OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

A Computational Framework for Analyzing Stochasticity in Gene Expression

Marc S. Sherman[1,2], **Barak A. Cohen**[2]*

1 Computational and Molecular Biophysics, Washington University in St. Louis, St. Louis, Missouri, United States of America, **2** Center for Genome Sciences, Department of Genetics, Washington University in St. Louis, St. Louis, Missouri, United States of America

Abstract

Stochastic fluctuations in gene expression give rise to distributions of protein levels across cell populations. Despite a mounting number of theoretical models explaining stochasticity in protein expression, we lack a robust, efficient, assumption-free approach for inferring the molecular mechanisms that underlie the shape of protein distributions. Here we propose a method for inferring sets of biochemical rate constants that govern chromatin modification, transcription, translation, and RNA and protein degradation from stochasticity in protein expression. We asked whether the rates of these underlying processes can be estimated accurately from protein expression distributions, in the absence of any limiting assumptions. To do this, we (1) derived analytical solutions for the first four moments of the protein distribution, (2) found that these four moments completely capture the shape of protein distributions, and (3) developed an efficient algorithm for inferring gene expression rate constants from the moments of protein distributions. Using this algorithm we find that most protein distributions are consistent with a large number of different biochemical rate constant sets. Despite this degeneracy, the solution space of rate constants almost always informs on underlying mechanism. For example, we distinguish between regimes where transcriptional bursting occurs from regimes reflecting constitutive transcript production. Our method agrees with the current standard approach, and in the restrictive regime where the standard method operates, also identifies rate constants not previously obtainable. Even without making any assumptions we obtain estimates of individual biochemical rate constants, or meaningful ratios of rate constants, in 91% of tested cases. In some cases our method identified all of the underlying rate constants. The framework developed here will be a powerful tool for deducing the contributions of particular molecular mechanisms to specific patterns of gene expression.

Citation: Sherman MS, Cohen BA (2014) A Computational Framework for Analyzing Stochasticity in Gene Expression. PLoS Comput Biol 10(5): e1003596. doi:10.1371/journal.pcbi.1003596

Editor: Alexandre V. Morozov, Rutgers University, United States of America

Received November 22, 2013; **Accepted** March 17, 2014; **Published** May 8, 2014

Funding: This work was funded by the NIH Genome Analysis Training Program grant 5T32HG000045-13 (http://www.nih.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cohen@genetics.wustl.edu

Introduction

Stochasticity in transcription and translation produces fluctuations in both RNA [1–4], and protein [5–16]. On a population level these fluctuations manifest as a distribution of RNA and protein counts across cells. Both protein and RNA distributions are thought to contain information about the molecular processes governing transcription and translation, though how much information is unclear [17]. Learning the mechanistic details of a gene's expression from its stochastic signature requires (1) a method to measure cell-to-cell expression variability experimentally, (2) an explanatory stochastic model that simulates this variability *in silico*, and (3) a method for fitting the model parameters from experimental data.

There are consensus methods for accomplishing tasks (1) and (2). Investigators routinely measure protein expression stochasticity by recording reporter gene fluctuations with flow cytometry [5,11,14,16,18] or microscopy [1,3,8,19–21]. The result is a protein count distribution that reflects cell-to-cell variation in gene expression. To simulate this cell-to-cell variation *in silico*, investigators developed a stochastic model of gene expression (Fig. 1), which has proven to be an effective abstraction of the central dogma [1,3,5,8,10,14,20,22]. This model is parameterized by six central dogma rate constants (CDRCs) that govern a gene's ON (t_{on}) and OFF (t_{off}) transitions, transcription from the active state (k_m), translation of RNAs (k_p), and degradation of RNA (d_m) and protein (d_p). With a specific set of CDRCs the gene expression model depicted in Fig. 1 can be simulated with the Gillespie algorithm [23] to produce the corresponding protein count distribution [18,24–29].

Task (3), fitting the model of stochastic gene expression to protein distributions, has no general solution. One possible approach would be to test candidate CDRC sets by Gillespie-simulating their corresponding distributions until a CDRC set that best approximates an experimentally measured distribution is identified. Several investigators have shown the utility of this approach, however, the systems being modeled comprise RNA expression only, where the number of molecules is low [4,30]. The well-documented inefficiency of the Gillespie algorithm at even modestly high molecule counts [31–36] renders this approach untenable for parameter estimation of protein distributions, where the average protein expression of single genes exceeds 12000 proteins per cell [37].

An alternative approach is to analytically solve for the shape of the protein distribution as a function of the CDRCs. This approach has yielded some elegant solutions; however, (1) the

Author Summary

Proteins, the molecular machines encoded by our genes, serve essential roles in every living cell. Investigators were therefore surprised to find widely variable levels of a particular protein within populations of genetically identical cells. This variation in protein level, called stochasticity, arises from the chemical nature of the processes that underlie protein production. The "central dogma" of biology dictates that the DNA encoding a particular gene transmits information via RNA to molecular factories called ribosomes in order to create proteins. Each step in transcription and translation introduces some variation, or stochasticity, into the production of the protein. In the current work, we tackled how one might learn more about the machinery responsible for creating proteins by the character of the stochasticity in the central dogma process. We find that many different mechanisms can explain any given stochastic protein signature. Even though there were many explanations for any particular pattern of stochasticity, the set of explanations still inform on how a given gene creates its protein. Our mathematical and computational framework will permit others to better understand how the machinery that expresses genes works. This, in turn, will enable investigators to better predict how a given mutation is likely to affect gene expression.

analytical solutions generally involve hypergeometric or gamma functions, themselves ill-suited for parameter estimation, and more problematically (2) each solution makes specific assumptions about a gene's expression [17,22,25,28,38–40]. Some methods are valid only when protein count is high [25,39], while most require RNA degradation to be orders of magnitude faster than protein degradation [22,25,28,38,40]. Several only apply when genes do not have an OFF state [9,25] or the ON-OFF transitions are rapid [39]. Three approaches do not model RNA fluctuations [17,38,40]. The result is a fundamental limitation in the applicability of these methods, since it is usually unknown beforehand whether a particular gene conforms to the basic assumptions of these methods.

Even an assumption-free analytical solution to the protein distribution may not adequately solve the fitting problem. Munsky *et al* determined that degeneracy in the solution space of CDRCs means that ratios of the CDRCs, but not the CDRCs themselves, can be extracted from steady state protein distributions, and further recommends temporal measurements for pinpointing individual CDRC values [41]. Similarly, Ingram, Stumpf and Stark demonstrated that many different combinations of CDRCs can give rise to the same translational burst distribution [33]. The authors suggested supplementing the burst distribution with the steady state protein distribution, but were hindered in part by the inefficiency of the Gillespie algorithm [33]. Any method for determining the CDRCs that underlie protein distributions must account for the expected degeneracy of CDRCs that can produce a given protein distribution. Given this degeneracy, it is an open question how much information any given protein distribution contains about the CDRCs that underlie its shape. Mechanistic information about the processes that produce an observed protein distribution will most likely come from analyzing ensembles of solutions that fit a particular protein distribution.

In this work we address how much information is contained in protein distributions. The principal result is an assumption-free solution to the steady state protein distribution. Our approach consists of two parts: (1) analytical solutions to the first four moments of the protein distribution, and (2) an efficient, exhaustive fitting algorithm that returns ensembles of CDRC sets that map to a particular set of moments. The main power of our approach is that it returns all CDRC solution sets that are consistent with a given protein distribution. These solution set ensembles were always informative. Even in cases where we observe degeneracy in both the individual CDRCs and their ratios, the set of solutions provides mechanistic information about gene behavior, for example distinguishing between genes undergoing transcription bursts from those that transcribe constitutively.

We compared our method directly to the Friedman analytical solution [25]. While the model solved analytically by Friedman *et al* does not incorporate an OFF state like other solutions [22,28,29], only the Friedman result enables rapid estimation of the causative CDRC ratios. Investigators recently took advantage of this result to infer k_m/d_p and k_p/d_m from protein distributions measured by flow cytometry [42–44]. We found that in the restrictive regime where the Friedman assumptions hold, our method not only identifies both Friedman ratios, but often obtains estimates for quantities inaccessible to the Friedman method, including the average RNA count, k_p/d_p, and in some cases even individual CDRCs.

Without assuming any regime, we find that supplementing our fitting algorithm with experimentally determined limits for each CDRC permits some CDRCs to be obtained individually. More

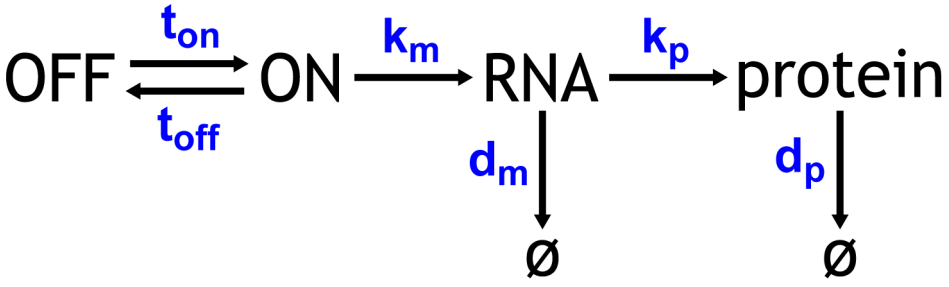

Figure 1. Central dogma model of gene expression with gene ON and OFF states.
doi:10.1371/journal.pcbi.1003596.g001

often, CDRC ratios are well conserved among the ensemble of solution outcomes. We can identify at least one CDRC in 40% of regimes tested, and at least one CDRC ratio in 91%. Our methodology provides a general, assumption-free framework for extracting information from protein distributions. We anticipate that our approach will be a powerful tool for quantitatively characterizing the molecular machinery that underlies gene expression.

Results

Computing moments from Central Dogma Rate Constants

The stochastic behavior of the central dogma model (Fig. 1) is exactly described by the chemical master equation (CME), Eq. 1 & 2.

$$\begin{aligned}\frac{dP_0(m,q)}{dt} &= t_{off}P_1(m,q)+d_m(m+1)P_0(m+1,q)\\ &+k_pmP_0(m,q-1)+d_p(q+1)P_0(m,q+1)\\ &-(md_m+mk_p+d_pq+t_{on})P_0(m,q)\end{aligned} \quad (1)$$

$$\begin{aligned}\frac{dP_1(m,q)}{dt} &= t_{on}P_0(m,q)+d_m(m+1)P_1(m+1,q)\\ &+k_pmP_1(m,q-1)+d_p(q+1)P_1(m,q+1)\\ &+k_mP_1(m-1,q)\\ &-(k_m+md_m+mk_p+d_pq+t_{off})P_1(m,q)\end{aligned} \quad (2)$$

Here we separated the CME into two equations that describe the system in its ON (P_1) and OFF (P_0) states. $P_0(m,q)+P_1(m,q)$ is the probability that a single cell will contain m RNAs and q proteins. Thus, the probability space is a joint distribution for all possible combinations of $[0,\infty)$ RNAs and $[0,\infty)$ proteins for each promoter state.

The challenge is to use the CME to determine the CDRCs that underlie an experimentally determined protein distribution. With the CME model and a set of input CDRC parameters, a protein distribution is determined by either Gillespie-simulating the CME, or by solving the CME analytically. However, as discussed above, numerical simulations are prohibitively expensive, and the analytical solutions are valid only under restrictive assumptions.

To address this challenge we chose a hybrid approach that takes advantage of the moments of a protein distribution as descriptors of the protein distribution's shape. A significant body of work establishes the relationship between the CDRCs and the first two moments, mean and variance [5,9,10,13,15,25,26,28,39,45]. However the mean and variance alone do not sufficiently characterize the shape of an experimentally measured protein distribution [24,41,46]. We therefore extended previous work by solving for the third (skewness) and fourth (kurtosis) steady state, central moments of protein distributions as functions of the six CDRCs.

To analytically derive protein skewness and kurtosis, we adopted the approach of Sánchez and Kondev in which the i-th moment is computed by multiplying both Eq. 1 and 2 by q^i and summing over all possible q [45]. Instead of assuming instantaneous geometric distributions of protein, however, we explicitly modeled protein translation. When translation is treated in this way, the j-th protein moment $\langle q^j\rangle$ depends on all RNA-protein covariate moments whose sum equals j. The result is the following expansion for the i-th and j-th moments and covariant moments between the RNA and protein distributions.

$$\begin{aligned}\frac{d\langle m_0^iq_0^j\rangle}{dt} &= t_{off}\langle m_1^iq_1^j\rangle-t_{on}\langle m_0^iq_0^j\rangle\\ &+d_m\langle m_0(m_0-1)^iq_0^j+m_0^{i+1}q_0^j\rangle\\ &+k_p\langle m_0^{i+1}(q_0+1)^j-m_0^{i+1}q_0^j\rangle\\ &+d_p\langle m_0^i(q_0(q_0-1)^j-q_0^{j+1})\rangle\end{aligned} \quad (3)$$

$$\begin{aligned}\frac{d\langle m_1^iq_1^j\rangle}{dt} &= t_{on}\langle m_0^iq_0^j\rangle-t_{off}\langle m_1^iq_1^j\rangle\\ &+k_m\langle(m_1+1)^iq_1^j-m_1^iq_1^j\rangle\\ &+d_m\langle m_1(m_1-1)^iq_1^j+m_1^{i+1}q_1^j\rangle\\ &+k_p\langle m_1^{i+1}(q_1+1)^j-m_1^{i+1}q_1^j\rangle\\ &+d_p\langle m_1^i(q_1(q_1-1)^j-q_1^{j+1})\rangle\end{aligned} \quad (4)$$

The result is a set of 28 linear ODE partial moment equations whose time derivatives we set to zero and solved simultaneously to obtain protein skewness and kurtosis. A complete derivation is provided in the Supplement.

The equilibrium equations for mean (Eq. 5) and variance (Eq. S66) agree exactly with previously published analytical results [47]. Equations for skewness and kurtosis are provided as implementations in both MATLAB (2012a, The MathWorks, Natick, MA) and C++ (Data S2). We checked these equations against Gillespie-simulated protein distributions generated using diverse sets of CDRCs. We then asked whether the sample moments of these simulated protein distributions agreed with our analytical moments. In all cases we found that the Gillespie simulations converged on our analytical solutions. An example convergence is shown in Fig. S1. Thus, we can express the first four moments of a protein distribution as functions of the CDRCs.

Finding CDRC sets from moments

The results so far are exact equations relating the CDRCs to the moments of protein distributions. To determine the CDRCs that underlie an experimentally measured protein distribution we must solve the inverse problem; we must implement a fitting method that takes the moments of a protein distribution as input, and returns the best estimate of the causative CDRCs. When considering estimation procedures, we took under consideration a previous result that steady state protein distributions alone cannot contain enough information to directly fit CDRCs [41]. Corroborating that observation, Ingram *et al* found that even constraining on both translational burst size and one of the degradation rates, d_m or d_p, revealed degeneracy in the remaining unknown CDRCs [33]. Since standard fitting approaches behave erratically when solution spaces comprise flat or valley shaped minima, we were motivated to consider alternatives more robust to parameter space degeneracy.

The approach we took is to exhaustively test all of six-dimensional CDRC space within the physiological ranges for each CDRC. Physiological ranges for each CDRC were drawn from various genome-wide analyses of *S. cerevisiae*. The degradation and synthesis rates' ranges were set to be $d_p^{(min)}=.00045s^{-1}$, $d_p^{(max)}=1s^{-1}$, $d_m^{(min)}=.000628s^{-1}$, $d_m^{(max)}=3.1s^{-1}$, $k_m^{(min)}=.01s^{-1}$, $k_m^{(max)}=200s^{-1}$, $k_p^{(min)}=.5s^{-1}$, and

$k_p^{(max)} = 55s^{-1}$ [11,48]. From the few *in vivo* measurements of t_{on} and t_{off} [8,49] we extrapolated several orders of magnitude, setting these ranges to be $t_{on}^{(min)}, t_{off}^{(min)} = .001s^{-1}$, and $t_{on}^{(max)}, t_{off}^{(max)} = 50s^{-1}$.

Our brute force fitting routine gains efficiency by leveraging the fact that the expression for the mean is simple (Eq. 5). Given physiological ranges for each CDRC and an equation for the mean, each choice for the value of one CDRC analytically limits the values of yet-to-be-assigned CDRCs. When the algorithm reaches the sixth and last CDRC, only one value of this CDRC satisfies the mean equation. In this way we reduce the complexity of our CDRC search by at least one dimension. The exact algorithmic and mathematical details of this Analytically Constrained Exhaustive Search (ACES) routine are a central result of this work, and are presented in Materials and Methods and continued in the Supplement. We also provide an ACES implementation in C++ (Data S2).

The ACES algorithm is efficient. If we examine 83 guesses across the range for each CDRC, the ACES algorithm tests 83^6, or 327 billion possible parameter sets, and returns all solutions that satisfy the input moments within some error tolerance. Here we chose a cutoff of $<1\%$, though we envision adjusting the tolerance to be consistent with measured experimental error in the moments. A typical ACES execution takes approximately one minute at a resolution of 83. Testing such a large number of CDRC sets is only possible because (1) calculating the moment objective functions is fast, (2) we check the most efficiently calculated objective function first (variance), and the least efficient objective function (kurtosis) last, thus our most expensive function is called least frequently, and (3) we only test candidate CDRC sets that we already know satisfy the mean moment equation exactly. It is (3) that contributes most substantially to the efficiency of ACES.

Generating a library testing all parameter regimes

With a set of equations in hand that relate the CDRCs to the moments of protein distributions, and an algorithm that uses those equations to fit CDRCs to sample moments, we now have a method for determining the CDRCs that underlie an experimentally measured protein distribution. To evaluate the utility of our method, we tested its ability to recover CDRC values from protein distributions defined by known CDRC sets.

We chose CDRC sets to test keeping in mind that a key drawback with other methods is that they only apply in specific parameter regimes. Our method makes no such assumptions, and so should be applicable in every regime of CDRC values. To test "every regime" in a course-grained manner, we constructed a representative library of CDRC sets. Because each of the six CDRCs can vary over several orders of magnitude, we selected five values spaced evenly in log space across the physiological range for each CDRC (Table S1). Our initial library contains all possible combinations of the 5 values for each of the 6 CDRCs, or $5^6 = 15625$ CDRC sets. We then removed all CDRC sets where the associated mean protein count was less than 17 or greater than 100,000, leaving 8053 CDRC sets. The lower limit was chosen as the lower limit of detection for fluorescent protein molecules, since our method is only applicable when fluorescence is measurable experimentally [50]. The upper limit was chosen arbitrarily to capture $>98\%$ of *Saccharomyces cerevisiae* genes [37].

To assess how well ACES recovers CDRCs across all parameter regimes, we applied ACES to each of the 8053 input CDRC sets. For each CDRC set in the library, we computed the moments of its protein distribution using our analytical solutions, and gave these moments to our ACES algorithm as input. ACES then returned a list of solutions, where each solution is a set of values for each of the six CDRCs (Fig. 2). Every solution returned by ACES produces moments that match the input moments with $<1\%$ relative error for each moment.

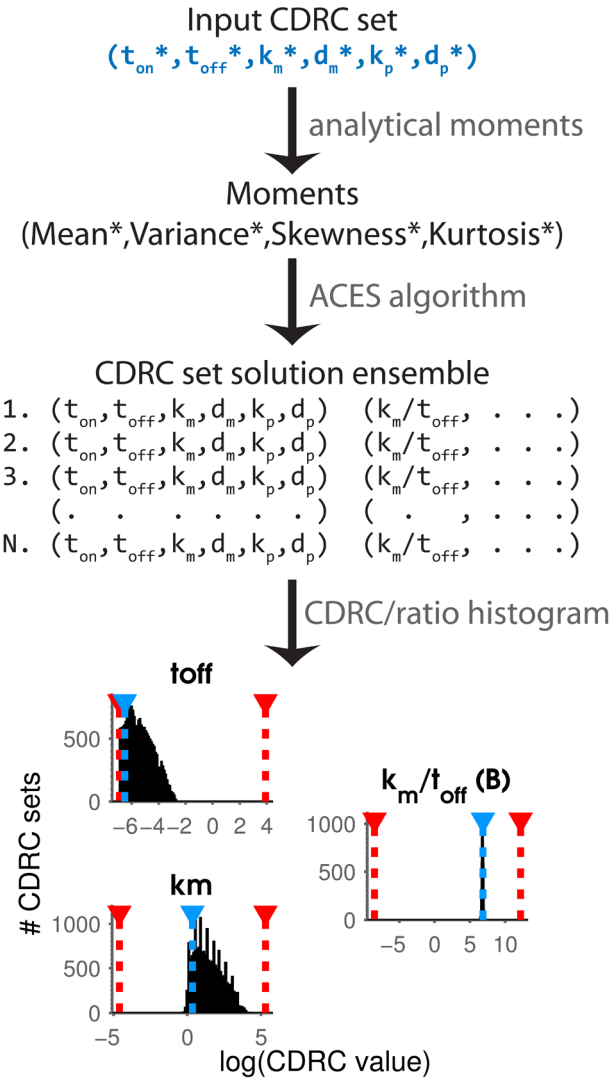

Figure 2. CDRC estimation pipeline. At the top, a library member input CDRC set (blue,*) is used to compute its corresponding moments (*). ACES takes as input only the moments, and returns all CDRC sets that correspond to those moments (the "ensemble"). To visualize this result, a histogram is generated for each CDRC or ratio column-wise. In the resulting histograms, physiological range of the parameter (red, dashed), the input value (blue dashed), and the ensemble solution values (black) are plotted for the tested parameter space (x-axis). On the bottom left, histograms for the CDRCs k_m and t_{off} span half their physiological range, but their ratio is constrained to a single value (right, burst size).
doi:10.1371/journal.pcbi.1003596.g002

The number of solutions returned by the method depended on the number of subdivisions probed by the ACES algorithm. At a resolution of 83 (each CDRC tested at values ranging over 83 divisions across its physiological range), we find that 7555 of 8053 library member inputs result in at least one solution within the tolerated error level. Of the remaining 498, increasing the resolution to 127 resulted in solutions for all but 48 input sets, and took approximately 3 minutes per run at this resolution. The remaining 48 were extremely resistant to increases in resolution, however, simply reversing the order in which ACES tests

parameters allowed us to find solutions for all of these remaining input sets at low resolution (See Supplement, section S3.2).

We find multiple candidate solutions for every input set. The number of solutions linked to any given CDRC set varied widely, with some producing dozens of solutions, and others generating tens of thousands. The median number of output solutions for a CDRC input set was 29286 solutions. This seems like a large number, but when put in the context that each ACES run tests hundreds of billions, or trillions of parameters sets, this number of solutions represents only a tiny fraction of tested CDRC sets. We considered two possible explanations for the multiplicity of solution sets obtained for every input set. (1) A previous analytical result suggests that CDRC ratios but not their individual values are recoverable from the steady state distribution [41]. Thus, every combination of CDRCs conserving a particular ratio will come out of our solution set. Although this explanation likely contributes to the degeneracy we observed, a competing explanation is that (2) four distribution moments inadequately describe the shape of protein distributions. We sought to distinguish between these two possibilities.

Maximally dissimilar CDRC sets with the same moments produce the same distribution. We asked how well moments approximated their corresponding protein count distribution by Gillespie-simulating examples of output CDRC sets, along with their input set, and comparing the resulting distributions. Given the abundance of output sets for every input set, we sought specific cases where the ensemble of output sets was least informative. We chose the 100 worst fit library members, where neither CDRCs nor ratios were well captured, and further identified CDRC output sets within each solution ensemble that were the farthest apart in log-Euclidean space. For example, the following two CDRC sets give rise to almost identical mean, variance, skewness, and kurtosis: (a) $t_{on}=0.003439$, $t_{off}=0.030224$, $k_m=0.11892$, $d_m=0.000628$, $k_p=0.53938$, $d_p=0.39368$, (b) $t_{on}=43.9773$, $t_{off}=45.7037$, $k_m=7.5668$, $d_m=3.1$, $k_p=22.1439$, $d_p=1$. Almost every CDRC value between these two CDRC sets differs by several orders of magnitude. We then Gillespie-simulated both CDRC sets and their associated input set to generate complete protein distributions. We found that the protein distributions for all 100 "maximally distant pairs" matched each other and their input distributions (Figs. S2–S6). An example of a maximally distant CDRC pair and its reference distribution are plotted in Fig. 3A.

To quantify the differences between distributions, we computed the Jensen-Shannon divergence [51] between distributions, which is the symmetrized version of the Kullback-Leibler directed divergence used by Shahrezaei *et al* for similar purposes [22]. Jensen-Shannon divergence (d_{JS}) values have an intuitive interpretation; if P_1 represents one probability distribution, and P_2 the second, $1/d_{JS}(P_1,P_2)$ roughly estimates the number of samples one would have to draw to determine from which distribution (P_1 or P_2) the values were taken [52]. Thus, Jensen-Shannon divergence ranges between 0, indicating that two distributions are identical, to 1, characterizing a pair of distributions as non-overlapping. We found that maximally different CDRC solution sets produce identical protein distributions (Fig. 3B). The median Jensen-Shannon divergence between protein distributions derived from maximally different CDRC solutions sets and the reference distribution is .0043, a number which is very close to the median divergence between Gillespie-simulated replicates of the reference distribution, .0033. This result indicates that independent replicates of one parameter set are essentially indistinguishable from distributions generated from maximally distant parameter sets that share the same moments.

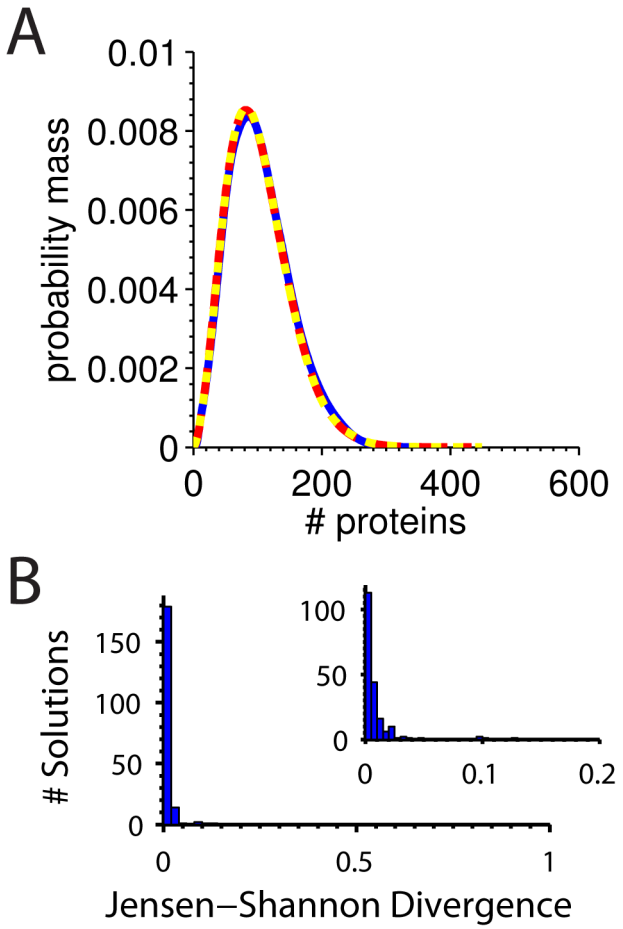

Figure 3. Distributions that correspond to CDRC sets with identical moments. A Two CDRC solution sets maximally different from one another in log-Euclidean space are plotted in (solid) blue ($t_{on}=0.0029963$, $t_{off}=0.023062$, $k_m=0.22529$, $d_m=0.00068275$, $k_p=0.5$, $d_p=0.18675$) and (long-dashed) red ($t_{on}=6.2045$, $t_{off}=50$, $k_m=60.3593$, $d_m=3.1$, $k_p=47.2619$, $dp=1$). These candidate CDRC solution sets map from ACES fitting of library member 6328, plotted in (short-dashed) yellow from CDRC set ($t_{on}=2.7301$, $t_{off}=33.333$, $k_m=133.33$, $d_m=2.0667$, $k_p=13.867$, $dp=0.66667$) **B** Distribution of Jensen-Shannon divergences of maximally-distant CDRC candidate solution sets from their reference CDRC input set. (Inset) Zoom on the x-axis.

doi:10.1371/journal.pcbi.1003596.g003

We then asked whether the 5th sample moment could distinguish between distributions that shared identical first four moments (Supplement). We found that among the maximally distant sets, their fifth moments were indistinguishable. Taken together, these results confirm that the many solutions we obtain from ACES arise solely from CDRC degeneracy and not from the inadequacy of the first four moments to capture distribution shape.

Protein distribution shape informs on molecular mechanism. With a rigorous way of identifying the ensemble of CDRC sets that correspond to a particular protein distribution, we surveyed how well these ensembles inform on the molecular mechanisms that underlie the shape of protein distributions. For each CDRC output set, we computed ratios that represent physically relevant quantities, including: P_{on} (probability that a promoter is active, $t_{on}/(t_{on}+t_{off})$), S (RNA synthesis rate, $P_{on}k_m$), B (burst size, number of RNAs produced per ON duration, k_m/t_{off}), F (burst frequency, frequency of promoter transitions

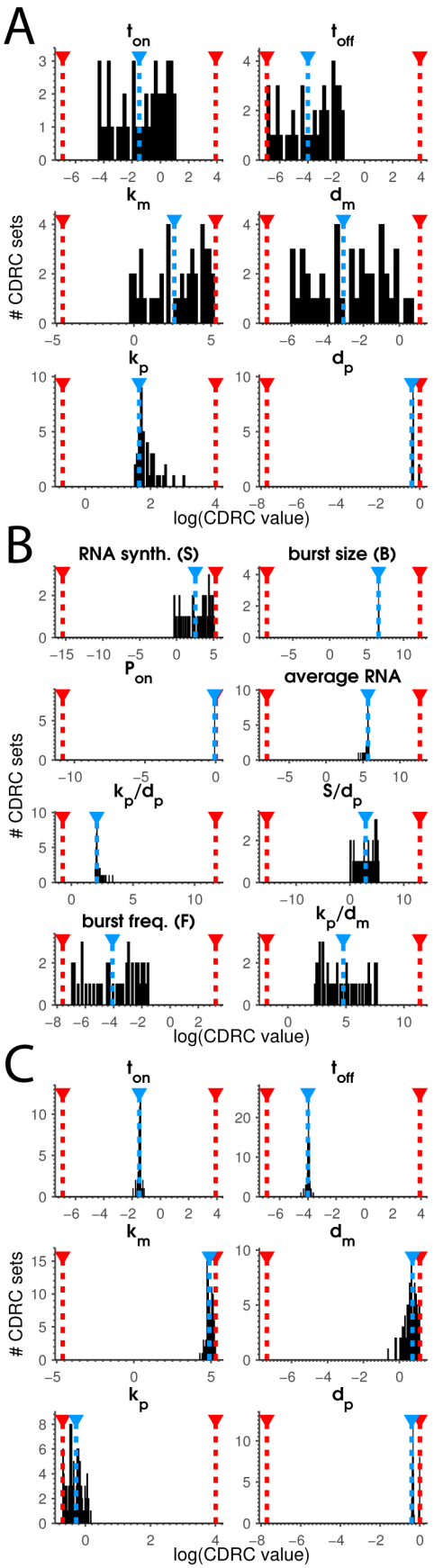

Figure 4. Fitting CDRCs from moments. Plotted is a histogram of CDRC solutions separated by parameter. Blue lines show the value of the input CDRC parameter, and red lines denote each CDRC's minimum and maximum value. **A** CDRC solution sets for fitting library member 3515 with CDRC input values ($t_{on}=.22s^{-1}$,$t_{off}=.018s^{-1}$,$k_m=13.7s^{-1}$, $d_m=.044s^{-1}$,$k_p=5.24s^{-1}$,$d_p=.66s^{-1}$), demonstrate only some parameters are well fit from four moments in this example. **B** Solution CDRC ratios from the same output set in (A) (library member 3515) reveal some CDRC ratios are well identified. **C** CDRC solution sets for fitting library member 3585, with input CDRC values ($t_{on}=.22s^{-1}$, $t_{off}=.018s^{-1}$,$k_m=133.3s^{-1}$,$d_m=2.07s^{-1}$,$k_p=.75s^{-1}$,$d_p=.66s^{-1}$), demonstrate all parameters are well fit from four moments in this example.
doi:10.1371/journal.pcbi.1003596.g004

between the ON and OFF states, $(t_{on}+t_{off})/(t_{on}t_{off})$), $\langle \mathrm{RNA} \rangle$ (average count of RNAs, S/d_m), k_p/d_p (ratio of average protein count to average RNA count), k_p/d_m (translational burst size [25]), and S/d_p (number of RNAs produced per cell cycle when protein degradation is dominated by dilution and there are no ON-OFF transitions [25]).

To develop our intuition, we plotted individual CDRC set output ensembles as histograms (Figs. 2, 4). Consistent with others' predictions [33,41], we found that many different individual CDRC values map to the same distribution, while some CDRC ratios are exactly conserved in the solution ensemble. This is illustrated at the bottom of Fig. 2, where histograms for k_m , t_{off} , and B reveal the range and frequency of each CDRC value in the solution set. While k_m and t_{off} may assume many different values, their ratio (k_m/t_{off}) is constrained across every solution, indicating that the transcriptional burst size is a defining feature of this protein distribution. The solution ensemble for all CDRCs in a different example reveals that some values for each CDRC are not allowed by the distribution shape (Fig. 4). For example, t_{on} is restricted to the middle range of its possible values. We plotted histograms of the CDRC ratios in Fig. 4B from the same input set as in Fig. 4A, revealing that several ratios are exactly inferred from the distribution shape (P_{on} , B, $\langle \mathrm{RNA} \rangle$). Contrary to our expectations, ACES also returned many examples of individual CDRCs themselves being well captured by the distribution shape. For example, Fig. 4A illustrates that d_p is well-estimated. In rare cases every CDRC solution clustered exactly on top of the associated input CDRC value (Fig. 4C). Equally rarely we discovered library members where neither CDRCs nor ratios were reasonably estimated (Fig. S7).

Examining each CDRC, or CDRC ratio ensemble as a histogram was the most informative way of viewing our results, however, this approach was not amenable to analyzing a library of 8053 members. To better grasp what ACES learned about CDRCs or ratios across the library, we developed a metric of fitness computed for each CDRC parameter. We tried many different metrics, and ultimately chose median log distance (MD) as our metric of fitness. To compute MD, we first calculated the log distances of each solution value for each CDRC compared to the known input value. That is, for the i^{th} solution, $\mathrm{log\text{-}dist}_i=|\log(\mathrm{CDRC}_{input})-\log(\mathrm{solution}_i)|$. We then take the median of this list of distances as the MD value. Thus, lower MD values correspond to closer clustering of output values on the true solution. We emphasize that this metric is for survey purposes only–the value of ACES is in producing all possible CDRC solutions consistent with a distribution.

With this purpose in mind, we arbitrarily defined a successful fit as an MD$<$.75, corresponding to the median solution for a particular CDRC or ratio being less than an order of magnitude away from the true (input) parameter value. We find that the MD

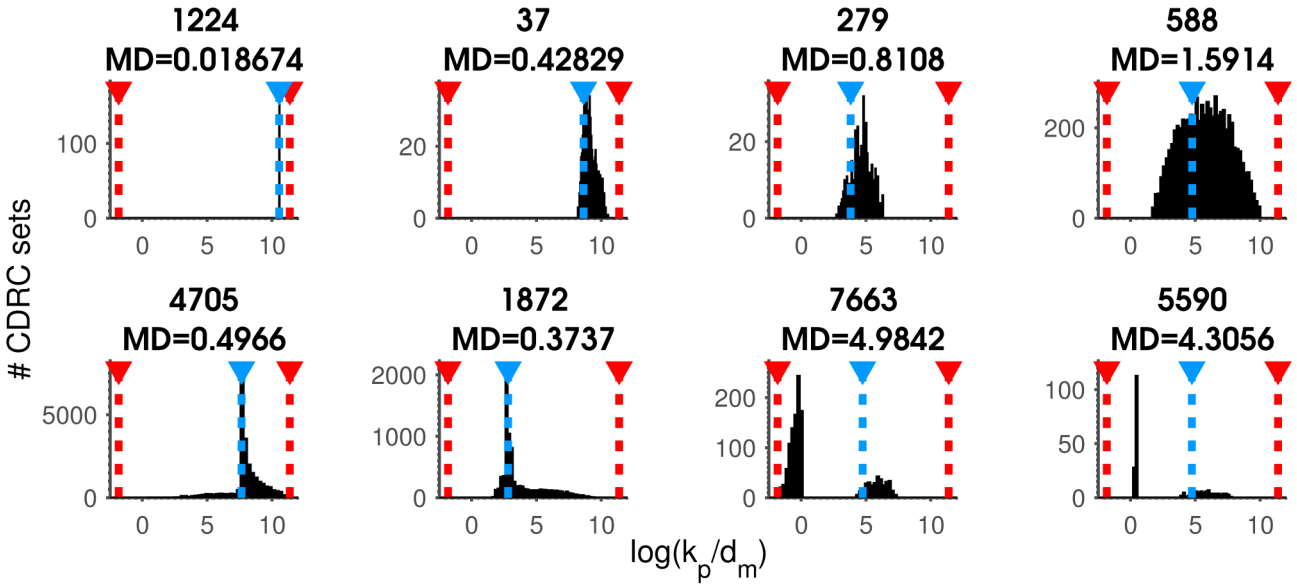

Figure 5. MD as a metric of fitness. Solution histograms for the CDRC ratio k_p/d_m. Titles show the library input set number above, and the corresponding MD for the histogram below. (Top row) Increasing MD values typically correspond to increasing distribution widths. (Bottom row, left 2 panels) Poorly estimated ratios with misleadingly low MD values. (Bottom row, right 2 panels) Informative CDRC ratio histograms with misleadingly high MD values.
doi:10.1371/journal.pcbi.1003596.g005

metric generally aligns with the shape of a solution histogram; very small MD values indicate a spike of solutions immediately on top of the true solution, while increasing MD values correspond smoothly to increasingly wide distributions about the true solution (Fig. 5, top row). Distribution shapes not well captured by the MD metric include cases where the majority of solutions map to the true solution, while a minority of solutions appear elsewhere (Fig. 5, bottom row, left two panels), as well as cases where the CDRC or ratio may only take on two values, but the values are very different (Fig. 5, bottom row, right two panels).

With reasonable confidence in our survey measure, we computed the MD for every CDRC and ratio in each solution output set. The results of this analysis for every library input set are recorded in Data S1, while a summary of the results broken down by CDRC or ratio are presented in column MVSK, Table 1. Overall we can obtain at least one CDRC value or ratio in 91% of the library members. Although only 89 distributions allow inference of all six CDRCs, our analysis revealed 3276 of the CDRC input sets contain at least one correct CDRC fit. The breakdown in CDRCs identified was as follows: 13.1% for t_{on}, 7.1% for t_{off}, 6.8% for k_m, 16.2% for d_m, 21.1% for k_p, 18.2% for d_p. Among ratios, ACES identified P_{on} most frequently, even when the individual values of t_{on} and t_{off} were poorly fit. 2396 input parameter sets correctly identify P_{on} (MD = .25 ± .06) when both t_{on} and t_{off} solution distributions were wide (MD>.75). On average, we fit .82 CDRCs and 2.55 ratios per library member, which we found to be remarkably informative given the overall degeneracy in protein distributions.

CDRC inference improves when degradation rates are known

One way to reduce the complexity of CDRC-space is to experimentally measure some of the CDRCs that underlie a given protein distribution. In the context of this *in silico* study, measuring a CDRC corresponds to giving ACES one or multiple CDRCs at the outset. We analyzed the performance of ACES when given the values of the RNA and protein degradation rates, either separately or together. We chose the degradation rates because they are the easiest CDRCs to measure experimentally [11,48,53].

We re-ran our 8053 library of input sets giving ACES d_m, d_p, or both at the outset. As before, ACES returned solution sets for every input set, though ACES typically returned solutions on the order of a few seconds rather than 1–3 minutes at comparable resolutions.

Knowing d_p or d_m *a priori* improved ACES' fit of the remaining CDRCs and CDRC-ratios. When no CDRCs were known, .8/6 CDRCs and 2.55/8 ratios were fit per CDRC input set on average (column MSVK, Table 1). When d_m was given, 1.78/5 CDRCs and 3.98/8 ratios were fit on average (column MSVK_{dm}). When d_p was given, these numbers are 1.80/5 CDRCs and 3.99/8 ratios (column MSVK_{dp}). Constraining ACES with d_m versus d_p improved estimation of two to three other CDRCs or CDRC ratios. Both degradation rates indiscriminately improved estimation of the remaining unknowns, though estimation of t_{off} improved only modestly, while estimation of the synthesis rate (S) improved the most (columns Δ_{dm}, Δ_{dp}, Table 1).

When ACES was given both degradation rate constants beforehand, we observed a dramatic improvement in CDRC and CDRC-ratio estimation (columns MSVK_{dm-dp}, Δ_{dm-dp}, Table 1). The parameters k_p, S, $\langle\text{RNA}\rangle$, k_p/d_p, S/d_p, and k_p/d_m were essentially always measurable. On average 2.21/4 CDRCs and 5.68/8 ratios were fit in this library. Only t_{off}, and two ratios that depend on t_{off} (B and F) were not significantly improved by providing either or both degradation rate, an observation that will be explored further in a later section. Overall, our results suggest that experimentally determining d_m and d_p will greatly improve the estimation of the remaining CDRCs from experimentally measured protein distributions.

Contribution of skewness and kurtosis to CDRC fitting

Higher moments contain diminishing information about the shape of a probability distribution. To determine if skewness and

Table 1. Summary of CDRC fitting across input CDRC libraries.

	MVSK	$\mathrm{MVSK_{dm}}$	$\mathrm{MVSK_{dp}}$	$\mathrm{MVSK_{dm,dp}}$	MV	MVS	Δ_{dm}	Δ_{dp}	$\Delta_{\mathrm{dm,dp}}$	Δ_{MV}	Δ_{MVS}
t_{on}	1.88	1.52	1.54	0.92	2.28	2.02	−0.36	−0.33	−0.96	0.40	0.14
t_{off}	3.08	2.90	2.83	2.55	4.03	3.28	−0.19	−0.25	−0.53	0.94	0.20
k_m	2.06	1.67	1.67	1.15	2.80	2.30	−0.40	−0.39	−0.92	0.74	0.24
d_m	1.64	-	1.15	-	2.44	1.82	-	-0.49	-	0.80	0.18
k_p	1.16	0.67	0.70	0.32	1.31	1.23	−0.49	−0.46	−0.84	0.15	0.07
d_p	1.58	1.05	-	-	2.47	1.69	−0.53	-	-	0.90	0.11
S	1.58	0.83	0.91	0.32	2.45	1.83	−0.75	−0.68	−1.26	0.86	0.24
B	2.56	2.22	2.32	2.09	3.93	2.93	−0.33	−0.23	−0.47	1.38	0.38
P_{on}	1.22	1.09	1.11	0.89	1.87	1.39	−0.13	−0.11	−0.33	0.65	0.17
$\langle RNA \rangle$	1.29	0.83	0.70	0.32	2.53	1.50	−0.46	−0.59	−0.97	1.24	0.21
k_p/d_p	1.29	0.83	0.70	0.32	2.53	1.50	−0.46	−0.59	−0.97	1.24	0.21
S/d_p	1.26	0.67	0.91	0.32	2.22	1.55	−0.59	−0.35	−0.94	0.96	0.29
F	2.37	2.16	2.14	1.86	2.76	2.49	−0.22	−0.23	−0.51	0.39	0.11
k_p/d_m	1.26	0.67	0.91	0.32	2.22	1.55	−0.59	−0.35	−0.94	0.96	0.29
$\langle \#\mathrm{fit,CDRCs} \rangle$	0.82/6	1.78/5	1.80/5	2.21/4	0.22/6	0.64/6	0.96	0.97	1.39	−0.60	−0.18
$\langle \#\mathrm{fit,ratios} \rangle$	2.55	3.98	3.99	5.68	0.27	1.86	1.43	1.44	3.13	−2.28	−0.69

All columns correspond to average MD fit values across the same CDRC library members, fit under different conditions. MVSK corresponds to fitting on all four moments (mean, variance, skewness, and kurtosis), while MV and MVS correspond to fitting on mean-variance only, and mean-variance-skewness only, respectively. All values are reported in MD. Subscripted d_m or d_p correspond to columns where those CDRCs were given to ACES *a priori*. Δ columns correspond to MD values in columns 2 through 6 minus the MD value for the reference column, MVSK, in column 1. $\langle \#\mathrm{fit,CDRCs} \rangle$ is the average number of CDRCs fit per input set according to our cutoff of MD$<$.75. In the relevant columns, CDRCs given as input are not included in the count of correct CDRCs. $\langle \#\mathrm{fit,ratios} \rangle$ then corresponds to the number of CDRC ratios fit by the same critieria.

doi:10.1371/journal.pcbi.1003596.t001

kurtosis contribute to ACES' fitting of the CDRCs, we performed two experiments on the exact same library of CDRC input sets as above. In the first, we modified ACES to fit only on the mean and the variance (column MV, Table 1), and in the second we fit on the first three moments (column MVS, Table 1).

We found that when ACES fits on mean and variance, on average only .22/6 CDRCs and .27/8 CDRC ratios were extracted per input set. When ACES was given the mean, variance, and skewness, it returned on average .64 CDRCs and 1.86 CDRC ratios for each input set. Thus, skewness contributed approximately .42 additional CDRCs and 1.59 CDRC ratios, while kurtosis added a modest but significant .18 CDRCs and .69 CDRC ratios per input. We conclude that skewness and kurtosis significantly contribute to the parameter fitting process, effectively quadrupling the number of CDRCs determinable per set, and increasing the number of CDRC ratios determinable by just under ten-fold.

Distinguishing between bursting and non-bursting genes

Because a major challenge in the field is to identify the *cis*- and *trans*-acting machinery responsible for correlated transcriptional bursts, we asked whether our method distinguishes between CDRC sets that demonstrate bursting from sets that do not. First, since our library contains every possible CDRC regime, we attempted to identify which input sets exhibit transcriptional bursting. Second, we asked how well the burst parameter is detected by ACES, and whether certain regimes are more or less amenable to measuring the burst size from a protein distribution.

The Fano factor of the RNA distribution ($F_{RNA}=RNA_{var}/RNA_{mean}$) distinguishes between constitutively active genes ($F_{RNA}\approx 1$) from bursting genes ($F_{RNA}>1$) [4,26]. This is because in the limit of $t_{on}\gg t_{off}$, or if t_{on} and t_{off} are both high, the RNA distribution's variance approaches its mean, and the Fano factor goes to 1 indicating constitutive RNA production. When we plot how well the burst parameter is fit versus F_{RNA}, we find that near $F_{RNA}\approx 1$ the average MD in all libraries is very high, but drops precipitously when F_{RNA} is slightly greater than 1 (Fig. S8A). Average MD of the burst parameter then slowly increases with increasing F_{RNA}. To gain some insight into this trend, we plotted how well the burst parameter was fit versus the value of the burst parameter itself (Fig. S8B). We observe that estimation of the burst parameter incrementally improves with increasing burst size until reaching an optimum corresponding to bursts of ones to tens of molecules. Above this, increasing burst size makes the burst parameter increasingly difficult to ascertain.

These data suggest that the high Fano factor regime corresponds to transcription of hundreds to thousands of transcripts per ON event, followed by long periods of gene quiescence. In this large burst regime, the burst parameter is difficult to infer precisely from the distribution. However, the ensemble of candidate solutions for B in these situations proves to be informative. Plotted in Fig. S9 are the forty highest Fano factor library members' burst parameter histograms. While some demonstrate exact inference of the burst parameter (305–307, 234–237), and others exhibit widely varying burst parameter values (637–645), all but one (638) of the forty reveal a burst size confined to $B\gg 1$, consistent with the behavior of the gene (Fig. S9). We conclude again that even when individual CDRCs or their ratios are not exactly inferred from the output of ACES, the shape of the solution ensemble conveys mechanistic information about the underlying gene.

Comparison to related methods

A number of analytical results map CDRCs or their ratios to protein distribution shape [17,22,25,28,38–40]. However, only Friedman, Cai and Xie's work demonstrating that protein counts are gamma-distributed [25] has been co-opted for solving the inverse problem: using distribution shape to infer the parameters [42–44]. Their result allows one to directly compute $V/M=k_p/d_m$ and $M^2/V=k_m/d_p$ from a gamma distribution's measured mean (M) and variance (V). We compared ACES' performance estimating the same ratios in the same regime as the one assumed by the Friedman model.

First, we screened our library for CDRC input sets where $d_m/d_p>10$, average protein count >1000, and $t_{on}\gg t_{off}$ ($P_{on}\approx 1$). These restrictions correspond to the Friedman assumptions that 1) $d_m\gg d_p$, 2) one can use a continuous approximation for protein count, and 3) the gene is always active. 251 input sets (4%) in our library satisfy these criteria. We examined the output solutions ACES returned for these specific input sets. Because ACES makes no assumptions about parameter regime, it returns solutions that conform to the Friedman assumptions as well as many additional solutions that do not. To compare ACES with the Friedman approach on equal footing, we first only considered ACES solutions that were consistent with the Friedman criteria. This set of solutions accurately identifies both Friedman ratios; the solutions cluster exactly on the true solution. By relative error, the ACES solution slightly outperforms Friedman's analytical estimation of both ratios (Fig. S11A,B): for k_m/d_p, the median relative error in the estimate was .024, versus a median error of .076 for the Friedman estimator. In addition to accurately recovering the Friedman ratios, ACES also recovered other CDRCs or ratios not obtainable from the Friedman model (Fig. S10B). Average RNA count was the most commonly identified parameter besides k_p/d_m and k_m/d_p. This exercise serves as a control for our method, demonstrating equivalence between our method and the Friedman analytical solution, in the regime assumed by the Friedman model.

However, ACES' utility derives from not assuming any particular parameter regime. We therefore compared the solutions that did not conform to the Friedman regime to the solutions that did conform. Even the solutions that did not conform to the Friedman assumption are highly enriched for successful fitting of the Friedman ratios k_p/d_m and k_m/d_p. Of 251 input sets, 222 have an MD<.75. Both ratios share an average MD value of .33, as compared to the overall library MD average of 1.25. Of the 29 input sets for which these ratios were fit with an MD >.75, all of them demonstrate bimodality in histograms of both ratios, suggesting that solutions outside the Friedman regime can produce gamma distributed protein distributions. The value of ACES is demonstrated by its ability to accurately fit the Friedman ratios even in regimes that explicitly violate the Friedman assumptions.

We explored this point further and discovered that the model of gene expression studied in this report (Fig. 1) readily generates gamma distributions, most of which do not correspond to the Friedman regime. There were 995 CDRC sets in our library that generate gamma distributed protein distributions (See examples, Fig. S11C,D). Of these, only 346 demonstrate accurate inference of the k_m/d_p and k_p/d_m ratios using the Friedman model. In Fig. S11E, ACES' fit of the distribution is compared to the Friedman fit (magenta triangles), revealing that while ACES' solution ensemble contains the correct answer, the Friedman solution settles on the incorrect value for both ratios. These results emphasize that having a gamma distribution is necessary but not sufficient for fitting the Friedman ratios, and reinforce that ACES operates robustly outside of the Friedman regime.

Discussion

Protein count distributions are readily measured from clonal populations of reporter-gene bearing cells [11,54], yet we lack a reliable computational framework for abstracting molecular mechanisms from distribution shape. Here we presented an efficient, assumption-free approach to inferring molecular mechanism from protein distribution shape. Though the full analytical solution of the protein distribution remains elusive, we were able to solve for exact expressions of the distributions' higher order moments, extending previous work which stopped at variance [5,9,10,13,15,25,26,28,39,45]. We found that the higher order moments skewness and kurtosis contributed significantly to the fitting problem; indeed, when fitting only on the mean and variance, we were rarely able to identify a CDRC or CDRC ratio in our representative library of test sets (column MV, Table 1).

We found that four moments accurately reproduce the complete protein distribution, even in the worst case scenario when very different CDRC sets map to the same four moments (Figs. 3, S2–S6). This is particularly exciting given the challenges of solving anything but the simplest chemical master equation model, as the moment-matching approach presented herein, while inelegant, is perfectly general. Even in cases where moments do not exactly capture distribution shape, moment matching could allow investigators to rule out the vast majority of candidate solutions, limiting the use of computationally intensive Gillespie simulations to cases where a given parameter set is likely to be correct.

ACES, our algorithm for linking parameters to moments, proved to be efficient enough to identify *all* CDRC sets consistent with each of 8053 input sets. To build confidence in our results, we checked whether ACES' CDRC ensembles agreed with a previous analytical result, which states that in a specific regime, gene expression is gamma distributed and the ratios k_m/d_p and k_p/d_m are directly calculable from the mean and variance [25]. We found that when limited to this regime, ACES' not only agreed with the Friedman results, but also frequently identified other expression-related quantities, such as the average RNA expression level, $\langle \mathrm{RNA} \rangle$ and k_p/d_p. In many cases, assuming the Friedman regime allowed direct inference of most of the remaining CDRCs and ratios (compare Fig. S10A to S10B), quantities which cannot be obtained by the Friedman model.

CDRC sets operating in the Friedman regime represented only 4% of the possible regimes we studied in our library, reinforcing the importance of bringing to bear an unrestricted framework for analyzing protein distributions. Opening our analysis to the whole library, we discovered several trends. First, about one in eight regimes recapitulate a gamma distribution, but only about one in three of gamma distributed input sets fall in the Friedman regime. In other words, model agreement with a distribution does not prove model validity. This recalls a similar outcome from Zopf *et al*, where the standard ON-OFF model studied here (Fig. 1) accurately models total population distributions, but fails to capture subpopulation distributions partitioned by cell cycle phase [21]. Both results provide motivation for continued refinement of stochastic gene expression models, which will benefit from the parameter estimation approach suggested in this work.

Second, we find that the exact nature of degeneracy in a solution set informs on mechanisms that underlie the shape of the corresponding protein distribution. Every library member corresponds to a unique ensemble of solutions. Sometimes, solution histograms encompass the complete range of a given CDRC or CDRC ratio, while often the range of a parameter is confined to a fraction of its possible space. Sometimes ACES reveals the only CDRC or CDRC ratio value consistent with a distributions shape. All results are informative. In some cases, ACES revealed bimodality in a particular parameter, suggesting that the degeneracy arises from uncertainty in one particular CDRC (Fig. 5). In other examples, we found that ACES reveals burst histograms consistent with slow ON-OFF transitioning behavior, even without always exactly identifying the burst parameter (Fig. S9). Even knowing when a distribution cannot rule out any value of a CDRC parameter or CDRC ratio, as in Fig. S7, informs an investigator that additional information is needed to further constrain the fitting problem.

Third, some ratios and CDRCs are more readily abstracted from the distribution than others. While the ON-OFF transition CDRCs were particularly challenging to infer, especially t_{off}, we found that t_{on} and t_{off} pairs often varied while conserving one value of P_{on}. If this observation bears out in experimental validation, accessibility to P_{on} suggests a novel way to parameterize another class of models, the so-called thermodynamic model of *cis*-regulation [55–57]. At the core of a thermodynamic model is a partition function that divides transcriptionally active promoter states by all molecular states, and is equal to P_{on}. Though ACES readily identified a variety of CDRC ratios besides P_{on} we were surprised to find individual CDRC values fit, despite evidence that CDRCs cannot be obtained from steady state distributions alone [41]. We speculate that explicitly confining the search to the physiological domain of each parameter permitted estimation of some CDRCs. Importantly, bounding CDRCs but fitting only on mean and variance did not permit much CDRC estimation, indicating that it is the full protein distribution in conjunction with parameter ranges that enables ACES to estimate individual CDRCs. Independent of CDRC or ratio, further constraining ACES with measures of one or both of the degradation rates naturally improves parameter estimation across the board (Table 1).

Although we explored CDRC estimation only when restricting d_m and d_p, the algorithm accepts user-defined constraints accounting for uncertainty in both the CDRCs and the moments. For example, an investigator may have experimentally determined the 95% confidence interval for a particular parameter. Confining the ACES search to this interval requires only replacing the physiological range with measured bounds. Similarly, measurement of distribution moments will be imperfect, likely with higher moments admitting more significant error. Again, one can adjust ACES' tolerance for each moment. This flexibility means that the worst outcome manifests as CDRC values spanning their entire range; an investigator need not fear overfitting or spurious convergence.

More importantly perhaps, ACES fits will aid investigators in model selection. ACES fit outcomes fall into three categories: (1) ACES fails to discover a candidate CDRC set (2) ACES finds CDRC solutions spanning most or all of the range of each parameter, or (3) ACES narrows down some or all CDRCs to a limited range. As discussed above, high experimental error or not enough constraints can lead to outcome (2). Both (1) and (3) are potentially instructive regarding model selection. In (1), either moment error tolerance was more strict than measurement error, ACES was not run with a sufficiently high enough resolution, or the model does not adequately explain the data. The first two possibilities are easily ruled out, and if excluded indicate the ON-OFF model insufficiently characterizes a particular set of moments, and therefore the distribution. Several examples in the literature highlight alternative models, including the possibility of a refractory OFF state [8], multiple ON states [30], and even transcription rates that depend on the cell cycle [21]. This last example illustrates how cellular state, called extrinsic noise [5,14,27], intertwines with the intrinsic-only stochasticity captured

by our approach. While a variety of methods have been developed for measuring and minimizing the impact of extrinsic noise [11,12,58–60], Zopf *et al* reinforce the idea that extrinsic can be made intrinsic by explicitly incorporating fluctuating inputs into a given model [21,24,29,60]. Though we focused solely on the ON-OFF model, the paradigm suggested here of solving for and exhaustively fitting to moments is perfectly viable for any linear stochastic model, including those that admit fluctuating inputs or comprise more elaborate promoter architectures. Whether the additional CDRCs implicit to more detailed models can be inferred from stationary distributions is an open question. The significant degeneracy we encountered with the ON-OFF model in this work suggests that significant constraints will be necessary for such distributions to be informative.

Having some CDRCs or CDRC ratios well fit while others span wider ranges, as in (3), suggest either the data are not sufficiently constrained to distinguish each parameter, or the model is too complex. We saw this latter scenario in our data; a variety of choices of t_{on} , t_{on} , and k_m can conspire to drive transcription as a Poisson process at rate S. However, by testing for constraint of various CDRC ratios, these patterns arise naturally in our framework. We saw numerous examples of model reduction. Encouragingly, we rediscovered the Friedman solution, a two parameter model characterized by k_p/d_m and k_m/d_p [25]. Their parameterization posits protein production as the product of the number of transcripts produced per cell division (k_m/d_p), assuming that dilution dominates protein loss, and the number of proteins produced in the lifetime of an RNA (k_p/d_m). In addition, we found alternate two-parameter formulations, for example, $\langle \mathrm{RNA} \rangle$ and k_p/d_p , suggesting that in some regimes this pair of ratios ascribes shape to the distribution, while the Friedman ratios do not. Of a large number of test sets where P_{on} was fit, 387 cases identified *only* P_{on}. Though we did not directly interrogate the ratio $(k_m k_p)/(d_m d_p)$, fitting P_{on} and given the protein mean guarantees that this four-CDRC ratio is also fit in those test sets, providing yet another two parameter model. Thus, the only undesirable consequence of assuming a more complex model is increased computational time, a potential roadblock that we did not encounter in the current study. Indeed, the more general model collapses into an intriguing number of unique simpler models that might otherwise not be intuited by an investigator (Data S1).

ACES will enable a more mechanistic understanding of gene regulation. We envision it will be most informative when used in pairwise studies of reporter gene activity, for example, by measuring reporter gene distributions before and after knocking out a histone modifying enzyme, or a transcription factor binding site in a promoter. Historically these manipulations reveal only whether a *cis*- or *trans*- acting factor activates or represses gene activity. By incorporating data from changes in stochasticity, we expect to refine our understanding of regulation by ascertaining which CDRC, or set of CDRCs, are regulated by a particular sequence element or protein. Importantly, even if the exact values of CDRCs cannot be obtained from a distribution, changes in the range of a particular parameter still provide mechanistic insight into the effects of specific genetic perturbations. Previous studies incorporating noise into their analysis reveal insights such as these, for example, that the mammalian *cis*-regulatory CCAAT box element chiefly modulates the k_m and t_{on} CDRCs, but has little impact on t_{off} [8].

Given that four moments capture the full shape of the protein expression distribution, and that we can rapidly determine all CDRC sets consistent with these moments, this approach shows great promise in learning mechanistic information from stochasticity in protein expression. We applied our framework to one particular steady state protein distribution (Fig. 1), but the general approach is amenable to analyzing any first-order chemical master equation model of protein or RNA distributions, RNA-protein joint distributions, and even distributions evolving over time. We expect the application of ACES to improve our understanding of the fundamental processes that govern gene expression.

Materials and Methods

ACES algorithm overview

For the first parameter k_m, we select its first value, for example, $k_m^* = k_{m-min}$. Given the mean expression level from a population of cells (μ_p), the equation for how this mean relates to the CDRCs,

$$\mu_p = \frac{t_{on}}{t_{on}+t_{off}} \frac{k_m k_p}{d_m d_p} \tag{5}$$

and the CDRC minima and maxima, $[d_p^{(min)}, d_p^{(max)}]$, $[d_m^{(min)}, d_m^{(max)}]$, $[k_p^{(min)}, k_p^{(max)}]$, $[k_m^{(min)}, k_m^{(max)}]$, $[t_{on}^{(min)}, t_{on}^{(max)}]$, $[t_{off}^{(min)}, t_{off}^{(max)}]$, we can potentially further constrain the boundaries on the next CDRC to be checked. If the next parameter to be selected is k_p, then we can check the boundaries on k_p given all the information we have to this point. That is, the new $k_p^{(min)}$ might be greater than the physiological $k_p^{(min)}$ given that k_m is fixed at k_m^*, and the mean expression level is fixed at μ_p^*. Substituting in the known values up to this point into Eq. 5, we get:

$$\mu_p^* = \frac{t_{on}}{t_{on}+t_{off}} \frac{k_m^* k_p}{d_m d_p} \tag{6}$$

Then solving for k_p we get:

$$k_p = \frac{t_{on}+t_{off}}{t_{on}} \frac{d_m d_p \mu_p^*}{k_m^*} \tag{7}$$

The result is a monotonic function for all variables; thus, k_p's minimum and maximum values can be obtained, subject to the fixed parameters to this point (k_m^* and μ_p^*) and the physiological minima and maxima of the remaining free parameters:

$$k_p^{*(min)} = \frac{t_{on}^{(max)}+t_{off}^{(min)}}{t_{on}^{(max)}} \frac{d_m^{(min)} d_p^{(min)} \mu_p^*}{k_m^*} \tag{8}$$

$$k_p^{*(max)} = \frac{t_{on}^{(min)}+t_{off}^{(max)}}{t_{on}^{(min)}} \frac{d_m^{(max)} d_p^{(max)} \mu_p^*}{k_m^*} \tag{9}$$

These candidate minimum and maximum values are not guaranteed to be greater than or less than, respectively, the physiologically determined bounds for k_p. To take this into account, we accept $k_p^{*(min)}$ if it is greater than $k_p^{(min)}$, and we accept $k_p^{*(max)}$ if it is less than $k_p^{(max)}$. Thus the new boundaries define a range for k_p that is less than or equal to the physiological range defined at runtime. In the next step, the algorithm subdivides the new k_p range with resolution R^*, shrinking the resolution R by the fraction of parameter space lost by accepting a narrower CDRC

range. The algorithm then selects the first value of k_p in this new range, and repeats the steps just outlined above with the next parameter.

Because the mean equation is linear, at the end of this process the last parameter can be directly solved for. At that point, we have a set of CDRCs k_m^*, k_p^*, d_m^*, d_p^*, t_{on}^*, and t_{off}^* which, by the nature of how each parameter was selected, already exactly equal the first moment equation (Eq. 5). We then plug in this candidate CDRC set into our variance solution; if the calculated variance also matches within 1% of the measured variance, we test skewness with the same criteria, and then again with kurtosis. If the 2nd through 4th moments all agree with the input (measured) moments, this CDRC set is recorded as a solution.

The full pseudocode is provided in the Supplement, and an implementation of the algorithm is also provided in C++ (Data S2).

Log Euclidean distance

Log Euclidean distances are computed as follows. For moment space, log-euclidean distance between two moment sets X and Y, with elements mean (M), variance (V), skewness (S), and kurtosis (K) would be:

$$LED(X,Y) = \sqrt{log(\frac{M_X}{M_Y})^2 + log(\frac{V_X}{V_Y})^2 + log(\frac{S_X}{S_Y})^2 + log(\frac{K_X}{K_Y})^2} \quad (10)$$

For CDRCs, the equation would be logs of ratios of CDRCs rather than moments, and there would be six summands rather than four. We used log distance so that a CDRC–or a moment–changing from .01 to .1 contributes the same weight as another parameter changing from 100 to 1000. Thus, very different scale parameters, or moments, can be compared.

Jensen-Shannon Divergence

Jensen-Shannon Divergence for probability distributions A and B (JSD(A,B)) is defined as

$$JSD(A,B) = \frac{1}{2}KL(A||M) + \frac{1}{2}KL(B||M) \quad (11)$$

given the definitions for Kullback-Leibler divergence (KL) and the mixture distribution (M)

$$KL(X||Y) = \sum_{i=0}^{\infty} ln(\frac{X_i}{Y_i})X_i \quad (12)$$

$$M = \frac{1}{2}\sum_{i=0}^{\infty} A_i + B_i \quad (13)$$

Supporting Information

Data S1 ACES library data. ACES was run for each of the 8053 library members, generating a list of solutions. The summary data for each ACES run is described is included in the excel spreadsheet. Data S1 is the source of the library summary results reported in Table 1. A more detailed explanation of each datasheet and column can be found in the README datasheet within Data S1.
(XLSX)

Data S2 ACES algorithm and moment solution routines. This archive contains both the ACES algorithm routine written in C++ and a Matlab implementation of the moment solutions. See README.txt for more detail.
(ZIP)

Text S1 Supporting information. This supplement documents the moments derivation and validation, continues the explanation of the ACES algorithm, and contains the supporting figures referenced throughout the text.
(PDF)

Acknowledgments

The authors are grateful to G. Rieckh and the Cohen lab for critical review of the manuscript.

Author Contributions

Conceived and designed the experiments: MSS BAC. Performed the experiments: MSS. Analyzed the data: MSS. Wrote the paper: MSS BAC.

References

1. Larson DR, Zenklusen D, Wu B, Chao Ja, Singer RH (2011) Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science (New York, NY) 332: 475–8.
2. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. Cell 123: 1025–36.
3. Zenklusen D, Larson DR, Singer RH (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. Nature structural & molecular biology 15: 1263–71.
4. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. PLoS biology 4: e309.
5. Raser JM, O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. Science (New York, NY) 304: 1811–4.
6. Blake WJ, KAErn M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. Nature 422: 633–7.
7. Li GW, Xie XS (2011) Central dogma at the single-molecule level in living cells. Nature 475: 308–15.
8. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, et al. (2011) Mammalian genes are transcribed with widely different bursting kinetics. Science (New York, NY) 332: 472–4.
9. Thattai M, van Oudenaarden a (2001) Intrinsic noise in gene regulatory networks. Proceedings of the National Academy of Sciences of the United States of America 98: 8614–9.
10. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. Nature genetics 31: 69–73.
11. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441: 840–6.
12. Volfson D, Marciniak J, Blake WJ, Ostroff N, Tsimring LS, et al. (2006) Origins of extrinsic variability in eukaryotic gene expression. Nature 439: 861–4.
13. Chabot JR, Pedraza JM, Luitel P, van Oudenaarden A (2007) Stochastic gene expression out-ofsteady-state in the cyanobacterial circadian clock. Nature 450: 1249–52.
14. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science (New York, NY) 297: 1183–6.
15. Sanchez A, Garcia HG, Jones D, Phillips R, Kondev J (2011) Effect of promoter architecture on the cell-to-cell variability in gene expression. PLoS computational biology 7: e1001100.
16. Mogno I, Vallania FLM, Mitra RD, Cohen BA (2010) TATA is a modular component of synthetic promoters. Genome research 20: 1391–7.
17. Zechner C, Ruess J, Krenn P, Pelet S, Peter M, et al. (2012) Moment-based inference predicts bimodality in transient gene expression. Proceedings of the National Academy of Sciences of the United States of America 109: 8340–5.

18. Blake WJ, Balázsi G, Kohanski Ma, Isaacs FJ, Murphy KF, et al. (2006) Phenotypic consequences of promoter-mediated transcriptional noise. Molecular cell 24: 853–65.
19. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. Science (New York, NY) 307: 1962–5.
20. Dar R, Razooky B (2012) Transcriptional burst frequency and burst size are equally modulated across the human genome. Proc Natl Acad Sci U S A109:17454–9.
21. Zopf CJ, Quinn K, Zeidman J, Maheshri N (2013) Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression. PLoS computational biology 9: e1003161.
22. Shahrezaei V, Swain PS (2008) Analytical distributions for stochastic gene expression. Proceedings of the National Academy of Sciences of the United States of America 105: 17256–61.
23. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22: 403–434.
24. Shahrezaei V, Ollivier JF, Swain PS (2008) Colored extrinsic fluctuations and stochastic gene expression. Molecular systems biology 4: 196.
25. Friedman N, Cai L, Xie X (2006) Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. Physical Review Letters 97: 1–4.
26. So LH, Ghosh A, Zong C, Sepúlveda La, Segev R, et al. (2011) General properties of transcriptional time series in Escherichia coli. Nature genetics 43: 554–60.
27. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proceedings of the National Academy of Sciences of the United States of America 99: 12795–800.
28. Elgart V, Jia T, Fenley AT, Kulkarni R (2011) Connecting protein and mRNA burst distributions for stochastic models of gene expression. Physical biology 8: 046001.
29. Lei J (2009) Stochasticity in single gene expression with both intrinsic noise and fluctuation in kinetic parameters. Journal of theoretical biology 256: 485–92.
30. Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, et al. (2013) Systematic identification of signal-activated stochastic gene regulation. Science (New York, NY) 339: 584–7.
31. Turner TE, Schnell S, Burrage K (2004) Stochastic approaches for modelling in vivo reactions. Computational biology and chemistry 28: 165–78.
32. Cao Y, Gillespie D, Petzold L (2005) Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. Journal of Computational Physics 206: 395–411.
33. Ingram PJ, Stumpf MPH, Stark J (2008) Nonidentifiability of the source of intrinsic noise in gene expression from single-burst data. PLoS computational biology 4: e1000192.
34. Chatterjee A, Vlachos DG, Katsoulakis Ma (2005) Binomial distribution based tau-leap accelerated stochastic simulation. The Journal of chemical physics 122: 024112.
35. Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. The Journal of chemical physics 124: 044104.
36. Cao Y, Gillespie DT, Petzold LR (2007) Adaptive explicit-implicit tau-leaping method with automatic tau selection. The Journal of chemical physics 126: 224101.
37. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737–41.
38. Peccoud J, Ycart B (1995) Markovian modeling of gene-product synthesis. Theoretical Population Biology 48: 222–234.
39. Kepler TB, Elston TC (2001) Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophysical journal 81: 3116–36.
40. Mugler A, Walczak A, Wiggins C (2009) Spectral solutions to stochastic models of gene expression with bursts and regulation. Physical Review E 80: 041921.
41. Munsky B, Trinh B, Khammash M (2009) Listening to the noise: random fluctuations reveal gene network parameters. Molecular systems biology 5: 318.
42. Shalem O, Carey L, Zeevi D, Sharon E, Keren L, et al. (2013) Measurements of the Impact of 3 End Sequences on Gene Expression Reveal Wide Range and Sequence Dependent Effects. PLoS Computational Biology 9: e1002934.
43. Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, et al. (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nature genetics 44: 743–50.
44. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, et al. (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science (New York, NY) 329: 533–8.
45. Sánchez A, Kondev J (2008) Transcriptional control of noise in gene expression. Proceedings of the National Academy of Sciences of the United States of America 105: 5081–6.
46. Huh D, Paulsson J (2011) Non-genetic heterogeneity from stochastic partitioning at cell division. Nature genetics 43: 95–100.
47. Paulsson J (2005) Models of stochastic gene expression. Physics of Life Reviews 2: 157–175.
48. Miller C, Schwalb B, Maier K, Schulz D, Dümcke S, et al. (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Molecular systems biology 7: 458.
49. Chubb JR, Trcek T, Shenoy SM, Singer RH (2006) Transcriptional pulsing of a developmental gene. Current biology : CB 16: 1018–25.
50. Garcia HG, Lee HJ, Boedicker JQ, Phillips R (2011) Comparison and calibration of different reporters for quantitative analysis of gene expression. Biophysical journal 101: 535–44.
51. Endres D, Schindelin J (2003) A new metric for probability distributions. Information Theory, IEEE. . . 49: 1858–1860.
52. Tkačik G, Callan CG, Bialek W (2008) Information capacity of genetic regulatory elements. Physical Review E 78: 011910.
53. Bernstein Ja, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color uorescent DNA microarrays. Proceedings of the National Academy of Sciences of the United States of America 99: 9697–702.
54. Carey LB, van Dijk D, Sloot PMa, Kaandorp Ja, Segal E (2013) Promoter sequence determines the relationship between expression level and noise. PLoS biology 11: e1001528.
55. Sherman MS, Cohen BA (2012) Thermodynamic State Ensemble Models of cis-Regulation. PLoS Computational Biology 8: e1002407.
56. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, et al. (2005) Transcriptional regulation by the numbers: applications. Curr Opin Genet Dev 15: 116–124.
57. Shea MA, Ackers GK (1985) The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. J Mol Biol 181: 211–230.
58. Singh A, Razooky BS, Dar RD, Weinberger LS (2012) Dynamics of protein noise can distinguish between alternate sources of gene-expression variability. Molecular systems biology 8: 607.
59. Komorowski M, Finkenstädt B, Rand D (2010) Using a single fluorescent reporter gene to infer half-life of extrinsic noise and other parameters of gene expression. Biophysical journal 98: 2759–69.
60. Hilfinger A, Paulsson J (2011) Separating intrinsic from extrinsic fluctuations in dynamic biological systems. Proceedings of the National Academy of Sciences of the United States of America 108: 12167–72.