Controlling *E. coli* gene expression noise

Kyung Hyuk Kim, Kiri Choi, Bryan Bartley, Herbert M. Sauro.

Abstract

Intracellular protein copy numbers show significant cell-to-cell variability within an isogenic population due to the random nature of biological reactions. Here we show how the variability in copy number can be controlled by perturbing gene expression. Depending on the genetic network and host, different perturbations can be applied to control variability. To understand more fully how noise propagates and behaves in biochemical networks we developed stochastic control analysis (SCA) which is a sensitivity-based analysis framework for the study of noise control. Here we apply SCA to synthetic gene expression systems encoded on plasmids that are transformed into *Escherichia coli*. The objective of the study was to show that we could differentially control the noise and mean levels of molecular concentrations in biological networks. We show that (1) dual control of transcription and translation efficiencies provides the most efficient way of noise-vs.-mean control. (2) The expressed proteins follow the gamma distribution function as found in chromosomal proteins. (3) Bursting size and frequency are strongly correlated, implying that transcription efficiency can affect transcript lifetimes and/or translation efficiency. (4) Lastly, genetic encoding in plasmids amplifies intrinsic noise of gene expression, showing that the two-promoter state model, commonly used to describe chromosomal gene expression, may need to be modified.

Keywords

Synthetic biology, gene expression noise, stochasticity, noise control, two state model, stochastic control analysis

Kyung Hyuk Kim, Kiri Choi, Bryan Bartley, and Herbert M. Sauro are with the Department of Bioengineering, University of Washington, Seattle, WA 98195 USA. The e-mail address of Kyung H. Kim is kkim@uw.edu.
Controlling *E. coli* gene expression noise

I. INTRODUCTION

CELL-TO-CELL variability in protein copy numbers within isogenic populations are typically observed in various types of cells due to underlying random biochemical reaction processes [1], [2], [3]. The variability can lead to noise-induced cellular phenotypes such as cellular differentiation [3], multiple stability [4], and either sensitivity enhancement or suppression [5], [6]. Here we investigate the ability to differentially control the noise and mean levels of gene expression in *E. coli*. The approach we use is based on stochastic control analysis (SCA)[8], [7], a body of theory we developed and reported in previously publications.

SCA is a sensitivity analysis framework, that is a direct extension of metabolic control analysis [9], [10] to the stochastic regime [8]. This approach is based on a local sensitivity analysis that can be applied to study first-order effects of finite-size perturbations. SCA can identify which parameters in stochastic systems – here, gene regulatory circuits – need to be varied by how much to achieve a desired control aim. This includes orthogonal control of noise levels with respect to mean levels, and simultaneous changes in noise and mean levels in the same or opposite directions for the same or different protein species. SCA can provide control efficiency and strength to identify the most effective control schemes that are experimentally relevant [7]. Here, we apply SCA experimentally to *E. coli* genetic systems and achieve differential noise control in vivo.

In this paper, gene circuits are encoded on plasmid backbones, which are inserted into *E. coli* MG1655. We show that encoding circuits in plasmids amplifies intrinsic circuit dynamics. In the *E. coli* transcriptome study [11], the extrinsic noise was found in many cases to completely suppress the intrinsic noise when transcription factor copy numbers were larger than 10 (refer to Fig. 2B in [11]). This means that all dynamics faster than cell doubling time such as transcription-translation processes is significantly averaged out. Therefore, to study these processes, it is necessary to use fast-responsive probes [13] or to come up with methods that amplify the intrinsic processes. Encoding genetic systems of interest in plasmids will be shown to amplify the intrinsic processes and this allows us to investigate *E. coli* gene expression with flow cytometry and fluorescence microscopy without resorting to single-molecule fluorescence microscopy [11], [12].

II. STOCHASTIC CONTROL ANALYSIS: REVIEW

SCA [7], [8] is a local sensitivity analysis based on control coefficients, which are defined approximately as percentage change in a response signal ($y$) divided by the percentage change in a system parameter ($p$):

$$C_p^y = \frac{p \cdot \delta y}{y \cdot \delta p} = \frac{d \log y}{d \log p}.$$  

We note that the slope in the log-log plot of $p$ vs. $y$ corresponds to $C_p^y$. The response signal can be the mean or noise levels of mRNAs or proteins. The parameters can include transcription and translation efficiencies, degradation rates of mRNAs and proteins, dilution rate due to cell growth, and reaction rates of transcription-factor binding and unbinding from promoter regions, etc. Another important quantity in SCA, is the control vector, each element of which corresponds to a control coefficient for a given response signal ($y$):

$$C_p^y = (C_p^{y_1}, C_p^{y_2}, \ldots, C_p^{y_N}),$$

where $N$ defines the number of parameters (dimension of the parameter space) that will be varied to control the value of $y$. In this paper, we are mostly interested in dual control of transcription and translation efficiencies, i.e., $N = 2$. One of the important properties of the control vector is that its inner-product with a parameter perturbation vector $\delta p$ becomes the amount of change in the response signal $\delta y$,

$$C_p^y \cdot \delta p = \frac{\delta y}{y}.$$  

We can quantify which parameter value, and by how much it should be controlled to achieve specific control aims. For example, consider a case where the noise level of a protein needs to be reduced by 9%, while its mean level should remain the same. Here, the noise level ($n$) is defined by the variance divided by the squared mean value, i.e., squared coefficient of variation. Two control vectors for the noise and mean levels, $C_n^p$ and $C_m^p$, need to be computed based on a given mathematical model. System parameters need to be perturbed while satisfying

$$\frac{\delta n}{n} = C_n^p \cdot \delta p = -0.09$$  

and

$$\frac{\delta m}{m} = C_m^p \cdot \delta p = 0.$$  

The perturbation vector $\delta p/p$ satisfying these two equations can be solved, but the solutions can be infinite. In that case, it is important to select the optimal control scheme (perturbation vector) among the possible solutions. For this, the control efficiency and strength were introduced [7]. Based on these two quantities, one can choose desired control schemes that are appropriate to systems of interest with the maximum control strength and/or efficiency.

III. SCA FOR A SINGLE GENE EXPRESSION CASSETTE

We constructed plasmid systems that express green fluorescent protein (GFP) under *lac*-promoters in *E. coli* (Fig. 1A). The plasmid copy number in a single cell fluctuates in time because the a set of plasmids are randomly partitioned during cell division and are synthesized in a stochastic fashion. Thus, the copy number of *lac*-promoters per cell fluctuates, which will be discussed later in the two-state model. For simplicity,
we will assume that the plasmid copy number is tightly controlled, i.e., constant, at the first level of approximation. The total number of *plac* will be the sum of the number of inactive and active *lac*-promoters (Fig. 1B), which will be set to a constant, *N* *p*. We call this the two-state model. The plasmid backbone that we used is pGA3K3 with the replication origin, p15A (*N* *p* = 10 – 30). Based on this two-state model, we computed control vectors for the mean and noise levels of GFP fluorescence as shown in Fig. 1C (refer to [7] for the control vector computation).

The computed control coefficients show that noise can be controlled efficiently by varying both *α* *m* and *γ* *m*. *C* *α* *m* = −1.00 and *C* *γ* *m* = 0.97, indicating that with an increase in *α* *m* for example by 10%, *n* will reduce by 10% (this is a first-order approximation, because control coefficients are defined locally). Similarly, with an increase in *γ* *m* by 10%, *n* will increase by 9.7%. With similar change in *α* *m* or *γ* *m*, the mean level can also be efficiently controlled; *C* *α* *m* = 1 and *C* *γ* *m* = −1, indicating that the mean level increases and decreases by 10% to 10% increase of *α* *m* and *γ* *m*, respectively.

**IV. MEAN LEVEL CONTROL**

From the computed control coefficients, the mean protein levels can be controlled without changing the noise level (with a minor change, ~10 folds less than the change in the mean levels) by varying either *α* *p* and *γ* *p*. To confirm this theoretical prediction, we changed the translation efficiency *α* *p* by using a library of both ribosome binding sites (RBSs) and spacer sequences as shown in Fig. 2. Among them, four different spacers – TACTAG, AAAAAA=(A)6, (A)10, and (A)13 – that are placed between B0034 and the start codon showed distinct GFP expression levels when *plac* is fully active ([IPTG] = 1 mM). Here, the introduced spacer sequences are presumed to change ribosome binding affinity, in particular, translation initiation – the limiting step for a translation rate [14], [15].

Based on our flow cytometry data, the mean level was successfully varied by using different spacer sequences as shown in Fig. 3 and Fig. 4A and C. We compared two different cases: Points A and B, and Points C and D in Fig. 3. As shown in Fig. 4A and C, the rescaled probability density functions (pdfs) were overlapped with excellent accuracy. This scale invariance confirms that the noise levels of both points are the same.

**Scale invariance in the gamma distribution:** Furthermore, the observed invariance implies a special property that we need to consider carefully. This invariance property is satisfied by the gamma distribution function as shown in the Materials and Method section when the bursting size is rescaled together. This implies that the difference between the system parameters of Points A and B is only the bursting size, which is consistent to our perturbation experiment. This result supports that the observed distribution function is the gamma distributions (confer to [16], [17] about claims for other types of distribution functions). To confirm this, we fit the GFP pdfs to the gamma distribution functions as shown in Fig. 5A. For the cases that the background fluorescence is well separated from GFP signals, we confirmed that the pdfs follow the gamma distributions well.

**V. NOISE LEVEL CONTROL**

As discussed above, the noise level can be efficiently controlled by varying *α* *m* and *γ* *m*. However, when these parameters are changed, the mean level also changes with the same fold difference but in the opposite direction; for example, in Fig. 1C, *C* *α* *m* and *C* *γ* *m* are −1.00 and 1.00, meaning that when *α* *m* is increased by x%, the noise level decreases by x%, while the mean level increases by x%. Thus, to change the noise level without changing the mean level, we must vary at least two different parameters simultaneously. Since the mean level can be controlled almost independently of the noise level by changing *α* *p*, we will vary both *α* *p* and *α* *m* to compensate...
Fig. 3. Scaling relationship between noise and mean levels: Four different cases of spacer sequences between the ribosome binding site BBa_B0034 and the start codon are shown. The same symbol represents the same spacer with different [IPTG]. The noise levels (squared coefficient of variation) are inversely proportional to the mean level. For comparison, a line with a slope $-1$ is drawn. The contribution to the noise level by background fluorescence signals was taken out (Materials and Methods).

VI. SCALING RELATIONSHIP BETWEEN THE NOISE AND MEAN LEVELS

Figure 3 shows that the noise level is inversely proportional to the mean level:

$$ n = \frac{c}{m}, $$

where $c$ is a constant. It is known that the two-state model satisfies this inverse relationship [18], [11], [19]. With the plasmid copy number fixed at any positive integer values, the constant $c$ is shown to be independent of $N_p$ (refer to its derivation in the Supplementary Note of [18]). This indicates that the plasmid copy number $N_p$ should not shift the plot for the noise vs. mean levels. However, we observed this shift to the right, indicating the value of $c$ is somehow increased. Thus, the observed shift implies that the two-state model may need to be modified by taking into account the stochastic fluctuations in $N_p$. Alternatively, the observed shift could be due to increased bursting size; ribonuclease activity may be saturated, leading to increased lifetime of transcript.
We need to further investigate whether the proposed model is incomplete or whether there is any non-trivial biological correlation between parameter values such as $N_p$ and $b$.

The observed shift indicates that the intrinsic gene expression processes were amplified and their dynamics can be observed out of the sea of strong extrinsic noise. Otherwise, the intrinsic noise would have been buried by the extrinsic noise as observed for typical *E. coli* transcription factors with their copy numbers higher than $\sim 10$ [11, 20] (confer to the study on yeast transcriptome [18], where intrinsic noise is strong enough to be observed for most transcription factors because the extrinsic noise level is lower than *E. coli* due to its longer doubling time). The minimum noise level that we observed in the GFP signals is slightly larger than 0.1, which is consistent with the *E. coli* transcriptome study [11].

Bursting frequency $a$ and bursting size $b$ were estimated from our flow cytometry data by fitting the gamma distribution to the observed GFP pdfs (Materials and Methods). Figures 5B and C show the ideal case that varying IPTG concentrations (Materials and Methods) contributed to the observed GFP pdfs (Materials and Methods). The noise level was corrected by using the property that the observed variance ($\text{Variance}_o$) is the sum of the GFP variance ($\text{Variance}_p$) and the background signal variance ($\text{Variance}_b$) under the assumption that the GFP signals are statistically independent of the background signals. More precisely, the noise level of GFP signals, defined by the square coefficient of variation, can be obtained by

$$CV^2 = \frac{\text{Variance}_o - \text{Variance}_b}{(\text{Mean}_o - \text{Mean}_b)^2}.$$  

where the subscripts $o$ and $b$ denote observed and background signals, respectively.

D. Nonlinear Regression

The mean level was corrected with a simple subtraction. The noise level was corrected by using the property that the observed variance ($\text{Variance}_o$) is the sum of the GFP variance ($\text{Variance}_p$) and the background signal variance ($\text{Variance}_b$) under the assumption that the GFP signals are statistically independent of the background signals. More precisely, the noise level of GFP signals, defined by the square coefficient of variation, can be obtained by

$$CV^2 = \frac{\text{Variance}_o - \text{Variance}_b}{(\text{Mean}_o - \text{Mean}_b)^2}.$$  

where the subscripts $o$ and $b$ denote observed and background signals, respectively.

VII. CONCLUSIONS

We showed that encoding gene circuits in plasmids amplifies intrinsic circuit dynamics, so that intrinsic gene expression in *E. coli* can be observed with flow cytometry without resorting to single-molecule microscopy. In addition, stochastic control analysis was applied to identify efficient ways to control noise and mean levels of gene expression, showing that SCA can be applied to gene regulatory networks and other stochastic biological systems.

MATERIALS AND METHODS

A. GFP expression circuits and strains

All genetic components used in this manuscript are BioBrick parts, from which genetic circuits were constructed by using the Gibson assembly method [21]. The constructed circuits were integrated into a low-to-medium copy number plasmid pGA3K3 with a Kanamycin resistance gene and *Escherichia coli* MG1655 Z1 was transformed with the plasmids. The strain (lacI$^0$) constitutively overexpresses LacI from its chromosome.

B. Cell Growth and Flow Cytometry Measurements

*E. coli* strains were grown to OD600$\rightarrow$0.2 in 2mL Luria-Bertani (LB) media (Becton Dickinson) with kanamycin 50$\mu$g/mL at $37^\circ$C and 300rpm in a shaker. The cultures were diluted 1:200 into 200 $\mu$L prewarmed fresh M9 media (Teknova 2M1990) in 96-well plates (Costar 3904) with kanamycin 50$\mu$g/mL. 12 different IPTG concentrations (0nM, 0.02$\sim$1mM) were used for each well and grown to OD600=0.3-0.4 in a shaker ($37^\circ$C, 300rpm). For the flow cytometry measurements, the grown cultures were diluted 1:4 in 1xPBS. A Sony Biotechnology ec800 flow cytometer was used with a 525nm filter and a 488nm excitation laser for GFP fluorescence. 100,000 events were collected for each sample and gated by using a 2-D normal distribution (Biocductor flowCore norm2filter function with scale.factor=1) [22] within the R software environment as well as by using python package FlowCytometryTools (http://gorelab.bitbucket.org/flowcytometrytools/). To prevent well-well contamination we executed a Medium Flush cycle after each sample well. When computing the mean and noise levels of GFP signals, background fluorescence was taken care of by using the mean and noise levels for the case without IPTG for each different gene circuit.

C. Noise Level Correction

The mean level was corrected with a simple subtraction. The noise level was corrected by using the property that the observed variance ($\text{Variance}_o$) is the sum of the GFP variance ($\text{Variance}_p$) and the background signal variance ($\text{Variance}_b$) under the assumption that the GFP signals are statistically independent of the background signals. More precisely, the noise level of GFP signals, defined by the square coefficient of variation, can be obtained by

$$CV^2 = \frac{\text{Variance}_o - \text{Variance}_b}{(\text{Mean}_o - \text{Mean}_b)^2}.$$  

where the subscripts $o$ and $b$ denote observed and background signals, respectively.

$$p(N_{pr}; a, b) = p(c_{xa}; a, b) = \frac{(c_{xa})^{a-1}e^{-c_{xa}/b}}{\Gamma(a) b^a}$$

$$= c_{xa}^{-1} \frac{\Gamma(a) (b/c_{xa})^{a}}{\Gamma(a) (b/c_{xa})^{a}}$$

Here, $\Gamma$ is a gamma function,

$$a \equiv \frac{\alpha_m}{\gamma_p} \times [\text{Active Promoter Copy Number}]$$

is the number of mRNA produced per cell doubling time, called bursting frequency, and

$$b \equiv \frac{\alpha_p}{\gamma_m}$$

is the number of proteins produced during the mRNA lifetime, called bursting size. Therefore, the fluorescence intensity should also follow the gamma distribution if its corresponding copy number follows the gamma distribution, with the bursting size rescaled with $c_{xa}$. Nonlinear regression was carried out using the Scipy curve_fit function (http://www.scipy.org/), which employs the Levenberg-Marquardt algorithm for the least squares fitting to estimate $a$ and $b$. 


ACKNOWLEDGMENT

This work was supported by the National Science Foundations (NSF MCB 1158573).

REFERENCES


