

Steady-state fluctuations of a genetic feedback loop: an exact solution

R. Grima¹, D. R. Schmidt^{2,3}, and T. J. Newman⁴

¹*SynthSys Edinburgh, School of Biological Sciences,
University of Edinburgh, Edinburgh, EH9 3JR, UK*

²*Mathematical Biosciences Institute, The Ohio State University,
Columbus, OH 43210, USA*

³*Computational Biomathematics Laboratory, Case Western Reserve University,
Cleveland, OH 44106, USA*

⁴*Division of Cell and Developmental Biology,
College of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK*

Abstract

Genetic feedback loops in cells break detailed balance and involve bimolecular reactions; hence exact solutions revealing the nature of the stochastic fluctuations in these loops are lacking. We here consider the master equation for a gene regulatory feedback loop: a gene produces protein which then binds to the promoter of the same gene and regulates its expression. The protein degrades in its free and bound forms. This network breaks detailed balance and involves a single bimolecular reaction step. We provide an exact solution of the steady-state master equation for arbitrary values of the parameters, and present simplified solutions for a number of special cases. The full parametric dependence of the analytical non-equilibrium steady-state probability distribution is verified by direct numerical solution of the master equations. For the case where the degradation rate of bound and free protein is the same, our solution is at variance with a previous claim of an exact solution (Hornos et al, Phys. Rev. E **72**, 051907 (2005) and subsequent studies). We show explicitly that this is due to an unphysical formulation of the underlying master equation in those studies.

I. INTRODUCTION

Biochemical reaction networks underpin the robustness of cells to both internal and external perturbations. Feedback and non-linearities make network behaviour hard to understand, and thus mathematical modelling of the networks can provide useful insights. Copy numbers of gene products, such as proteins, are often relatively small [1, 2], which argues for a careful evaluation of the role of stochasticity. The importance of stochasticity is without doubt in gene expression, given that there are only one or two copies of most genes per cell [3]. Modelling of the stochastic dynamics of networks is typically more involved than sets of deterministic rate equations. Exact solutions have been obtained for reaction networks obeying detailed balance [4–6] and for those composed of first-order (unimolecular) reactions [7–11]. However these restrictions are not typical of biochemical processes inside living cells. Detailed balance conditions are characteristic of closed systems of reversible chemical reactions in thermal equilibrium conditions; they only hold for open systems in special cases [4]. Living cells are open biochemical systems which actively exchange matter with their surroundings and which possess non-equilibrium steady states, and hence it is clear that the principle of detailed balance will not generally hold for intracellular biochemical systems [12]. It is also a fact that most systems of interest involve a number of second-order (bimolecular) reactions such as substrate-enzyme interactions, and protein-DNA interactions.

Here we focus on perhaps the simplest example of a biochemical reaction network which overtly breaks detailed balance and which involves both unimolecular and bimolecular reaction steps. We consider a genetic regulatory network with a feedback loop, namely one in which the product of a gene binds to the promoter of that same gene, and regulates its expression. Furthermore the free and bound protein are assumed to be degraded via proteolysis. Note that while bound protein degradation is not as well known or obvious as free protein degradation, there are mechanisms which could mediate it, e.g. the ubiquitin-proteasome pathway can target and degrade parts of protein complexes [13, 14]. Similar feedback mechanisms as discussed above are ubiquitous in biology, appearing in such diverse contexts as metabolism [15], signaling [16], somitogenesis [17] and circadian clocks [18]. Naturally, given its simplicity, special cases of this model have already been the subject of a number of studies. Hornos et al [19] and subsequent follow up studies [20–22] have claimed an exact solution for the case where the rate of bound protein degradation is equal

to the rate of free protein degradation. Qian et al [23] have studied the case where the bound protein degradation rate is zero and developed an approximate solution of the master equation in the limits of slow and rapid switching of the gene between the unbound and bound states. A few studies have also considered variations of this simple model by allowing inhibition or activation by protein dimers [24, 25].

In this paper we obtain an exact solution of the master equation for arbitrary free and bound protein degradation rates in non-equilibrium steady state conditions. For the case of equal bound and free protein degradation rates, our solution differs markedly from the exact solution claimed by Hornos et al [19]; we explicitly show that the difference between the two solutions stems from the fact that the master equations studied in the latter work have no consistent physical interpretation and hence constitute an incorrect description of the biochemical processes at play.

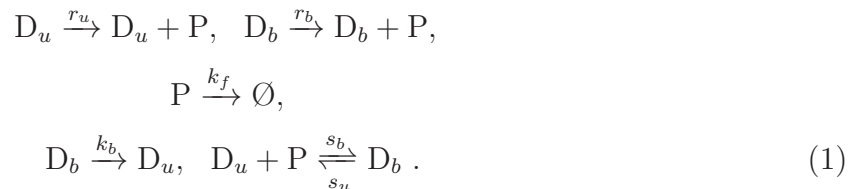
In the next section we define the model and write down a master equation formulation of the stochastic dynamics. In section III we present the exact solution using the generating function method. In section IV we study three special cases in which detailed balance does not hold and contrast with the case in which detailed balance holds. In section V we directly compare our exact solution with numerical solutions of the master equations and show their correctness. In section VI we present a careful comparison of our exact solution to previous studies. We summarise our results and conclude in section VII.

II. THE MODEL AND THE MASTER EQUATION

We consider a single gene and its accompanying promoter region. Self-regulation means that the protein corresponding to the gene can bind to the promoter region and thereby affect the transcription and translation processes. Following previous work, we do not explicitly consider the transcription process, and the intermediate mRNA. Rather we model the process by following only the number of free proteins and the state of the promoter, namely, whether it is bound or unbound. It is important for us to carefully define the processes in order to make completely transparent how these will be encoded into a master equation. By ‘free proteins’ we mean proteins that have been created from transcription/translation of the gene in question, and which are neither bound to the promoter nor degraded. We only allow one protein to be bound to the promoter region at any given time. We do not

consider dimerization of free proteins.

We define two conditional probabilities: $P_0(n, t)dt$ is the probability that in the time interval $(t, t + \delta t)$ there are n free proteins and the promoter is unbound, and $P_1(n, t)dt$ is the probability that in the interval $(t, t + \delta t)$ there are n free proteins and the promoter is bound. In the latter case, there are in fact $n + 1$ proteins in the system, n of which are free, and one of which is bound to the promoter. The transcription process will be altered if the promoter is bound, and so the rate of production of proteins will depend on the state of the promoter region. This dependence of production rate on the promoter state breaks detailed balance and makes analytic solution of this problem non-trivial. We define r_u and r_b to be the production rates of protein given than the promoter region is unbound or bound respectively. We define k_f to be the degradation rate of free proteins. We define k_b to be the degradation rate of the bound protein. Allowing k_b to be non-zero breaks detailed balance even if $r_u = r_b$ and, again, makes analytic solution non-trivial, but not impossible as we shall see. Lastly, we define s_b to be the binding rate per protein to the promoter region, and s_u to be unbinding rate from the promoter region. Note, we assume that if the bound protein is degraded it is removed from the system: explicitly the total number of proteins will decrease by one, the total number of free proteins will remain unchanged, and the state of the promoter will change from bound to unbound. These processes and their accompanying rates are schematically illustrated in the following reaction scheme, where D_u , D_b , and P represent the unbound DNA, the bound DNA, and the free proteins, respectively:



Assuming that each process is an independent Poisson process allows us to encode the dynamics using master equations [26]. The master equations for P_0 and P_1 have the following

forms

$$\begin{aligned}
\frac{d}{dt}P_0(n, t) &= r_u (P_0(n - 1, t) - P_0(n, t)) \\
&+ k_f ((n + 1)P_0(n + 1, t) - nP_0(n, t)) \\
&+ k_b P_1(n, t) + s_u P_1(n - 1, t) - s_b n P_0(n, t) ,
\end{aligned} \tag{2}$$

$$\begin{aligned}
\frac{d}{dt}P_1(n, t) &= r_b (P_1(n - 1, t) - P_1(n, t)) \\
&+ k_f ((n + 1)P_1(n + 1, t) - nP_1(n, t)) \\
&- k_b P_1(n, t) - s_u P_1(n, t) + s_b (n + 1) P_0(n + 1, t) .
\end{aligned} \tag{3}$$

We have formatted these equations to make as clear as possible the relation of the terms to the corresponding molecular processes. Each line of the above equations corresponds to the processes described on the corresponding line of reaction scheme (1). The first line of each equation refers to processes originating from the gene, i.e., production of the protein. The second line of each equation refers to processes in the cytosol of the cell, i.e., degradation of free proteins. The third line of each equation refers to processes occurring on the promoter, i.e., degradation of the bound protein, and binding and unbinding of individual proteins on the promoter.

There is no need to write special boundary conditions for these equations, so long as we impose $P_0(n, t) = 0$ and $P_1(n, t) = 0$ for $n < 0$. Explicitly, if we insert $n = 0$ into the above equations we find

$$\frac{d}{dt}P_0(0, t) = -r_u P_0(0, t) + k_f P_0(1, t) + k_b P_1(0, t) , \tag{4}$$

$$\frac{d}{dt}P_1(0, t) = -r_b P_1(0, t) + k_f P_1(1, t) - k_b P_1(0, t) - s_u P_1(0, t) + s_b P_0(1, t) , \tag{5}$$

which correctly describe the time evolution of $P_0(0, t)$ and $P_1(0, t)$.

In the next section we present the exact solution of these equations in the steady-state using the generating function method.

III. EXACT SOLUTION

It is convenient to work with a dimensionless time variable, and so we choose to scale time by the rate of degradation of free proteins k_f , i.e., $\tau = k_f t$. We define the dimensionless

rates

$$\rho_u = r_u/k_f, \rho_b = r_b/k_f, \theta = k_b/k_f, \sigma_u = s_u/k_f, \sigma_b = s_b/k_f . \quad (6)$$

For future reference we define the convenient parameters

$$\Sigma_b = 1 + \sigma_b, R = \rho_u - \rho_b \Sigma_b . \quad (7)$$

In dimensionless variables, the master equations (2) and (3) take the form

$$\begin{aligned} \frac{d}{d\tau} P_0(n, \tau) &= \rho_u (P_0(n-1, \tau) - P_0(n, \tau)) \\ &+ ((n+1)P_0(n+1, \tau) - nP_0(n, \tau)) \\ &+ \theta P_1(n, \tau) + \sigma_u P_1(n-1, \tau) - \sigma_b n P_0(n, \tau) , \end{aligned} \quad (8)$$

$$\begin{aligned} \frac{d}{d\tau} P_1(n, \tau) &= \rho_b (P_1(n-1, \tau) - P_1(n, \tau)) \\ &+ ((n+1)P_1(n+1, \tau) - nP_1(n, \tau)) \\ &- \theta P_1(n, \tau) - \sigma_u P_1(n, \tau) + \sigma_b (n+1) P_0(n+1, \tau) . \end{aligned} \quad (9)$$

We will solve these equations using the generating function method [27, 28]. We define the generating functions via

$$G_0(z) = \sum_{n=0}^{\infty} z^n P_0(n) , \quad (10)$$

$$G_1(z) = \sum_{n=0}^{\infty} z^n P_1(n) . \quad (11)$$

Henceforth, we will work in the steady-state, and set $dP_0(n)/d\tau = 0$ and $dP_1(n)/d\tau = 0$. Summing the master equations (8) and (9) over n with a weight of z^n , one finds the coupled pair of first-order differential equations

$$\rho_u(z-1)G_0 - (z-1)G_0' + (\theta + \sigma_u z)G_1 - \sigma_b z G_0' = 0 , \quad (12)$$

$$\rho_b(z-1)G_1 - (z-1)G_1' - (\theta + \sigma_u)G_1 + \sigma_b G_0' = 0 . \quad (13)$$

The obvious way to proceed is to write G_1 in terms of G_0 and G_0' in Eq. (12) and then substitute into Eq. (13). However, this leads to a second-order differential equation for G_0 which is not of the Riemann type, and so cannot be solved in terms of hypergeometric functions.

The less obvious way to proceed is as follows. We differentiate Eq. (12) to get an equation involving G_0 , G'_0 , G''_0 , G_1 and G'_1 , and then use (12) again to eliminate G_0 in favour of G'_0 and G_1 . This leads us to an equation involving G'_0 , G''_0 , G_1 and G'_1 . Finally by means of Eq. (13), G'_0 and G''_0 can be expressed in terms of G_1 and its derivatives. This leads us to a second-order differential equation for G_1 which reads

$$A(z)G''_1 + B(z)G'_1 + C(z)G_1 = 0 , \quad (14)$$

with

$$A(z) = 1 - \Sigma_b z , \quad (15)$$

$$B(z) = (\rho_u + \rho_b \Sigma_b)z - ((1 + \theta)\Sigma_b + \sigma_u + \rho_u + \rho_b) , \quad (16)$$

$$C(z) = \rho_u(\theta + \sigma_u + \rho_b) + \rho_b \Sigma_b - \rho_u \rho_b z . \quad (17)$$

Since Eq. (14) has linear coefficients in z it can be transformed to the differential equation for the confluent hypergeometric function. On writing $G_1(z) = e^{az} \tilde{G}_1(bz + c)$, and substituting into Eq. (14) one can determine a , b , and c , which provides the following solution

$$G_1(z) = e^{\rho_b z} \tilde{G}_1(w) , \quad (18)$$

where

$$w = R \frac{(\Sigma_b z - 1)}{\Sigma_b^2} . \quad (19)$$

and $\tilde{G}_1(w)$ satisfies Kummer's equation [29] (i.e., the confluent hypergeometric differential equation)

$$w\tilde{G}_1'' + (\beta - w)\tilde{G}_1' - \alpha\tilde{G}_1 = 0 , \quad (20)$$

with

$$\alpha = \theta + \frac{\sigma_u(\rho_u - \rho_b)}{R} , \quad (21)$$

and

$$\beta = 1 + \theta + \frac{1}{\Sigma_b} \left(\sigma_u + \rho_u - \frac{\rho_u}{\Sigma_b} \right) . \quad (22)$$

Eq. (20) admits two independent solutions, the Kummer function $M(\alpha, \beta, w)$ and the Tricomi function $U(\alpha, \beta, w)$. The latter is inadmissible as a solution for \tilde{G}_1 , as we require that $P_1(n) \rightarrow 0$ for $n \rightarrow \infty$ and that the sum over n of $P_1(n)$ is finite. Thus, we have the exact solution of the generating function in the form

$$G_1(z) = Ae^{\rho_b z} M(\alpha, \beta, w) , \quad (23)$$

where A is a normalization constant. Referring to Eq. (13) we see that knowledge of $G_1(z)$ enables us to find an exact expression for $dG_0(z)/dz$. Substituting Eq. (23) into Eq. (13) and using the transformation properties of the Kummer function, we find

$$\frac{dG_0(z)}{dz} = Ae^{\rho_b z} \left[\frac{\alpha}{(\Sigma_b - 1)} M(\alpha + 1, \beta, w) - \frac{\sigma_u \rho_b}{R} M(\alpha, \beta, w) - \frac{R}{\Sigma_b^2} \frac{\alpha}{\beta} M(\alpha + 1, \beta + 1, w) \right]. \quad (24)$$

It is difficult to extract an explicit solution for $G_0(z)$ by integrating this expression, and this is consistent with the fact that the second-order differential equation for G_0 is not of the Riemann form. However, as we shall see, this is not an impediment to finding explicit expressions for $P_0(n)$.

We now proceed to obtain the probability distributions $P_0(n)$ and $P_1(n)$. The probability distribution $P_1(n)$ can be retrieved from the generating function via

$$P_1(n) = \frac{1}{n!} \frac{d^n}{dz^n} G_1(z) \Big|_{z=0}. \quad (25)$$

Substituting Eq. (23) in the above equation leads us to

$$P_1(n) = \frac{A}{n!} \sum_{m=0}^n C_m^n \rho_b^{n-m} \left(\frac{R}{\Sigma_b} \right)^m \frac{(\alpha)_m}{(\beta)_m} M(\alpha + m, \beta + m, w_0), \quad (26)$$

where we have defined $w_0 = w(0) = -R/\Sigma_b^2$, the Pochhammer symbol $(a)_n = \Gamma(a + n)/\Gamma(a) = a(a + 1) \dots (a + n - 1)$, with $(a)_0 = 1$, and the combinatorial symbol $C_m^n = n!/m!(n - m)!$.

Using the analogous expression to Eq. (25) for G_0 and P_0 , we can obtain $P_0(n)$ for $n \geq 1$ by differentiating Eq. (24), with respect to z , $(n - 1)$ times, dividing by $n!$ and setting $z = 0$. With some use of the transformation equations for the Kummer functions we find the compact expression

$$P_0(n) = \frac{A}{n!} \sum_{m=0}^{n-1} C_m^{n-1} \rho_b^{n-1-m} \left(\frac{R}{\Sigma_b} \right)^m \frac{(\alpha)_m}{(\beta)_m} \left[\frac{\Sigma_b}{(\Sigma_b - 1)} (m + \alpha) M(\alpha + m + 1, \beta + m, w_0) - \left(m + \alpha + \frac{\sigma_u \rho_b}{R} \right) M(\alpha + m, \beta + m, w_0) \right], \quad (27)$$

for $n \geq 1$. This just leaves $P_0(0)$, which can be found directly from the master equation (8) on setting $n = 0$ in the steady-state:

$$\rho_u P_0(0) = P_0(1) + \theta P_1(0). \quad (28)$$

On substituting the exact forms for $P_1(0)$ and $P_0(1)$ from Eqs. (26) and (27) respectively, we find

$$P_0(0) = A \left[\frac{\Sigma_b}{(\Sigma_b - 1)} \frac{\alpha}{\rho_u} M(\alpha + 1, \beta, w_0) - \frac{\sigma_u}{R} M(\alpha, \beta, w_0) \right]. \quad (29)$$

Note, this last expression requires $\rho_u > 0$ for $P_0(0)$ to be finite. In fact, this condition is required for the existence of a non-empty steady-state. If $\rho_u = 0$ then there is an absorbing state of zero proteins, and the steady-state in that case will correspond to an empty system.

The evaluation of A requires us to normalise by calculating the sum over n of $P_0(n)$ and $P_1(n)$

$$\sum_{n=0}^{\infty} (P_0(n) + P_1(n)) = 1. \quad (30)$$

The sum over $P_1(n)$ is straightforward, since it is just $G_1(1)$. However, the sum over $P_0(n)$ cannot be reduced to a simple form as we do not have an explicit expression for $G_0(z)$. One can perform the sum over n of $P_0(n)$ using the explicit expression (27), but this cannot be reduced beyond definite integrals over Kummer functions whose form is apparently unknown. As such, the normalisation constant A is most easily found by numerically computing the sums over n of the explicit expressions for $P_0(n)$ and $P_1(n)$. The probability that there are n proteins in steady-state conditions, $P(n)$, is then given by the sum of $P_0(n)$ and $P_1(n)$. Ratios of moments such as the Fano factor, i.e. the variance of fluctuations divided by the mean number of molecules, do not depend on A and hence explicit expressions can always be written down for such quantities.

We finish this section by noting that the exact expressions, Eqs. (26), (27) and (29), are valid for $R \neq 0$. The singular case $R = 0$ requires special attention and is treated in Appendix A.

IV. SPECIAL CASES

In this section, we provide results for four special cases which can be grouped into two classes: (i) the case of detailed balance ($\rho_u = \rho_b$ and $\theta = 0$), in which Poisson statistics are expected to hold, (ii) three cases in which detailed balance does not hold: $\rho_b = 0$, $\alpha = 0$ and $\alpha = \beta$. In each of these cases one can write down simple expressions for the normalisation constant of the probability distribution and hence one can also obtain explicit expressions for all the moments of the distribution. In what follows we calculate the dependence of the

fraction of time for which the promoter is bound on the average number of free proteins. We denote the former quantity by f_{off} , which is defined explicitly by

$$f_{\text{off}} = \sum_{n=0}^{\infty} P_1(n) , \quad (31)$$

and denote the average number of free proteins by $\langle n \rangle$, which is defined explicitly by

$$\langle n \rangle = \sum_{n=0}^{\infty} n(P_0(n) + P_1(n)) . \quad (32)$$

We will find that the functional dependence $f_{\text{off}}(\langle n \rangle)$ under non-detailed balance conditions can differ markedly from the detailed balance case.

A. Detailed balance conditions

The self-regulating gene, described purely in terms of the states representing the number of free proteins, does not generally satisfy detailed balance. This has two causes: (i) the degradation of bound protein, and (ii) the differing rates at which proteins are produced depending on the state of the promoter. Thus, detailed balance can be restored by taking $\theta = 0$ and $\rho_u = \rho_b$ (see for example [30] for a general discussion of detailed balance in a gene regulation context). In this case, production of protein is independent of the state of the promoter, and degradation occurs only in the free protein pool. As such, detailed balance holds, and, indeed, the distribution of free protein is trivially Poisson. A well known, yet non-trivial, observation is the simple relationship between f_{off} and $\langle n \rangle$ in this case, which satisfies the Hill function [31]

$$f_{\text{off}} = \frac{\langle n \rangle}{\langle n \rangle + \frac{s_u}{s_b}} . \quad (33)$$

This provides a test of our exact solution which we now confirm. We equate the scaled production rates to each other, defining in the process $\rho \equiv \rho_u = \rho_b$. We first note from Eq. (21) that when $\rho_u = \rho_b$ and $\theta = 0$, then $\alpha = 0$. The Kummer function $M(\alpha = 0, \beta, w) = 1$, and so, from Eq. (23) we have

$$G_1(z) = Ae^{\rho z} . \quad (34)$$

It is straightforward to integrate Eq. (24) for G_0 in this case, and one finds

$$G_0(z) = A \frac{\sigma_u}{\rho \sigma_b} e^{\rho z} + B , \quad (35)$$

where B is an integration constant. This constant can be fixed by imposing the condition (28) and one finds $B = 0$. Normalisation fixes A , and one can then retrieve the explicit solutions for the probability distributions:

$$P_0(n) = \frac{e^{-\rho}}{\left(1 + \frac{\rho\sigma_b}{\sigma_u}\right)} \frac{\rho^n}{n!}, \quad (36)$$

$$P_1(n) = \frac{e^{-\rho}}{\left(1 + \frac{\sigma_u}{\rho\sigma_b}\right)} \frac{\rho^n}{n!}. \quad (37)$$

Note, the distribution of free proteins, $P_0(n) + P_1(n) = e^{-\rho}\rho^n/n!$: a Poisson distribution as anticipated. Summing these distributions over n with a weight of n , the average number of proteins is easily found to be $\langle n \rangle = \rho$. Now the fraction of time for which the promoter is bound, f_{off} , is equal to the probability that the gene is ‘off’, which is obtained by summing Eq. (37) over n . Expressing the latter in terms of $\langle n \rangle$ and reverting to unscaled rate parameters we obtain Eq. (33), as required.

We finish this section by noting that the deterministic model of the genetic feedback loop predicts that f_{off} has a Hill function dependence on $\langle n \rangle$ for all parameter values (see Appendix B). Hence for the detailed balance case, the predictions of the stochastic and deterministic models are in agreement.

B. Non-detailed balance conditions

1. The case $\rho_b = 0$

This is the case of strong transcriptional repression since the production of the gene in the bound state is zero. In this case, we see from Eq. (24) that the expression for $dG_0(z)/dz$ is a sum of Kummer functions, which can be directly integrated. On doing so, and, again, utilising the transformation formulae for Kummer functions, one finds the compact result

$$G_0(z) = \frac{A}{\rho_u(\Sigma_b - 1)} \left[(\theta + \sigma_u)\Sigma_b M(\alpha + 1, \beta, w) - \sigma_u(\Sigma_b - 1)M(\alpha, \beta, w) \right]. \quad (38)$$

In principle, an unknown constant B should be added to this expression; however, one can fix this constant by utilising the relation (28) and one finds that $B = 0$. Directly from Eq. (23), we have

$$G_1(z) = AM(\alpha, \beta, w), \quad (39)$$

where we have the simpler expressions $\alpha = \theta + \sigma_u$ and $w(z) = \rho_u(\Sigma_b z - 1)/\Sigma_b^2$; the parameter β is still given by Eq. (22).

It is now straightforward to determine the normalisation constant A by imposing condition (30), which is equivalent to $G_0(1) + G_1(1) = 1$. One finds

$$A = \rho_u(\Sigma_b - 1)\tilde{A}, \quad (40)$$

where

$$\tilde{A} = \left[(\theta + \sigma_u)\Sigma_b M(\alpha + 1, \beta, w_1) + (\rho_u - \sigma_u)(\Sigma_b - 1)M(\alpha, \beta, w_1) \right]^{-1}, \quad (41)$$

and $w_1 = w(1) = \rho_u(\Sigma_b - 1)/\Sigma_b^2$.

The explicit forms for the probability distributions can be obtained directly from the general formulae (26), (27), and (29), or from direct evaluation from the explicit generating functions (38) and (39). In either case, one obtains

$$P_0(n) = \frac{\tilde{A}}{n!} \left(\frac{\rho_u}{\Sigma_b} \right)^n \frac{(\alpha)_n}{(\beta)_n} \left[(\alpha + n)\Sigma_b M(\alpha + n + 1, \beta + n, w_0) - \sigma_u(\Sigma_b - 1)M(\alpha + n, \beta + n, w_0) \right], \quad (42)$$

and

$$P_1(n) = \frac{A}{n!} \left(\frac{\rho_u}{\Sigma_b} \right)^n \frac{(\alpha)_n}{(\beta)_n} M(\alpha + n, \beta + n, w_0), \quad (43)$$

where in this case $w_0 = w(0) = -\rho_u/\Sigma_b^2$. Note that Eq. (42) for $P_0(n)$ is valid for all $n \geq 0$.

The fraction of time for which the promoter is bound, f_{off} , is given by $G_1(1)$, and has the form

$$f_{\text{off}} = \rho_u(\Sigma_b - 1)\tilde{A}M(\alpha, \beta, w_1). \quad (44)$$

The mean number of free proteins, regardless of the state of the gene is given by $G'_0(1) + G'_1(1)$ and has the form

$$\langle n \rangle = \rho_u \alpha \tilde{A} \left[\Sigma_b M(\alpha + 1, \beta, w_1) - (\Sigma_b - 1)M(\alpha, \beta, w_1) \right]. \quad (45)$$

Unlike the detailed balance case, it does not seem generally possible to write f_{off} as a function of $\langle n \rangle$. The behavior for small and large $\langle n \rangle$ can however be easily deduced. It is clear that in the limit of large $\langle n \rangle$, f_{off} approaches one. The small $\langle n \rangle$ behavior can be inferred by a series expansion of Eqs. (44) and (45) in powers of ρ_u

$$f_{\text{off}} = \frac{\sigma_b \rho_u}{\theta \Sigma_b + \sigma_u} + O(\rho_u^2), \quad (46)$$

$$\langle n \rangle = \frac{(\theta + \sigma_u)\rho_u}{\theta \Sigma_b + \sigma_u} + O(\rho_u^2). \quad (47)$$

From these expressions, one can deduce that

$$f_{\text{off}} \simeq \frac{\sigma_b \langle n \rangle}{\theta + \sigma_u}. \quad (48)$$

The intermediate $\langle n \rangle$ behavior is however unknown. The dependence of f_{off} with $\langle n \rangle$ is most easily explored by numerically evaluating Eqs. (44) and (45) for various values of ρ_u . In Fig. 1 we show two plots generated in this manner, one for σ_b small and the other for σ_b large, both with $\theta = 0$. In each case we compare with the Hill function

$$f_{\text{off}}^* = \frac{\langle n \rangle}{\langle n \rangle + \frac{\theta + \sigma_u}{\sigma_b}}. \quad (49)$$

Note that this function's small and large $\langle n \rangle$ dependence are the same as those of the exact solution. Furthermore this function is the prediction of the deterministic model (see Appendix B). From Fig. 1, we see that the Hill function is a good approximation to the actual function for small σ_b ; in the opposite limit of large σ_b , the two functions are considerably different for intermediate $\langle n \rangle$. In this case the exact solution shows a piecewise linear form, with the linear dependence of f_{off} approximately holding until the function 'breaks' at its threshold value of unity.

As a last comment in this subsection, defining by $\langle n \rangle_0$ the average of n conditioned on the promoter being unbound, i.e.

$$\langle n \rangle_0 = \frac{\sum_{n=0}^{\infty} n P_0(n)}{\sum_{n=0}^{\infty} P_0(n)}, \quad (50)$$

we find the simple form

$$\langle n \rangle_0 = \frac{\rho_u \tilde{A}(\theta + \sigma_u) M(\alpha, \beta, w_1)}{(1 - f_{\text{off}})}. \quad (51)$$

Comparing this with Eq.(44) we have

$$\langle n \rangle_0 = \frac{(\theta + \sigma_u)}{(\Sigma_b - 1)} \frac{f_{\text{off}}}{(1 - f_{\text{off}})}, \quad (52)$$

which can be inverted to obtain the curious result

$$f_{\text{off}} = \frac{\langle n \rangle_0}{\langle n \rangle_0 + \frac{(\theta + \sigma_u)}{\sigma_b}}. \quad (53)$$

Thus, an equation resembling the Hill function is found, but the average of n is replaced by the average of n conditioned on the promoter being unbound (i.e. the gene being switched

on). In fact, one can show that this relationship holds quite generally (i.e. for arbitrary values of ρ_b) by summing the master equation (8) for $P_0(n)$ over n and using the definitions of f_{off} and $\langle n \rangle_0$.

2. The case $\alpha = 0$

We saw in the subsection on detailed balance conditions that when $\alpha = 0$ the Kummer function is a constant, and G_1 then simplifies to an exponential function, giving Poisson statistics for $P_1(n)$ (and also for $P_0(n)$ in that case). The cases considered in this subsection and the next are two further cases in which the exact solution for G_1 reduces to a pure exponential function. Note, we have previously defined $R = \rho_u - \rho_b \Sigma_b$. We will not be using R for this case and the next one, as it is helpful to see the role of ρ_u and ρ_b explicitly.

In setting $\alpha = 0$, but without imposing the two conditions for detailed balance, we have

$$G_1(z) = Ae^{\rho_b z} , \quad (54)$$

and Poisson statistics for $P_1(n)$. The condition $\alpha = 0$ can be written

$$\theta = \frac{\sigma_u(\rho_u - \rho_b)}{(\rho_b \Sigma_b - \rho_u)} . \quad (55)$$

To hold, this relation requires a quite severe constraint on the range of ρ_u , namely $\rho_b \leq \rho_u < \rho_b \Sigma_b$. Given the form of G_1 in Eq. (54), one can use Eq. (24) to retrieve the simple exponential form for G_0 and the normalisation condition and Eq. (28) to fix the two arbitrary constants that arise. After some algebra one finds

$$P_0(n) = \frac{e^{-\rho_b}}{\left(1 + \frac{(\rho_b \Sigma_b - \rho_u)}{\sigma_u}\right)} \frac{\rho_b^n}{n!} , \quad (56)$$

$$P_1(n) = \frac{e^{-\rho_b}}{\left(1 + \frac{\sigma_u}{(\rho_b \Sigma_b - \rho_u)}\right)} \frac{\rho_b^n}{n!} . \quad (57)$$

The mean number of free proteins has the particularly simple form $\langle n \rangle = \rho_b = r_b/k_f$.

In this and the next subsection the special cases impose non-trivial relationships between parameters (see Eqs (55) and (59)). As such, these solutions are valid on hypersurfaces in parameter space. Because of this non-trivial relationship between parameters, one cannot change $\langle n \rangle$ through variation of a single parameter, and as such the functional relationship $f_{\text{off}}(\langle n \rangle)$ is of limited experimental interest, and thus we do not report such results here.

3. The case $\alpha = \beta$

In setting $\alpha = \beta$ we can take advantage of the fact that $M(\alpha, \alpha, w) = e^w$ [29]. Thus, using Eqs. (19) and (23) we have

$$G_1(z) = A' e^{\rho_u z / \Sigma_b}, \quad (58)$$

and, consequently, Poisson statistics for $P_1(n)$. Using Eqs. (21) and (22), the condition $\alpha = \beta$ translates to fixing σ_u in terms of ρ_u, ρ_b , and Σ_b , independent of θ , as follows

$$\sigma_u = \frac{(\rho_u - \rho_b \Sigma_b)(\Sigma_b^2 + \rho_u(\Sigma_b - 1))}{\rho_u \Sigma_b (\Sigma_b - 1)}. \quad (59)$$

Note, this condition requires $\rho_u > \rho_b \Sigma_b$, and so this condition has no overlap with the condition $\alpha = 0$ studied above. The implicit reason for this stems from the definitions of α and β : while the former can take a value of zero, the latter is always greater than 1.

One can now substitute the expression for $G_1(z)$ into Eq. (24), and find the explicit form for $G_0(z)$

$$G_0(z) = \frac{A'}{(\Sigma_b - 1)} \left[\left(\frac{\rho_u - \rho_b \Sigma_b}{\rho_u} \right) z + \frac{\Sigma_b}{\rho_u} \left(\theta + \frac{(\rho_u - \rho_b \Sigma_b)}{\rho_u (\Sigma_b - 1)} \right) \right] e^{\rho_u z / \Sigma_b}. \quad (60)$$

Note, the arbitrary constant which arises from integrating Eq. (24) is found to be zero on application of the condition (28). Given the z dependence of the prefactor to the exponential, $P_0(n)$ will not have a purely Poisson form.

After some algebra, one finds the explicit forms for the probability distributions:

$$P_0(n) = \frac{A'}{(\Sigma_b - 1)} \left[\left(\frac{\rho_u - \rho_b \Sigma_b}{\rho_u} \right) \frac{(\rho_u / \Sigma_b)^{n-1}}{(n-1)!} + \frac{\Sigma_b}{\rho_u} \left(\theta + \frac{(\rho_u - \rho_b \Sigma_b)}{\rho_u (\Sigma_b - 1)} \right) \frac{(\rho_u / \Sigma_b)^n}{n!} \right], \quad (61)$$

$$P_1(n) = A' \frac{(\rho_u / \Sigma_b)^n}{n!}, \quad (62)$$

where

$$A' = \frac{\rho_u^2 (\Sigma_b - 1)^2 e^{-\rho_u / \Sigma_b}}{\Sigma_b [(\rho_u - \rho_b \Sigma_b) + \rho_u (\Sigma_b - 1)(\theta + \rho_u - \rho_b)]}. \quad (63)$$

Note, Eq. (61) holds for $n = 0$ with the understanding that $1/(-1)! = 0$.

Using these distributions, we find

$$\langle n \rangle = \rho_u \left[\frac{(\rho_u - \rho_b \Sigma_b) \Sigma_b + \rho_u (\Sigma_b - 1)(\theta + \rho_u - \rho_b)}{(\rho_u - \rho_b \Sigma_b) + \rho_u (\Sigma_b - 1)(\theta + \rho_u - \rho_b)} \right]. \quad (64)$$

Note that $\langle n \rangle > \rho_u$, and hence $\langle n \rangle > \rho_u / \Sigma_b$ which is the parameter in the Poisson-like distributions.

V. NUMERICAL VALIDATION OF THE EXACT SOLUTION

In this section we numerically solve the master equations, Eqs. (8)-(9) and compare with the exact solutions obtained in Section III. $2N$ difference equations are generated by substituting $n = 0, 1, 2, \dots, N - 1$ in Eqs. (8)-(9) with the time derivative set to zero. The boundary conditions are set to $P_0(-1) = P_1(-1) = P_0(N) = P_1(N) = 0$, taking into account also the absence of probability flux into $P_0(N)$ and $P_1(N)$. This set of difference equations is solved simultaneously for $P_0(0), P_1(0), P_0(1), P_1(1), \dots, P_0(N - 1), P_1(N - 1)$. Note that the exact solution corresponds to N equals positive infinity. Of course practically we are only interested in obtaining the probability distribution solution to some desired accuracy and hence it is sufficient to solve the equations for a large enough positive integer N . This should be chosen large enough such that the probability distribution solution $P_0(n) + P_1(n)$ smoothly decays to zero as n approaches N .

We use the latter method with $N = 500$ to obtain the dependence of the steady-state probability distribution solution of the master equations, Eqs. (8)-(9), on the five non-dimensional parameters $\theta, \sigma_b, \sigma_u, \rho_b$ and ρ_u . The same is obtained by means of the analytical solutions given by Eqs. (26), (27) and (29). The results from the two methods are compared in Fig. 2 where the open circles show the numerics and the crosses show the analytical solution. Note that in all cases, $P(n)$ goes to zero for n much less than 500 and hence artificial boundary effects due to finite N should be negligible; indeed we verified that the probability distribution solution obtained with $N = 1000$ is indistinguishable from the one obtained with $N = 500$. Note that the numerics and the analytical solution are in perfect agreement which indeed verifies the correctness of the main result of this paper, namely the exact solution given by Eqs. (26), (27) and (29).

It is interesting that in (a), (b), as we gradually increase θ and σ_b respectively, we observe a transition from a bimodal to a unimodal probability distribution while in (c) we see the reverse transition as the parameter ρ_u is increased. The bimodal character of the distribution is particularly interesting since the deterministic model of the genetic feedback loop does not exhibit bistability (see Appendix B). The mechanism behind the origin of bimodality and the transition from bimodal to unimodal behavior can be inferred as follows. In cases (a)-(c), the peaks of the bimodal distribution occur at $n \simeq \rho_u$ and at $n \simeq \rho_b$ while the single peak of the unimodal distribution occurs at one of these two (depending on parameter values). Now the

processes $D_u \xrightarrow{r_u} D_u + P$, and $P \xrightarrow{k_f} \emptyset$ considered by themselves lead to a Poisson distribution for the number of protein molecules with peak at $n \simeq \rho_u$ in steady-state conditions; similarly a peak at $n \simeq \rho_b$ can be associated with the processes $D_b \xrightarrow{r_b} D_b + P$, and $P \xrightarrow{k_f} \emptyset$. Hence it is clear that bimodality occurs whenever the gene switches slowly between its bound and unbound states which leads to a switch between the two sets of reactions discussed above. Inspection of the reaction mechanism Eq. (1) shows that the switching rates increase with θ , σ_b and σ_u . Hence if for some parameter set we have bimodality, increasing any one of the aforementioned three parameters will lead to a switch from bimodal to unimodal behavior; these are cases (a) and (b). If we have unimodality then the distribution will of course stay unimodal upon variation of one of the three parameters; this is case (d). Slow transitions between unbound and bound states are necessary but not sufficient to induce bimodality; the protein production rates of the two gene states must be sufficiently different such that the peaks of the Poisson distributions associated with each state are well separated; this is case (c). Cases (a)-(d) are ones in which the genetic feedback loop is negative, i.e. $\rho_b \leq \rho_u$. In contrast cases (e) and (f) are for a positive feedback loop, i.e. $\rho_b > \rho_u$. As for the negative feedback case, both unimodal (case (e)) and bimodal behaviors (case (f)) are possible and their existence can be understood by the same switching mechanism elucidated above. For example the case $\theta = 0$ in (f) is bimodal with peaks at $n \simeq \rho_u$ and $n \simeq \rho_b$ indicating slow switching between the steady-states of the bound and unbound genes. Increasing θ leads to an increase in the switching rate from the bound to the unbound states explaining the increase in the size of the peak associated with the unbound state and the corresponding decrease of the peak size associated with the bound state.

VI. CRITIQUE OF A PREVIOUS “EXACT” SOLUTION

In this section we make a careful and explicit comparison of our master equations with those of Hornos et al [19] in the original paper claiming an exact solution to the problem of a self-regulating gene with the condition $k_f = k_b$ (i.e. $\theta = 1$). In that work, the variable n of the probability distributions represents the *total number* of proteins in the system, that is the number of free proteins plus (when the promoter is bound) the bound protein.

The following key provides the translation between the notation we have used in section

II and the notation of Hornos et al.

$$P_0(n, t) \longleftrightarrow \alpha_n(t) \quad (65)$$

$$P_1(n, t) \longleftrightarrow \beta_{n+1}(t) \quad (66)$$

$$r_u \longleftrightarrow g_\alpha \quad (67)$$

$$r_b \longleftrightarrow g_\beta \quad (68)$$

$$k_f \longleftrightarrow k_f \quad (69)$$

$$k_b \longleftrightarrow k_b \quad (70)$$

$$s_u \longleftrightarrow f \quad (71)$$

$$s_b \longleftrightarrow h \quad (72)$$

Note, in their original work, Hornos et al assume from the outset that the rates of degradation of free and bound protein are the same, and use the symbol k for this degradation rate. Here, for clarity in the handling of these two different processes, we allow the rates to be different, and use the symbols k_f and k_b consistent with the notation used in section II, setting them equal eventually.

With the key given above, the correct master equations (2) and (3) take the form

$$\begin{aligned} \frac{d}{dt}\alpha_n(t) &= g_\alpha (\alpha_{n-1} - \alpha_n) \\ &+ k_f ((n+1)\alpha_{n+1} - n\alpha_n) \\ &+ k_b\beta_{n+1} + f\beta_n - hn\alpha_n , \end{aligned} \quad (73)$$

$$\begin{aligned} \frac{d}{dt}\beta_n &= g_\beta (\beta_{n-1} - \beta_n) \\ &+ k_f (n\beta_{n+1} - (n-1)\beta_n) \\ &- k_b\beta_n - f\beta_n + hn\alpha_n . \end{aligned} \quad (74)$$

The first equation is valid for $n \geq 0$ while the second equation is valid for $n \geq 1$. The boundary conditions $P_0(-1, t) = 0$ and $P_1(-1, t) = 0$ imply the new boundary conditions $\alpha_{-1} = \beta_0 = 0$.

On setting the free and bound degradation rates to be equal $k_f = k_b = k$, the master

equations above take the form

$$\begin{aligned}
\frac{d}{dt}\alpha_n(t) &= g_\alpha (\alpha_{n-1} - \alpha_n) \\
&+ k((n+1)\alpha_{n+1} - n\alpha_n) \\
&+ k\beta_{n+1} + f\beta_n - hn\alpha_n ,
\end{aligned} \tag{75}$$

$$\begin{aligned}
\frac{d}{dt}\beta_n &= g_\beta (\beta_{n-1} - \beta_n) \\
&+ k(n\beta_{n+1} - (n-1)\beta_n) \\
&- k\beta_n - f\beta_n + hn\alpha_n .
\end{aligned} \tag{76}$$

We are now in a position to directly compare these master equations with those of Hornos et al. Equations (1) and (2) of that paper read (for $n > 0$)

$$\begin{aligned}
\frac{d}{dt}\alpha_n(t) &= g_\alpha (\alpha_{n-1} - \alpha_n) \\
&+ k((n+1)\alpha_{n+1} - n\alpha_n) \\
&+ f\beta_n - hn\alpha_n ,
\end{aligned} \tag{77}$$

$$\begin{aligned}
\frac{d}{dt}\beta_n &= g_\beta (\beta_{n-1} - \beta_n) \\
&+ k(n\beta_{n+1} - (n-1)\beta_n) \\
&+ k(\beta_{n+1} - \beta_n) - f\beta_n + hn\alpha_n .
\end{aligned} \tag{78}$$

Obviously, in the second and third lines of Eq. (78) one can combine the terms with a prefactor of k to give $k((n+1)\beta_{n+1} - n\beta_n)$, but we have separated these terms to emphasise the comparison to the correct equation (76) in which, as stressed in section II, each line of the equation corresponds to processes occurring on the gene, the cytosol, and the promoter respectively. In order to conserve probability, Hornos et al are obliged to write a separate equation for $n = 0$ which reads

$$\frac{d}{dt}\alpha_0(t) = -g_\alpha\alpha_0 + k(\alpha_1 + \beta_1) , \tag{79}$$

which is consistent with Eq. (75) for $n = 0$ and using the boundary condition $\alpha_{-1} = 0$. They also enforce the boundary condition $\beta_0 = 0$. The same master equations, Eqs. (77) and (78), appear also in several follow up papers [20–22].

Let us denote the master equations (75) and (76), as ME(GSN), and the master equations (77) and (78) as ME(H-ET-AL). Clearly ME(GSN) and ME(H-ET-AL) are different, and given they purport to describe the same process, they cannot both be correct. The difference between the two sets of equations is in how degradation of the bound protein is described. In our master equations ME(GSN), which were derived from Eqs. (2) and (3) by use of the key, if the bound protein is degraded, the system returns to the unbound state, i.e., that process connects the two conditional probabilities α_n and β_{n+1} . In ME(H-ET-AL), degradation of the bound protein keeps the system in the bound state. How is this possible? Literally, the ME(H-ET-AL) master equations are describing bound protein degradation as the following composite of processes: degradation of the bound protein, with instantaneous rebinding of a protein to the promoter from the free protein pool, such that the state β_{n+1} transitions to the state β_n . This composite of processes occurs with a constant rate k , independent of the number of free proteins. Clearly this composite of processes is not the intended instantiation of the self-regulating gene, and furthermore, it is difficult to imagine any physical process that could operate in this manner. One could concoct an imaginary process using a Maxwell Demon, a tiny animated figure, who sits by the promoter with a free protein always to hand, to instantaneously latch onto the promoter should the bound protein be degraded, but such constructions belong to philosophical discussions of irreversibility, and are not relevant to the biological question at hand.

One might argue that the manner in which the bound protein degradation is handled has little bearing on the form of the distributions. To test this, we used the method expounded in Sec. V to numerically solve the two master equations, ME(GSN) and ME(H-ET-AL), in steady-state conditions for four different sets of parameter values. The results are shown in Fig. 3, where the open circles show the predictions of ME(GSN) while the solid lines show the prediction of ME(H-ET-AL). The parameters g_α , g_β , and k are, in all cases, equal to 80.0, 0.0, and 1.0 respectively, while $h = 0.001\Lambda$ and $f = 0.1\Lambda$ where Λ takes the values 0.01 in Fig. 3(a), 1 in Fig. 3(b), 10 in Fig. 3(c) and 100 in Fig. 3(d). Note that as Λ is varied from 0.01 to 100, the ME(H-ET-AL) predicts a transition from unimodal to bimodal probability distribution and back to unimodal distribution. However no such transitions are seen in ME(GSN). Note that the case $\Lambda = 1$ in Fig. 3(b) corresponds to the exact set of parameters used in the case $\omega = 0.1$ in Fig. 1 of the Hornos et al paper [19]. The good agreement between ME(GSN) and ME(H-ET-AL) for cases (a) and (d) can be explained as follows. In

case (a), the rate of protein binding to the promoter is very small, the gene is most of the time in the unbound state and hence bound protein degradation rarely occurs. In case (d), the gene switches between the unbound and unbound states very rapidly but its decay to the unbound state occurs primarily via the reaction $D_b \rightarrow D_u + P$ and only occasionally via the reaction $D_b \rightarrow D_u$. In both cases, bound protein degradation is a rare event compared to the other molecular processes and hence it follows that the form of the probability distribution is practically insensitive to whether bound protein degradation is correctly or incorrectly described. In cases (b) and (c), bound protein degradation events occur at a frequency comparable with that of other molecular processes, and hence its correct description using the ME(GSN) becomes crucial to obtaining the correct steady-state probability distribution. The discussion on the role of θ , particularly when it approaches unity, explains why the bimodal distribution is not to be expected in a correct formulation of this problem.

It is worth mentioning that a cursory comparison of our results with those of Hornos et al indicates some similarities, which are in some cases superficial. For example, Hornos et al find Kummer function solutions for the generating functions. Note, in the solution of ME(H-ET-AL), it is the generating function conditioned on the unbound promoter (G_0 in our notation) which is expressed in terms of a single Kummer function, while the generating function conditioned on the bound promoter (G_1 in our notation) is expressed in terms of a sum of Kummer functions. In our exact solution of the correct master equation ME(GSN) we find that G_1 is the generating function which has the simple form in terms of a single Kummer function, as shown in Eq.(23), while G_0 involves integrals of Kummer functions, and does not have a simple closed form expression. We also note that Figure 2 of Hornos et al shows a plot of f_{off} versus $\langle n \rangle$, which indeed continuously varies from the Hill form to a linear piece-wise form as σ_b is decreased, similar to our result shown in Figure 1. In the caption of Figure 2 of their paper, the curious relation between f_{off} and $\langle n \rangle_0$ is noted, similar to our Eq. (53), although the combination $(\theta + \sigma_u)/\sigma_b$ in our equation is replaced by σ_u/σ_b in theirs, indicating again the incorrect handling of protein degradation in that work.

Qian et al [23] studied the case where the bound protein does not degrade (i.e. $\theta = 0$). The coupled master equations, Eqs. 17(a)-17(b) in their paper, are the same as our master equations, Eqs. (73)-(74), with $k_b = 0$. Hence their master equations are correct for this special case. However they do not solve these equations exactly. Rather they derive an approximative solution in the limit that the gene switches between the D_u and D_b states

very rapidly and in the limit that the switching occurs very slowly. In the latter limit, i.e., the limit of small protein binding and unbinding rates to the promoter, they show that the probability distribution is bimodal if the production rates of the gene in the bound and unbound states are sufficiently different. This is confirmed by our exact solution, see the case $\theta = 0$ in Fig. 2 (a). We find that this bimodal behaviour disappears when θ is increased from 0 to 1 (see Fig. 2 (b) and (c)).

VII. CONCLUSIONS

In this paper we have presented an exact solution for the simplest model of a self-regulating gene. Our solution is valid for an arbitrary degradation rate of the bound protein, which we have denoted throughout the paper by the parameter θ in dimensionless units. The explicit solution for the probability distributions for the number of free proteins, conditioned on the gene's promoter being bound or unbound, are given by Eqs. (26) and (27) respectively. It is interesting that an exact solution for this general case can be found given that the model breaks detailed balance and includes a bimolecular reaction step, features not typical of the exact solutions reported in the literature [4–9, 11]. In particular, to the best of our knowledge, our exact solution is a first for a gene regulatory network with a feedback loop. As we have shown, the previous exact solution claimed by Hornos et al [19] for the special case in which bound and free protein degradation are equal is incorrect because it is based on master equations which possess no coherent physical interpretation. The only other exact solution known in the context of stochastic gene expression is that derived by Shahrezaei and Swain for a model of gene expression involving first-order processes describing transcription, translation, protein degradation and mRNA degradation but no feedback loop [11].

We anticipate that our exact solution will be useful to explore the dependence of common experimental measures of noise intensity, e.g., the coefficient of variation [32] and Fano factors [33], on various parameter values and on the nature of the feedback loop (repressing or activating). This may lead to insights into the mechanisms used by cells to regulate fluctuations in the protein concentrations, a topic of intense current research [34, 35]. Other interesting avenues of research, which we have briefly touched upon in this paper, are the investigation of the transition from unimodal to bimodal protein distributions, and of de-

viations from the Hill function describing activation or repression of the gene. Our exact solution will enable a more thorough analysis of these topics which could previously only be investigated by means of approximation methods in some restricted parameter regimes.

For chemical systems involving bimolecular reactions, the moment equations obtained from the master equation cannot generally be solved in closed form, and thus various approximations of the master equation have been developed. In particular for systems composed of unimolecular and bimolecular reactions, the time-evolution equation of the M^{th} central moment of the probability density function solution of the master equation is generally a function of the $(M+1)^{th}$ central moment. This implies an infinite hierarchy of coupled equations which cannot be generally solved [36], although see [37] for an exception. In the limit of intermediate or large numbers of molecules, various methods have been developed to obtain approximate expressions for the moments (examples of two widely used methods are the system-size expansion [26, 38–42] and moment-closure approximations [36, 43–46]). However the reliability and accuracy of these methods when applied to systems characterized by low copy number of molecules and bimolecular reaction steps has remained an outstanding question of practical interest. Hence we anticipate that our exact solution will also provide a useful benchmark with which to compare the gamut of approximation methods used to estimate the effect of noise in biochemical systems.

ACKNOWLEDGMENTS

Two of the authors (DRS and TJN) gratefully acknowledge the 2007-08 program on Mathematics of Molecular and Cellular Biology, organised by the Institute for Mathematics and its Applications, University of Minnesota, where this project was initiated. RG and TJN acknowledge support from SULSA (Scottish Universities Life Sciences Alliance). DRS acknowledges support from NSF agreement 0112050 through the Mathematical Biosciences Institute.

[1] Y. Ishihama, T. Schmidt, J. Rappsilber, M. Mann, F. Hartl, M. Kerner and D. Frishman, BMC genomics **9**, 102 (2008).

- [2] R. Grima and S. Schnell, *Essays in Biochemistry* **45**, 41 (2008)
- [3] A. Eldar and M. B. Elowitz *Nature* **467**, 167 (2010)
- [4] H. Haken, *Synergetics* (Springer-Verlag, 1978)
- [5] I. J. Laurenzi, *J. Chem. Phys.* **113**, 3315 (2000)
- [6] The principle of detailed balance asserts that for any pair of states n and n' , the probability of being in state n multiplied by the probability per unit time of jumping from n to n' is equal to the probability of being in state n' multiplied by the probability per unit time of jumping from n' to n . Hence detailed balance implies that pairs of terms in the master equation are separately zero whereas steady-state conditions require that the whole sum of terms is zero. The stronger condition of local equilibrium imposed by detailed balance leads to an exact solution of the master equation.
- [7] I. G. Darvey and P. J. Staff, *J. Chem. Phys.* **44**, 990 (1966)
- [8] C. Gadgil, C. H. Lee and H. G. Othmer, *B. Math. Biol.* **67**, 901 (2005)
- [9] W. J. Heuett and H. Qian, *J. Chem. Phys.* **124**, 044110 (2006)
- [10] T. Jahnke and W. Huisinga, *J. Math. Biol.* **54**, 1 (2007)
- [11] V. Shahrezaei and P. S. Swain, *Proc. Nat. Acad. Sci. USA* **105**, 17256 (2008)
- [12] H. Qian, *J. Phys. Chem. B* **110**, 15063 (2006)
- [13] M. Hochstrasser, *Annu. Rev. Genet.* **30**, 405 (1996)
- [14] E. S. Johnson, D. K. Gonda and A. Varshavsky, *Nature* **346**, 287 (1990)
- [15] E. E. Selkov, *Eur. J. Biochem.* **4**, 79 (1968)
- [16] A. Goldbeter, *Nature* **253**, 540 (1975)
- [17] J. Lewis, *Curr. Biol.* **13**, 1398 (2003)
- [18] J. J. Tyson, C. I. Hong, C. Dennis Thron and B. Novak, *Biophys. J.* **77**, 2411 (1999)
- [19] J. E. M. Hornos, D. Schultz, G. C. P. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, *Phys. Rev. E* **72**, 051907 (2005)
- [20] D. Schultz, J. N. Onuchic, and P. G. Wolynes, *J. Chem. Phys.* **126**, 245102 (2007)
- [21] A. F. Ramos, G. C. P. Innocentini, F. M. Forger, J. E. M. Hornos, *IET Sys. Biol.* **4**, 311 (2010)
- [22] A. F. Ramos, G. C. P. Innocentini, J. E. M. Hornos, *Phys. Rev. E.* **83**, 062902 (2011)
- [23] H. Qian, P-Z Shi and J. Xing, *Phys. Chem. Chem. Phys.* **11**, 4861 (2009)
- [24] R. Bundschuh, F. Hayot and C. Jayaprakash, *J. Theor. Biol.* **220**, 261 (2003)

- [25] H. Feng, B. Han and J. Wang, *J. Phys. Chem. B* **115**, 1254 (2011)
- [26] N. G. van Kampen, *Stochastic processes in physics and chemistry* (Elsevier, 1992)
- [27] D. A. McQuarrie, *J. Appl. Prob.* **4**, 413 (1967)
- [28] C. W. Gardiner, *Handbook of Stochastic Methods for physics, chemistry and the natural sciences* (Springer, 2004)
- [29] M. Abramowitz, M. and I. A. Stegun, *Handbook of mathematical functions* (Dover, 1972)
- [30] O. G. Berg, J. Paulsson, M. Ehrenberg, *Biophys. J.* **79**, 2944 (2000)
- [31] A. Fersht, *Structure and mechanism in protein science* (WH Freeman, 1998)
- [32] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel, and N. Barkai, *Nature Genetics* **38**, 636 (2006)
- [33] Y. Taniguchi, P. J. Choi, G-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, *Science* **329**, 533 (2010)
- [34] J. M. Raser and E. K. O'Shea, *Science* **309**, 2010 (2005)
- [35] C. V. Rao, D. M. Wolf and A. P. Arkin, *Nature* **420**, 231 (2002)
- [36] C. A. Gomez-Uribe and G. C. Verghese, *J. Chem. Phys.* **126**, 024109 (2007)
- [37] M. O. Stefanini, A. J. McKane, and T. J. Newman, *Nonlinearity* **18**, 1575 (2005)
- [38] J. Elf and M. Ehrenberg, *Genome research* **13**, 2475 (2003)
- [39] A. J. McKane, J. D. Nagy, T. J. Newman, and M. O. Stefanini, *J. Stat. Phys.* **128** 165 (2007)
- [40] R. Grima, *BMC Sys. Biol.* **3**: 101 (2009)
- [41] R. Grima, *J. Chem. Phys.* **133**, 035101 (2010)
- [42] R. Grima, *Phys. Rev. E* **84**, 056109 (2011)
- [43] J. Goutsias, *Biophys. J.* **92**, 2350 (2007)
- [44] L. Ferm, P. Lotstedt and A. Hellander, *J. Sci. Comput.* **34**, 127 (2008)
- [45] M. Ullah and O. Wolkenhauer, *J. Theor. Biol.* **260**, 340 (2009)
- [46] R. Grima, *J. Chem. Phys.* **136**, 154105 (2012)

Appendix A: The case $R=0$

The exact solution presented in Eqs. (19), (21), (22) and (23) assumes that the parameter $R = \rho_u - \rho_b \Sigma_b$ is different to zero. The case $R = 0$ requires a separate analysis, either by taking the Kummer function solution (23) and taking a careful limit, or else by reexamining

the second-order differential equation (14) explicitly for $R = 0$. For illustrative purposes, we demonstrate the second method here, and derive an exact expression for $G_1(z)$ and $P_1(n)$. These calculations can be extended to derive the exact form for $P_0(n)$, if desired by the reader.

Starting with Eq.(14) we set $R = 0$ by writing $\rho_b = \rho_u/\Sigma_b$. We make the double transformation

$$G_1(z) = \exp(\rho_u/\Sigma_b) \tilde{G}_1(z) , \quad (\text{A1})$$

and

$$u = (z - 1/\Sigma_b)^{1/2} , \quad (\text{A2})$$

to obtain the differential equation for $\tilde{G}_1(u)$ which has the form

$$\frac{d^2 \tilde{G}_1}{du^2} + \frac{(2\gamma + 1) d\tilde{G}_1}{u du} - \lambda \tilde{G}_1 = 0 , \quad (\text{A3})$$

where

$$\gamma = \theta + \frac{\sigma_u}{\Sigma_b} + \frac{\rho_u(\Sigma_b - 1)}{\Sigma_b^2} , \quad (\text{A4})$$

and

$$\lambda = \frac{4\rho_u\sigma_u(\Sigma_b - 1)}{\Sigma_b^2} . \quad (\text{A5})$$

Eq. (A3) can be directly related to the differential equation for the Bessel function [29], and we have

$$\tilde{G}_1(z) = Au^{-\gamma} I_\gamma(\lambda^{1/2}u) , \quad (\text{A6})$$

where A is a normalisation constant and I_γ is the modified Bessel function. Note, $I_{-\gamma}$ is an independent solution (for γ not equal to an integer), but can be discarded, since its asymptotic properties lead to a non-normalisable probability distribution.

In order to derive an explicit form for $P_1(n)$ we use Eq. (25), and the differentiation formula [29]

$$\left(\frac{1}{v} \frac{d}{dv}\right)^n v^{-\gamma} I_\gamma(v) = v^{-\gamma-n} I_{\gamma+n}(v) . \quad (\text{A7})$$

This provides the final result

$$P_1(n) = \frac{A'}{n!} \sum_{m=0}^n C_m^n \left(\frac{\rho_u}{\Sigma_b}\right)^{n-m} \left(\frac{\lambda}{2}\right)^m \left(\frac{\Sigma_b}{\lambda}\right)^{(m+\gamma)/2} J_{\gamma+m} \left((\lambda/\Sigma_b)^{1/2}\right) , \quad (\text{A8})$$

where A' is a normalisation constant. The reason the Bessel function J appears in P_1 rather than the modified Bessel function I , is due to the argument of Eq. (A6) becoming imaginary when $z = 0$.

Appendix B: Mean-field theory

We consider a fictitious experiment in which the cell membranes of a large number of identical cells enclosed in some reaction volume Ω are dissolved such that the genes and proteins from each individual cell can interact with that from every other cell. The well-mixed dynamics of this reaction system is deterministic (due to the large number of cells) and the state of the system at any point in time is described by two variables: the concentration of unbound DNA molecules, ϕ_u , and of the protein, ϕ . Note that the concentration of bound DNA molecules is $\phi_b = \phi_T - \phi_u$ (where ϕ_T is the concentration of bound and unbound DNA) and is hence not an independent variable. The conventional mass-action rate equations for the concentrations can be directly deduced by inspection of the reaction scheme Eq. (1) and are given by

$$\partial_t \phi_u = k_b(\phi_T - \phi_u) - k\phi_u\phi + s_u(\phi_T - \phi_u), \quad (\text{B1})$$

$$\partial_t \phi = r_u\phi_u + r_b(\phi_T - \phi_u) - k_f\phi - k\phi_u\phi + s_u(\phi_T - \phi_u). \quad (\text{B2})$$

Note that the bimolecular rate constant k is not s_b but rather is equal to $s_b\Omega$. This is since s_b is a transition rate with units of inverse time and hence is in reality equal to the macroscopic rate constant k (with units of volume divided by time) divided by the reaction volume. Note also that these equations are based on the implicit assumption that the covariance of fluctuations in the number of molecules of any pair of species is zero, i.e. fluctuations are not important.

In order to compare with the single gene results derived in the main text, we first multiply the above equations by the reaction volume Ω and then set this volume equal to the cellular volume, i.e. $\Omega\phi_T = 1$, which leads to mean-field equations for the average molecule numbers of unbound DNA $\langle n_u \rangle$ and of protein $\langle n \rangle$ in a cell:

$$\partial_t \langle n_u \rangle = k_b(1 - \langle n_u \rangle) - s_b \langle n_u \rangle \langle n \rangle + s_u(1 - \langle n_u \rangle), \quad (\text{B3})$$

$$\partial_t \langle n \rangle = r_u \langle n_u \rangle + r_b(1 - \langle n_u \rangle) - k_f \langle n \rangle - s_b \langle n_u \rangle \langle n \rangle + s_u(1 - \langle n_u \rangle). \quad (\text{B4})$$

These equations admit a single steady-state solution for $\langle n_u \rangle$ and $\langle n \rangle$ implying that the deterministic model does not predict bistability.

Within this approach, the quantity $1 - \langle n_u \rangle$ can be interpreted as the fraction of time that the promoter is bound, f_{off} . Substituting in the latter equation, the steady-state solution

of Eq. (B3) for $\langle n_u \rangle$ and using the dimensionless rates as given by Eq. (6), one obtains the Hill equation

$$f_{\text{off}} = \frac{\langle n \rangle}{\langle n \rangle + \frac{\theta + \sigma_u}{\sigma_b}}. \quad (\text{B5})$$

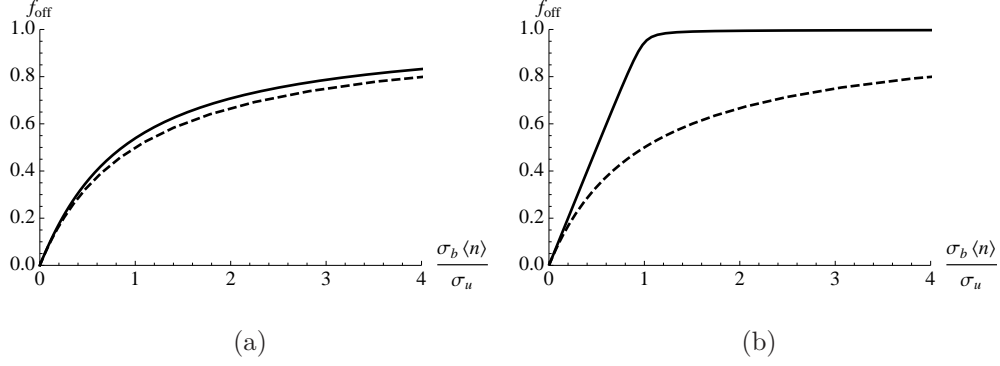


FIG. 1. Plot of f_{off} as a function of $\langle n \rangle$ for the case $\rho_b = 0$. The solid lines are generated by evaluating Eqs. (44) and (45) for varying values of ρ_u . We also plot, for comparison, the Hill function (dashed lines) given by Eq. (49), whose small and large $\langle n \rangle$ limits agree with those of the actual function (the one given by solid lines). Note that the Hill function is also the prediction of the deterministic model of the genetic feedback loop (Appendix B). The parameters are $\sigma_b = 0.01, \sigma_u = 2$ in (a) and $\sigma_b = 2, \sigma_u = 0.01$ in (b). In both cases $\theta = 0$. Note that the Hill function approximates well the actual function for small σ_b ; this is generally the case for all θ .

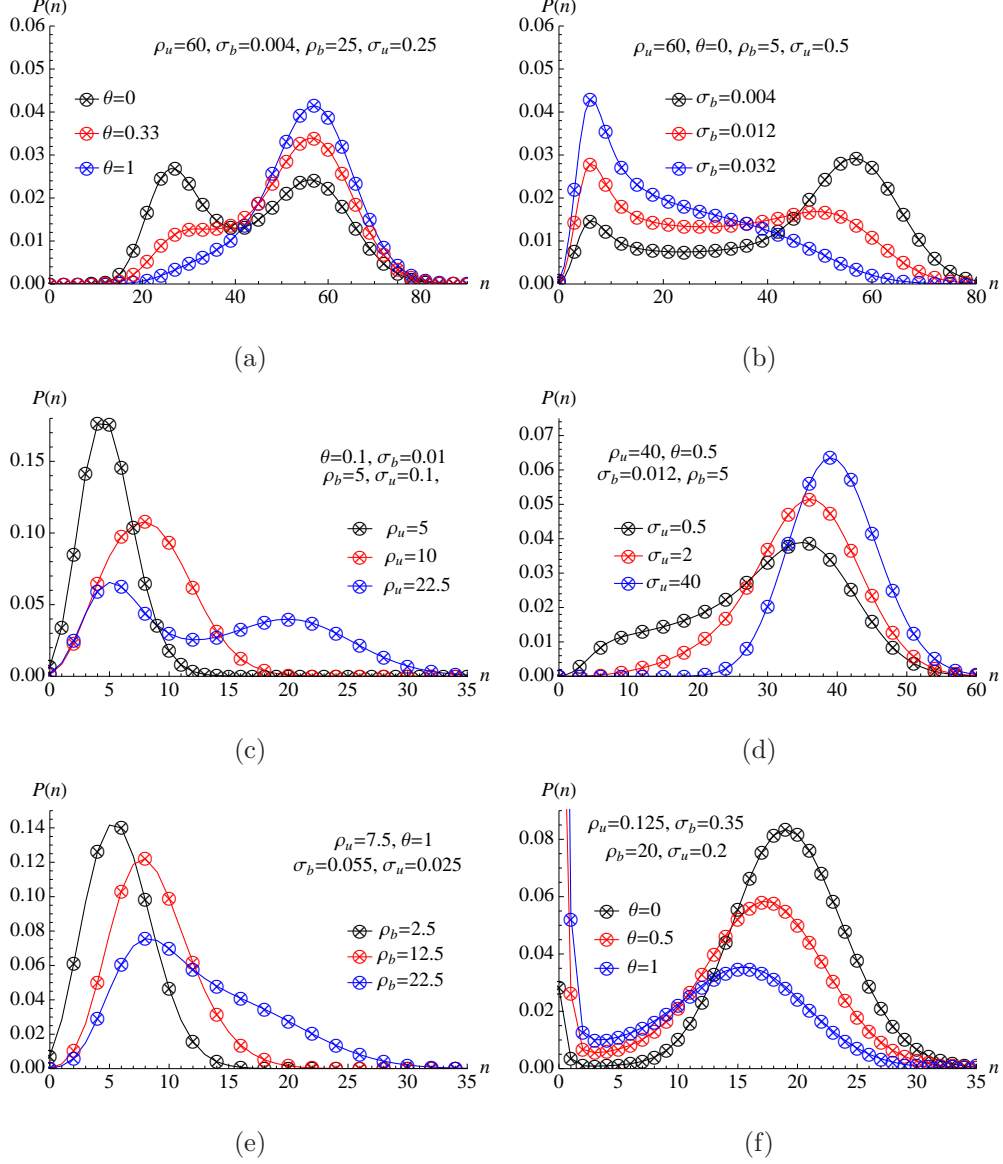


FIG. 2. Dependence of the steady-state probability distribution on the non-dimensional parameters θ (panel a and f), σ_b (panel b), ρ_u (panel c), σ_u (panel d), and ρ_b (panel e). The distributions are obtained by numerically integrating the master equations, Eqs. (8) and (9) (shown by the open circles), and by evaluating the analytical solutions, Eqs. (26), (27) and (29) (shown by the crosses). The perfect agreement of the two verifies that the latter equations are an exact solution of the master equation for the self-regulating gene. The solid lines are a guide to the eye.

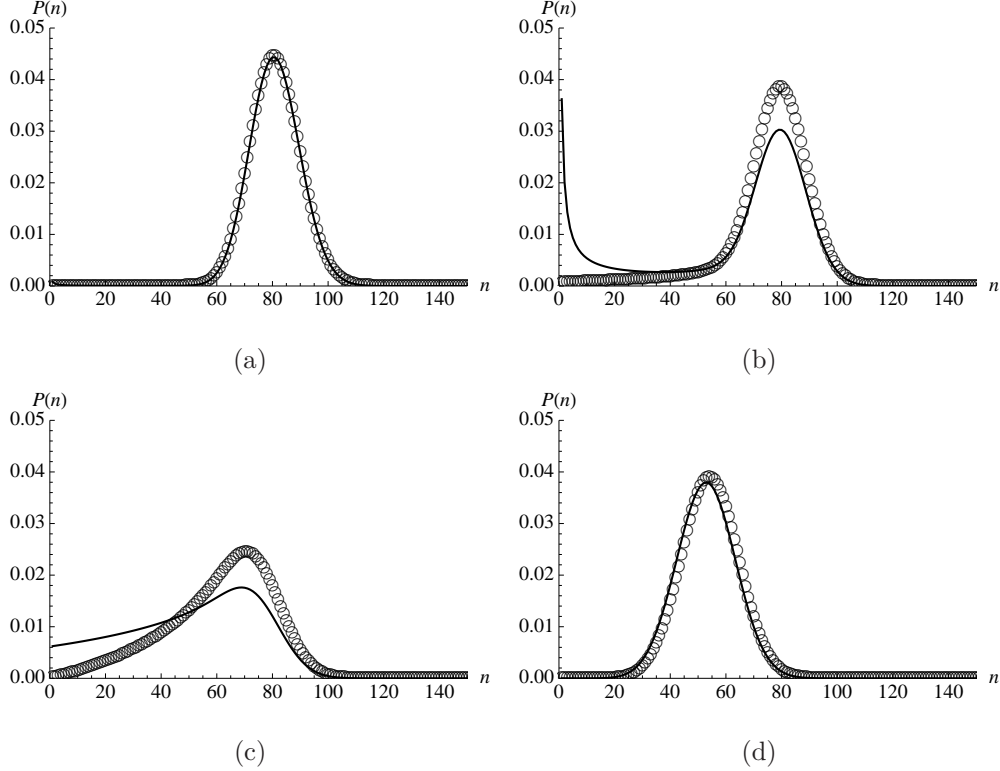


FIG. 3. Comparison of the steady-state probability distributions as predicted by the master equations ME(GSN) (open circles) and ME(H-ET-AL) (solid lines). The distributions are obtained by numerically integrating the two master equations to obtain $P(n) = \alpha_n + \beta_{n+1}$ as a function of n , the number of free proteins. The binding and unbinding rates of the protein to the promoter region, h and f respectively, take progressively larger values as we go from (a) to (d), while the rest of the parameters are fixed. Explicitly, g_α , g_β , and k are, in all cases, equal to 80.0, 0.0, and 1.0 respectively, while $h = 0.001\Lambda$ and $f = 0.1\Lambda$ where Λ takes the values 0.01 in (a), 1 in (b), 10 in (c) and 100 in (d). Note that the ME(H-ET-AL) predicts a transition from unimodal (a) to bimodal (b) and back to unimodal probability distribution (c) and (d), while the ME(GSN) predicts a unimodal distribution in all cases. This example shows that the incorrect handling of bound protein degradation by the ME(H-ET-AL) leads to qualitatively incorrect features of the steady-state probability distribution.