An analytical model of the effect of plasmid copy number on transcriptional noise strength

William and Mary iGEM 2015

1 Introduction

In the early phases of our experimental design, we considered incorporating our two fluorescent reporters into a single plasmid. However, under this method the fluorescence readout for each channel from each cell would represent the cumulative output of many identical but independent noisy functions.

The interactions between these independent signals would lead to a damping of the noise in our measurements– it is a well-known result that for identical signals with with an uncorrelated zero-mean Gaussian noise term, the signal-to-noise ratio scales proportionally to $1/\sqrt{n}$ where n is the number of independent copies of the signal [1].

On the other hand, fluctuations of plasmid copy number around its steady-state value would be an added source of noise in our measurements. Although the dual-reporter system filters out extrinsic noise by considering the correlation between the two fluorescence signals, the derivation of the equation which Elowitz *et al* 2002 used to decompose the observed noise into intrinsic and extrinsic components explicitly accounted for the gene copy number within a cell [2][3]. Without a reliable way to separate the additional copy number fluctuation-induced noise effects from the other noise terms, we would be unable to use this pre-established theory to analyze our measurements for noise.

Hence we found the need to develop a model that explicitly describes the simultaneously reducing and enhancing impacts of plasmid copy number on our fluorescence signals.

2 The Basis for the Model

We found through a search of the literature that Jones *et al* 2014 derived a simple equation that accounts for the effect of chromosome replication on transcription rate in *E. coli* [4] :

$$\frac{\mathrm{Var}[m]}{\langle m\rangle} = \frac{\langle m^2\rangle_1 - \langle m\rangle_1^2}{\langle m\rangle_1} + \frac{f(1-f)}{1+f}\langle m\rangle_1 \tag{1}$$

where the Fano Factor, $\mathrm{Var}[m]/\langle m\rangle$, is their measure of transcriptional noise. Here m is a random variable which indicates the number of mRNA copies present in the cell, and $\mathrm{Var}[\cdot]$ and $\langle\cdot\rangle$ are used to denote the variance and mean, respectively, across the population. $\langle m\rangle_i$ is the expected value of m given that there exist i copies of the gene in the cell. f represents the fraction of time that a given cell contains two chromosomes, and hence two copies of the gene.

This result is a powerful one, stating that the fluctuations in gene copy number always confer an additive term onto the Fano Factor of transcripts in the cell (notice that the first term is equivalent to the Fano Factor of mRNA copy number in a cell that always has one gene copy). The Fano Factor, though it is not the way we define noise in our data analysis, can be interpreted as noise strength or the fold change of the noise of the process compared to that of an equivalent Poisson process [4][5]. Unfortunately, Jones *et al*'s derivation only accounts for cells that contain either 1 or 2 gene copies. Given that the copy number of pSB1C3 can range from 100-300 per cell [6], we needed to expand their model to accommodate any number of gene copies.

3 Expansion to n Gene Copies

We begin by defining some new random variables to ease the generalization of Jones *et al*'s work: let M be a random variable which represents the mRNA copy number in the cell, R a random variable which represents the net mRNA production of one gene copy at a given time, and G a random variable which represents the number of *additional* gene copies in the cell (ie. the number of gene copies -1). M is distributed

$$M = \begin{cases} R, & G = 0 \\ R^{(2)}, & G = 1 \\ \cdots, & \cdots \\ R^{(n)}, & G = n-1 \end{cases}$$

for any integer n, and

$$R^{(i)} = \sum_{j=1}^{i} R_j$$

with the R_j identical and mutually independent and having mean and variance equal to $\mathrm{E}[R]$ and $\mathrm{Var}[R]$, respectively. Then the n-gene copy equation is

$$\frac{\mathrm{Var}[M]}{\mathrm{E}[M]} = \frac{\mathrm{Var}[R]}{\mathrm{E}[R]} + \frac{\mathrm{Var}[G]}{1+\mathrm{E}[G]}\mathrm{E}[R] \tag{2}$$

where $\mathrm{E}[\cdot]$ denotes the expected value of the variable over the population. In particular, $1 + \mathrm{E}[G]$ is the expected value of the actual number of plasmids in the cell.

Proof. First, recall from the properties of mutually independent random variables that

$$\mathrm{E}[R^{(i)}] = \mathrm{E}\left[\sum_{j=1}^{i} R_j\right] = \sum_{j=1}^{i} \mathrm{E}[R_j] = \sum_{j=1}^{i} \mathrm{E}[R] = i\mathrm{E}[R]$$

and

$$\mathrm{Var}[R^{(i)}] = \mathrm{Var}\left[\sum_{j=1}^{i} R_j\right] = \sum_{j=1}^{i} \mathrm{Var}[R_j] = \sum_{j=1}^{i} \mathrm{Var}[R] = i\mathrm{Var}[R].$$

Now, by the Law of Total Expectation,

$$\begin{aligned}
\mathrm{E}[M] &= \mathrm{E}_G[\mathrm{E}[M|G]] \\
&= \mathrm{E}_G\left[\left\{\begin{array}{ll} \mathrm{E}[R], & G=0 \\ 2\mathrm{E}[R], & G=1 \\ \cdots & \cdots \\ n\mathrm{E}[R], & G=n-1 \end{array}\right\}\right] \\
&= [1-p_G(1)-p_G(2)-\cdots-p_G(n-1)]\mathrm{E}[R] + \\
&\quad [2p_G(1)\mathrm{E}[R]+3p_G(2)\mathrm{E}[R]+\cdots+np_G(n-1)]\mathrm{E}[R] \\
&= [1+p_G(1)+2p_G(2)+\cdots+(n-1)p_G(n-1)]\mathrm{E}[R] \\
&= \left(1+\sum_{g\in\mathfrak{G}} gp_G(g)\right)\mathrm{E}[R] \\
&= (1+\mathrm{E}[G])\mathrm{E}[R] \qquad (3)
\end{aligned}$$

where $\mathfrak{G}$ is the support of G.

Using the Law of Total Variance, we can also obtain

$$\begin{aligned}
\mathrm{Var}[M] &= \mathrm{E}_G[\mathrm{Var}[M|G]] + \mathrm{Var}_G[\mathrm{E}[M|G]] \\
&= \mathrm{E}_G\left[\left\{\begin{array}{ll} \mathrm{Var}[R], & G=0 \\ 2\mathrm{Var}[R], & G=1 \\ \cdots & \cdots \\ n\mathrm{Var}[R], & G=n-1 \end{array}\right\}\right] + \mathrm{Var}_G\left[\left\{\begin{array}{ll} \mathrm{E}[R], & G=0 \\ 2\mathrm{E}[R], & G=1 \\ \cdots & \cdots \\ n\mathrm{E}[R], & G=n-1 \end{array}\right\}\right] \\
&= [1+p_G(1)+2p_G(2)+\cdots+(n-1)p_G(n-1)]\mathrm{Var}[R] + \\
&\quad \mathrm{E}_G\left[\left(\left\{\begin{array}{ll} \mathrm{E}[R], & G=0 \\ 2\mathrm{E}[R], & G=1 \\ \cdots & \cdots \\ n\mathrm{E}[R], & G=n-1 \end{array}\right\}\right)^2\right] - \mathrm{E}_G\left[\left\{\begin{array}{ll} \mathrm{E}[R], & G=0 \\ 2\mathrm{E}[R], & G=1 \\ \cdots & \cdots \\ n\mathrm{E}[R], & G=n-1 \end{array}\right\}\right]^2 \\
&= (1+\mathrm{E}[G])\mathrm{Var}[R] + \mathrm{E}_G\left[\left\{\begin{array}{ll} \mathrm{E}[R], & G=0 \\ 4\mathrm{E}[R], & G=1 \\ \cdots & \cdots \\ n^2\mathrm{E}[R], & G=n-1 \end{array}\right\}\right] - \\
&\quad (1+p_G(1)+\cdots+(n-1)p_G(n-1))^2\mathrm{E}[R]^2 \\
&= (1+\mathrm{E}[G])\mathrm{Var}[R] + [1+3p_G(1)+8p_G(2)+\cdots+(n-2)^2p_G(n-1)]\mathrm{E}[R]^2 - \\
&\quad (1+\mathrm{E}[G])^2\mathrm{E}[R]^2 \\
&= (1+\mathrm{E}[G])\mathrm{Var}[R] + \left(1+\sum_{g\in\mathfrak{G}} 2gp_G(g)+\sum_{g\in\mathfrak{G}} g^2p_G(g)\right)\mathrm{E}[R]^2 - \\
&\quad (1+2\mathrm{E}[G]+\mathrm{E}[G]^2)\mathrm{E}[R]^2 \\
&= (1+\mathrm{E}[G])\mathrm{Var}[R] + [1+2\mathrm{E}[G]+\mathrm{E}[G^2]-(1+2\mathrm{E}[G]+\mathrm{E}[G]^2)]\mathrm{E}[R]^2 \\
&= (1+\mathrm{E}[G])\mathrm{Var}[R] + (\mathrm{E}[G^2]-\mathrm{E}[G]^2)\mathrm{E}[R]^2 \\
&= (1+\mathrm{E}[G])\mathrm{Var}[R] + \mathrm{Var}[G]\mathrm{E}[R]^2. \qquad (4)
\end{aligned}$$

Substituting (3) and (4) into the definition of the Fano Factor, we obtain

$$\begin{aligned}\frac{\mathrm{Var}[M]}{\mathrm{E}[M]} &= \frac{(1+\mathrm{E}[G])\mathrm{Var}[R]+\mathrm{Var}[G]\mathrm{E}[R]^2}{(1+\mathrm{E}[G])\mathrm{E}[R]} \\ &= \frac{\mathrm{Var}[R]}{\mathrm{E}[R]}+\frac{\mathrm{Var}[G]}{1+\mathrm{E}[G]}\mathrm{E}[R]\end{aligned}$$

which is equivalent to (2), our n-gene copy equation.

□

Note that letting $G \sim \mathrm{Bernoulli}(f)$ in (2) allows us to exactly reproduce (1).

4 Properties of the Model

We can run a number of checks to bolster our confidence in the model. First, fluctuations in copy number at low copy numbers should have a larger impact on the noise than fluctuations at high copy numbers– a change from 1 to 2 plasmids is a 100 percent increase in gene copy, whereas a change from 100 to 101 is only a 1 percent increase. If we set $G \sim$ Discrete Uniform(a, b) for simplicity, we find that plasmids which are allowed to fluctuate between 0 and 10 extra copies confer a much higher value to the noise strength (Figure 1) than plasmids which are allowed to fluctuate between 10 and 20 extra copies (Figure 2). This dampening is consistent across a variety of common distribution choices.

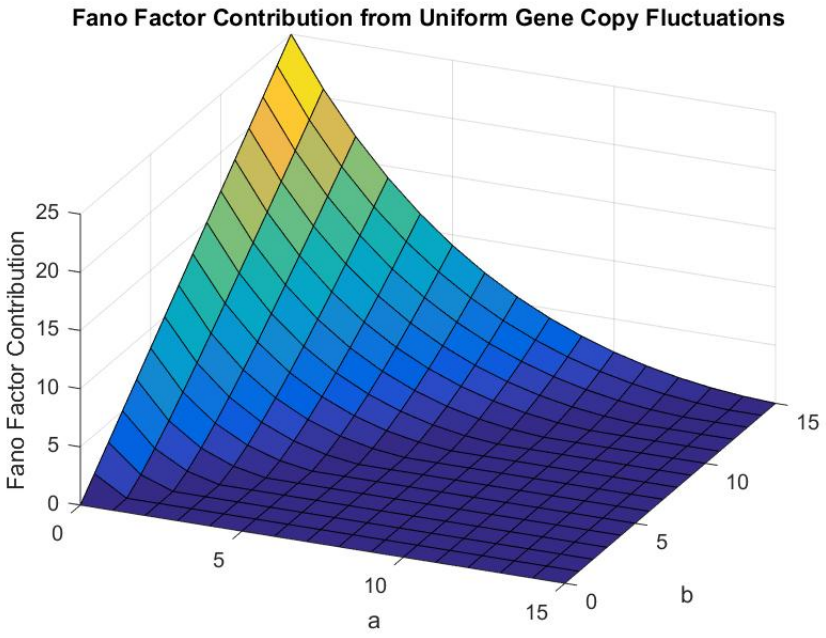

Figure 1: Fano Factor contributions from plasmid fluctuations for various distributions of the family $G \sim$ Discrete Uniform(a, b), ranging through $a, b \in [0, 15]$. $\mathrm{E}[R] = 10$.

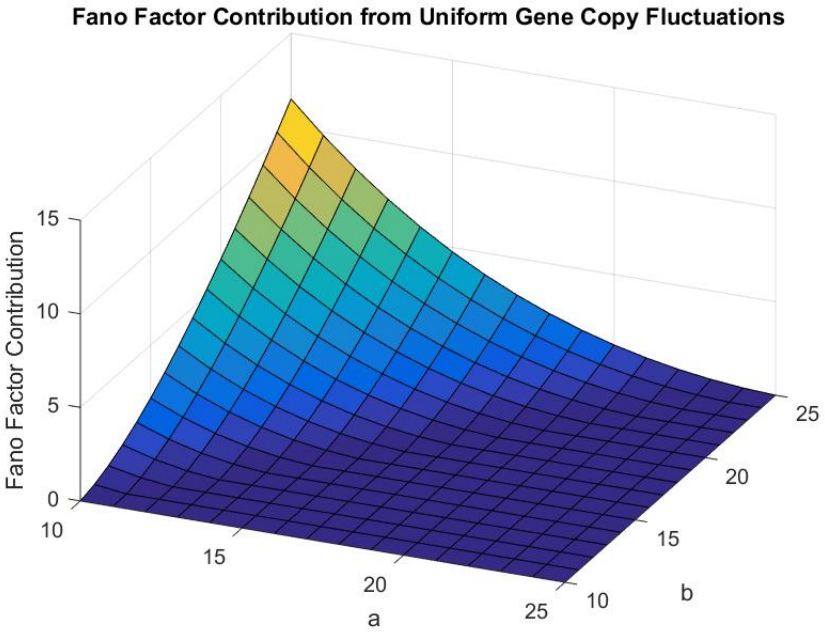

Figure 2: Fano Factor contributions from plasmid fluctuations for various distributions of the family $G \sim$ Discrete Uniform(a, b), ranging through $a, b \in [10, 25]$. $\mathrm{E}[R] = 10$.

We can also examine the extent to which the choice of the distribution of G affects the noise conferred by the plasmid fluctuations. If we set $G \sim$ Bernoulli(f), we see for various values of $\mathrm{E}[R]$ that the largest contribution to noise strength occurs near $f \approx 0.5$, ie. when fluctuations in gene copy number happen more frequently (Figure 3).

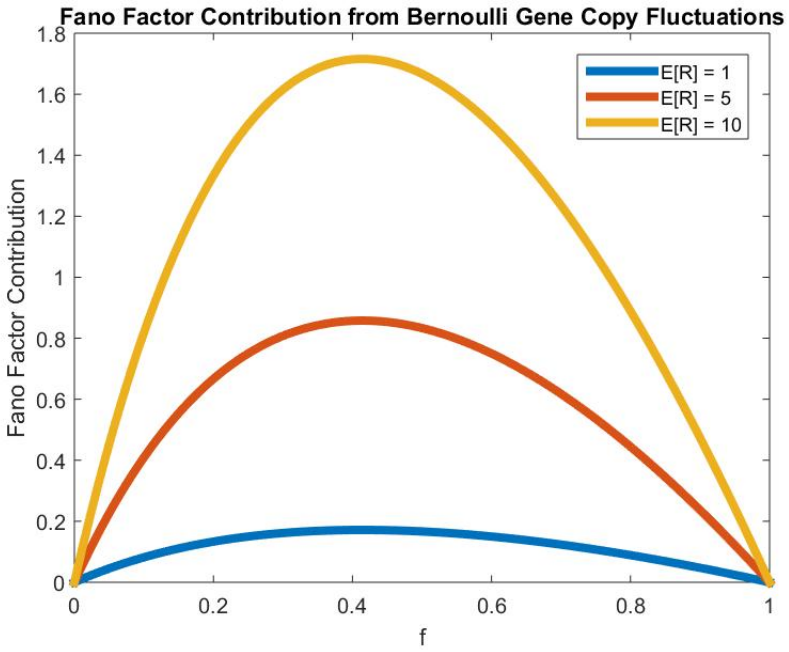

Figure 3: Fano Factor contributions from plasmid fluctuations for various distributions of the family $G \sim \text{Bernoulli}(f)$, ranging through $f \in [0, 1]$. Values of $\text{E}[R]$ chosen to correspond with Figure S6 of [4].

Generalizing the distribution of G, however, we find that this effect persists– if G is normally distributed so that the average plasmid copy number is 200, the Fano Factor contribution increases monotincally as the distribution widens. If we allow $\sigma = 49$ so that the any values in the published 100-300 copies-per-cell range fall within 2 standard deviations of the mean, the additional noise strength is so high that it questions the validity of the assumption (Figure 4).

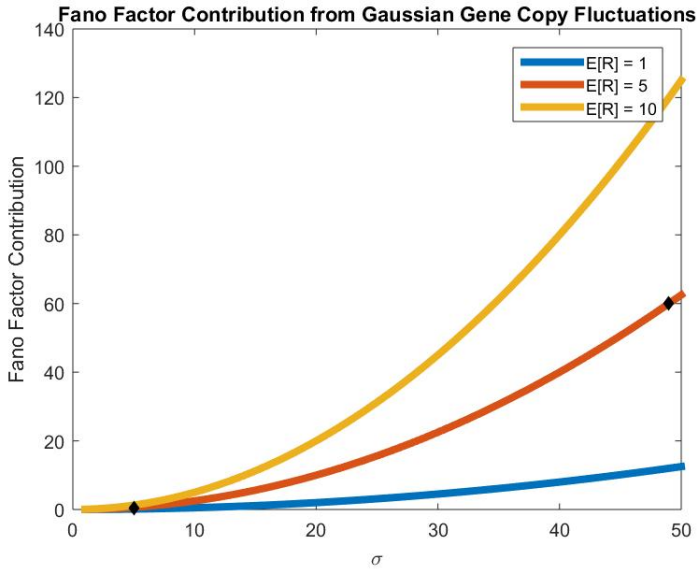

Figure 4: Fano Factor contributions from plasmid fluctuations for various distributions of the family $G \sim N(199, \sigma)$, ranging through $\sigma \in [1, 50]$. E$[R]$ values chosen to correspond to Figure 3. Black diamonds correspond to distributions in Figure 5.

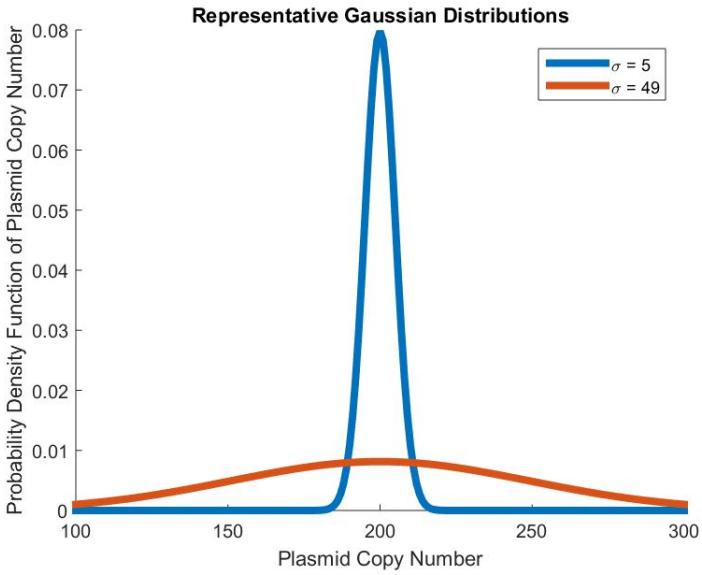

Figure 5: A portion of the distributions corresponding to black diamonds in Figure 4. The $\sigma = 5$ distribution of G induces much less noise than the $\sigma = 49$ distribution.

5 Conclusions

From our model it is clear that the additive effect of gene copy number fluctuations on noise strength that Jones *et al* observed for up to 2 gene copies persists for any number of possible gene copy numbers. However, we can see that if we want to analyze the magnitude of this additive term for a given mean transcription level, it is sufficient, but also necessary, to know the first two moments of the distribution of G.

Although the distributions of steady-state copy number are considered to be narrow Gaussian around the mean in most wild-type plasmids [7], the distributions for the pSB1X3-type plasmids are not well-described.

While we felt it would be relatively simple to disentangle the effect of signal damping on the noise values we would obtain from our measurements, the confounding factors arising from the fluctuations of the plasmids themselves would require an explicit input of the distribution of G into our model. Because our model revealed that the choice of distribution so heavily determines the strength of the additional noise, we decided that we could not justifiably assume any distribution for G without additional information which was beyond the scope of our project. To ensure the validity of our experiment's results, we chose not to proceed with the plasmid-based reporter system.

References

[1] `https://en.wikipedia.org/wiki/Signal_averaging`

[2] Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swian, P.S. Stochastic Gene Expression in a Single Cell. (2002) *Science* **297**, 1183-1186.

[3] Swain, P.S., Elowitz, M.B., and Siggia, E.D. Intrinsic and extrinsic contributions to stochasticity in gene expression. (2002) *PNAS* **99**, 12795-12800.

[4] Jones, D.L., Brewster, R.C., and Phillips, R. Promoter architecture dictates cell-to-cell variability in gene expression. (2014) *Science* **346**, 1533-1536.

[5] Arriaga, E.A. Determining biological noise via single cell analysis. (2009) *Anal Bioanal Chem* **393**, 73-80.

[6] `http://parts.igem.org/Part:pSB1C3`

[7] del Solar, G. and Espinosa, M. Plasmid copy number control: an ever-growing story. (2000) *Molecular Microbiology* **37**, 492-500.