

# Biochemical Complexity Drives Log-Normal Variation in Genetic Expression

Jacob Beal<sup>1</sup> ✉

<sup>1</sup> Raytheon BBN Technologies, Cambridge, MA 02138, USA

✉ E-mail: jakebeal@ieee.org

ISSN 1751-8644

Received on 17Feb2017

Revised on ???

Accepted on ???

doi: 000000000

www.ietdl.org

**Abstract:** Cells exhibit a high degree of variation in levels of gene expression, even within otherwise homogeneous populations. The standard model to describe this variation centers on a gamma distribution driven by stochastic bursts of translation. Stochastic bursting, however, cannot account for the well-established behavior of strong transcriptional repressors. Instead, it can be shown the very complexity of the biochemical processes involved in gene expression drives an emergent log-normal distribution of expression levels. Emergent log-normal distributions can account for the observed behavior of transcriptional repressors, are still compatible with stochastically constrained distributions, and have important implications for both analysis of gene expression data and the engineering of biological organisms.

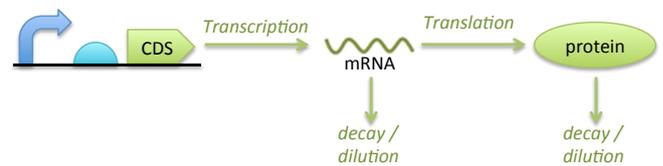
## 1 Introduction

One of the key challenges in understanding and engineering biological organisms is the high degree of cell-to-cell variation commonly observed in gene expression, even within otherwise homogeneous populations of cells. Even with strong genetic expression, the observed range of cell-to-cell variation is often on the same order as the ranges over which expression can be regulated or engineered. As such, cell-to-cell variation in gene expression is a factor that needs to be taken into account in engineering the behavior of biological organisms, and it is important to have a well-founded quantitative model of the nature and origins of cell-to-cell variation.

The current dominant model for cell-to-cell variation in otherwise homogeneous populations is a combination of temporal fluctuations due to “intrinsic noise” from the stochasticity of chemical reactions and “extrinsic noise” that comes from differences in the environment in which those reactions occur (see, e.g., [1–4]). The chemical reactions involved in gene expression are certainly stochastic, based both on the fundamental nature of chemical interactions and on a number of experimental observations of the activity of single molecules (e.g., [5–7]). Applying stochastic analysis to an abstract model of transcription and translation (Figure 1) then leads to an expected gamma distribution of protein expression levels [8]. This model thus explains high cell-to-cell variation as the result of a “bursty” process of translation driven by sparse transcription events. As one might expect from the law of large numbers, however, with higher rates of transcription this stochastic model predicts that there should be little cell-to-cell variation.

Experimental study of *E. coli* has indeed found that the gamma distribution fits well for a wide range of observed natural protein expression levels [2]. For proteins with an average of more than about 10 molecules per cell, however, stochastic bursting cannot explain the observed variation [2]. Under the standard model, then, the explanation of variation for many of the most important systems in the cell is left to fall back on rate variations caused by “extrinsic noise,” an ad hoc definition with no mathematically grounded mechanism and many competing definitions and models (consider, for example, the variety found in [1, 3, 4, 9–12]). This situation is particularly problematic for the engineering of biological organisms, which typically relies quite heavily on strongly expressed genes.

The canonical model in Figure 1, of course, is well-known to be a vast oversimplification of the complex processes involved in genetic expression. A number of attempts have been made to elaborate the basic stochastic bursting model by inclusion of additional complexity in the “intrinsic” stochastic elements and/or the “extrinsic” black



**Fig. 1:** Stochastic analysis of the abstract transcription / translation model depicted above predicts a gamma distribution of protein expression, as derived in [8].

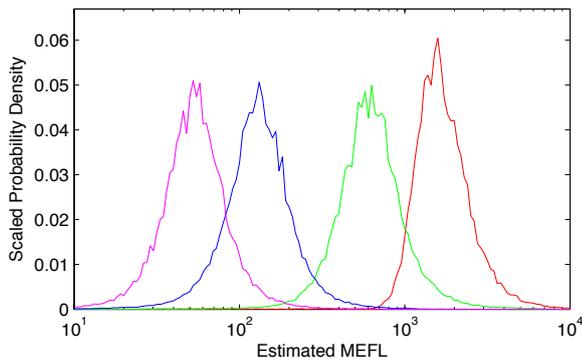
box of rate modulations (e.g., [12–16]). None, however, either provides a stochastic model sufficient to account quantitatively for the high degree of variation present with strong genetic expression nor an “extrinsic” model with mechanistic grounding.

This paper proposes a fundamentally different approach, based on the emergent properties of complex reaction networks. In particular, we show that, absent dominating factors, the process of genetic expression is expected to produce a log-normal distribution. Any significant deviation from log-normal then indicates one or more mechanisms being in an extreme condition such that its distribution becomes dominant. When such dominating factors exist, the model predicts a distribution of cell-to-cell variation equal to a product of a log-normal distribution with the dominating factor distribution. Thus, stochastic bursting becomes a special case in which an extremely low transcription rate modulates the log-normal baseline; another example this paper discusses is stochastic transfection creating a bimodal log-normal distribution.

In the remainder of the paper, Section 2 introduces the strong transcriptional repressors that will be used as a primary test case and demonstrates that their variation cannot be explained by transcriptional stochasticity. Section 3 then develops the log-normal model of complex biochemical processes and shows that it conforms well with strong transcriptional repressors. Section 4 explores the implications of embracing the log-normal model, and finally Section 5 summarizes results and discusses future directions.

## 2 Stochastic Bursting Cannot Explain Variation in Strong Transcriptional Repressor Devices

Strong transcriptional regulatory devices, such as TetR/pTet, LacI/pLac, and AraC/pBAD, were amongst the first systematically



**Fig. 2:** Cells exhibit a high degree of variation in genetic expression. For example, this figure shows several representative distributions of expression taken from the LmrA repressor data for Cello [19]: fully repressed (magenta), unexpressed (red), and two intermediate levels (blue and green). Note that the distributions have a width on the same order as the entire range of regulation and that they are roughly symmetric on a logarithmic scale.

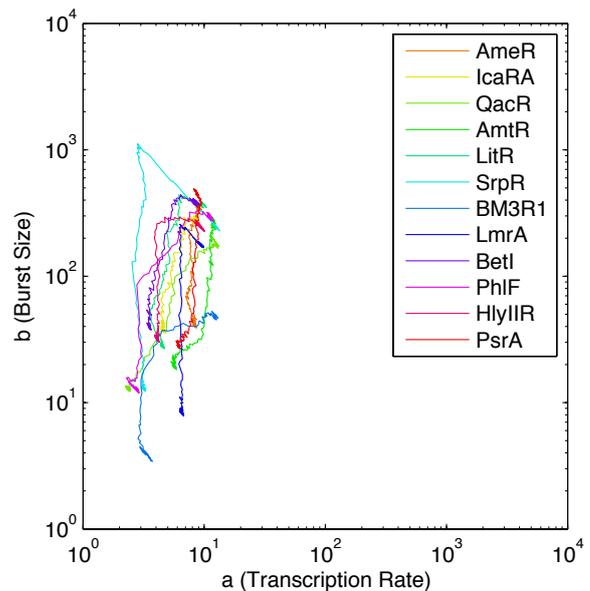
engineered biological components [17, 18] and are still critical and widely used components in the engineering of biological systems. The range of expression between fully “on” and fully “off” states for such devices can be quite large, often in the range of 10-fold to 1000-fold difference of expression levels, depending on the particular devices, host, and context. The degree of cell-to-cell variation, however, is often not strongly affected by such changes in expression level and as such we shall see that such variation cannot be accounted for by stochastic bursting.

As an example of strong devices used for effective engineering of biological circuits, let us consider the regulated expression levels of twelve *E. coli* engineered transcriptional repressor devices from Cello [19]\* Each device has an associated “transfer curve” providing a function from input expression level to output distribution of expression levels, constructed for [19] by smoothing input and output distributions collected from twelve measurements for each device across a wide range of induction levels. The units of the original curves are only relative, so for better intuition regarding their meaning, discussion in this paper will be presented with linear rescaling to an approximate estimate of Molecules of Equivalent FLuorescein (MEFL) based on the values in [20]. Note, however, that this does not affect any of the mathematical analyses in this paper, which are all invariant to linear scaling.

Figure 2 shows several expression distributions representative of repressors in the Cello data set. In particular, note that the distributions have a width similar to the entire range of regulation and that they are roughly symmetric on a logarithmic scale. Similar log-symmetric distributions across a large range of regulation are not peculiar to Cello, but a pattern of variation observed frequently across a wide range of engineered biological systems (see, for example [21–23]). Wide, roughly log-symmetric distributions such as these are fairly typical for “well-behaved” strong transcriptional regulators in bacteria (and in many other contexts as well). Moreover, the scale of the cell-to-cell variation is remarkably similar across the full range of devices and input levels: the geometric standard deviation of expression ranges only from a minimum of 1.5-fold to a maximum of 2.6-fold across all twelve devices and three orders of magnitude difference in geometric means.

Such log-symmetric distributions can be fit well by the gamma distribution (as determined in [2]), but can also be fit well by log-normal distributions (hence the use of geometric statistics above), and by a number of other distributions, including Weibull, generalized gamma, etc., all of which can be quite difficult to distinguish

\*In particular, the collection of transfer curves from the implementation at <https://github.com/CIDARLAB/cello/>, as of commit 27f6354f41cd2997610e79e2a41ded61f2c3fa91



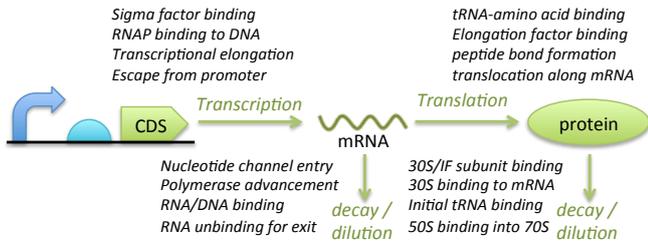
**Fig. 3:** Gamma distribution parameters  $a$  and  $b$  fit for expression and transfer curves from [19]: transcription rate  $a$ , which should change dramatically, is confined to a small and fairly constant range, while translation burst size  $b$ , which should be unaffected, changes across multiple orders of magnitude

between simply by looking at the structure of a distribution [24–27]. With a mechanistic model of the origins of a distribution, however, we can readily test whether changes in the configuration of the biological system are matched by predicted changes in the corresponding parameters of the distribution.

For the stochastic bursting model, the gamma distribution parameters  $a$  and  $b$  correspond with transcription rate and translation burst size, respectively. Since a transcriptional repressor device modulates transcription but not translation, there should be little change in  $b$  but dramatic change in  $a$ : when the repressor is expressing at a low level, there should be a low  $a$  value and a broad distribution of variation, but when the repressor is expressing at a high level, there should be a high  $a$  value and a much tighter distribution of variation. Intuitively, if strong expression is driven by high rates of transcription (as we know it is for transcriptional repressors), the law of large numbers should make the unevenness of bursts largely insignificant.

Figure 3 shows the results of fitting the Cello devices to a gamma function, per the stochastic bursting model. Here we see almost the exact opposite of what is predicted by the stochastic bursting model: while transcription rate does change slightly with expression, it is a very weak change that does not match the orders of magnitude difference known to occur with such repressor devices. Instead, the orders of magnitude change are attributed by the model to changes in translation burst size, which also does not match any known mechanism. As the mechanisms of transcriptional regulation are quite well-established experimentally, we must conclude that stochastic bursting can account for at most a small fraction of the observed cell-to-cell expression variation in the Cello transcriptional repressor devices.

Enhanced stochastic bursting models with bursty transcription or more complicated transcriptional interactions can amplify the amount of variation expected for particular species concentrations, but still cannot address the fundamental problem: if variation is due primarily to some form of stochastic bursts of translation, then when the transcription rate rises across several orders of magnitude, the amount of variation must fall dramatically but, in general, does not. Moreover, given that the cell-to-cell variation in the devices we have tested is fairly typical (and in fact relatively tight compared to many), the failure of the stochastic bursting model to account for variation appears likely to extend to any other device that does not



**Fig. 4:** Some of the many complex biochemical reactions that take place as part of transcription and translation and are abstracted away from simple models such as the one shown in Figure 1.

have very low levels of transcription, particularly given the finding in [2] of a large and consistent level of variation for proteins with more than about 10 copies per cell. We focus here on strong transcriptional regulation only because it is a decisive means of distinguishing stochastic bursting from other possible contributions to variation\*. We thus must turn elsewhere for a mechanistic explanation of the remarkable regularity of expression variation across orders of magnitude difference in expression level.

### 3 Complexity of Genetic Expression Implies Log-Normal Variation

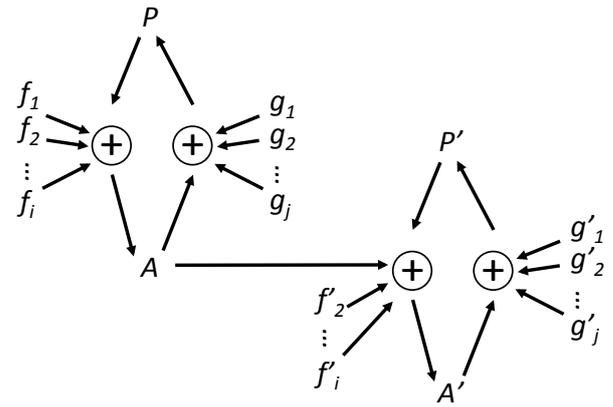
As has already been well recognized (e.g., [1]), the standard stochastic bursting models abstract away most of the biochemical complexity of genetic regulation and expression. Figure 4 illustrates this with annotation of the standard model from Figure 1 with a few of the many complex processes that are omitted, most of which are also omitted from all of the previously proposed enhancements of the stochastic bursting model. Abstracting away this complexity is reasonable when extremely low rates of transcription cause stochastic bursting to dominate other sources of cell-to-cell variation. When stochastic bursting does not dominate, however, some combination of these other factors must be taken into account.

Consider that the operation of nearly every biochemical reaction involved in determining genetic expression should be expected to vary at least slightly from one cell to another due to differences in the state of individual cells, such as their size, health, available pools of various resources, etc. This might seem to indicate that it is hopeless to attempt to model cell-to-cell variation without a full accounting of the expected distributions of each of these many models. That very complexity, however, leads to an emergent property that provides a different route to abstraction with a strong mechanistic and mathematical grounding.

Unlike most non-biological chemical reactions, many biological processes in the cell follow a composable catalytic pattern (Figure 5). In this pattern, a transition of some molecular species from passive state  $P$  to active state  $A$  is driven by a chemical reaction with some set of catalytic factors  $f_1, f_2, \dots, f_i$ . Molecules in active state  $A$  then transition back to  $P$  again either spontaneously or catalyzed by reaction with another set of factors  $g_1, g_2, \dots, g_j$ . For example, RNA polymerase serves as a catalyst for the transformation of nucleotides into mRNA, and ribosomes and tRNA serve as catalysts for the transformation of amino acids into protein. Looking further into the details, one can find yet more catalytic processes on a similar pattern, such as how transcription factors and sigma factors serve as a catalyst for binding RNA polymerase to DNA, initiation factors combine to serve as a catalyst for binding ribosomes to mRNA, tRNAs are charged by aminoacyl tRNA synthetases, etc.

By the standard laws of chemical reactions, we can express the expected rate equation for the active state  $A$  in such a pattern as

\*Note that independent vs. correlated expression (c.f. [28]) does not actually distinguish stochastic bursting, it distinguishes correlated vs. uncorrelated components of genetic expression effects.



**Fig. 5:** Many biological processes follow a composable catalytic reaction pattern, in which reaction with a number of chemical species drives transition of some molecular species from a passive state  $P$  (e.g., unbound RNA polymerase) to an active state  $A$  (e.g., RNA polymerase bound to a particular transcription initiation region) and back again. The active state may then serve as a catalyst in other patterns (e.g., bound RNA polymerase catalyzing transition of nucleotides into a particular mRNA species).

follows:

$$\frac{dA}{dt} = \left( \rho \cdot [P] \cdot \prod_{k=1}^i [f_k] \right) - \left( \lambda \cdot [A] \prod_{k=1}^j [g_k] \right) \quad (1)$$

where  $\rho$  and  $\lambda$  are reaction rate constants (if some reactants are higher order, e.g., acting as dimers, they can appear as more than one  $f_k$ ).

For purposes of this analysis, we will assume that (as is usually the case) the individual genetic expression process under consideration does not dominate cellular resources, and thus that input levels are not changing rapidly in a feedback relation with this particular process. As a result, even though available reactants may be affected by sequestration [29], we can treat the system as quasi-static for purposes of analysis. Under this assumption, the expected equilibrium concentration of  $A$  may be computed as:

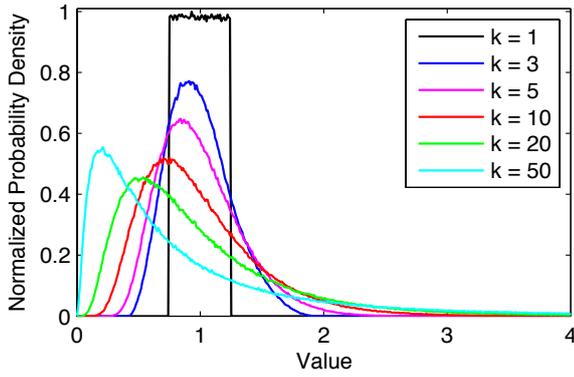
$$0 = \left( \rho \cdot [P] \cdot \prod_{k=1}^i [f_k] \right) - \left( \lambda \cdot [A] \prod_{k=1}^j [g_k] \right) \quad (2)$$

$$[A] = \frac{\rho}{\lambda} \cdot \frac{[P] \cdot \prod_{k=1}^i [f_k]}{\prod_{k=1}^j [g_k]} \quad (3)$$

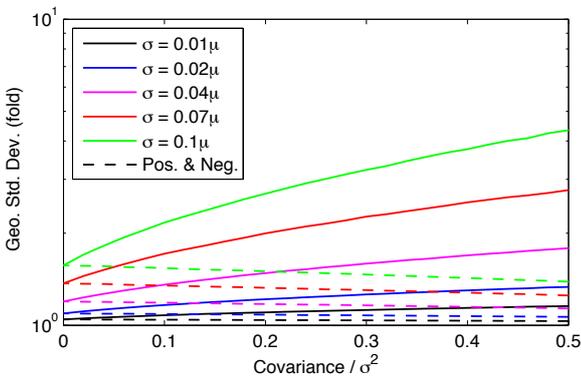
$$[A] = \frac{\rho}{\lambda} \cdot [P] \cdot \prod_{k=1}^i [f_k] \cdot \prod_{k=1}^j [g_k]^{-1} \quad (4)$$

Critically, notice that since the concentration of  $A$  is proportional to a product of input concentrations (and their inverses), then if we consider the inputs as distributions over possible values, then the distribution of  $A$  will be proportional to a product of many distributions (distributions of inverses being distributions as well).

The same holds when we compose patterns of this sort together, such that  $A$  serves as one of the catalytic inputs driving the concentration of species  $A'$ , e.g., a bound RNA polymerase  $A$  acting as a catalyst for the conversion of nucleotides  $P'$  into an mRNA  $A'$ . Composing the equations for  $A$  and their equivalent for  $A'$  will then



**Fig. 6:** By the central limit theorem, any product of independent distributions converges to a log-normal distribution, as illustrated by these histograms of  $10^6$  samples of a product of  $k$  uniform distributions.



**Fig. 7:** Positive correlations (solid lines) spread the log-normal distribution significantly, while mixed positive and negative (dashed lines) tighten it somewhat, as illustrated by the geometric standard deviation of  $10^5$  samples from a product of 20 multivariate normally distributed variables with various standard deviations  $\sigma$  and covariance.

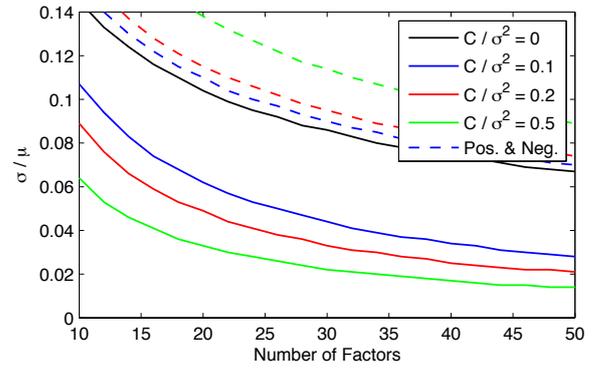
have an equilibrium concentration:

$$[A'] = \frac{\rho'}{\lambda'} \cdot [P'] [A] \cdot \prod_{k=1}^{i'} [f'_k] \cdot \prod_{k=1}^{j'} [g'_k]^{-1} \quad (5)$$

$$[A'] = \frac{\rho' \rho}{\lambda' \lambda} [P'] [P] \prod_{k=1}^i [f_k] \prod_{k=1}^j [g_k]^{-1} \prod_{k=1}^{i'} [f'_k] \prod_{k=1}^{j'} [g'_k]^{-1} \quad (6)$$

The same holds for catalytic inputs driving the opposite transition and for multiple compositions: in every case, the ultimate equation is a product of many concentrations, such that the distribution of the product will be a product of many input distributions.

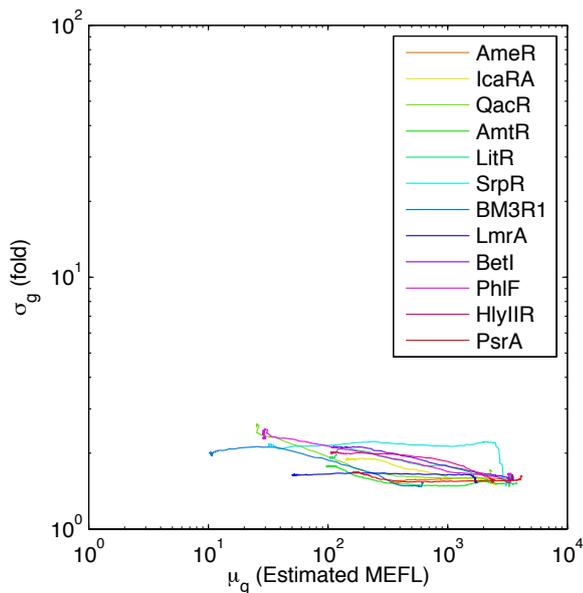
With this, we now have an answer as to the expected distribution of genetic expression. The central limit theorem, which shows that sums of independent random variables converge to a normal distribution, also shows that products of independent random variables converge to a log-normal distribution (multiplication being equivalent to addition on a log scale). Figure 6 illustrates the rapidity of this convergence, showing the histogram of  $10^6$  samples of a product of  $k$  uniform distributions on the interval  $[0.75, 1.25]$ , normalized against the maximum count across all of the plotted histograms. Notice how the distributions spread rapidly and asymmetrically with increasing  $k$ , such that even with fairly small  $k$  the distribution is quite smooth and asymmetric, conforming closely to the ideal pattern.



**Fig. 8:** As the number of multiplicative factors rises, the amount of variation per factor necessary to account for a certain level of variation falls, as illustrated by the minimum  $\sigma/\mu$  that provides 1.6-fold variance in a geometric standard deviation of  $10^5$  samples from a product of a variable number of multivariate normally distributed variables with various levels of positive (solid lines) or mixed positive and negative (dashed lines) covariance.

In cells, of course, many of these factors will not be independent, but will have some degree of correlation. For example, a larger or healthier cell is likely to have more of many types of resources, and differences in chemical environment are likely to have correlated effects on many different reactions. As long as correlation is not too great, the convergence of the distribution to log-normal will still hold, and the distribution will either spread significantly or tighten somewhat based on the degree of correlation. To illustrate the strength of such effects, Figure 7 shows the effect of correlation on the geometric standard deviation computed from  $10^5$  samples of a product of 20 multivariate normally distributed random variables at various levels of standard deviation and correlation. In particular, standard deviation  $\sigma$  is sampled at various levels from  $0.01\mu$  to  $0.1\mu$ , covariance between each pair of variables ranges from 0 to  $0.5\sigma^2$  in steps of  $0.02\sigma^2$ , and correlations are either all positive or are mixed, such that the first half of variables are negatively correlated with the second half of the variables. As expected, positive correlations spread the distribution significantly, while the addition of negative correlations tightens the distribution slightly. Qualitatively, however, correlation in distributions makes no significant change to the core result.

Furthermore, because the distribution emerges from the interaction of many factors, the expectation of log-normal distributions is also likely to remain robust to changes in our understanding of the processes of genetic expression. As long as there are a sufficient number of factors with a multiplicative relationship, a log-normal distribution will emerge, and the more factors are involved and the more that they are positively correlated, the less variation is required from each factor in order to predict a given observed level of variation. To illustrate this, Figure 8 shows how increasing the number of factors decreases the amount of cell-to-cell variation needed in individual factors in order to produce a given level of variance. Geometric standard deviation is computed for  $10^5$  samples of a product of multivariate normally distributed random variables at various levels of positive or mixed correlation as for Figure 8, but in this case  $\sigma$  ranges from  $0.01\mu$  to  $0.015\mu$  in steps of  $0.001\mu$  and the number of variables ranges from 10 to 50 in steps of 2. The figure plots the first  $\sigma/\mu$  for which the geometric standard deviation is greater than 1.6-fold, a typical level of variation for strong expression in the Cello data set from Section 2. As expected, the more variables there are, the less cell-to-cell variation in individual factors is needed in order to account for a given level of observed variation, but the impact per variable is small and decreases with additional variables. Note also that in general the amount of perturbation necessary to achieve significant amount variation is quite small indeed—only a few percent. All together, these results mean that elaborations and changes in models of genetic expression are unlikely to significantly affect the main result.



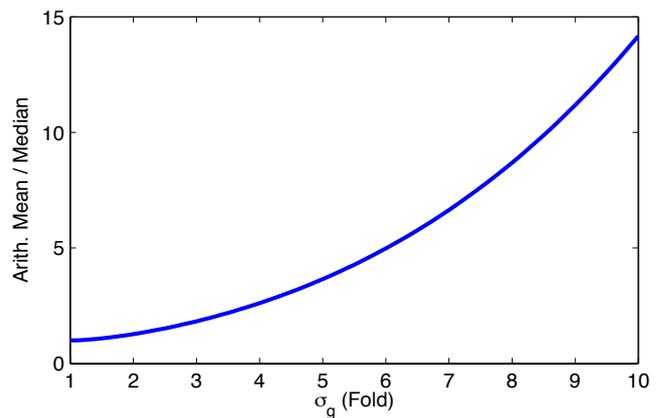
**Fig. 9:** Log-normal distribution parameters  $\mu_g$  and  $\sigma_g$  fit for expression and transfer curves from [19]: the geometric mean  $\mu_g$  ranges broadly, as predicted since it is controlled by strong regulation specific to the rate of transcription initiation, while the geometric standard deviation  $\sigma_g$ , which is expected to arise mainly from many small variations in factors not affected by transcriptional regulation, shows only a slight decrease with increased transcription rate.

We should thus expect the baseline for cell-to-cell distribution of any gene product to be log-normal. If certain factors have an unusually high degree of variability, however, those factors will dominate and the distribution of expression levels should instead be expected to be a product of the dominant distribution with a log-normal distribution. For example, with very low transcription rates variability from stochastic bursting should indeed dominate other factors—and the distributions produced are quite similar to those from a pure gamma distribution with slightly different parameters.

As a further test of this model, we apply the log-normal model to the Cello repressor devices from Section 2 [19], computing the geometric mean and geometric standard deviation for each device and input level. In this case, since these devices are transcriptional repressors, expected to act primarily by greatly changing the rate at which RNA polymerase binds and initiates transcription, we should expect a radical change in the geometric mean across the range of input. Since this is only one aspect of the complex process of genetic expression, however, there should be little change in the standard deviation, though there may be some increased variation at low expression levels if a particular device represses transcription tightly enough to bring the device into the range where stochastic bursting becomes significant. Figure 9 shows that the results of applying the log-normal model to fit the Cello data is as expected: the geometric mean ranges broadly, while the geometric standard deviation tightens slightly with increased transcription rate—as might be expected if a high transcription rate is removing a stochastic component of variation. There is too much variability between devices, however, to attempt to separate and quantify such an effect.

#### 4 Implications of Log-Normal Distribution

The most immediate implication of the convergence to log-normal distributions is that the cell-to-cell variation of gene products should be interpreted in terms of its relationship to log-normal distributions, rather than normal distributions or (except in case of very low expression) gamma distributions. For example, when computing or presenting a statistical summary of single-cell expression data, the geometric mean  $\mu_g$  and geometric standard deviation  $\sigma_g$  should be



**Fig. 10:** Gene expression data should be analyzed using geometric rather than arithmetic statistics: increased variance moves arithmetic mean sharply away from median behavior.

used in preference to the arithmetic mean  $\mu$  and arithmetic standard deviation  $\sigma$ . The difference between these statistics can be quite dramatic, depending on the size of the variation, as the mean and standard deviation of a log-normal distribution are, respectively:

$$\mu = e^{\mu_g + \sigma_g^2/2} \quad (7)$$

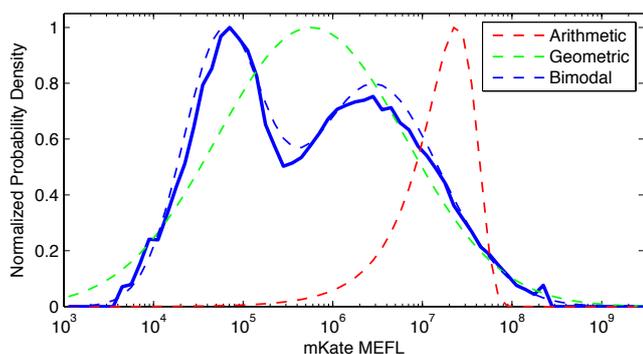
$$\sigma = e^{\sigma_g^2 + 2\mu_g} \cdot (e^{\sigma_g^2} - 1) \quad (8)$$

Figure 10 illustrates the magnitude of this impact when dealing with even a simple log-normal distribution: as the geometric standard deviation  $\sigma_g$  rises, misinterpretation of the high outliers causes the arithmetic mean to depart strongly from the peak of the distribution that it aims to capture. Geometric statistics, of course, capture a log-normal distribution precisely.

Geometric statistics should also be used for comparison across replicates or statistics, both for single-cell data (e.g., flow cytometry, microscopy) and for population data (e.g., plate readers, mass spectrometry). The reason is that the log-normal distribution model implies that perturbations in the state of a cell are likely to result in a log-normal distribution of differences between replicates, rather than a normal distribution, as is commonly assumed. Again, if arithmetic statistics are used, they will give too much weight to high outliers, distorting the results.

This becomes even more critical when interpreting data with more complex distributions, such as the sample shown in Figure 11. This shows the distribution of expression from a strong constitutive promoter transiently transfected into HEK293 mammalian cells in [30]. With such a broad distribution, arithmetic statistics bear little resemblance to the observed distribution. Geometric statistics are better, but with the log-normal model we can recognize that this distribution is the product of typical log-normal expression variation with a bimodal distribution of transfection levels: a tight low component from cells where transfection has been effectively unsuccessful and a broader high component from an apparently log-normal distribution of transfection levels.

Finally, geometric distributions may also have implications for the engineering of genetic expression levels and regulation. The value of diversity in microbial populations has been recognized before, for example for parallelizing experiments within a sample (e.g., [31, 32]) or for differentiating behavior in order to better survive environmental stress [33]. With the log-normal model, we can go further and exploit the fact that the higher the variance of a log-normal distribution, the more its integral is dominated by the high tail. As a result, when synthesizing a chemical product from a population of cells, a population with the same median and a higher degree of variance will outperform one with a lower degree of variance. In optimizing chemical production from engineered biological organisms, then, in many cases it may be valuable to deliberately increase cell-cell variation rather than to attempt to control it. Such



**Fig. 11:** Log-normal model enables better statistical interpretation of data, as in this comparison of normal (red), log-normal (green), and bimodal log-normal (blue) distribution fits to a sample from [30].

strategies may be applicable to many other circumstances as well, in any application where the extremes can dominate the performance of the aggregate.

## 5 Summary and Discussion

This paper has presented a foundational model of the origins of cell-to-cell variation in gene expression, mathematically connecting the well-known complexity of the underlying biochemical mechanisms to the common observation of broad log-normal population distributions. The log-normal distribution model presented accounts for the behavior of strong transcriptional repressors, which cannot be explained by prior models based on stochastic bursting, while simultaneously remaining compatible with the results of those prior models. The results presented here imply that, in general, gene expression data should be interpreted with geometric rather than arithmetic statistics. Furthermore, in engineering biological systems, in some cases it may be counterintuitively advantageous to embrace and amplify cell-to-cell variation rather than attempting to control it.

Looking to the future, while this paper has focused on genetic expression, and its examples have been only of simple protein expression, the results should be more generally applicable. Expression of nucleic acid products, such as gRNA and miRNA, should be susceptible to the same analysis implying a baseline log-normal distribution, though the degree of variation is likely to be smaller when there are less mechanisms involved. Complementarily, post-processing stages such as glycosylation, splicing, or cleavage are also likely to remain log-normal but with increased variation from interaction with additional mechanisms. Log-normal convergence may also be applicable to a wide variety of other complex biochemical processes structured as catalytic cascades. Finally, it may be worth noting that in this case the complexity of biological systems results in simplicity rather than intractability, and that taking such aggregate-based perspectives may turn out to be valuable in dealing with other instances of biological complexity as well.

## Acknowledgments

This work was supported in part by National Science Foundation Expeditions in Computing Program Award #1522074 as part of the Living Computing Project. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

## 6 References

1 Tsimring, L.S.: ‘Noise in biology’, *Reports on Progress in Physics*, 2014, **77**, (2), pp. 026601

2 Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., et al.: ‘Quantifying *e. coli* proteome and transcriptome with single-molecule sensitivity in single cells’, *science*, 2010, **329**, (5991), pp. 533–538

3 Swain, P.S., Elowitz, M.B., Siggia, E.D.: ‘Intrinsic and extrinsic contributions to stochasticity in gene expression’, *Proceedings of the National Academy of Sciences*, 2002, **99**, (20), pp. 12795–12800

4 Hilfinger, A., Chen, M., Paulsson, J.: ‘Using temporal correlations and full distributions to separate intrinsic and extrinsic fluctuations in biological systems’, *Physical review letters*, 2012, **109**, (24), pp. 248104

5 Cai, L., Friedman, N., Xie, X.S.: ‘Stochastic protein expression in individual cells at the single molecule level’, *Nature*, 2006, **440**, (7082), pp. 358–362

6 Abbondanzieri, E.A., Greenleaf, W.J., Shaevitz, J.W., Landick, R., Block, S.M.: ‘Direct observation of base-pair stepping by rna polymerase’, *Nature*, 2005, **438**, (7067), pp. 460–465

7 Aitken, C.E., Petrov, A., Puglisi, J.D.: ‘Single ribosome dynamics and the mechanism of translation’, *Annual review of biophysics*, 2010, **39**, pp. 491–513

8 Friedman, N., Cai, L., Xie, X.S.: ‘Linking stochastic dynamics to population distribution: an analytical framework of gene expression’, *Physical review letters*, 2006, **97**, (16), pp. 168302

9 Shahrezaei, V., Ollivier, J.F., Swain, P.S.: ‘Colored extrinsic fluctuations and stochastic gene expression’, *Molecular systems biology*, 2008, **4**, (1), pp. 196

10 Hilfinger, A., Paulsson, J.: ‘Separating intrinsic from extrinsic fluctuations in dynamic biological systems’, *Proceedings of the National Academy of Sciences*, 2011, **108**, (29), pp. 12167–12172

11 Bowsher, C.G., Swain, P.S.: ‘Identifying sources of variation and the flow of information in biochemical networks’, *Proceedings of the National Academy of Sciences*, 2012, **109**, (20), pp. E1320–E1328

12 Lim, Y.R., Kim, J.H., Park, S.J., Yang, G.S., Song, S., Chang, S.K., et al.: ‘Quantitative understanding of probabilistic behavior of living cells operated by vibrant intracellular networks’, *Physical Review X*, 2015, **5**, (3), pp. 031014

13 Shahrezaei, V., Swain, P.S.: ‘Analytical distributions for stochastic gene expression’, *Proceedings of the National Academy of Sciences*, 2008, **105**, (45), pp. 17256–17261

14 Pedraza, J.M., Paulsson, J.: ‘Effects of molecular memory and bursting on fluctuations in gene expression’, *Science*, 2008, **319**, (5861), pp. 339–343

15 Munsky, B., Neuert, G., van Oudenaarden, A.: ‘Using gene expression noise to understand gene regulation’, *Science*, 2012, **336**, (6078), pp. 183–187

16 Kumar, N., Singh, A., Kulkarni, R.V.: ‘Transcriptional bursting in gene expression: analytical results for general stochastic models’, *PLoS Comput Biol*, 2015, **11**, (10), pp. e1004292

17 Elowitz, M., Leibler, S.: ‘A synthetic oscillatory network of transcriptional regulators’, *Nature*, 2000, **403**, (6767), pp. 335–338

18 Weiss, R.: ‘Cellular Computation and Communications using Engineered Genetic Regulatory Networks’ [PhD].

MIT. Cambridge, MA, USA, 2001

- 19 Nielsen, A.A., Der, B.S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E.A., et al.: 'Genetic circuit design automation', *Science*, 2016, **352**, (6281), pp. aac7341
- 20 Iverson, S.V., Haddock, T.L., Beal, J., Densmore, D.M.: 'Cidar moclo: improved moclo assembly standard and new e. coli part library enable rapid combinatorial design for synthetic and traditional biology', *ACS synthetic biology*, 2015, **5**, (1), pp. 99–103
- 21 Carr, S.B., Beal, J., Densmore, D.M.: 'Reducing dna context dependence in bacterial promoters', *PLOS ONE*, 2017, **12**, (4), pp. e0176013
- 22 Li, Y., Jiang, Y., Chen, H., Liao, W., Li, Z., Weiss, R., et al.: 'Modular construction of mammalian gene circuits using tale transcriptional repressors', *Nature chemical biology*, 2015, **11**, (3), pp. 207–213
- 23 Bonnet, J., Yin, P., Ortiz, M.E., Subsoontorn, P., Endy, D.: 'Amplifying genetic logic gates', *Science*, 2013, **340**, (6132), pp. 599–603
- 24 Dumonceaux, R., Antle, C.E.: 'Discrimination between the log-normal and the weibull distributions', *Technometrics*, 1973, **15**, (4), pp. 923–926
- 25 Bain, L.J., Engelhardt, M.: 'Probability of correct selection of weibull versus gamma based on likelihood ratio', *Communications in statistics-theory and methods*, 1980, **9**, (4), pp. 375–381
- 26 Wiens, B.L.: 'When log-normal and gamma models give different results: a case study', *The American Statistician*, 1999, **53**, (2), pp. 89–93
- 27 Dick, E.: 'Beyond ?lognormal versus gamma?: discrimination among error distributions for generalized linear models', *Fisheries Research*, 2004, **70**, (2), pp. 351–366
- 28 Bar.Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., et al.: 'Noise in protein expression scales with natural protein abundance', *Nature genetics*, 2006, **38**, (6), pp. 636–643
- 29 Del Vecchio, D., Ninfa, A.J., Sontag, E.D.: 'Modular cell biology: Retroactivity and insulation', *Molecular Systems Biology*, 2008, **4**:161
- 30 Davidsohn, N., Beal, J., Kiani, S., Adler, A., Yaman, F., Li, Y., et al.: 'Accurate predictions of genetic circuit behavior from part characterization and modular composition', *ACS synthetic biology*, 2014, **4**, (6), pp. 673–681
- 31 Zhang, C., Tsoi, R., You, L.: 'Addressing biological uncertainties in engineering gene circuits', *Integrative Biology*, 2016, **8**, (4), pp. 456–464
- 32 Beal, J., Weiss, R., Yaman, F., Davidsohn, N., Adler, A. 'A method for fast, high-precision characterization of synthetic biology devices'. (MIT, 2012. MIT-CSAIL-TR-2012-008. technical Report: MIT-CSAIL-TR-2012-008 <http://hdl.handle.net/1721.1/69973>
- 33 Martins, B.M., Locke, J.C.: 'Microbial individuality: how single-cell heterogeneity enables population level strategies', *Current opinion in microbiology*, 2015, **24**, pp. 104–112