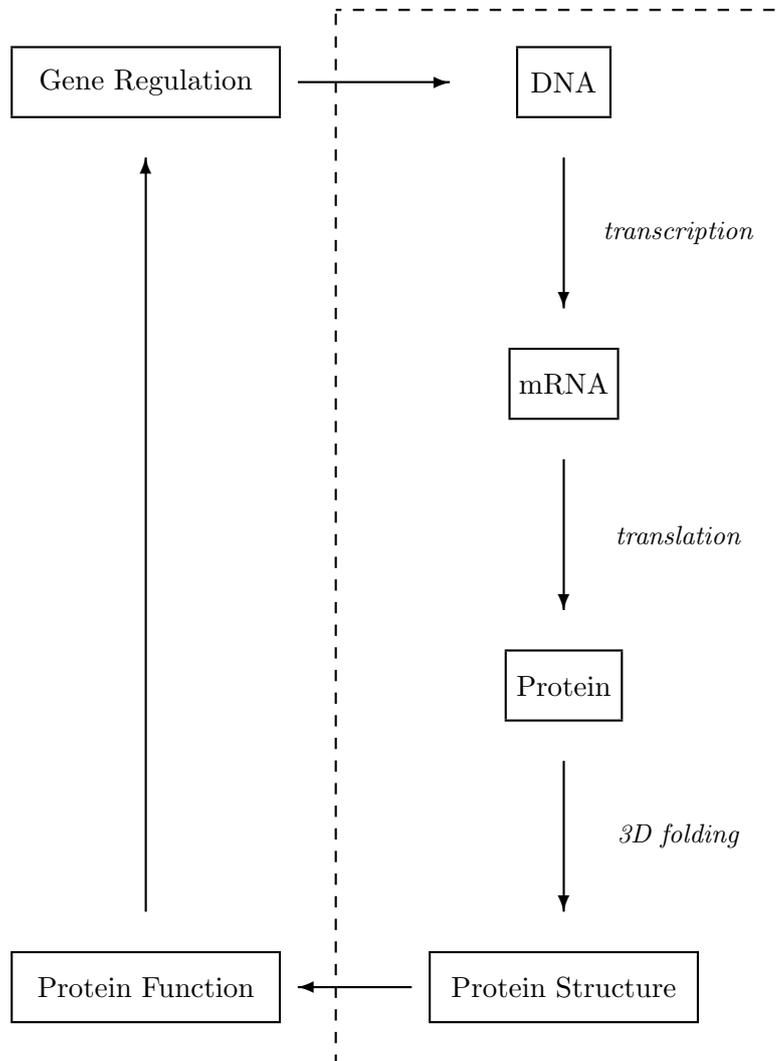
CENTRAL DOGMA

Figure 1. The central dogma of genetics: once a gene has been activated by the gene regulatory network it is *expressed* in the cell i.e. the gene's DNA is *transcribed* into mRNA which is in turn *translated* into a protein sequence made up of amino acids. The protein *folds* into a 3D structure. In response to the needs of the cell and the demands of the external environment, the proteins perform functions. One of those functions is gene regulation.

The first draft of the human genome was released in February 2001[20, 34]. At the same time that the human genome was being assembled, methods of sequence comparison were being developed, with which gene sequences could be compared to one another and their similarity assessed in a consistent manner across large databases [26, 31]. With the success

of tools such as BLAST and dynamic programming, we now have efficient algorithms for this purpose [2, 18], although fine-tuning of these tools continues.

2   Modelling gene regulatory networks

With the recent explosion of genomic data, one of the major concerns of bioinformatics is how genes are regulated and how their products interact within cellular networks. In a complex cell, gene products and external substrates regulate the genes that are expressed in that cell. See Figure 1. Some gene products promote other genes, usually depending on the concentration of the gene product. Other products repress the expression of genes. In some cases two genes compete for expression resulting in a population of cells distributed between the competing states (see Section 6). Yet other genes are expressed/repressed due to an outside parameter such as temperature (the sex-determining genes of the crocodile [5]) or UV light (the *Bacteriophage λ* [6]).

One of the major advances in experimental biology was the development of microarrays [21, 4]. Microarrays aim to measure the mRNA expression levels of many thousands of genes in a single experiment. Unfortunately there is a high level of noise in the mRNA data, and the interpretation of this data remains a difficult statistical challenge. Another problem is that the high cost of microarrays often precludes extensive replication. Furthermore, to infer a dynamic picture of gene regulation requires a time series of microarray experiments [32, 29]. Such data is very difficult and expensive to obtain – certainly it is beyond the reach of many laboratories. As time-series experiments become more affordable, these data are likely to drive the top-down approach to gene regulatory networks in which inferences are drawn from the gene expression within the cell.

This paper discusses some of the key aspects of the bottom-up modelling of gene regulation. The challenge here is to model the complex genetic componentry that enables a cell to switch genes on and off at the correct time [13, 8]. Smaller "toy" models have been developed to describe gene promotion and repression [15], and these components can be combined into more complex models [3, 19]. In some cases, artificial genetic machines based upon well-understood genetic components have been constructed and their behavior has been analyzed in a more controlled environment [17, 10]. To some extent, we can model the interplay between key genes, proteins and external substrates [1]. Some biologically stable states and bistable systems can be modelled using stochastic differential equations [3, 6]. Many of the commonly occurring genetic components can be modelled in this way [19].

There are two main problems in modelling gene regulation. The first of these is that, as the systems become more complex, the computational problems grow quickly. In this paper we show some examples of small systems that are well-understood. In some cases such as the *Bacteriophage λ* (phage λ) in Section 6, a switching mechanism depends upon a small number of key chemical species. But generally, at the level of the whole cell, the possible number of interactions increases dramatically.

The second, possibly larger problem is that although the reaction rates of key biological processes are likely to be measured in some form or other, they are not always known to the level of detail required for a rigorous mathematical model. The rates or probabilities with which the various chemical species interact with each other are functions of the entire state of the organism and the environment. It is a real challenge to quantify the biology at the experimental level [13]. Direct experimentation in the laboratory is the most reliable way to determine biological function or interactions, but the experiments need to be repeated many times under varying conditions if we are to obtain the rates of the reactions as reliable
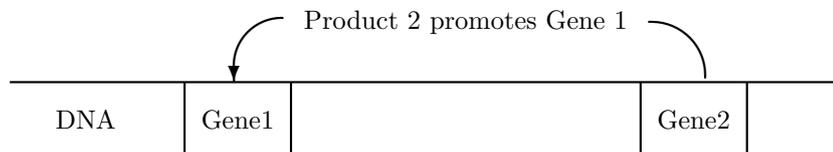
Figure 2. Gene Promotion: The product of Gene 2 promotes the expression of Gene 1.
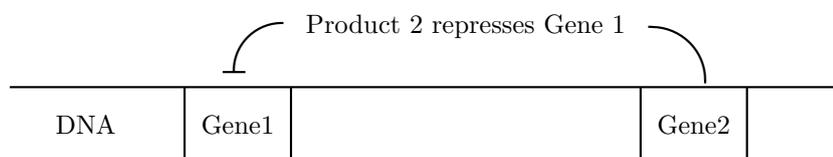

Figure 3. Gene repression: The product of Gene 2 represses the expression of Gene 1.

functions of the chemical species involved. This type of quantitative experiment is often prohibitively expensive.

It may be worth pointing out here that mathematical modelling may not always be the top priority to a researcher in biology whose breakthrough papers might be more descriptive. For all of these reasons, this whole area of research (modelling biological networks) is still in an early stage of development and many challenges lie ahead. At all levels of *genome-to-phenome* analysis mathematical and computational modelling improves incrementally as bioinformatics evolves alongside experimental technology. The most interesting problems in bioinformatics are driven by the biology.

3 Gene promotion and repression

Gene promotion and repression are the two basic building blocks of gene regulation. The biological mechanisms of gene promotion are complex but in the case where one gene is regulated by another, a positive feedback loop occurs (Fig. 2). Gene repression is the corresponding negative feedback loop (Fig. 3).

Interestingly enough, the negative feedback mechanism of gene repression occurs frequently in gene networks because it is often "easier" for nature to evolve a mechanism to switch a gene off than it is to evolve a new gene. Many genetic controls, such as the sex-determining genes in mammals, are complex combinations of off switches [14, 30].

A simple model of gene promotion and repression has four possible gene states:

$$\text{State } 1 = [0,0] \qquad (\text{gene 1 off, gene 2 off})$$
$$\text{State } 2 = [1,0] \qquad (\text{gene 1 on, gene 2 off})$$
$$\text{State } 3 = [0,1] \qquad (\text{gene 2 on, gene 1 off})$$
$$\text{State } 4 = [1,1] \qquad (\text{gene 1 on, gene 2 on})$$

These states are represented in the state diagram in Fig. 4. All possible transitions between the states are shown. The α_r, $r = 1, \ldots, 4$ are propensities for the gene to be switched on
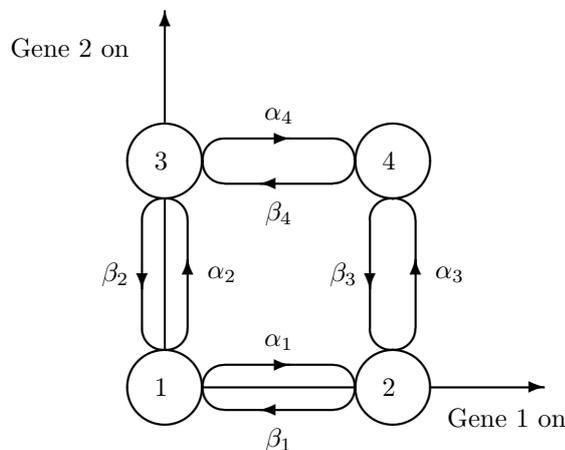
Figure 4. Generalized network for gene repression and promotion between genes 1 and 2 consists of all possible transitions between four states.

and the β_r are propensities for a gene to be switched off. A precise definition of propensity will be given in the next section. In a larger network, the above diagram can be generalized to a hypercube in which the α_r are the forward propensities i.e. gene promotion and the β_r are the backward propensities i.e. gene repression.

The probability of moving between these states depends upon the α_r and β_r. These in turn depend upon the state. This is the most difficult part of mathematically modelling gene regulation – to quantify the probability of moving between gene states given that the transition propensities will depend upon not only the gene states but also the protein states (which are not shown here).

4   Stochastic master equation model of gene regulation

The stochastic formulation of gene regulation is based on an assumption that the underlying physical processes are Markovian. An efficient tool for dealing with Markovian processes is the stochastic master equation [33]. In this formalism, a regulatory network is typically represented in a state space, elements of which describe the states or abundances of a finite number of chemical species which may be made up of any combination of genes, RNA, proteins or substrates. The gene states for example may be defined as gene on/off i.e. the gene is/is not being expressed in the cell. Alternatively, we may wish to specify that a protein or enzyme is attached/not attached to a promoter or operator site, and the protein states may be presence/absence of a protein or protein levels measured in numbers of molecules or in concentrations. Of course if continuous concentrations of proteins are used the number of states will be uncountable, so either some discretisation is necessary, or a continuum limit of the stochastic master equation must be taken, leading to a partial differential equation resembling the diffusion equation.

Consider a system that can be in any one of a finite number of states $n = 1, \ldots, N$, and capable of making transitions $r = 1, \ldots, R$ between states. The system is further assumed to be Markovian, that is, the probability of making a transition at any given time depends only on the state of the system at that time and not on its history. We represent the system by a directed graph with N nodes and R arcs. Associated with each transition is a *propensity*

$\alpha_r > 0$. If r is the transition from state m to state n, the probability of making the transition r in the time interval $[t, t+dt)$, conditional on being in state m at time t, is $\alpha_r dt$. Given an initial probability distribution among the states of $\mathbf{p}(0) = (p_1(0), \ldots, p_N(0))^{\mathrm{T}}$, with $p_n > 0$ and $\sum_n p_n = 1$, the system evolves in time to a distribution $\mathbf{p}(t) = (p_1(t), \ldots, p_N(t))^{\mathrm{T}}$

Returning to the example of promotion and repression represented by Fig. 4, we obtain from the above definitions the set of equations

$$p_1(t+dt) = p_1(t)(1 - \alpha_1 dt - \alpha_2 dt) + p_2(t)\beta_1 dt + p_3(t)\beta_2 dt$$
$$p_2(t+dt) = p_2(t)(1 - \beta_1 dt - \alpha_3 dt) + p_1(t)\alpha_1 dt + p_4(t)\beta_3 dt$$
$$p_3(t+dt) = p_3(t)(1 - \beta_2 dt - \alpha_4 dt) + p_1(t)\alpha_2 dt + p_4(t)\beta_4 dt$$
$$p_4(t+dt) = p_4(t)(1 - \beta_3 dt - \beta_4 dt) + p_2(t)\alpha_3 dt + p_3(t)\alpha_4 dt, \tag{1}$$

which rearranges to give

$$\frac{d\mathbf{p}}{dt} = - \begin{pmatrix} \alpha_1 + \alpha_2 & -\beta_1 & -\beta_2 & 0 \\ -\alpha_1 & \beta_1 + \alpha_3 & 0 & -\beta_3 \\ -\alpha_2 & 0 & \beta_2 + \alpha_4 & -\beta_4 \\ 0 & -\alpha_3 & -\alpha_4 & \beta_3 + \beta_4 \end{pmatrix} \mathbf{p}(t). \tag{2}$$

This example easily generalises to the equation describing the evolution of $\mathbf{p}(t)$ for an arbitrary network, namely

$$\frac{d\mathbf{p}}{dt} = -A\mathbf{p}(t), \tag{3}$$

where the matrix A is given in terms of propensities. If α_{mn} is the propensity associated with the arc running from node m to node n, then

$$A = \sum_{m,n} \alpha_{mn} \Pi_{mn}, \tag{4}$$

where the matrices Π_{mn} are defined by

$$(\Pi_{mn})_{kl} = (\delta_{mk} - \delta_{nk})\,\delta_{ml}. \tag{5}$$

Note that each column of A sums to zero. One easily checks that this is equivalent to saying that total probability is conserved: $d\left(\sum_n p_n(t)\right)/dt = 0$.

For a specified initial condition, and with constant propensities, the solution to Eq. (3) is given formally by

$$\mathbf{p}(t) = e^{-tA}\mathbf{p}(0). \tag{6}$$

Assuming that A has a complete set of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$, and writing $\mathbf{p}(0) = \sum_{i=1}^n a_i\mathbf{v}_i$, gives the general solution in the computationally more useful form

$$\mathbf{p}(t) = \sum_{i=1}^n a_i e^{-\lambda_i t}\mathbf{v}_i. \tag{7}$$

5   A simple example: The gene cascade

Gene cascades occur in situations where a set of genes, often co-located on the genome in *operons*, act in such a way that the product of each gene activates the expression of its successor gene, thus enabling a staged release of several gene products. For example, a gene cascade occurs in the phage λ during a late stage of *lysis* when the genes for forming the head and tail of the phage are sequentially activated. Protein cascades also occur [16] in which proteins regulate the activity of other proteins, without resorting to genetic regulation.
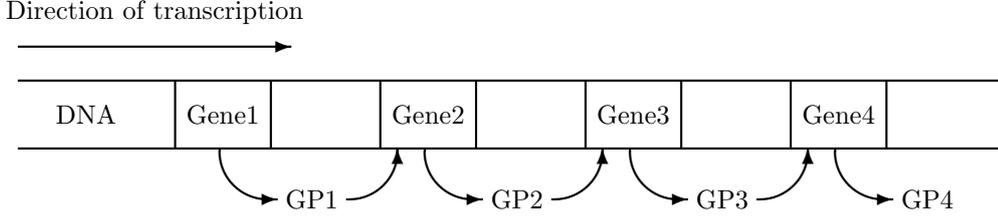
Direction of transcription

Figure 5. Gene cascade: Initially Gene 1 is expressed in the cell. Once the product of Gene 1 (GP1) is present in sufficient quantity it activates Gene 2. Gene 2 begins to produce GP2. Once the quantity of GP2 is above a threshold, Gene 3 is activated, followed by Gene 4. The result of this process is a staged release of gene products.
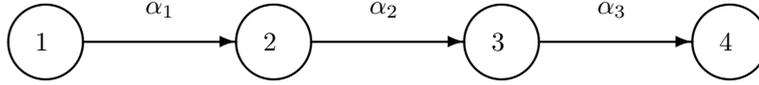

Figure 6. Network for the gene cascade shown in Fig. 5 is a chain of four states.

In this section we show a simplified model of the states for a gene cascade in which the product of one gene enhances the production of another gene. In a gene cascade, often the genes are situated close to each other on the DNA in "operon" form i.e. one after the other on the genome. But the operon structure is not necessary for our model of this network component. In our example in (Fig. 5), Gene 1 is initially expressed in the cell. Once the product of Gene 1 (GP1) is present in sufficient quantity it activates Gene 2, which begins to produce GP2. Once the quantity of GP2 is above a threshold, Gene 3 is activated, followed by Gene 4. The result of this process is a staged release of gene products 1, 2, 3 and 4.

The simplified gene network corresponding to the example in Fig. 5 consists of four possible gene states shown in Fig. 6:

$$\text{State } 1 = [1, 0, 0, 0] \qquad \text{(gene 1 on)}$$
$$\text{State } 2 = [1, 1, 0, 0] \qquad \text{(genes 1 and 2 on)}$$
$$\text{State } 3 = [1, 1, 1, 0] \qquad \text{(genes 1, 2 and 3 on)}$$
$$\text{State } 4 = [1, 1, 1, 1] \qquad \text{(genes 1, 2, 3 and 4 on)}$$

Using the formalism described in the last section, the cascade of N genes is modelled by the equations

$$\frac{dp_1}{dt} = -\alpha_1 p_1,$$
$$\frac{dp_n}{dt} = -\alpha_n p_n + \alpha_{n-1} p_{n-1}, \qquad n = 2, \ldots, N-1$$
$$\frac{dp_N}{dt} = \alpha_{N-1} p_{N-1}. \tag{8}$$

This model is of course a gross simplification of the real world, but serves to illustrate the type of modelling employed. In a more complete model the mRNA-producing steps and the transitions between the various stages of the central dogma would be included. A very detailed stochastic model of the mRNA and protein producing mechanisms in the phage λ, for example, is given in [24].

In the somewhat artificial case that all the propensities are equal, this set of equations (8) admits an easy analytic solution. Setting $\alpha_1 = \alpha_2 = \ldots = \alpha_{N-1} = \alpha$, one can check that the solution with the initial condition $\mathbf{p}^T(0) = (1, 0, \ldots, 0)$ is

$$p_n(t) = \frac{e^{-\alpha t}(\alpha t)^{n-1}}{(n-1)!}, \qquad n = 1, \ldots, N-1,$$

$$p_N(t) = 1 - \sum_{n=1}^{N-1} p_n(t). \tag{9}$$

The solution describes the probability of finding only the first n genes activated at a given time t. For all genes except the final gene the probabilities have the form of a Poisson distribution with mean and variance αt, while for the final gene the probability is the sum of the tail of the Poisson distribution from N onwards.

Note also that, given a state n, the portion of time the system can be expected to be found in that state during the interval $[t, t+dt)$ is $p_n(t)dt$. This enables us to introduce a probability distribution over time (conditional on n) given by

$$\pi_n(t)dt = \frac{p_n(t)dt}{\displaystyle\int_0^\infty p_n(t)dt} = \frac{\alpha^n}{\Gamma(n)} e^{-\alpha t} t^{n-1} dt, \qquad n = 1, \ldots, N-1, \tag{10}$$

where the denominator ensures that the probability over all time is normalised to 1. The meaning of this distribution density is as follows. Suppose we observe that the system is currently in state n, then the probability that the time is currently in the interval $[t, t+dt)$ is $\pi_n(t)dt$. This gives us a measure of how long the cascade might take to progress to a given point. Eq. (10) is the well known Gamma distribution with mean n/α and standard deviation $\sqrt{n}/\alpha$. For large n it is well approximated by a gaussian distribution about the expected time n/α.

For this example $\pi_n(t)$ could also have been arrived at with the following observation: Because of the Markovian nature of the cascade, the time taken for each timestep is an exponential random variable. The time taken to reach the nth gene is therefore the sum of n identical and independent exponential random variables, and this is a Gamma random variable.

6   The switching mechanism of the Bacteriophage λ

An example of a more complex switching system is the biologically well-understood *Bacteriophage* λ (phage λ). After the phage λ invades a bacteria cell (*E. Coli*), it can enter into one of two alternative lifestyles called *lysogeny* and *lysis* [27, 23]. The lysogeny stage is a dormant stage in which the phage inserts its DNA into the host's DNA and passively reproduces with the host. When the host becomes stressed, the phage is more likely to go into lysis, in which it reproduces more phages, kills the host and spreads to other bacteria cells. The decision between lysis and lysogeny can be thought of as a switching mechanism. The stochastic switch is based upon a competition between the *cro* and *cI* genes.

Figure 7 shows how this competition uses the proteins Cro and CI (protein names are captalized) to repress the expression of the other gene. Figure 7(a) shows the three operators
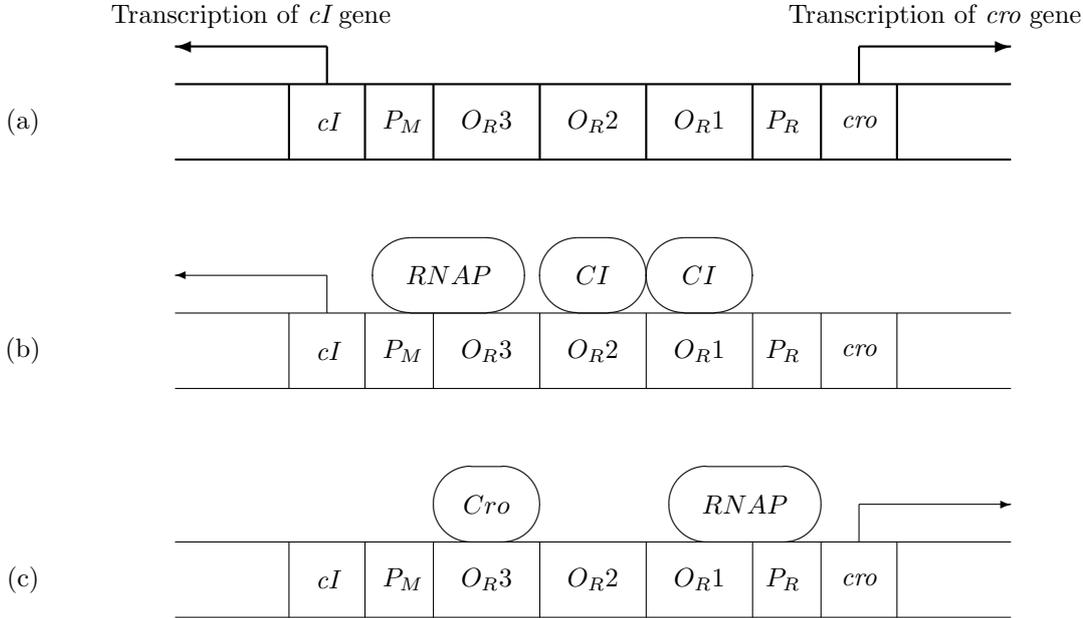
Transcription of cI gene

Transcription of cro gene

(a)

| | cI | P_M | O_R3 | O_R2 | O_R1 | P_R | cro | |

(b)

RNAP   CI   CI

| | cI | P_M | O_R3 | O_R2 | O_R1 | P_R | cro | |

(c)

Cro   RNAP

| | cI | P_M | O_R3 | O_R2 | O_R1 | P_R | cro | |

Figure 7. (a) The switching mechanism of the phage λ, in which the cI and cro genes stochastically compete for expression.(b) If the protein CI (encoded by the cI gene) succeeds in binding to O_R1 and O_R2, it represses the transcription of cro.(c) If the protein Cro (encoded by the cro gene) succeeds in binding to O_R3, it represses the transcription of cI.

to which the CI and Cro proteins are able to bind. CI binds preferentially to O_R1 and O_R2, but can also bind to O_R3, with a lower probability. In Figure 7(b) CI has succeeded in binding to O_R1 and O_R2, and it represses the transcription of cro. In Figure 7(c) Cro has successfully bound to O_R3, and it represses the transcription of cI.

The full story of the lysis/lysogeny decision mechanism is considerably more complex than the simplified version given here. See [27] for a full biological description, and [24] for a comprehensive stochastic model that includes the other proteins (CII, CIII, and N) involved in the process, as well as the production mRNA, its translation into proteins, the degradation rates of all chemical species and cell division.

The simulations in [24, 6] use the Gillespie algorithm [12, 6] to approximate the time spent by the phage λ in each of its forty possible switching states, based upon the (correct) assumption that the switching between these states is relatively fast. We are currently constructing a model which makes stochastic transitions between the 40 states and initiates a competition between the concentrations of the Cro and CI proteins. It is easily shown that there are only 164 possible transitions between the forty states (if we restrict the possible transitions to the case where one molecule binds/unbinds at a time). The propensities for these transitions can be derived based upon the thermodynamic model given in [7, 28] and the reaction rates given in [6].

7 Conclusion

The authors are currently working on further modelling of gene states of the *Bacteriophage* λ. From this we want to move onto other gene regulatory systems, and also to examine various more general aspects of biological pathway modelling. Specifically, what computational techniques can be used to model these networks more efficiently, what are the effects of combining the states of systems into more manageable aggregated states, and how the stochastic master equation can model biological systems at different levels of detail.

Mathematical models of regulatory networks aim to be predictive rather than descriptive. Analyses of the stability, bistability and robustness are possible once one has a sound model of the system, usually based upon stochastic processes and differential equations. The stochastic master equation models the evolving probability of the system occupying the entire state space, but another approach is to follow the life cycle of a single cell, as it makes a simulated choice between states, based upon their probability. Both of these approaches can be used to model a population of cells that have different individual fates, so that one can predict the proportion of the population that will be in one of several different phylogenic states (such as the competing lysis/lysogeny sub-populations of the λ phage). The eventual aim is to provide models that can lead back into experiment by predicting the proportions of such sub-populations, by predicting the upper and lower limits of unknown pathway parameters and rates, by modelling the behavior of systems under perturbation or by providing the quantitative reasoning behind existing biological systems.

References

[1] D. Adalsteinsson, D. McMillen and T.C. Elston, *Biochemical Network Stochastic Simulator (BioNetS): software for stochastic modeling of biochemical networks*, BMC Bioinformatics **5** (2004), 24.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, *Basic local alignment search tool*, J.Mol.Biol. **215** (1990), 403–410.

[3] A. Arkin, J. Ross and H.H. McAdams, *Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage λ-Infected Escherichia coli Cells*, Genetics **149** (1998), 1633–1648.

[4] P.O. Brown and D. Botstein, *Exploring the new world of the genome with DNA microarrays*, Nature Genetics **21** (1999), 33–37.

[5] J.J. Bull, *Sex determination in reptiles*, Quart. Rev. Biol. **55** (1980), 3–21.

[6] T. Tian and K. Burrage, *Bistability and Switching in the Lysis/Lysogeny Genetic Regulatory Network of Bacteriophage λ*, J. Theor. Biol. **227** (2004), 229–237.

[7] P.J. Darling, J.M. Holt and G.K. Ackers, *Coupled energetics of λ cro repressor self-assembly and site-specific DNA operator binding II: Cooperative interactions of cro dimers*, J. Mol. Biol. **139** (2000), 163–194.

[8] H. de Jong, *Modeling and simulation of genetic regulatory systems: a literature review*, J. Comp. Biol. **9** (2002), 67–103.

[9] R. Durbin, S.R. Eddy, A. Krough and G. Mitchison, *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press Cambridge 1998).

[10] M.B. Elowitz and S. Lelbler, *A synthetic ocsillatory network of transcriptional regulators*, Nature **403** (2000), 335–338.

[11] W.J. Ewens and G.R. Grant, *Statistical Methods in Bioinformatics: An Introduction* (Springer New York 2000).

[12] D.T. Gillespie, *Approximate accelerated stochastic simulation of chemically reacting systems*, J. Chem. Phys **115** (2001), 1716–1733.

[13] A. Gilman and A. Arkin, *Genetic "Code": Representations and dynamical models of genetic components*, Annu. Rev. Genomics Hum. Gent. **3** (2002), 341–369.

[14] J.A.M. Graves and S.M. Gartler, *Mammalian X chromosome inactivation; testing the hypothesis of transcriptional control*, Somat. Cell Molec. Genet. **12** (1986), 275–280.

[15] J. Hasty, D. McMillen, F. Isaacs and J.J. Collins, *Computational studies of gene regulatory networks: in numero molecular biology*, Nature **2** (2001), 268–278.

[16] C.F. Huang and J.E. Ferrell Jr., *Ultrasensitivity in the mitogan-activated protein kinase cascade*, Proc. Natl. Acad. Sci. USA **93** (1996), 10078–10083.

[17] E.M. Judd, M.T. Laub and H.H. McAdams, *Toggles and oscillators: new genetic circuit designs*, BioEssays **22** (2000), 507–509.

[18] S. Karlin and S.F. Altschul, *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes*, Proc. Natl. Acad. Sci. USA **87** (1990), 2264–2268.

[19] T.B. Kepler and T.C. Elston, *Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations*, Biophysical Journal **81** (2001), 3116–3136.

[20] E.S. Lander et al, *Initial Sequencing and analysis of the human genome*, Nature **409** (2001), 860–921.

[21] E.S. Lander, *Array of hope*, Nature Genetics **21** (1999), 3–4.

[22] A.M. Lesk, *The Unreasonable Effectiveness of Mathematics in Molecular Biology*, in: The Mathematical Intelligencer 2000 (Springer-Verlag New York 2000), 28–37.

[23] B. Lewin, *Genes VIII* (Paramus New Jersey 2004).

[24] H.H. McAdams and A. Arkin, *Stochastic mechanisms in gene expression*, Proc. Natl. Acad. Sci. USA **94** (1997), 814–819.

[25] H.H. McAdams and L. Shapiro, *Circuit Simulation of Genetic Networks*, Science **269** (1995), 650–656.

[26] S.B. Needleman and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, J. Mol. Biol. **48** (1970), 443–453.

[27] M. Ptashne, *A Genetic Switch: Phage λ and Higher Organisms*, second edition (Cell Press Cambridge Massachusetts 1992).

[28] J. Reinitz and J.R. Vaisnys, *Theoretical and experimental analysis of the phage lambda genetic switch missing levels of co-operativity*, J. Theor. Biol. **145** (1990), 295–318.

[29] E. Segal, H. Wang and D. Koller, *Discovering molecular pathways from protein interaction and gene expression data*, Bioinformatics **19** Suppl. 1 (2003), i264–i272.

[30] A.H. Sinclair, J.W. Foster, J.A. Spencer, D.C. Page, M. Palmer, P.N. Goodfellow and J.A.M. Graves, *Sequences homologous to ZFY, a candidate human sex-determining gene, are autosomal in marsupials*, Nature **336** (1988), 780–783.

[31] T.F. Smith and M.S. Waterman, *Identification of common molecular subsequences*, J. Mol. Biol. **147** (1981), 195–197.

[32] J. Tegner, M.K.S. Yeung, J. Hasty and J.J. Collins, *Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling*, Proc. Natl. Acad. Sci. USA **100** (2003), 5944–5949.

[33] N.G. van Kampen, *Stochastic Processes in Physics and Chemistry*, (North-Holland New York 1981).

[34] C.V. Venter et al, *The Sequence of the Human Genome*, Science **291** (2001), 1304–1351.

[35] M.S. Waterman, *Introduction to Computational Biology* (Chapman & Hall London 1995).

[1]Centre for Bioinformation Science, Australian National University, Canberra ACT 0200

[2]Mathematical Sciences Institute, Australian National University, Canberra ACT 0200

[3]John Curtin School of Medical Research, Australian National University, Canberra ACT 0200

[4]ARC Centre for Bioinformatics, University of Queensland, St. Lucia, QLD 4072

E-mail: Hilary.Booth@anu.edu.au