Mathematical Foundations of Variation in Gene Expression

*Jacob Beal**

**Raytheon BBN Technologies, Cambridge, MA, USA*

Keywords: gene expression, variation, lognormal distribution

1 Introduction

A key challenge in engineering biological organisms is the high degree of cell-to-cell variation commonly observed in gene expression. The inherently discrete and stochastic nature of the chemical reactions that underly gene expression has been proposed as an explanation for the highly asymmetric distributions that are frequently observed [1], with bursts of expression leading to a Gamma distribution. While this may explain the behaviour of systems with very low expression, it is insufficient to account for the high degree of cell-to-cell variation that is typically still observed even with strong expression (e.g., more than 2-fold standard deviation with a mean of many millions of molecules in [2]). In essence, with strong expression there are typically so many molecules involved that the law of large numbers will generally render the impact of chemical stochasticity largely insignificant.

Stochasticity, however, is only one potential source of variation in observed gene expression levels from cell to cell. A typically much stronger source of variation is the indisputable fact that cells have at least small variations in their state (size, health, available resources, etc.). These small variations are then amplified by composition of chemical reactions to produce a broad log-normal distribution of expression levels at all levels of expression.

2 Log-Normal Variation in Gene Expression

To understand the impact of cell state on gene expression levels, we first need to abandon typical abstractions used for modelling gene expression. Typically, models focus on only the few parameters needed to model differences in observed gene expression levels, such as promoter strength and transcription factor binding and unbinding. In fact, of course, the processes of transcription and translation, as well as their regulation, are fantastically chemically complex processes involving transcriptional and translational machineries, nucleotides, tRNA, ATP, etc. These details are typically abstracted away, however, since they are supporting cellular machinery that is not actively being engineered. If we were to write out the full chemical equation for gene expression, it would thus have a rate term that looks something like:

$$P_{\text{expression}} = P_1 P_2 P_3 P_4 \ldots P_k \qquad (1)$$

If each rate term has a small degree of variation, then the overall rate distribution is a product of their distributions.

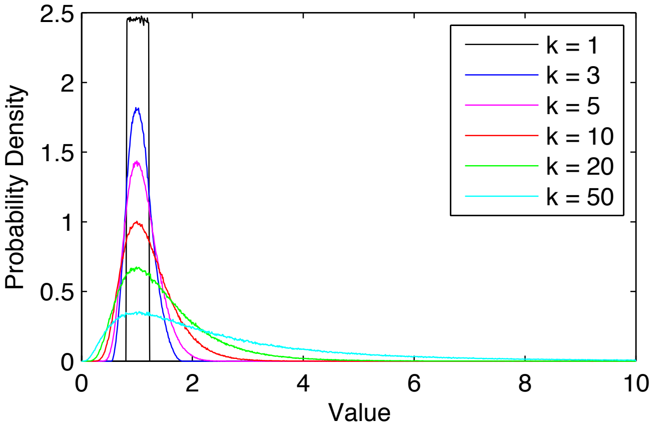

Figure 1: Products of distributions converge to log-normal

The same central limit theorem that shows that sums of independent random variables converge to a normal distribution also shows that that their products converge to a log-normal distribution (multiplication being equivalent to addition on a log scale). Moreover, even a relatively small number of such product terms can produce a quite broad log-normal distribution, as illustrated in Figure 1.

3 Implications of Log-Normal Distribution

The immediate implication of the convergence to log-normal distributions is that we should expect to see log-normal distributions in gene expression (as we do), and in fact in any cellular process dominated by complex chemical reactions. In addition, this may have implications for the engineering of gene regulation networks: the higher the variance of a log-normal distribution, the more its integral is dominated by the high tail: thus, when optimizing products from populations of cells, in many cases it may be valuable to deliberately increase cell-cell variation rather than to attempt to control it.

Acknowledgements

This work is partially supported by NSF Award #1522074.

References

[1] N. Friedman, L. Cai, and X.S. Xie, "Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression," *Physical Review Letters,* **97,** 168302 (2006).

[2] J. Beal, et al. "Model-Driven Engineering in Gene Expression from RNA Replicons," *ACS Syn. Bio.*, **4(1),** pp. 48-56 (2015).