Quantitatively Characterizing Genetic Behavior from Stochasticity in Gene Expression

Marc S. Sherman[1] and Barak A. Cohen[1]

***Short Abstract* — Predicting the behavior of a gene from its component parts is a key challenge in synthetic biology. The mathematical models that enable investigators to predict gene expression output intrinsically depend on the rate constants that underlie that gene's expression. We developed a computational approach that allows *in vivo* estimation of these rate constants from stochasticity in gene expression. We find that the steady state distribution of protein expression often contains enough information to infer transcriptional, translational, and degradational rate constants, or their ratios. We expect our approach to enable a more mechanistic understanding of the relationship between sequence and expression.**

***Keywords* — stochasticity in gene expression, synthetic biology**

I. Background

A central challenge in synthetic biology is to engineer novel genes with a desired behavior. Strikingly however, the projects and databases that collect and catalog genetic parts characterize these components descriptively rather than quantitatively [1,2]. It is therefore unsurprising that the most successful synthetic biology ventures—for example, in optimizing biofuel production [3]—favor random mutagenesis approaches over rational design.

Rational gene design requires investigators to write down a model that takes into account the fundamental processes that drive gene transcription, translation, and degradation of RNA and protein [4]. Each process is described by a set of rate constants. Measuring the relevant rate constants that govern a particular promoter, untranslated region, or other genetic element would enable *de novo* construction of genes with a desired behavior.

Stochasticity in RNA and protein production appears to contain mechanistic information about each process. The rates of the RNA transcription, protein translation, and RNA and protein degradation are encoded in the shape of a protein expression distribution [5]. However, extracting that information from the shape of protein distributions has proven challenging for two reasons. First, analytical solutions relating protein distribution shape to the underlying rate constants make specific assumptions about the rate constant regimes a gene must operate in. The validity of these assumptions is difficult to establish *a priori*, and often involves assuming something about the very quantities being measured. Second, the gold-standard and assumption-free method for simulating stochasticity in gene expression is numerical and inefficient, rendering it unsuitable for parameter-estimation.

II. An Efficient and Assumption-free Approach

To this end, we investigated a moments-based approach for estimating the fundamental gene expression rate constants from the shape of protein distributions. The first and second moments of the expression distribution have previously been used to infer properties of gene expression [5]. Here we investigated using the first four moments.

A. Four moments completely capture distribution shape

We were able to analytically solve the random-telegraph model of gene expression for its first four moments. We find that four moments is enough to exactly capture a distribution's shape, as measured against reference distributions generated by Gillespie-algorithm simulation.

B. Distribution shape informs on rate constants

99% of distributions contain enough information to infer at least one rate constant or rate constant ratio. On average we can extract 3.5 rate constants or ratios per distribution.

III. Conclusions

Our framework allows investigators to infer the molecular rate constants that govern gene expression from the shape of protein distributions. We expect quantitative characterization of genetic elements in this way will enable rational design of synthetic genes.

References

[1] BioBricks Foundation web site, http://biobricks.org/
[2] Registry of Standard Biological Parts web site, http://partsregistry.org
[3] Sticklen MB (2008). Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nature Reviews Genetics* **9**, 433-443.
[4] Sherman MS and Cohen BA. (2012). Thermodynamic State Ensemble Models of *cis*-Regulation. *PLoS Computational Biology*, **8** (3).
[5] Friedman N, Cai, L, and Xie X (2006). Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Physical Review Letters*, **97** (16), 1–4.

[]Acknowledgements: This work was funded by NIH grant 5T32HG000045-13.

[1] Center for Genome Sciences, Department of Genetics, Washington University in St. Louis, St. Louis, MO, United States. E-mail: marc.sherman@wustl.edu and cohen@genetics.wustl.edu