1 Abstract

The following investigation explores the effect of changes in the regulatory motif of gene expression on means and variances of mRNA and protein copy numbers. Gillespie's algorithm was used to numerically simulate transcription and translation. A version of the Fluctuation Dissipation Theorem, with drift and diffusion matrices derived from finding first and second moment equilibria of the multivariate Master Equation governing the system dynamics was used to estimate the covariance matrix for mRNA and protein. It was found that for a simple linear motif, the predictions of the Fluctuation Dissipation Theorem accurately represented the simulation dynamics, however a high protein variance led to the conclusion that there was a lack of robustness. Introduction of negative autoregulation in the system improved its robustness, but there were limitations in the model predictions given the need for linearisation. Finally, a delay in the system - as a result of the introduction of translation to an unmatured protein, followed by its folding to give the final product - was shown to further improve system robustness when variance of the unmatured protein was not immediately propagated to the final product, but the success of this form of motif must also consider the trade-off with its adaptation speed.

2 Introduction

Protein expression is a fundamental process in the life of a biological system, serving the purpose of the structural, mechanical and biochemical elements of the cell or life form. Synthetic biology is a rapidly advancing field, with technologies projected to affect many aspects of our lives in the immediate future, from the food we eat, to the drugs that we are given when we are sick [2] [29]. Interdependency of proteins in a system's underlying network plays a crucial role in an organism's ability to survive in a fluctuation environment, be that through maintaining a state of homeostasis or rapidly adapting to the changes. Developing a rigorous understanding of the dynamics which these systems use to achieve evolutionarily advantageous characteristics can help us to design robust genetic circuits which can, for example, be applied to help crops survive climate change [5], grow personalised tissues in regenerative medicine [11], and design computational tools which can provide automated systems for these designs themselves [20].

It has been observed that there is a typical trade-off between adaptability of an organism and its incurred energetic costs in gene expression [10], therefore it is important to understand the regulatory motifs which can be adopted to optimise the system dynamics. Biological control systems use a combination of linear and non-linear dynamics to try to achieve this optimum [27], hence this investigation explores the effects of replacing a linear gene expression network with negative autoregulation, and the introduction of a delay.

Further to this, abundances of species copy numbers such as mRNA and proteins can vary over a wide range of orders of magnitude [26]. At a large order of magnitude, the system can be described through the use of continuum dynamics, whereas at the small scale, the governing differential equations can be reduced to discrete 'birth' and 'death' events of a particular species, with the propensities of these governed by (typically non-linear) probability distributions. This investigation focuses on relatively low-copy number expression, hence the analytical solving of the system dynamics is intractable in the majority of scenarios. Instead, a numerical simulation is carried out, using Gillespie's Algorithm [8], allowing the stochastic process to be analysed qualitatively and quantitatively.

3 Methods

Gillespie's Algorithm and the subsequent analysis was carried out using a Python script (see Appendix A). Further detail on the derivations of mean values and linerisations for the purpose of finding the matrices

for applying the Fluctuation Dissipation Theorem can be found in Appendix B.

4 Results and Discussions

4.1 Simple transcription and translation

4.1.1 Effect of production and degradation rates on the mean and variance of mRNA copy number

A simple gene transcription model was first used to explore the dynamics of the mRNA copy number in response to a change in production and degradation rates. This used the linear model:

$$m \xrightarrow{\lambda} m + 1 \tag{1}$$

$$m \xrightarrow{\beta\ \mathrm{m}} m - 1 \tag{2}$$

in which λ represents a spontaneous production rate of mRNA, and β a degradation rate - of which the propensity of degradation depends on the current mRNA copy number. A constant production rate of mRNA may arise in a cell as a result of an inducible promoter, dependent on the state of the surrounding environment, whilst the linear degradation rate can be explained, for example, by intracellular enzyme binding, where β is dependent on the enzyme concentration [7], [14].

The rate of change of mean mRNA copy number can simply be evaluated by taking the time derivative of the first moment of the standard Master Equation which governs the stochastic birth-death process. In a similar manner, we can use the standard formula for a variable's auto covariance and the time derivative of the second moment of the Master Equation to derive the variance of mRNA copy number. At equilibrium, this gives the mean copy number and variance to be:

$$\langle m \rangle = \frac{\lambda}{\beta} \tag{3}$$

$$\sigma_{mm} = \langle m \rangle = \frac{\lambda}{\beta} \tag{4}$$

as expected, given that the process follows a Poisson distribution.

Using the parameters $\lambda = 1$ and $\beta = 0.01$, we can expect the values for the mean and variance of mRNA copy number to both be 100. Figure 1 provides evidence that this is indeed the result that is observed. For this particular run of the Gillespie algorithm simulation of transcription, 50000 iterations gave rise to a simulated mean of 99.98 and variance of 102.68, confirming the validity of this stochastic method in the one-dimensional case.

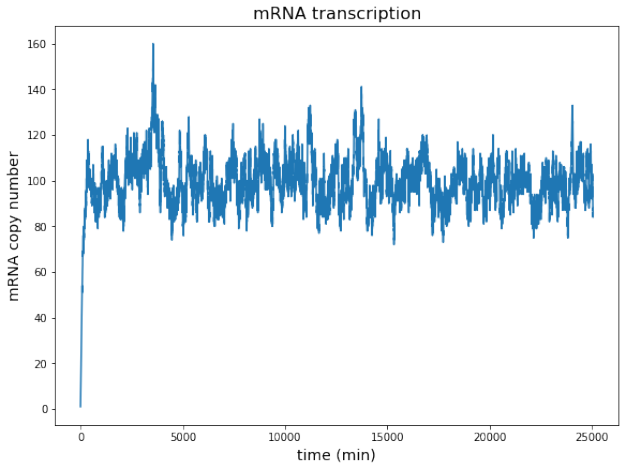


Figure 1: Simulation of mRNA transcription, with $\lambda = 1$, $\beta = 0.01$

This is further supported from observations of the effects of varying λ and β. The plots of Figure 2 were generated by holding each parameter constant in turn whilst varying the other. The evidence supports the theory, with a proportional relationship observed at λ is varied and an inversely proportional relationship observed as β is varied. The variability of the variance about a line of best fit when λ is altered is greater

than that of the mean. This may be a result of more numerical instability in the variance, as a result of the dependency on time spent at each state in the algorithmic model, however the exact details of the cause of this are beyond the scope of this investigation [19].

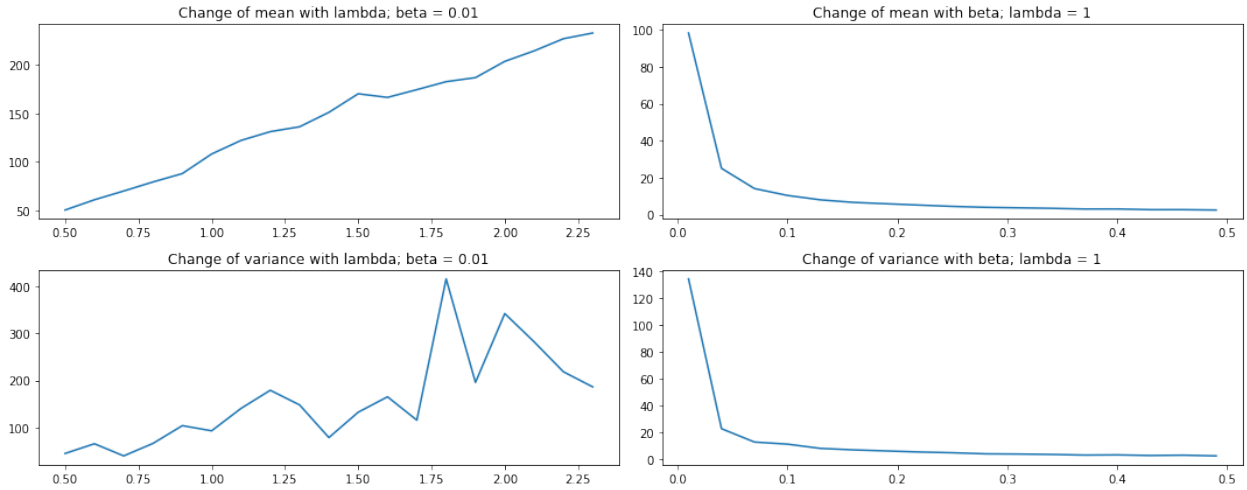


Figure 2: The variation of means and variances of mRNA copy number as λ and β vary in turn

4.1.2 The Fluctuation Dissipation Theorem in a linear gene expression system

In the following simulations, each iteration of the algorithm represents a minute in real-time dynamics of an E. coli cell, the elements of the state vectors represent species' copy numbers, and hence the rates used are scaled as per minute. Depending on the status of the surrounding environment, mRNA production is at an elongation value between 12 and 55 nucleotides per second [23], [16], [21], [4]. Given that there are around 1000 to 2000 nucleotides per mRNA strand (and therefore 300 to 600 amino acids per protein, with the ratio of nucleotides:amino acids as 3:1) [4], [15], [3], this leads to mRNA production being in the range of 0.36 to 3.3 transcripts per minute. Amino acid chain elongation typically occurs at a rate between 4 and 16 amino acids per second [23], [21], [6], [9], [24], [25], giving a production rate between 0.4 and 3.2 proteins per minute. mRNA strand degradation typically takes between 3 and 10 minutes, giving a rate of 0.1 to 0.33 per minute [28], [1], whilst typical protein half lives and cell division times give their degradation rates in the range 0.007 to 0.03 per minute [21], [16], [12], [22]. Often, degradation rates scale with production rates in a cell, and this too was represented as far as practically possible in the simulations.

A simple linear translation model was next added to the gene expression simulation:

$$p \xrightarrow{\alpha\ \mathrm{x}} p + 1 \tag{5}$$

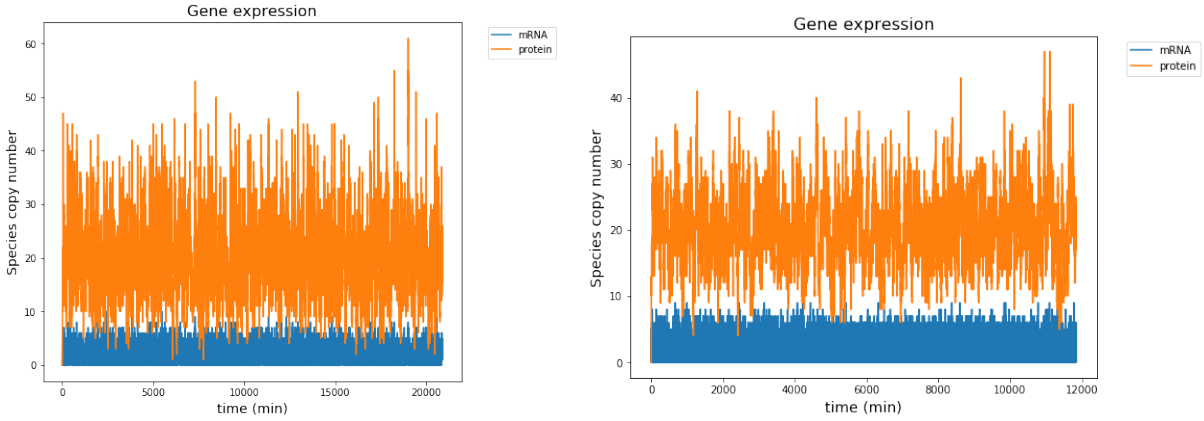
$$p \xrightarrow{\gamma\ \mathrm{p}} p - 1 \tag{6}$$

in which α and β represent production and degradation rates respectively. Production of protein in this form may be controlled by the one to one binding activity of a ribosome to mRNA strand, and the abundancies of tRNAs in proximity to the formed complex in order to synthesise the amino acid chain, whilst the linear degradation rate may be a result of active degradation of protein through enzymatic activity, or simply through dilution as a result of cell division [13].

The balance between the governing parameters for the dynamics of mRNA transcription and protein translation can have a great impact on a cell's functionality, often with a trade-off between adaptability and stability. If mRNA production and degradation is at a high rate, this could be energetically unfavourable for a cell, however too low and the cell may struggle to adapt to a rapidly-changing environment [14], as represented in Figure 3 whereby the decrease in protein copy number variance for faster mRNA dynamics shows a faster response to the fluctuating environment.

The first and second order moments of the Master equation governing the multivariate case can similarly be calculated as in the first case, with the reactions of each species at a given time step simply modelled as a reaction vector, as opposed to a single variable. As such, we can find the general forms for the differential equations of the mean vector and covariance matrix for a system of birth-death processes:

$$\frac{d\langle x\rangle}{dt} = \langle \sum_i r_i W_i(x)\rangle = \langle f(x)\rangle = \langle Ax + f_0\rangle \tag{7}$$



(a) mRNA parameters of $\lambda = 0.8, \beta = 0.4$ give a protein variance of around 55

(b) mRNA parameters of $\lambda = 2.8, \beta = 1.4$ give a protein variance of around 30

Figure 3: Simulations of gene expression, with mRNA copy number of 2 and protein copy number of 20. The production and degradation rates of mRNA are varied such as to observe the effects on protein copy number variance

$$\frac{d\Sigma}{dt} = A\Sigma + \Sigma A^T + \langle \sum_i r_i W_i(x) r_i^T \rangle = A\Sigma + \Sigma A^T + D \tag{8}$$

where x represents the vector of species (in this case $[m, p]^T$), r_i are the reaction vectors (representing a birth or a death process for each species), and W_i are the propensity functions. A and D are known as the drift and diffusion matrices respectively, which characterise the Fluctuation Dissipation Theorem at equilibrium (where Equation 7 and Equation 8 are both equal to zero), where all processes causing a drift out of a particular state balance the restoring force into that state.

In this linear case, Equation 7 can be written in the form:

$$\frac{d\langle x \rangle}{dt} = \langle f(x) \rangle = A\langle x \rangle + f_0 \tag{9}$$

from which we can derive the following drift and diffusion matrices:

$$A = \begin{bmatrix} -\beta & 0 \\ \alpha & \gamma \end{bmatrix} \tag{10}$$

$$D = \begin{bmatrix} \lambda + \beta\langle m \rangle_{eq} & 0 \\ 0 & \alpha\langle m \rangle_{eq} + \gamma\langle p \rangle_{eq} \end{bmatrix} \tag{11}$$

with $\langle m \rangle_{eq} = \frac{\lambda}{\beta}$ and $\langle p \rangle_{eq} = \frac{\alpha}{\gamma}\langle m \rangle_{eq} = \frac{\alpha\lambda}{\beta\gamma}$. This was solved both manually, and verified through the use of a Lyapunov equation solving function in the simulation, giving rise to the species' varicances and covariances of $\sigma_{mm} = \frac{\lambda}{\beta}$, $\sigma_{mp} = \left(\frac{\alpha}{\gamma+\beta}\right)\left(\frac{\lambda}{\beta}\right)$ and $\sigma_{pp} = \frac{\alpha\lambda}{\gamma\beta}\left(1 + \frac{\alpha}{\gamma+\beta}\right)$.

The use of the Fluctuation Dissipation Theorem as a means of predicting the equilibrium values for mRNA and protein copy number was verified and validated through comparisons of theoretical and simulated outcomes for mean copy numbers and covariance matrices across a variety of values of the constants governing the gene expression dynamics. The deviation between the theory and simulations was found to be very small, for example the system in Figure 3 exhibited the properties of 1

	Theoretical value	Simulated value
$\langle m \rangle$	2.00	2.50
$\langle p \rangle$	20.0	20.5
σ_{mm}	2	2.17
σ_{mp}	0.95	0.93
σ_{pp}	29.5	30.3

Table 1: Theoretical and simulated means and (co)variances found for simple gene expression, with parameters $\lambda = 2.8, \beta = 1.4, \alpha = 0.7, \gamma = 0.07$

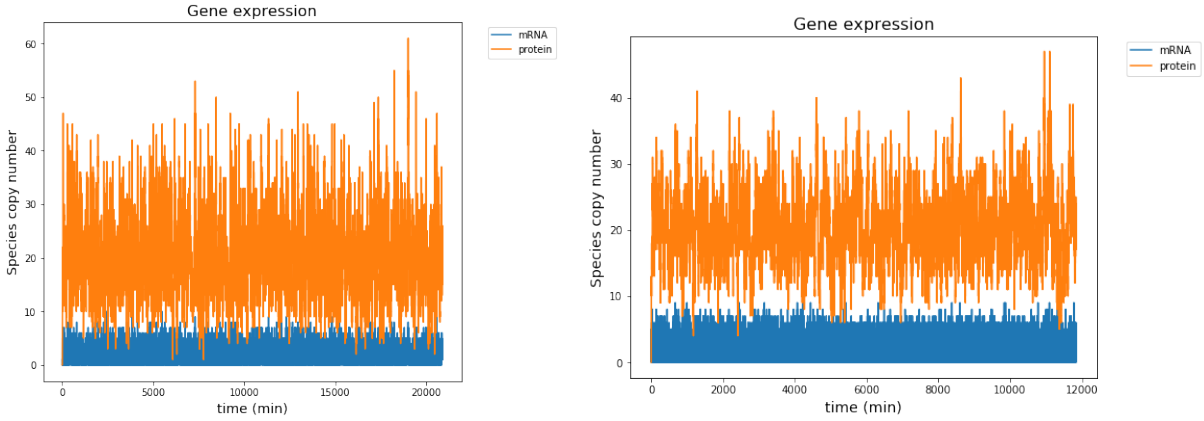
4.2 Self-regulation of transcription

In a further model, the constant transcription rate of mRNA was replaced with a self-regulating motif - a negatively-autoregulation of the gene expression by its protein, with Equation 5 becoming:

$$m \xrightarrow{\frac{2\lambda K}{K+p^h}} m+1 \tag{12}$$

which can be recognised as an inhibiting Hill function, where the positive integer h (the Hill coefficient) represents the number of protein molecules cooperatively acting to regulate the rate of transcription and K affects the distribution of the function such that at $p = \sqrt[h]{K}$, the mRNA production rate will be at half of its maximum (i.e. at λ in this case). Considering a simple plot of this function, it can be noted that increasing h at a constant value of K results in a much greater increase in steepness of the curve and decrease of the value of p needed for half maximum rate than increasing K at a constant value of h, indicating that the means and variances of the species' copy numbers will more heavily depend on the Hill coefficient, h.

The parameter values as used in Figure 3a were kept the same (at $\lambda = 0.8, \beta = 0.4, \alpha = 0.8, \gamma = 0.08$, whilst K and h were varied. Across the explored ranges of these two parameters, it was found that the optimum values of h and K to reduce the variance of protein copy number as far as possible whilst keeping its mean near-constant were 2 and around 110 respectively, where the variance was reduced from around 55 to around 20. This suggests that, for the values of the other parameters used, beyond a protein copy number of around 10, the mRNA production rate will be reduced below its constant rate for the simple case, λ. Such negative feedback motifs allow for the robustness of a genetic expression system to be improved (the reduction of the variance implying a faster reaction to environmental change), whilst not incurring extra energetic costs associated with having faster maximum mRNA production and degradation rates, as seen previously. There is evidence that such motifs are evolutionarily stable, given that their existence is observed more often than chance in the regulatory gene network of E. coli, and has been shown to improve the robustness of both natural and synthetic genetic circuits [18], [17].



(a) No self regulation gives a protein variance of around 55

(b) Self regulation parameters $K = 110, h = 2$ give a protein variance of around 21

Figure 4: Simulations of gene expression, with parameters of $\lambda = 0.8, \beta = 0.4, \alpha = 0.8, \gamma = 0.08$, with and without self regulation

Given that the propensity function of Equation 12 is non-linear, we cannot directly use Equations 7 and 8 to analytically calculate the means and covariance matrix of the species, however we may consider an approximation to this through a linearisation. For a large copy number of a species y (therefore having relatively small approximations), we can make the approximation:

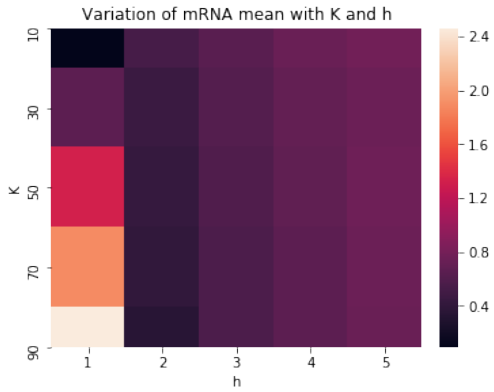
$$\langle f(y) \rangle \approx f(\langle y \rangle) \tag{13}$$

which if we call the Hill function $f(p)$ and assume a large protein copy number can allow us to find equilibrium mean mRNA and protein copy numbers of approximately $\langle m \rangle_{eq} = \frac{\gamma}{\alpha}\left(\frac{2\lambda K \alpha}{\beta\gamma}\right)^{\frac{1}{h+1}}$, and $\langle p \rangle_{eq} = \left(\frac{2\lambda K \alpha}{\beta\gamma}\right)^{\frac{1}{h+1}}$. Linearising about this mean vector allows us to find approximated forms of the drift and diffusion matrices for the self-regulated system:

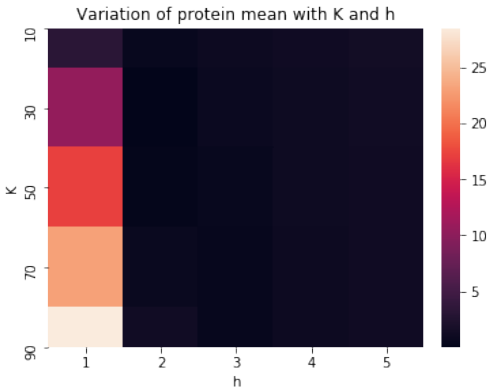
$$A = \begin{bmatrix} -\beta & -\frac{2\lambda K h \langle p \rangle_{eq}^{h-1}}{\left(K + \langle p \rangle_{eq}^{h}\right)^2} \\ \alpha & \gamma \end{bmatrix} \tag{14}$$

$$D = \begin{bmatrix} \frac{2\lambda K}{K + \langle p \rangle_{eq}^{h}} + \beta \langle m \rangle_{eq} & 0 \\ 0 & \alpha \langle m \rangle_{eq} + \gamma \langle p \rangle_{eq} \end{bmatrix} \tag{15}$$

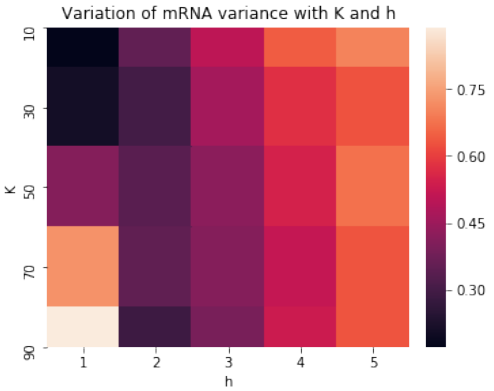
The validity of this linearisation was explored, by observing the differences between expected protein and mRNA means, variances and covariances with the variation of the parameters K and h. As suggested in Figure 5, the linearisation is the most unsuitable wherever $h = 1$, likely the result of ignoring first order terms in the protein mean when approximating its equilibrium value, and working on the assumption that the copy number is always large (see Appendix B). Aside from this, the general trend is that greatest differences are observed where h is large and K is relatively small, at which the Hill function is the most steep. The steeper the Hill function for negative regulation, the faster the system's response is to perturbations from equilibrium - linearising the system may fail to carry these dynamics through as the Hill function approaches a negative step.



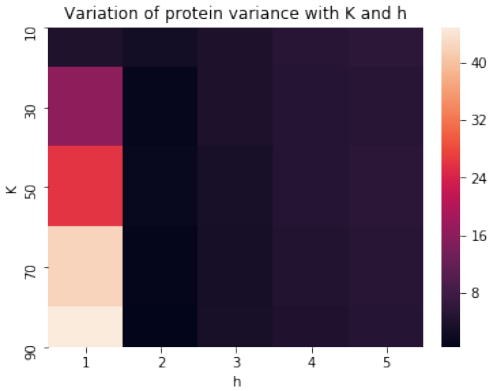
(a) Mean mRNA copy number difference



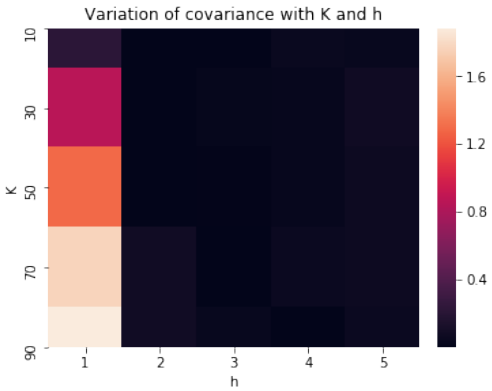
(b) Mean protein copy number difference



(c) mRNA variance difference



(d) Protein variance difference



(e) Covariance difference

Figure 5: Heat maps showing the differences between predicted means and (co)variances from the linearisation of the self-regulating motif for the Fluctuation Dissipation Theorem, and those observed in simulation

4.3 Protein maturation

A final adjustment to the model was introduced, replacing the single-step protein translation of Equation 1 with a two step process in which translation results in an unmatured version of the protein, which requires a folding process to recover the final protein species:

$$p' \xrightarrow{\alpha m} p' + 1 \tag{16}$$

$$(p', p) \xrightarrow{\tau^{-1} p'} (p' - 1, p + 1) \tag{17}$$

Varying τ between 1 and 100 (Figure 7), we find that the variance of the protein copy number decreases with decreasing negative gradient as τ increases.

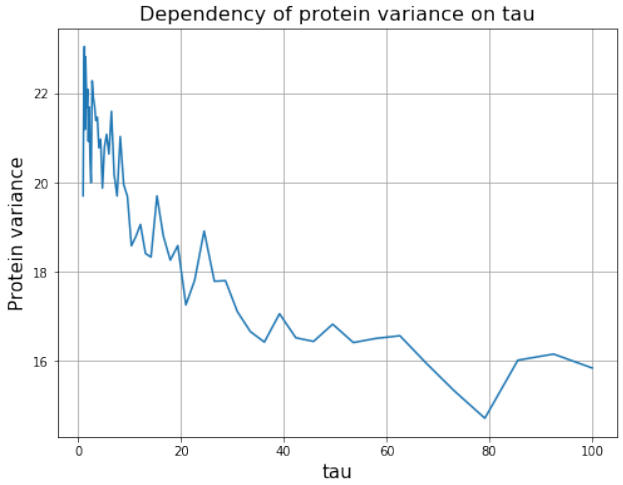
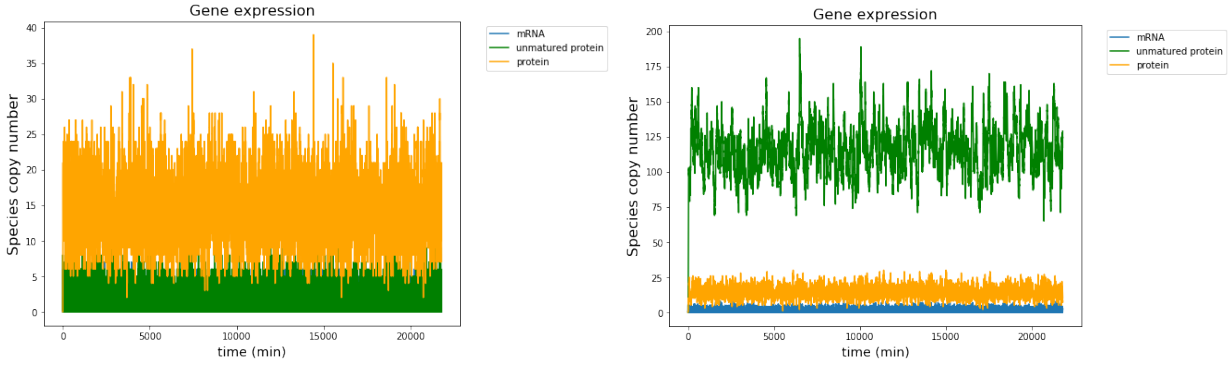


Figure 6: The change in protein number variance as τ increases from 1 to 100, with gene expression parameters $\lambda = 0.8, \beta = 0.4, \alpha = 0.8, \gamma = 0.08, K = 100, h = 2$

As τ represents the time delay of protein production as a result of the folding of the unmatured protein, we can expect that when this value is low (when the folding happens very quickly), the variance in the unmatured protein is propagated through strongly to the folded protein, whereas as this time delay increases, the folded protein essentially responds to some time averaged function of the unmatured protein number (rather than its instantaneous value), and therefore its variance will decrease. This further increases the robustness of the genetic circuit. However, introducing a time delay in the system puts into question its ability to adapt quickly to an external change when necessary, so trade-offs may be encountered here.



(a) $\tau = 1$ (b) $\tau = 100$

Figure 7: A comparison of the gene expression dynamics for $\tau = 1$ and $\tau = 100$, with gene expression parameters $\lambda = 0.8, \beta = 0.4, \alpha = 0.8, \gamma = 0.08, K = 100, h = 2$

5 Conclusions

This investigation provided evidence for the validity of using a stochastic gene expression model, at relatively low copy numbers of species, through a numerical simulation tusing Gillespie's Algorithm. It was

shown, from the calculations of moment equilibria of the Master Equation governing the system dynamics, that theoretical derivations of the species' means and variances closely matched those observed in simulation for the linear case. Predictions from the Fluctuation Dissipation Theorem proved accurate for the linear model, and could be supported up to a reasonable extent in the non-linear case of negative autoregulation. Further to this, it was shown that the use of autoregulation provided the system with more evolutionarily favourable characteristics, reducing the variance of the protein copy number whilst roughly maintaining the mean value found for the linear system. A time delay due to protein maturation (folding), also gave rise to this decrease in variance, however the trade-off between the system's stability and rate of adaptation in a fluctuating environment, as well as the typical environments a system would need to adapt to, must be carefully assessed when drawing conclusions about the system's robustness.

References

[1] Somenath Bakshi, Albert Siryaporn, Mark Goulian, and James C Weisshaar. Superresolution imaging of ribosomes and rna polymerase in live escherichia coli cells. *Molecular microbiology*, 85(1):21–38, 2012.

[2] Adam M Beitz, Conrad G Oakes, and Kate E Galloway. Synthetic gene circuits as tools for drug discovery. *Trends in Biotechnology*, 2021.

[3] Florian Brandt, Stephanie A Etchells, Julio O Ortiz, Adrian H Elcock, F Ulrich Hartl, and Wolfgang Baumeister. The native 3d organization of bacterial polysomes. *Cell*, 136(2):261–271, 2009.

[4] Hans Bremer, P Dennis, and Mans Ehrenberg. Free rna polymerase and modeling global transcription in escherichia coli. *Biochimie*, 85(6):597–609, 2003.

[5] Jennifer AN Brophy, Katie J Magallon, Lina Duan, Vivian Zhong, Prashanth Ramachandran, Kiril Kniazev, and José R Dinneny. Synthetic genetic circuits as a means of reprogramming plant roots. *Science*, 377(6607):747–751, 2022.

[6] David G Dalbow and RY Young. Synthesis time of β-galactosidase in escherichia coli b/r as a function of growth rate. *Biochemical Journal*, 150(1):13–20, 1975.

[7] Vlad Elgart, Tao Jia, and Rahul Kulkarni. Quantifying mrna synthesis and decay rates using small rnas. *Biophysical journal*, 98(12):2780–2784, 2010.

[8] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[9] Călin C Guet, Luke Bruneaux, Taejin L Min, Dan Siegal-Gaskins, Israel Figueroa, Thierry Emonet, and Philippe Cluzel. Minimally invasive determination of mrna concentration in single living bacteria. *Nucleic acids research*, 36(12):e73–e73, 2008.

[10] Jean Hausser, Avi Mayo, Leeat Keren, and Uri Alon. Central dogma rates and the trade-off between precision and economy in gene expression. *Nature communications*, 10(1):1–15, 2019.

[11] CP Healy and TL Deans. Genetic circuits to engineer tissues with alternative functions. *Journal of Biological Engineering*, 13(1):1–7, 2019.

[12] Charles E Helmstetter and Stephen Cooper. Dna synthesis during the division cycle of rapidly growing escherichia coli br. *Journal of molecular biology*, 31(3):507–518, 1968.

[13] Avram Hershko. The ubiquitin system for protein degradation and some of its roles in the control of the cell division cycle. *Cell Death & Differentiation*, 12(9):1191–1197, 2005.

[14] Monica P Hui, Patricia L Foley, and Joel G Belasco. Messenger rna degradation in bacterial cells. *Annual review of genetics*, 48:537–559, 2014.

[15] Ben Jones, Dov Stekel, Jon Rowe, and Chrisantha Fernando. Is there a liquid state machine in the bacterium escherichia coli? In *2007 IEEE Symposium on Artificial Life*, pages 187–191. Ieee, 2007.

[16] Stefan Klumpp and Terence Hwa. Growth-rate-dependent partitioning of rna polymerases in bacteria. *Proceedings of the National Academy of Sciences*, 105(51):20245–20250, 2008.

[17] David C Marciano, Rhonald C Lua, Christophe Herman, and Olivier Lichtarge. Cooperativity of negative autoregulation confers increased mutational robustness. *Physical review letters*, 116(25):258104, 2016.

[18] Agustino Martínez-Antonio, Sarath Chandra Janga, and Denis Thieffry. Functional organisation of escherichia coli transcriptional regulatory network. *Journal of molecular biology*, 381(1):238–247, 2008.

[19] S Mauch and M Stalzer. An efficient method for computing steady state solutions with gillespie's direct method. *The Journal of chemical physics*, 133(14):10B607, 2010.

[20] Anna Nowogrodzki et al. The automatic-design tools that are changing synthetic biology. *Nature*, 564(7735):291–292, 2018.

[21] Anand Pai and Lingchong You. Optimal tuning of bacterial sensing potential. *Molecular systems biology*, 5(1):286, 2009.

[22] Oleg Paliy and Thusitha S Gunasekera. Growth of e. coli bl21 in minimal media with different gluconeogenic carbon sources and salt contents. *Applied microbiology and biotechnology*, 73(5):1169–1172, 2007.

[23] Sergey Proshkin, A Rachid Rahmouni, Alexander Mironov, and Evgeny Nudler. Cooperation between translating ribosomes and rna polymerase in transcription elongation. *Science*, 328(5977):504–508, 2010.

[24] Marlena Siwiak and Piotr Zielenkiewicz. Transimulation-protein biosynthesis web service. *PLoS One*, 8(9):e73943, 2013.

[25] Michael Askvad Sørensen, Annaleigh Ohrt Fehler, and Sine Lo Svenningsen. Transfer rna instability as a stress response in escherichia coli: Rapid dynamics of the trna pool as a function of demand. *RNA biology*, 15(4-5):586–593, 2018.

[26] Boumediene Soufi, Karsten Krug, Andreas Harst, and Boris Macek. Characterization of the e. coli proteome and its modifications during growth and ethanol stress. *Frontiers in microbiology*, 6:103, 2015.

[27] Arno Steinacher, Declan G Bates, Ozgur E Akman, and Orkun S Soyer. Nonlinear dynamics in gene regulation promote robustness and evolvability of gene expression levels. *PloS one*, 11(4):e0153295, 2016.

[28] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *science*, 329(5991):533–538, 2010.

[29] Christopher A Voigt. Synthetic biology 2020–2030: six commercially-available products that are changing our world. *Nature Communications*, 11(1):1–6, 2020.