Supporting Text

Phenomenological Analysis of Lognormal-to-Gaussian Crossover of Protein Distribution

In the main paper the following expression was used to phenomenologically explain the crossover from a lognormal to a gaussian protein distribution:

$$\mathcal{P}(N) = \left(\frac{N + N_0}{\sqrt{2\pi}\sigma N N_0}\right) \exp\left[\frac{-\left(\ln\left(\frac{N}{N_0}\right) + \frac{N}{N_0} - \mu\right)^2}{2\sigma^2}\right] \qquad [1]$$

Fig. 9 shows various plots of expression (1) for different values of μ, with all other parameters fixed ($N_0 = 1500, \sigma = 1$). As the value of μ increases, the function makes a crossover from a lognormal to a gaussian form. The function is distinctly long-tailed for negative μ values, and also for small positive values. When μ is postive and large, the function is more symmetric.

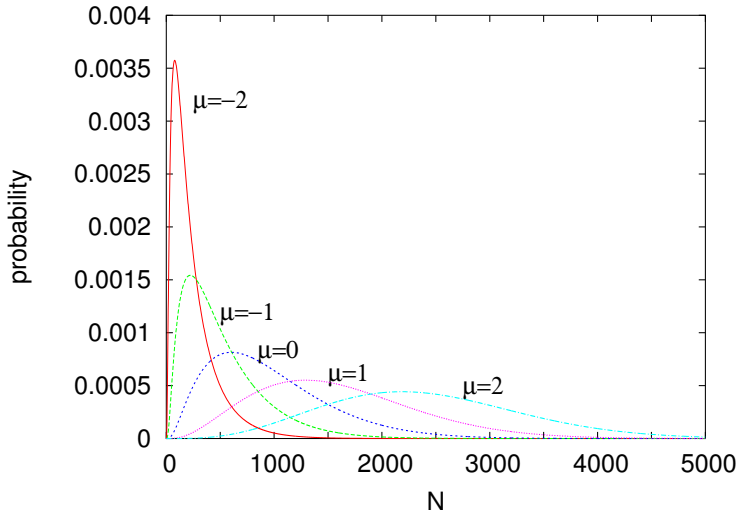
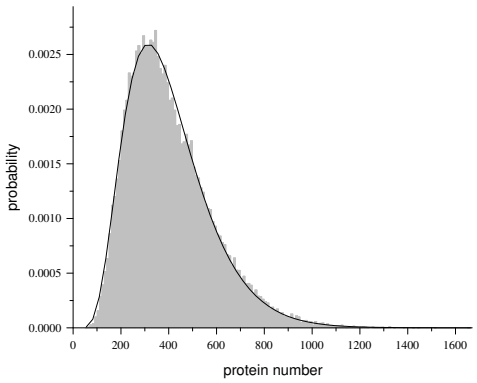


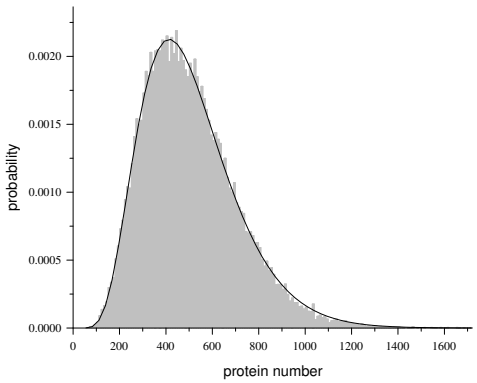
Fig. 9. Plots of expression (1) for various values of μ, with $N_0 = 1500, \sigma = 1$.

The fits of expression (1), with $N_0 = 1500$, to experimentally observed protein distributions from samples taken at 5, 7, 9, 11, 13 and 15 hours after innoculation are shown in Fig. 10. The fits were obtained from a nonlinear Levenberg-Marquardt least-squares fitting algorithm.

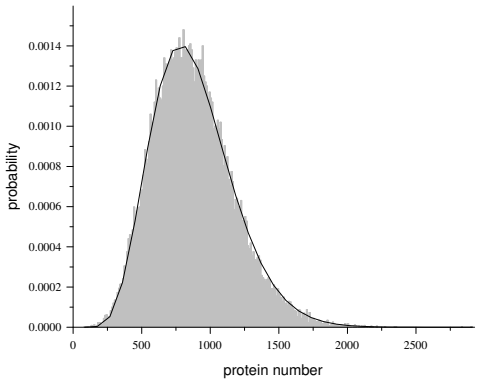
a) 5 hours



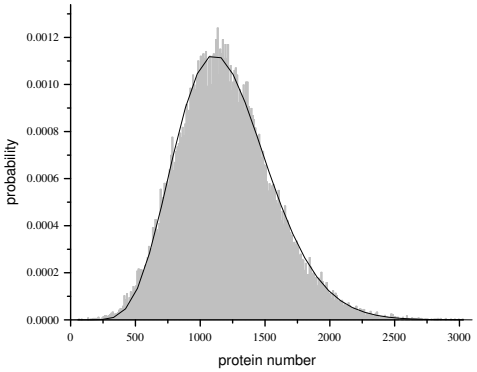
b) 7 hours



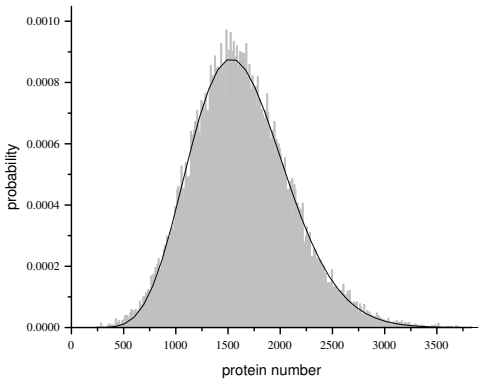
c) 9 hours



d) 11 hours



e) 13 hours



f) 15 hours

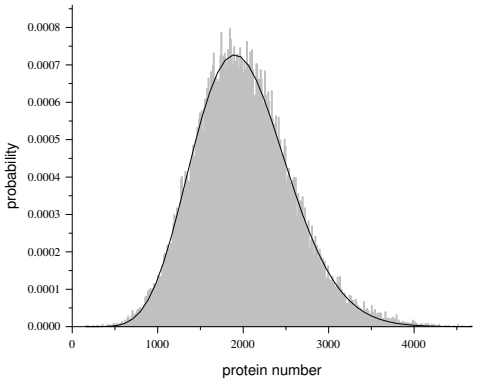


Fig. 10. Fit of expression (1), with $N_0 = 1500$, to experimental data. Best fit parameters: (a) $\mu = -1.145, \sigma = 0.551$; (b) $\mu = -0.828, \sigma = 0.547$; (c) $\mu = -0.002, \sigma = 0.536$; (d) $\mu = 0.537, \sigma = 0.547$; (e) $\mu = 1.129, \sigma = 0.0.594$; (f) $\mu = 1.586, \sigma = 0.649$.

Stochastic Simulation of a Chemical Model of Gene Expression

Operator Regulation of an Isolated Gene. We represent the process of expression of a single gene with one operator site using the following chemical reactions, based on ref. 1:

Repressor binding

- $D + R \rightleftharpoons \{D.R\} \quad k_{fwd} = 10^8 M^{-1} s^{-1}, k_{back} = 1 s^{-1}$

Transcription

- $D + RNAP \rightleftharpoons \{D.RNAP\} \quad k_{fwd} = 10^8 M^{-1} s^{-1}, k_{back} = 10 s^{-1}$
- $\{D.RNAP\} \rightarrow TrRNAP \quad k_{fwd} = 0.1 s^{-1}$
- $TrRNAP \rightarrow RBS + D + RNAP \quad k_{fwd} = 0.3 s^{-1}$

Translation

- $RBS \rightarrow \phi \quad k_{fwd} = 0.3 s^{-1}$
- $RBS + Rib \rightleftharpoons \{RibRBS\} \quad k_{fwd} = 10^8 M^{-1} s^{-1}, k_{back} = 2.25 s^{-1}$
- $\{RibRBS\} \rightarrow ElRib + RBS \quad k_{fwd} = 0.5 s^{-1}$
- $ElRib \rightarrow Protein \quad k_{fwd} = 0.015 s^{-1}$

Protein folding and decay

- $Protein \rightarrow FoldedProtein \quad k_{fwd} = (\ln 2/5400)\, s^{-1}$
- $Protein \rightarrow \phi \quad k_{fwd} = 10^{-5} s^{-1}$
- $FoldedProtein \rightarrow \phi \quad k_{fwd} = 10^{-5} s^{-1}$

Parameter values have been chosen from ref. [1], except the following: Protein folding time has been taken to match the GFP folding time, which is characterised by a half life of 90 minutes [2]. Further, protein decay times have been taken as very large because the GFP used in the experiments does not decay appreciably over the duration of the experiment [2]. The rate constants have been kept fixed for all reactions, except for the transition from closed to open complex which is a function of the strength of the promoter, and therefore, we have modeled promoters of different strengths, following ref. [3], by varying that rate constant.

The Gillespie Algorithm. Deterministic methods of simulating a set of chemical reactions are naturally not suited for a study of the effects of noise on such systems. Instead one has to use a method in which various sources of noise can be added. One such method is that invented by Gillespie [4] in which the probability per unit time for the occurance of a reaction is taken to be a product of a combinatorial factor, which is a function of the numbers of reactants, and the rate constant of the reaction (taking into account the volume of the cell, here assumed to be linearly increasing through the cell cycle, for second- and higher-order reactions). Thus, knowing the rate constants of all reaction and the number of each type of molecule present, one can assign a probability per unit time to each reaction. This list is then used to choose (a) the time at which the next reaction will occur, and (b) which of the possible reactions will occur at that chosen time. Once this is decided the time variable is incremented, and the numbers of molecules are updated to reflect the occurance of the chosen reaction.

Each cell cycle, of duration T, is implemented as follows: First, the Gillespie simulation is run for a time $t_D = 0.4T$, at which point the gene copy number n is doubled. The simulation is then run till time T when the cell volume and gene copy number are halved, other molecules are partitioned binomially and plasmid copy number noise is added. Then the process is iterated for the next cell cycle, with one daughter cell followed after each partitioning. The whole algorithm is diagrammed in Fig. 11. Similar algorithms have been used in studies of intrinsic and extrinsic contributions to noise in gene expression [3] and transcription with the lac promoter [1, 5].

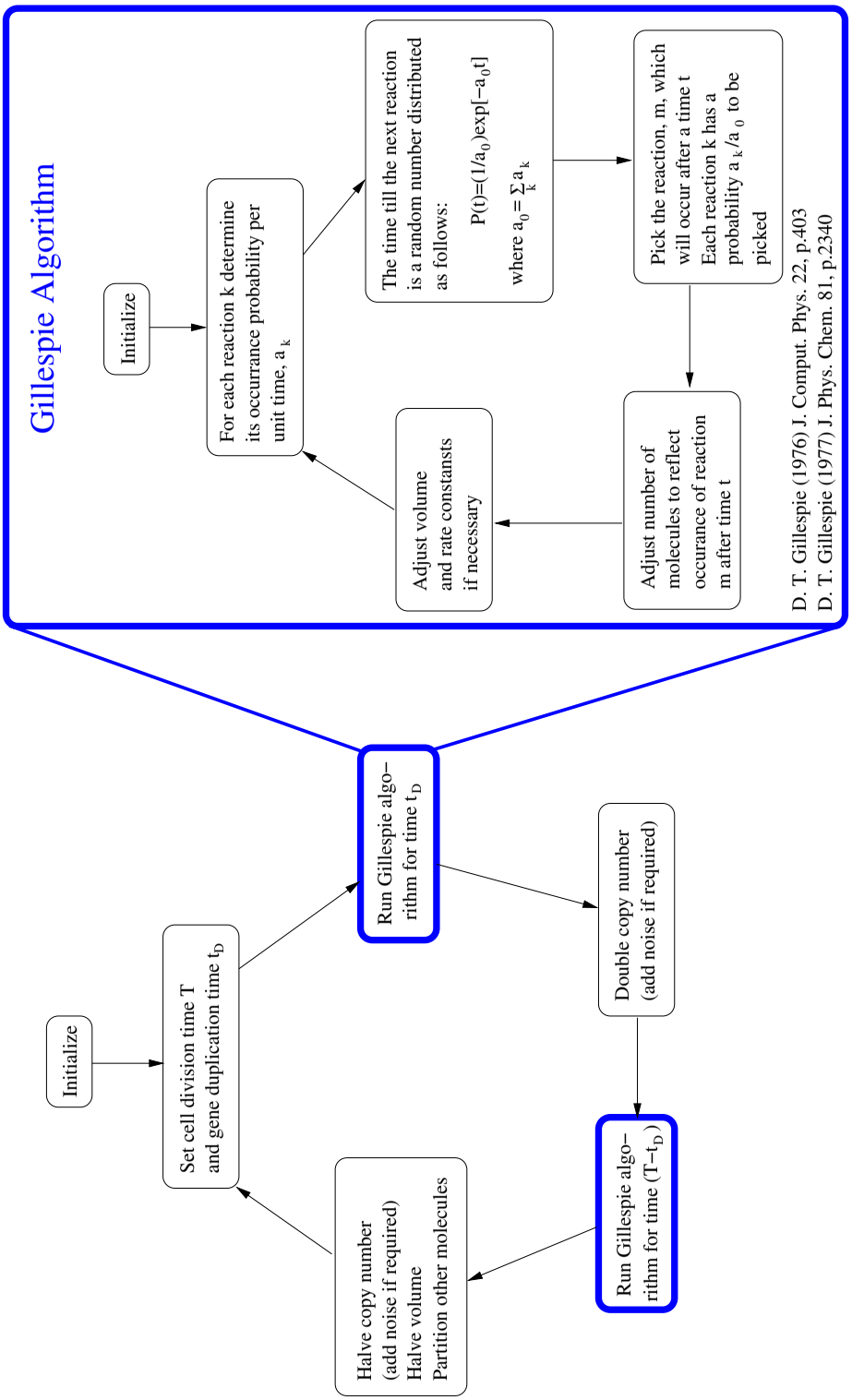


Fig. 11. Schematic diagram of the algorithm used for stochastic simulations of the chemical model of gene expression.

Analytic Expression for the Protein Distribution

We assume:

1. Over a cell cycle, of duration T, the protein number grows exponentially, $N = N_0 e^{\alpha t}$, with $\beta \equiv \alpha T$ being a constant.
2. Over the ensemble of runs N_0 is distributed normally with mean $\bar{N_0}$ and standard deviation σ.
3. Cells are not synchronized, so the observed distribution is an average over the whole cell cycle.

At a given time t, the distribution of N is related to the distribution of N_0:

$$\mathcal{P}(N) = \frac{\mathcal{P}(N_0)}{|dN/dN_0|} = \mathcal{P}(N_0)e^{-\alpha t}$$

Averaging over the whole cell cycle, the distribution of N is then:

$$\Rightarrow \mathcal{P}(N) = \frac{1}{T}\int_0^T \frac{dt}{\sqrt{2\pi}\sigma e^{\alpha t}} \exp\left[\frac{-\left(Ne^{-\alpha t} - \bar{N_0}\right)^2}{2\sigma^2}\right]$$

Substituting $y = \left(Ne^{-\alpha t} - \bar{N_0}\right)/\sqrt{2}\sigma$,

$$\mathcal{P}(N) = \frac{1}{(\alpha T)N\sqrt{\pi}} \int_{\left(Ne^{-\alpha T} - \bar{N_0}\right)/\sqrt{2}\sigma}^{\left(N - \bar{N_0}\right)/\sqrt{2}\sigma} dy\ e^{-y^2}$$

$$\Rightarrow \mathcal{P}(N) = \frac{\mathrm{Erf}\left[\frac{Ne^{-\beta} - \bar{N_0}}{\sqrt{2}\sigma}, \frac{N - \bar{N_0}}{\sqrt{2}\sigma}\right]}{(2\beta)N} \qquad [2]$$

where $\mathrm{Erf}(A, B) = \frac{2}{\sqrt{\pi}} \int_A^B e^{-x^2} dx$.

The mean of the above expression is:

$$\langle N \rangle \equiv \int_{-\infty}^{+\infty} N\ \mathcal{P}(N)\ dN = \bar{N_0}\frac{\left(e^{\beta} - 1\right)}{\beta}.$$

Higher moments are:

$$\langle N^2 \rangle = \left(\bar{N_0}^2 + \sigma^2\right)\frac{\left(e^{2\beta} - 1\right)}{2\beta},$$

$$\langle N^3 \rangle = \bar{N_0}\left(\bar{N_0}^2 + 3\sigma^2\right)\frac{\left(e^{3\beta} - 1\right)}{3\beta}.$$

Therefore, the variance of the distribution is:

$$\sigma_N^2 = \sigma^2\frac{(e^{2\beta} - 1)}{2\beta} + \frac{\bar{N_0}^2}{2\beta^2}[\beta(e^{2\beta} - 1) - 2(e^{\beta} - 1)^2].$$

When $\bar{N_0}$ is large, then

$$\frac{\sigma_N^2}{\langle N \rangle^2} = \frac{\beta}{2}\frac{\left(e^{2\beta} - 1\right)}{\left(e^{\beta} - 1\right)^2} - 1,$$

that is, the standard deviation of the distribution is proportional to the mean for a fixed β.

We first present evidence for assumption 2, that the distribution of N_0 is a gaussian. Fig. 12 shows the distributions of N values taken only from, respectively, the beginning, the middle and the end of each cell cycle. The first is thus the distribution of N_0 and is gaussian. If assumption 1 is correct, the other two distributions should also be gaussians but with larger means and variances. This is indeed true for

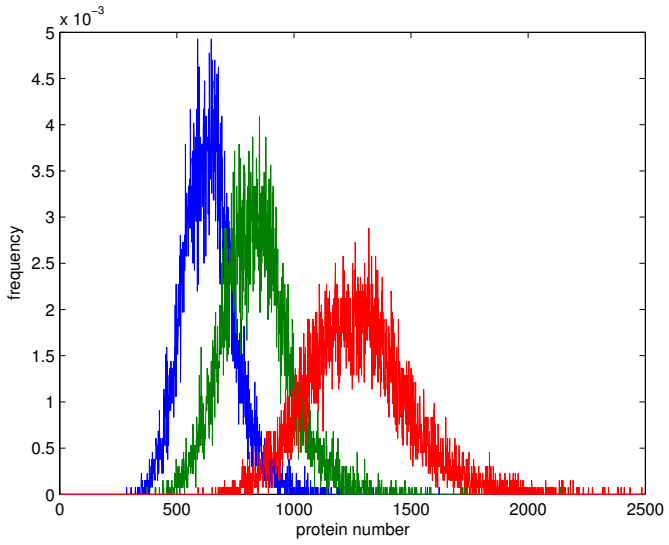


Fig. 12. Distributions of protein number at different times in the cell cycle from a simulation run with 100 repressors (as in Fig. 2d of main text). Blue: t=0s, beginning of cell cycle; Green: t=720s, middle of cell cycle; Red: t=1800s, end of cell cycle.

the distributions shown. Thus, assumptions 1 and 2 appear to be reasonable for simulation runs where the protein number is large.

Expression (2), with $\alpha = \ln 2/T$, i.e., $\beta = \ln 2$ gives an excellent fit to the protein distributions obtained from the simulation. In Fig. 13 the red dots show the protein number distribution for three runs with copy number noise 0, 0.2 and 0.4 respectively, and black curves plot expression (2) for $\bar{N_0} = 310$ and $\sigma = 40, 50, 65$ respectively.

In the experiment, the cells do not have a fixed cell division time. However, if the cell division time is changing sufficiently slowly it might be a reasonable approximation to assume that at any given point in the growth curve the system can, for a short period of time, be described by assuming that the cell division time is fixed. In order to do this we fit expression (2) to the experimental data in Fig. 14.

Expression (2) fits the data at all growth hours well, with a slight deviation for earlier times, for low protein numbers. This deviation is a result of the assumption, made for analytical convenience, that N_0 is distributed as a gaussian. When $\bar{N_0}$ is small and σ sufficiently large this allows a non-zero probability for negative N_0. This is what causes expression (2) to overestimate the protein distribution at the lower tail for earlier times in the growth curve.

Fig. 15 shows how the various parameters of the fit change as a function of growth hour. β decreases slightly with growth hour from 1.04 to 0.45. This decrease can be attributed to the expected decrease of all the transcription and translation rate constants as the bacterial culture leaves the exponentially growing phase and enters the stationary phase.

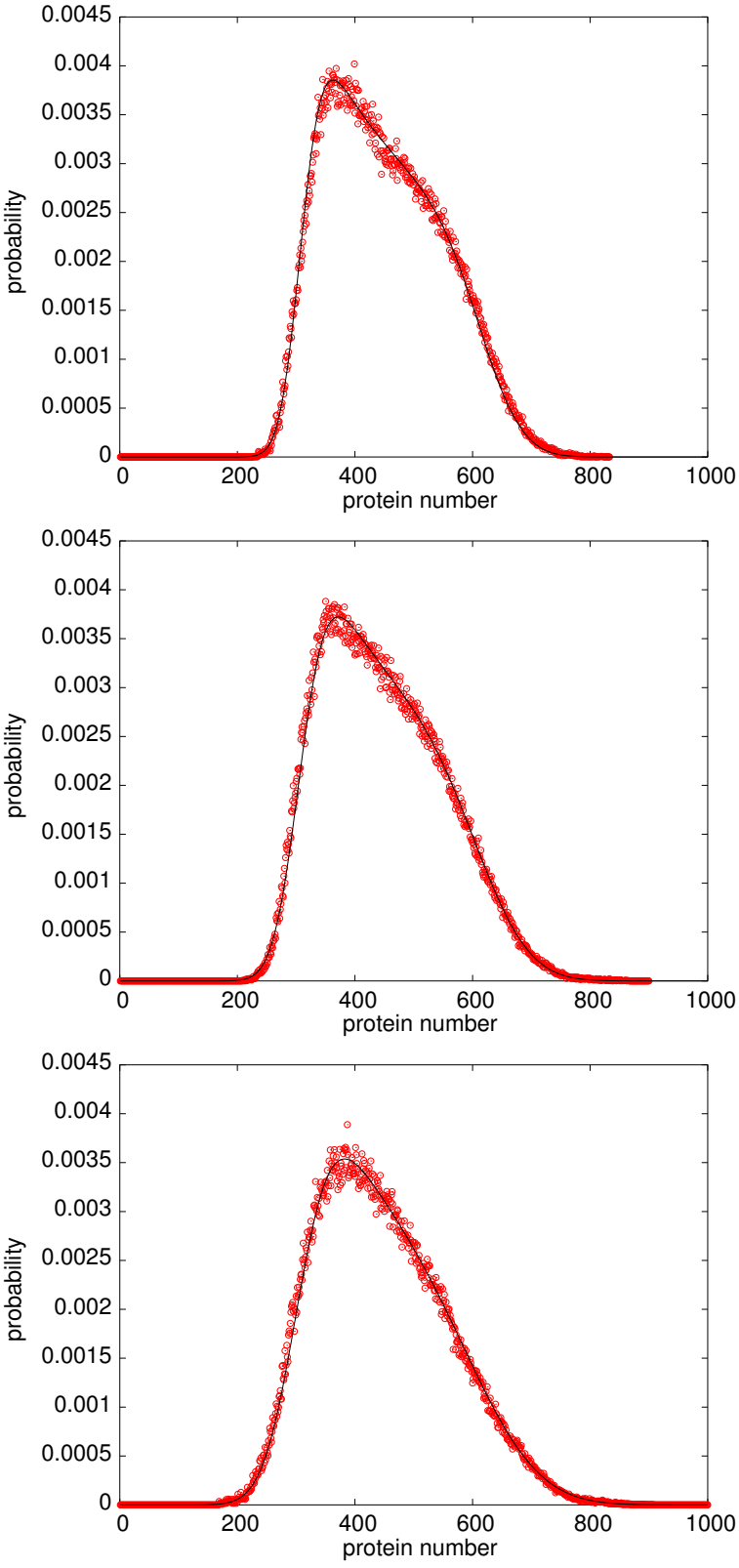


Fig. 13. Fit of expression (2) to simulation data. Red dots: protein number distribution for three runs with copy number noise 0, 0.2 and 0.4 respectively. Black lines: expression (2) for $\bar{N}_0 = 310$ and $\sigma = 40, 50, 65$ respectively.

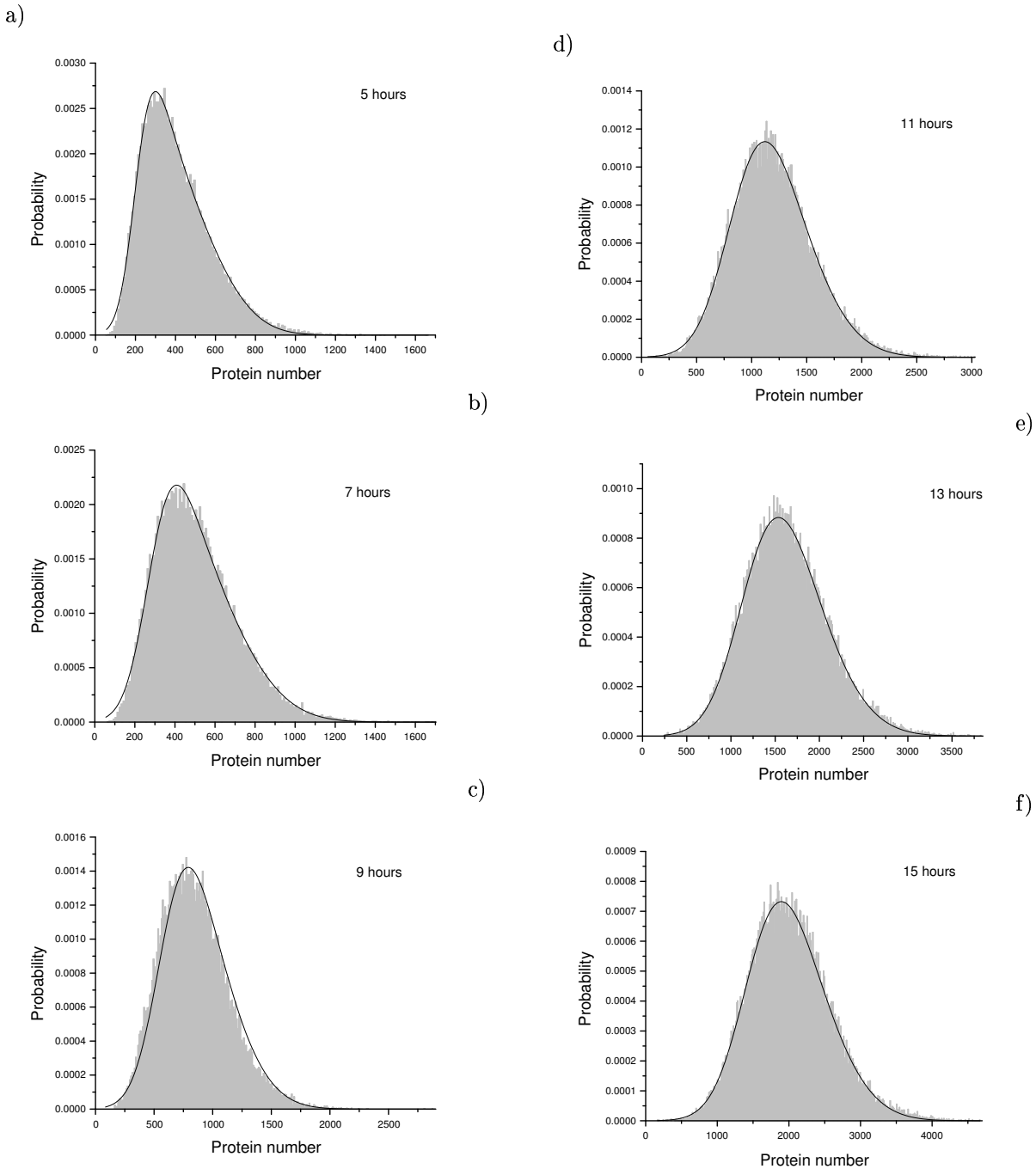
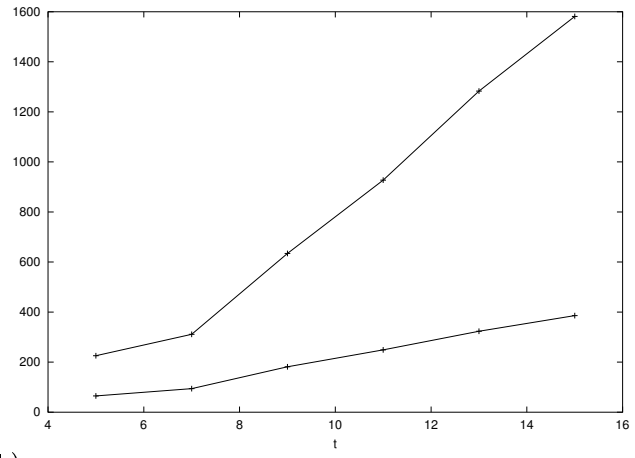
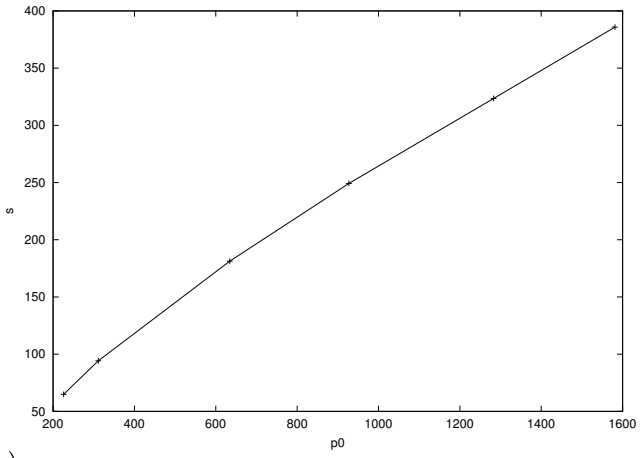


Fig. 14. Fit of expression (2) to experimental data. Fits were obtained using a nonlinear Levenberg-Marquardt leas-squares fitting algorithm. Best fit parameters: (a) $\bar{N}_0 = 226, \sigma = 65, \beta = 1.04$; (b) $\bar{N}_0 = 311, \sigma = 94, \beta = 0.88$; (c) $\bar{N}_0 = 634, \sigma = 181, \beta = 0.59$; (d) $\bar{N}_0 = 927, \sigma = 249, \beta = 0.48$; (e) $\bar{N}_0 = 1283, \sigma = 324, \beta = 0.45$; (f) $\bar{N}_0 = 1581, \sigma = 386, \beta = 0.45$.

a)



b)



c)

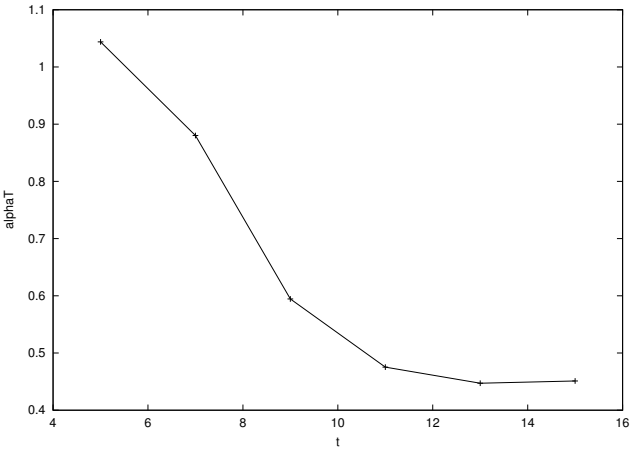


Fig. 15. Parameters of the fits. a) $\bar{N}_0$ and σ as a function of growth hour. b) σ vs. $\bar{N}_0$. c) $\beta \equiv \alpha T$ as a function of growth hour.

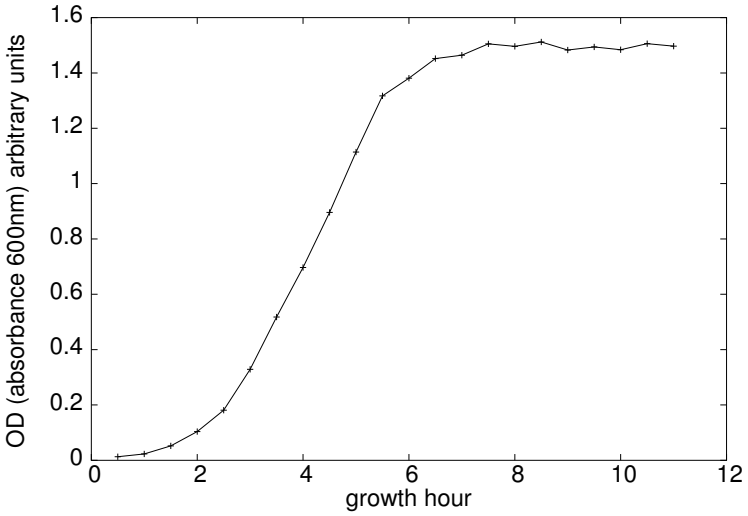


Fig. 16. Optical density of the bacterial culture as a function of time. This is a measure of the number of cells in the culture.

Simulations with Variable Cell Division Time

We have also done simulations where the cell division time is not fixed because that is the situation in the experiment. As a first approximation we assume that the probability per unit time for a cell to divide is proportional to the growth rate of the bacterial culture. The growth rate can be obtained from the slope of the curve in Fig. 16 which shows the optical density (at 600nm absorbance) of the bacterial culture as a function of time. We fitted $\ln(OD)$ using a sigmoidal function:

$$A_2 - \frac{(A_1 - A_2)}{1 + \exp\left[\frac{(t-x_0)}{\Delta x}\right]}$$

This yielded the following best fit parameters: $A_1 = -6.30, A_2 = 0.43, x_0 = 1.56hrs, \Delta x = 1.18hrs$. Taking the derivative:

$$\frac{1}{OD}\frac{d(OD)}{dt} = \frac{(A_2 - A_1)}{\Delta x}\frac{\exp\left[(t - x_0)/\Delta x\right]}{\left(1 + \exp\left[(t - x_0)/\Delta x\right]\right)^2}$$

We equate $\frac{1}{OD}\frac{d(OD)}{dt}$ with the cell division probability per unit and get the curve shown in Fig. 17. As expected, the probability density initially rises, reaches a maximum where the growth rate is largest, and then declines. Fig. 18 shows ten runs of the simulation where the cell division time is chosen according to this probability density. Fig. 19 displays various moments of the protein distributions obtained from these runs as a function of growth hour. Fig. 19b shows evidence of a crossover from a long-tailed to a gaussian distribution: the skewness is initially large but later approaches zero, while the kurtosis approaches 3, the value expected for a gaussian distribution. However, the standard deviation is not proportional to the mean (Fig. 19a).

In addition to the cell division time, the rate constants would also vary with growth hour, as mentioned previously. To approximate this effect we therefore allow the rate of one reaction to vary. The reaction chosen is the transition from the closed to the open DNA-RNA-polymerase complex ($\{D.RNAP\} \xrightarrow{k} TrRNAP.$) The value follows the growth rate (Fig. 17), initially increasing and then decreasing. In this case we found that the standard deviation is proportional to the mean (Fig. 20) but the protein distribution remains long-tailed.

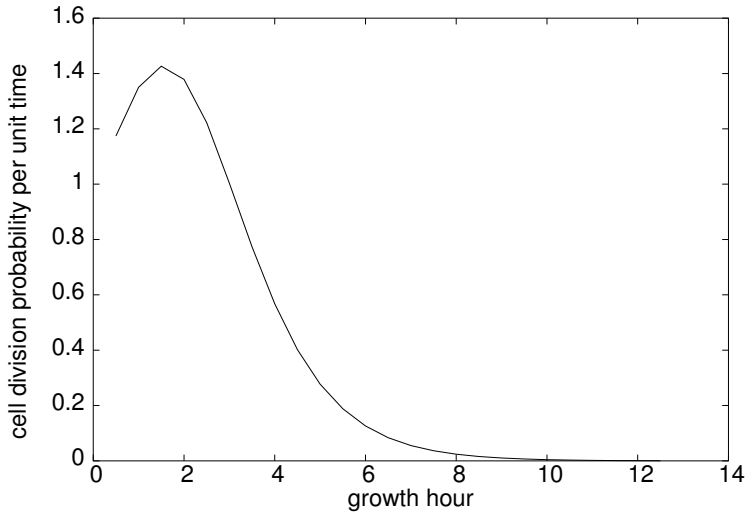


Fig. 17. Growth rate of the bacterial culture and the cell division probability per unit time for the varying cell division time simulations.

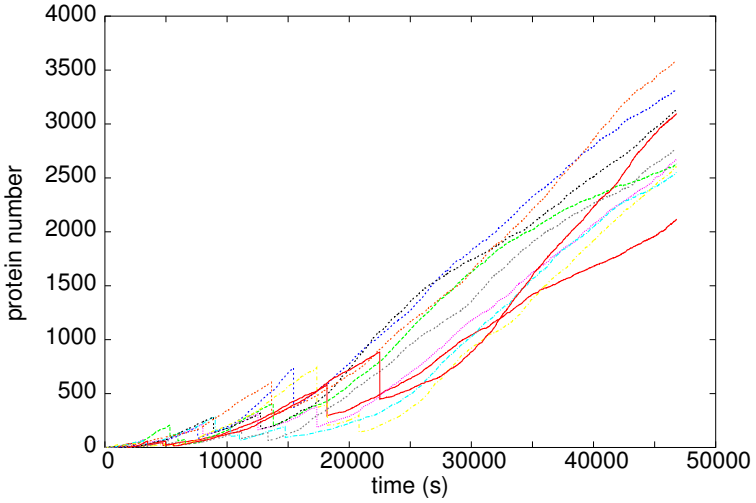


Fig. 18. Ten runs with cell division times chosen according to the probability per unit time displayed in Fig. 17.

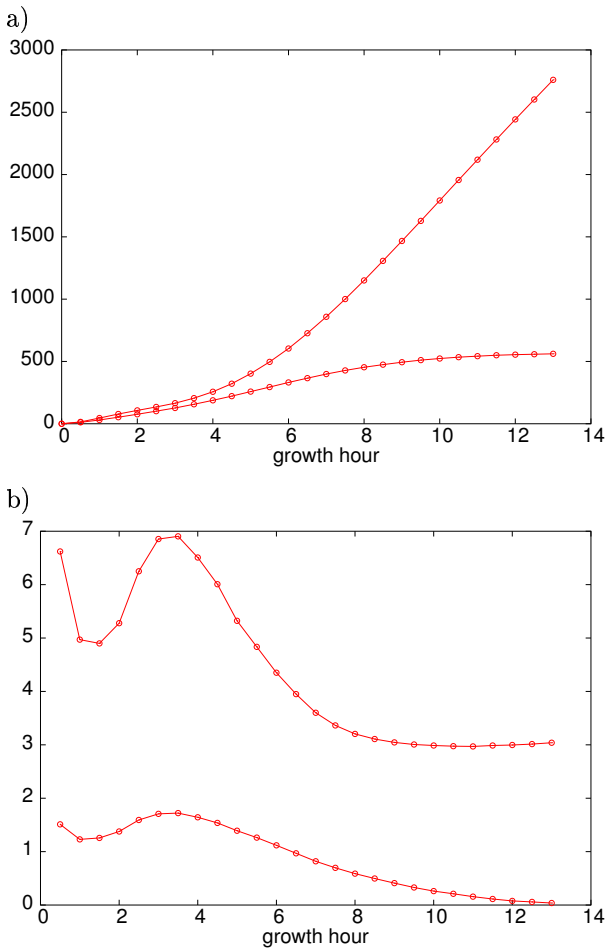


Fig. 19. Moments of the protein distriubtions at various growth hours for the variable cell division time runs. (a) mean (top) and standard deviation (bottom). (b) kurtosis (top) and skewness (bottom).

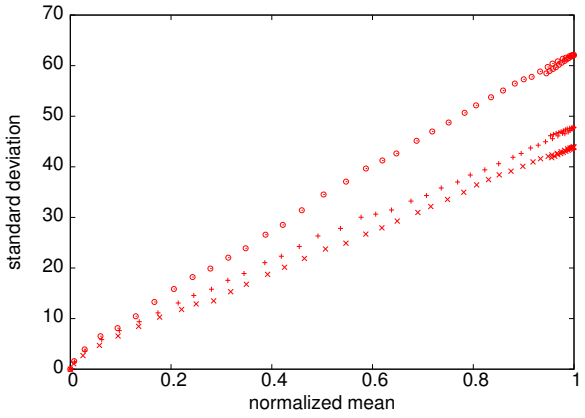


Fig. 20. Standard deviation vs. mean for three runs with different total repressor numbers, with variable cell division times and one time dependent rate constant.

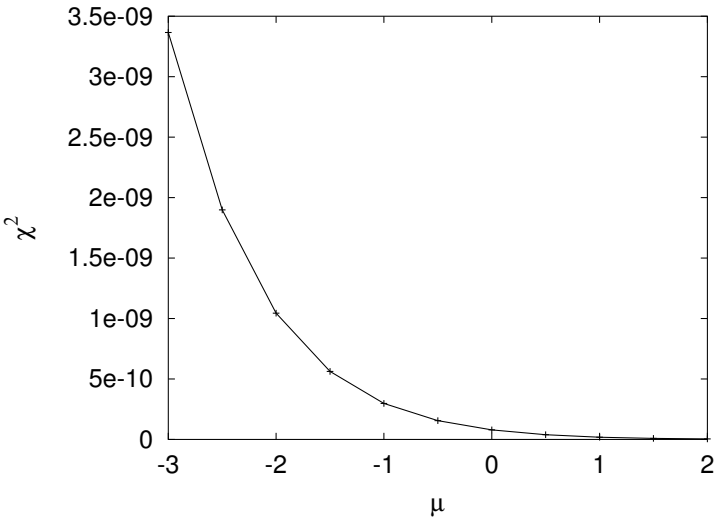


Fig. 21. χ^2 value for the fit of expression (2) to expression (1) with $N_0 = 1500, \sigma = 0.5$.

Comparison of Error Function Distribution with the Crossover Function

We have numerically compared the crossover function (1) with the error function distribution (2). We used a nonlinear Levenberg-Marquardt least-squares algorithm to fit equation (2) to equation (1), keeping N_0 and σ fixed. We find that expression (2) fits better for larger values of μ. In fact, there is a steady increase of χ^2, a measure of the goodness of fit, as μ is decreased, as shown in Fig. 21 for $N_0 = 1500, \sigma = 0.5$. Visually, the fits are good for $\mu \geq 0$. Below that the best fit error function distribution overestimates the crossover function at the left tail, while the right tail falls off faster than the crossover function. These observations have been made for $N_0 = 1500$ and various σ values in the range $0.5 - 0.65$ which covers the range of values of σ in the fits to experimental data (see figures 14 and 15). The fits are much worse for $\sigma = 1$ and larger.

Sensitivity of Long-Tailed and Gaussian Distributions to Switches

A switch is a logical element, which has an on-off response to its input. The following functional form has been used to describe the response function of a class of biological switches [6]:

$$f(N) = \frac{K^h + AN^h}{K^h + N^h}$$

This function has three paramters, the threshold K, the Hill coefficient h and the amplitude A; N is the input. The amplitude specifies the level of response when the input is large: $\lim_{N\to 0} f(N) = 1$ and $\lim_{N\to\infty} f(N) = A$. K is the value of the input at which the response is halfway between the lower and upper limits, i.e., $f(K) = (1 + A)/2$. The Hill coefficient h is a measure of the steepness of the transition from a low response to a high response; the higher the value of h, the steeper the transition. In the limit $h \to \infty$ the response becomes a step function with the step at K, $f(N) = \Theta(N - K)$, an ideal on-off switch.

We compare the sensitivity to switches of the following two distributions, the first a lognormal, with mean $\mu \exp(\sigma^2/2)$ and standard deviation $\mu\sqrt{\exp(2\sigma^2) - \exp(\sigma^2)}$, and the second a gaussian, with mean μ and standard deviation σ:

$$P_1(N) = \frac{1}{\sqrt{2\pi}\sigma N} \exp\left[\frac{-(\ln N - \ln \mu)^2}{2\sigma^2}\right]$$

$$P_2(N) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(N - \mu)^2}{2\sigma^2}\right]$$

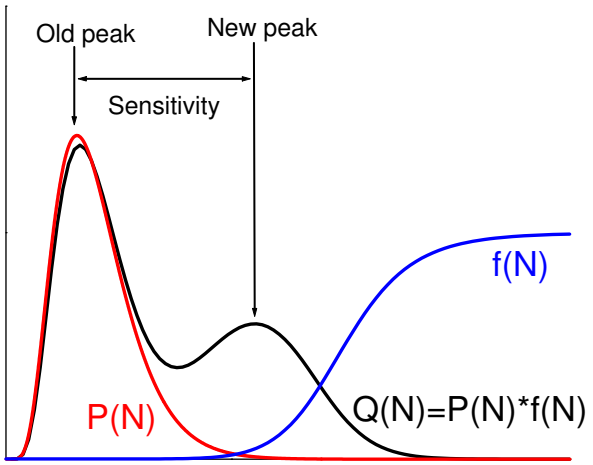


Fig. 22. Definition of sensitivity of the distribution $P(N)$ to the switch $f(N)$.

$P_1(N)$ has a single peak at position $N_0 = \mu \exp(-\sigma^2)$. $P_2(N)$ has a single peak at position $N_0 = \mu$. When $P_1(N)$ or $P_2(N)$ is multiplied by $f(N)$, the resultant function, which we denote $Q(N)$, will have either one or two peaks, depending on the parameters μ and σ as well as the switch parameters. If there is only one peak, at position N_1, we define the sensitivity to be

$$\text{sensitivity} = \frac{N_1 - N_0}{N_0}.$$

If there are two peaks we use the same formula, taking N_1 to be the position of the rightmost peak (Fig. 22.)

Variation of Sensitivity with Threshold. Fig. 23a shows how the sensitivity of the lognormal and gaussian distributions changes as the switch threshold, K, is varied, with other parameters fixed at $h = 4, A = 1000$.

Lognormal distribution. (see Fig. 23b,c) for small values of threshold, upto approximately 500, $Q(N)$ has a single peak. The position of the peak shifts to the right as the threshold is increased leading to an increase in sensitivity. At a little over $K = 500$, $Q(N)$ becomes bimodal, with a second peak appearing close to the peak of $P_1(N)$. The rightmost peak continues to move towards larger values of N as K is increased; the sensitivity keeps increasing. The sensitivity reaches a peak just below $K = 4000$ and then starts to decrease slightly as the rightmost peak moves to the left. Around $K = 4300$ the distribution again becomes unimodal with the rightmost peak disappearing and only the peak close to the peak of $P_1(N)$ remaining; the sensitivity discontinuously drops to zero.

Gaussian distribution. (see Fig. 23d,e) the distribution $Q(N)$ after multiplication of $P_2(N)$ by $f(N)$ is always unimodal. Initially the peak moves towards the right resulting in an increase in the sensitivity. However, after around $K = 1200$ the peak starts moving left and the sensitivity drops towards zero, initially slowly and later, just after $K = 2300$, very rapidly.

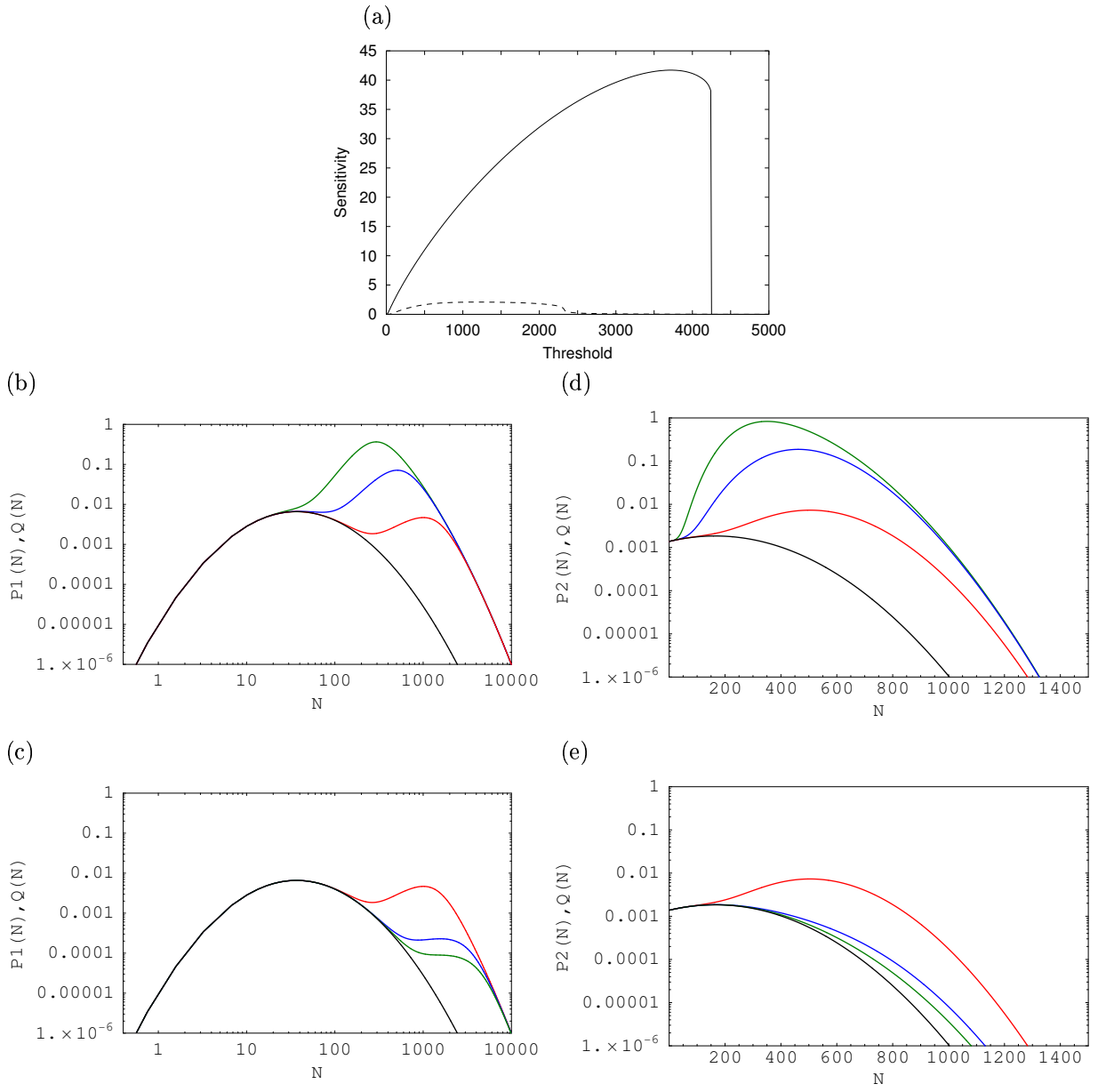


Fig. 23. (a) Sensitivity as a function of switch threshold for $P_1(N)$ (solid line) and $P_2(N)$ (dashed line); (b) $P_1(N)$ (black), $Q(N) = P_1(N)f(N)$ for $K = 300$ (green), $K = 600$ (blue), $K = 1500$ (red); (c) $P_1(N)$ (black), $Q(N) = P_1(N)f(N)$ for $K = 1500$ (red), $K = 3500$ (blue), $K = 4500$ (green); (d) $P_2(N)$ (black), $Q(N) = P_2(N)f(N)$ for $K = 300$ (green), $K = 600$ (blue), $K = 1500$ (red); (e) $P_2(N)$ (black), $Q(N) = P_2(N)f(N)$ for $K = 1500$ (red), $K = 3500$ (blue), $K = 4500$ (green).

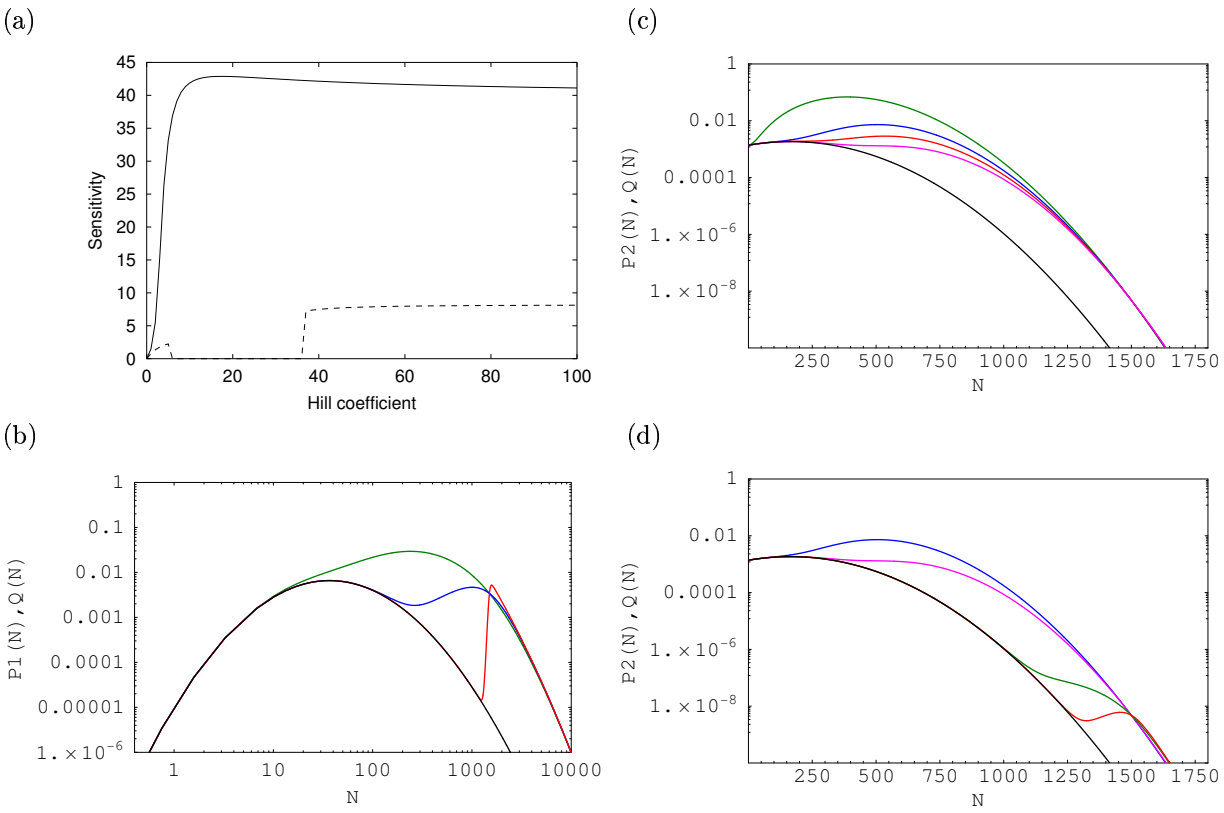


Fig. 24. (a) Sensitivity as a function of Hill coefficient for $P_1(N)$ (solid line) and $P_2(N)$ (dashed line); (b) $P_1(N)$ (black), $Q(N) = P_1(N)f(N)$ for $h = 2$ (green), $h = 4$ (blue), $h = 50$ (red); (c) $P_2(N)$ (black), $Q(N) = P_2(N)f(N)$ for $h = 2$ (green), $h = 4$ (blue), $h = 5$ (red), $h = 6$ (magenta); (d) $P_2(N)$ (black), $Q(N) = P_2(N)f(N)$ for $h = 4$ (blue), $h = 6$ (magenta), $h = 25$ (green), $h = 50$ (red).

Variation of Sensitivity with Hill Coefficient. Fig. 24a shows how the sensitivity of the lognormal and gaussian distributions changes as the Hill coefficient, h, is varied, with other parameters fixed at $A = 1000, K = 1500$.

Lognormal distribution. (see Fig. 24b) for $h = 0, 1, 2$, $Q(N)$ is unimodal. From $h = 3$ onwards it becomes bimodal. As h is increased the peak moves beyond the position of the threshold and the sensitivity increases beyond 40. However, as h is increased further, the sensitivity starts dropping and we expect it to asymptotically approach a value just under 40. The reason for this is that as $h \to \infty$ the switch response becomes a step function and the peak therefore settles at the switch threshold 1500. Therefore the sensitivity approaches $1500/36.788 - 1 \approx 39.77$.

Gaussian distribution. (see Fig. 24c,d) the gaussian distribution has a more involved behaviour. For $h = 0$–5, $Q(N)$ is unimodal and the peak shifts to the right as h is increased leading to an increase in the sensitivity. Then at $h = 6$ the peak abruptly shifts to the position of the old peak and the sensitivity drops to zero. However, $Q(N)$ remains unimodal for $h = 6$ and indeed higher h values. Eventually, when h is raised to 37 or more, $Q(N)$ becomes bimodal with a small second peak arising near the switch threshold. Thus, by our definition, the sensitivity jumps discontinously to a value close to $1500/164.87 - 1 \approx 8.1$. Note that this kind of discontinuous behaviour is a consequence of our always choosing the position of the rightmost peak irrespective of its height. Alternate definitions can be constructed, for instance where the rightmost peak is taken only if its height is suffieciently large. However, the definition we have chosen serves perfectly well to emphasise our point that long-tailed distributions are generally more sensitive to switches than a comparable gaussian.

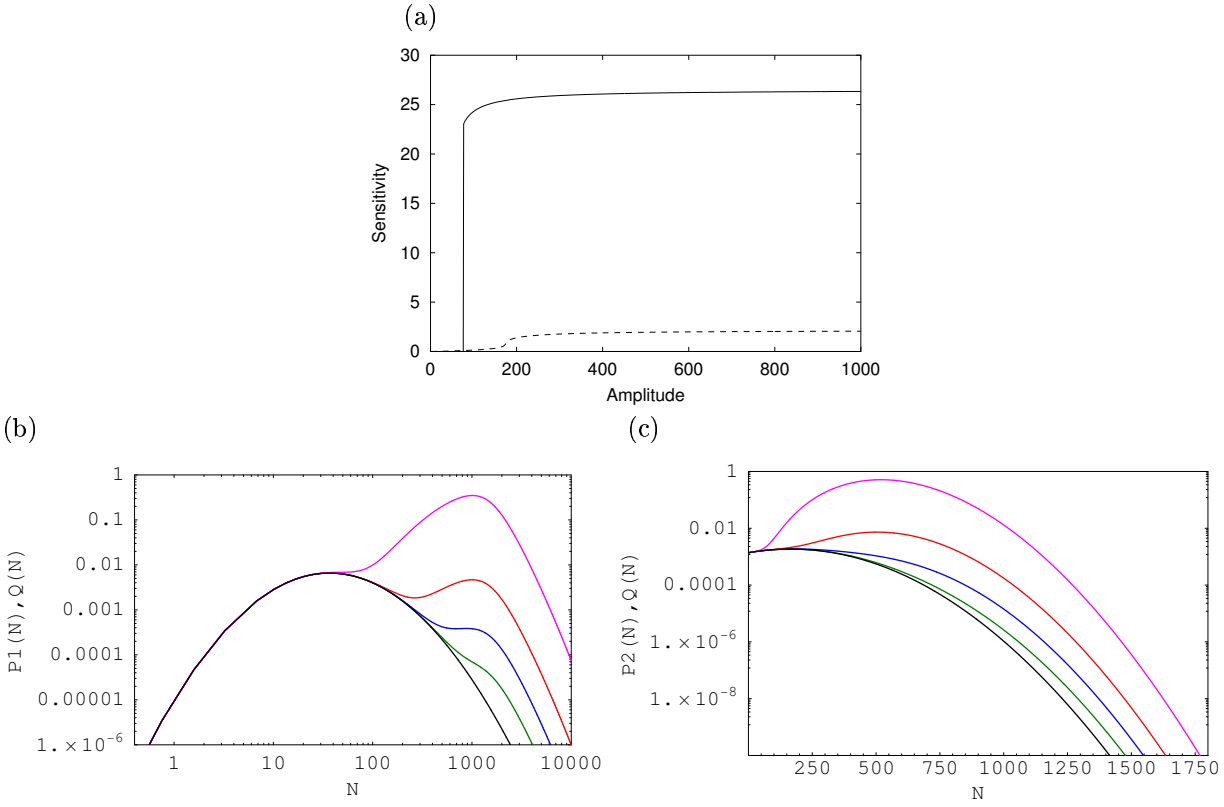


Fig. 25. (a) Sensitivity as a function of switch amplitude for $P_1(N)$ (solid line) and $P_2(N)$ (dashed line); (b) $P_1(N)$ (black), $Q(N) = P_1(N)f(N)$ for $A = 10$ (green), $A = 77$ (blue), $A = 1000$ (red), $A = 75000$ (magenta); (c) $P_2(N)$ (black), $Q(N) = P_2(N)f(N)$ for $A = 10$ (green), $A = 77$ (blue), $A = 1000$ (red), $A = 75000$ (magenta).

Variation of Sensitivity with Amplitude. Fig. 25a shows how the sensitivity of the lognormal and gaussian distributions changes as the switch amplitude, A, is varied, with other parameters fixed at $h = 4, K = 1500$.

Lognormal distribution. (see Fig. 25b) for small values of amplitude, upto $A = 76$, $Q(N)$ is unimodal with a peak at the position of the original peak. Therefore the sensitivity is zero. At $A = 77$ a second peak appears much to the right of the old peak and the sensitivity jumps to over 20. As A is increased further there is a small increase of the sensitivity, but it appears to saturate a little over 26.

Gaussian distribution. (see Fig. 25c) unlike the lognormal, the gaussian shows a more continuous behaviour with the peak shifting continuously to the right and eventually settling around $N = 500$. Thus the sensitivity rises continuously and then appear to saturate to a value just over 2.

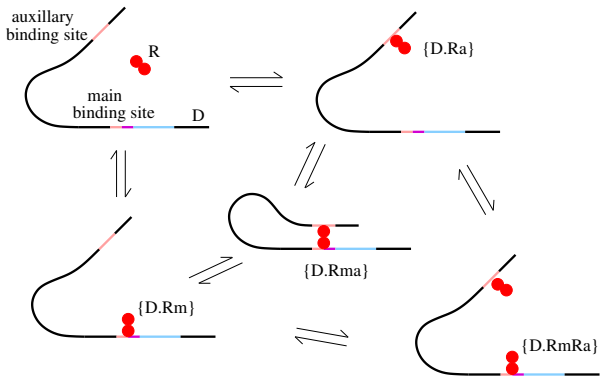


Fig. 26. With DNA looping, the system can be in one of five states.

DNA Looping

In order to model DNA looping, the Operator Binding module in the chemical model is replaced to allow the five possible states (based on [7]) shown in Fig. 26. Further, the Transcription module is modified to allow transcription to occur only from D and $\{D.Ra\}$ (if R is a repressor) or $\{D.Rma\}, \{D.Rm\}$ and $\{D.RmRa\}$ (if R is an enhancer). Henceforth we consider only the former case.

Repressor binding

- $D + R \rightleftharpoons \{D.Ra\}$
- $D + R \rightleftharpoons \{D.Rm\}$
- $\{D.Ra\} \rightleftharpoons \{D.Rma\}$
- $\{D.Rm\} \rightleftharpoons \{D.Rma\}$
- $\{D.Ra\} + R \rightleftharpoons \{D.RmRa\}$
- $\{D.Rm\} + R \rightleftharpoons \{D.RmRa\}$

Transcription

- $D + RNAP \rightleftharpoons \{D.RNAP\}$
- $\{D.RNAP\} \rightarrow TrRNAP$
- $TrRNAP \rightarrow RBS + D + RNAP$
- $\{D.Ra\} + RNAP \rightleftharpoons \{D.Ra.RNAP\}$
- $\{D.Ra.RNAP\} \rightarrow \{TrRNAP.Ra\}$
- $\{TrRNAP.Ra\} \rightarrow RBS + \{D.Ra\} + RNAP$

Translation

- $RBS \rightarrow \phi$
- $RBS + Rib \rightleftharpoons \{RibRBS\}$
- $\{RibRBS\} \rightarrow ElRib + RBS$
- $ElRib \rightarrow Protein$

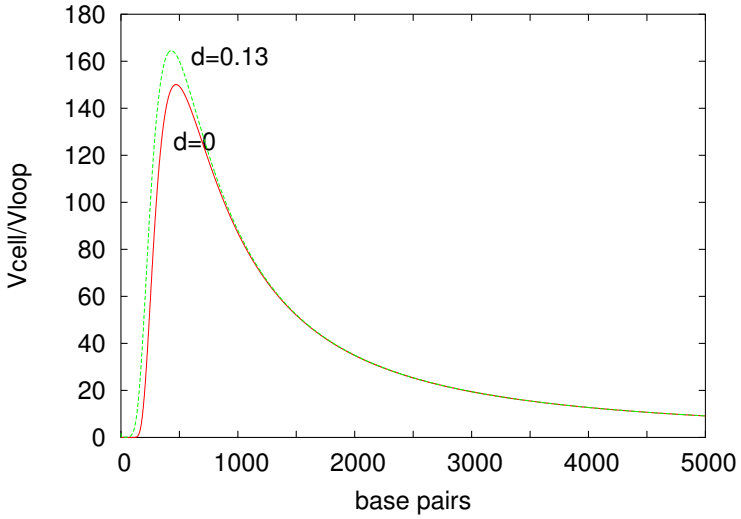


Fig. 27. V_{cell}/V_{loop} as a function of distance between the operator sites, in base pairs.

Protein folding and decay

- $Protein \rightarrow FoldedProtein$
- $Protein \rightarrow \phi$
- $FoldedProtein \rightarrow \phi$

Using the framework of ref. [7], when the repressor is bound to one of the operator sites (either the main or auxillary one) the change in free energy because of looping, ΔG_l, results in a multiplication of the association rate of the repressor to the other operator site by a factor $\exp(-\Delta G_l)$. Further, when the auxillary operator site is sufficiently strong, ΔG_l can be related to the effective increase in local concentration of the repressor: $\Delta G_l = -\ln(V_{cell}/V_{loop})$, where V_{cell} is the volume of the cell and V_{loop} is the effective volume in which the repressor is allowed to move once it is bound to the auxillary operator site. This effective volume is related to the local concentration, j_M, as follows: $V_{loop} = 1/(N_0 j_M)$, where N_0 is Avogardo's number. Ref. [8] gives the following expression for j_M for a double-stranded DNA polymer, as a function of distance, b in base pairs, between the operator sites:

$$j_M(b) = 2.7 \times 10^{-3} \times b^{-3/2} \times \exp\left(\frac{d-2}{1.2 \times 10^{-5} \times b^2 + d}\right) \quad \frac{\text{mol}}{\text{liter}} \qquad [3]$$

Interaction probabilities calculated using expression (3) with $d = 0.13$ have been found to agree well with simulation data for a reaction radius of 10nm [8]. We therefore use this expression to calculate j_M, hence V_{loop} and the reaction rates, which are required for the Gillespie simulation. Fig. 27 shows the dependence of V_{cell}/V_{loop} as a function of distance between the operator sites, in base pairs.

Fig. 28 shows the protein number distributions obtained for three runs with DNA looping, which differ in the distance between the two operator sites. All the distributions are long-tailed. Comparing distributions with and without DNA looping, we find that for distributions having comparable means, looping results in longer tails (see Fig. 29).

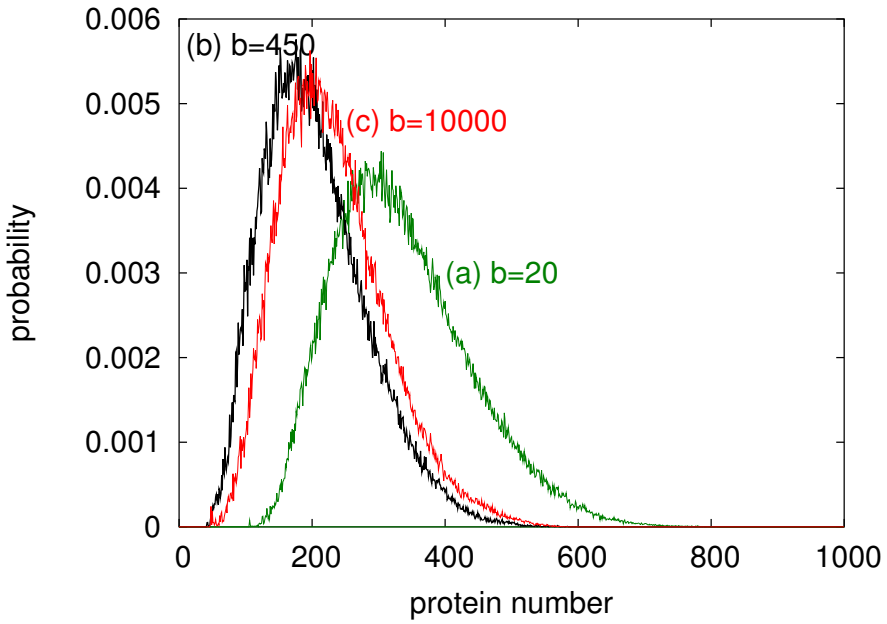


Fig. 28. Protein distributions obtained from runs with DNA looping.

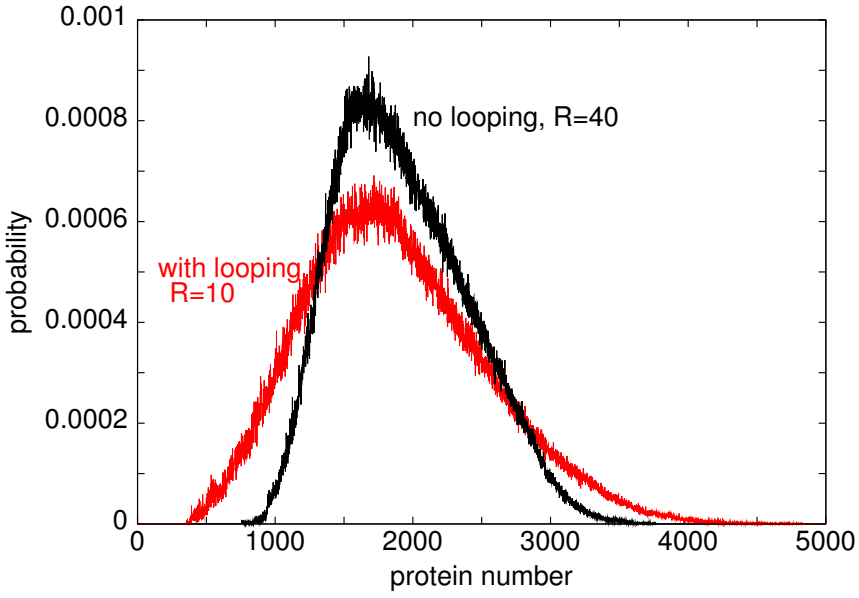
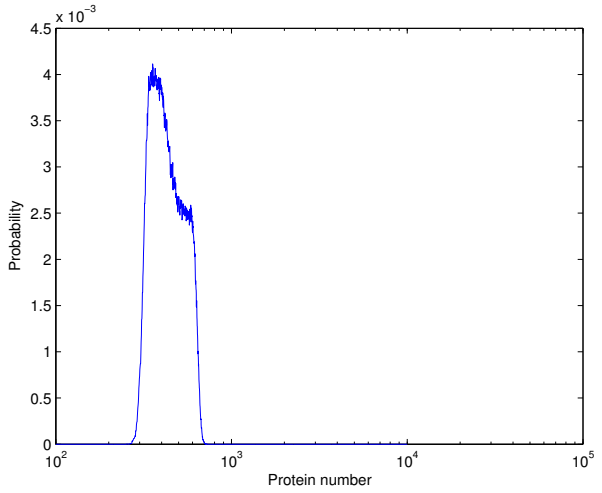
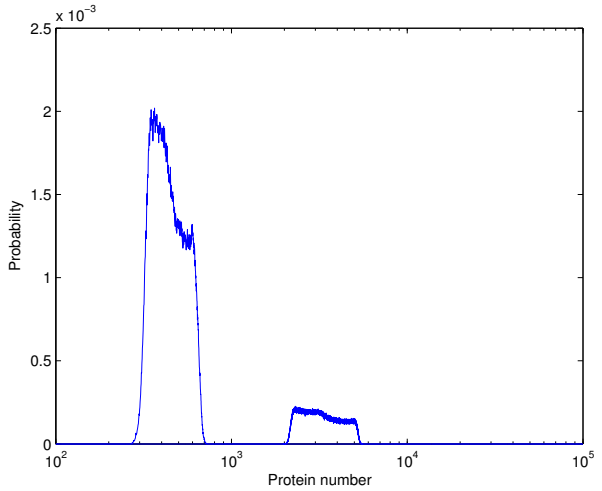


Fig. 29. Protein distributions from runs with and without looping. Repressor numbers have been chosen so that the means of the two distributions are almost identical.

(a)



(b)



(c)

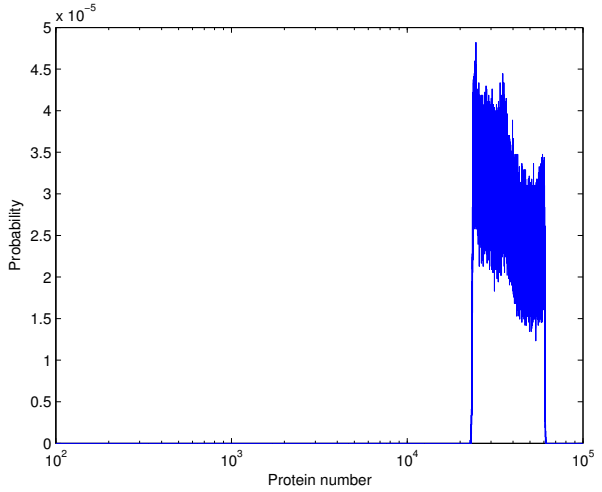


Fig. 30. Protein number distributions obtained from runs where the number of non-specific sites at which the tetramers of the folded protein could bind and release RNA polymerases were (a) zero, (b) 315 and (c) 3465.

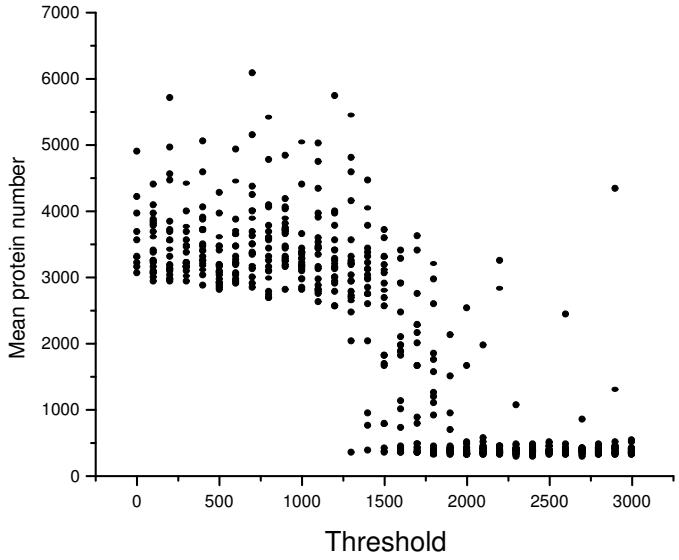


Fig. 31. Each data point is the mean protein number obtained from a run of the chemical model coupled to a switch via the RNA-polymerase numbers. The only parameter that changes across these 610 runs is the switch threshold.

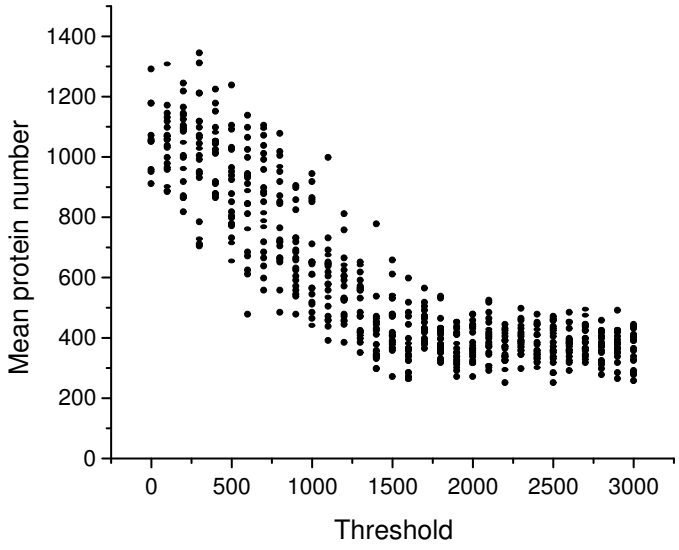


Fig. 32. Each data point is the mean protein number obtained from a run of the chemical model with DNA looping coupled to a switch via the RNA-polymerase numbers. The only parameter that changes across these 610 runs is the switch threshold.

Bistability due to Positive Feedback via a Switch

In the main text it was argued that the combination of long-tailed distributions and a cooperative switch in a positive feedback loop would lead to bimodal distributions, such as those seen in an autoregulatory circuit (Fig. 7d-f of main text.) Such a positive feedback might exist in that circuit because of the strong non-specific binding of the GFP-lacI fusion protein to DNA. In order to confirm that such a mechanism would behave like a cooperative switch, we studied stochastic simulations of the following system of reactions, which is a simplified version of the full chemical model with the addition of tetramerization of the folded protein and its binding to non-specific sites on the DNA, releasing RNA polymerases:

Tetramerzation and binding to nonspecific sites

- $4 \times FoldedProtein \rightleftharpoons T$
- $N + T \rightleftharpoons \{N.T\}$

Transcription and translation

- $D + RNAP \rightarrow D + RNAP + Protein; \quad RNAP = RNAP_0 + \{N.T\}$

Protein folding and decay

- $Protein \rightarrow FoldedProtein$
- $Protein \rightarrow \phi$
- $FoldedProtein \rightarrow \phi$

The binding and dissociation rates for the tetramer to the non-specific sites fixes the threshold of the effective switch and the number of non-specific sites fixes the amplitude of the switch by controlling the maximum number of RNA polymerases that are available for transcription. Fig. 30a shows the protein distribution obtained from runs where the number of non-specific sites are zero, i.e. there is no switch. When the number of non-specific sites is increased to 315 (which makes the amplitude of the effective switch $A = 10$ because we keep $RNAP_0 = 35$) the distribution becomes bimodal (Fig. 30b.) When the amplitude is further increased to $A = 100$ by increasing the number of non-specific sites to 3465, the distribution once again becomes bimodal (Fig. 30c.)

Fig. 31 shows the mean protein numbers obtained from a total of 610 runs of the chemical model coupled to a switch via a positive feedback: the RNA-polymerase number is multiplied by the response of a cooperative switch to the protein number. The response function of the switch is $(K^h + AN^h)/(K^h + N^h)$, where the amplitude, A, is 10 and the Hill coefficient, h, is 4 for each run. N stands for the protein number and the 610 runs sample a number of values of the switch threshold, K. A region of bistability is clearly visible in Fig. 31. By contrast, a similar figure for runs with DNA looping (Fig. 32) does not have a bistable region. This is probably a result of the longer-tailed distributions with DNA looping; it is likely that bistability would be observed for switches with higher Hill coefficients.

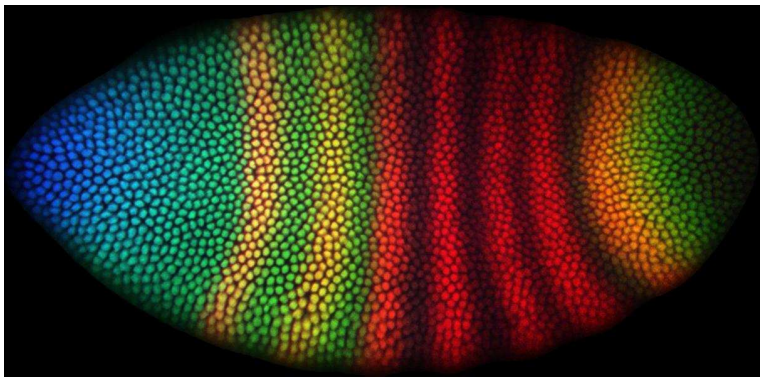


Fig. 33. Fluorescence image of gene expression patterns for 2249 nuclei in Drosophila embryo hx21 from the FlyEx database. The genes monitored are: even-skipped (red), hunchback (green) and bicoid (blue).

Table 1. A portion of the data for embryo hx21 available at the FlyEx database (see text).

Nucleus number	A-P coord.	D-V coord.	Red	Green	Blue
0	2.06	45.70	5.83	17.14	72.71
1	2.98	48.25	9.48	27.32	107.62
2	3.22	42.82	9.68	34.90	137.42
3	3.25	45.37	9.93	37.12	145.96
4	3.59	50.88	8.45	38.34	126.11
5	3.94	40.62	11.70	40.27	146.38
6	4.31	38.15	10.32	42.63	139.14
7	4.34	46.95	11.16	44.06	154.28
8	4.64	43.86	12.21	45.04	154.73
9	4.74	49.91	11.27	49.04	163.06
10	4.92	52.88	9.98	40.92	125.47

Bicoid and Hunchback in Early-Stage *Drosophila* Embryos

The FlyEx database (http://urchin.spbcas.ru/flyex) is a repository of segmentation gene expression data in the fruit fly Drosophila melanogaster. The database contains fluorescence images of gene expression patterns such as that shown in Fig. 33. The image shows an embryo named hx21 in cleavage cycle 14. The image has 2249 nuclei. Red indicates the level of even-skipped, green indicates hunchback and blue indicates bicoid levels in each nucleus. The database also provides the tabulated data for the intensities of the red, green and blue channels for each nucleus as well as the Anterior-Posterior (A-P) and Dorsal-Ventral (D-V) coordinates of each nucleus. A portion of this table for hx21 is shown in Table 1.

Fig. 34 shows the level of bicoid and hunchback as a function of the A-P coordinate. We concentrate on the nuclei which have A-P coordinates between 10% and 70% of the egg length. There are 1606 such nuclei. Fig. 35 shows a scatter plot of the bicoid vs. the hunchback fluorescence for each of these 1606 nuclei. This scatter plot indicates that bicoid acts as a switch which causes the hunchback gene to express in regions of the embryo where the bicoid concentration is above a critical threshold. The response function of the bicoid switch shown in Fig. 8 of the main text has been created from the data in Fig. 35 by suitable binning. Fig. 36 shows the histograms of bicoid and hunchback intensities for the 1606 nuclei mentioned above. The bicoid distribution is clearly unimodal, while the hunchback distribution is bimodal. The bimodal distribution of hunchback shown in Fig. 8 of the main text has been created from Fig. 36 by binning the data into bins of width 5 fluorescence counts. These features are seen in all other embryos in the FlyEx database for which bicoid and hunchback levels are recorded and which are in the same cleavage cycle.

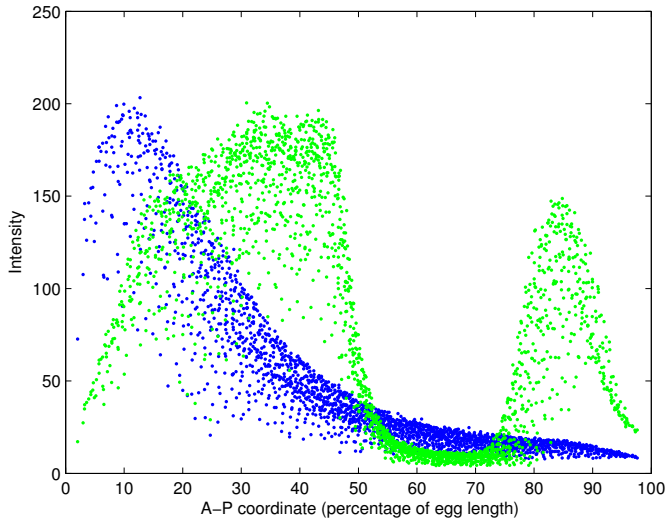


Fig. 34. Bicoid (blue) and Hunchback (green) intensities as a function of the Anterior-Posterior coordinate for 2249 nuclei in embryo hx21.

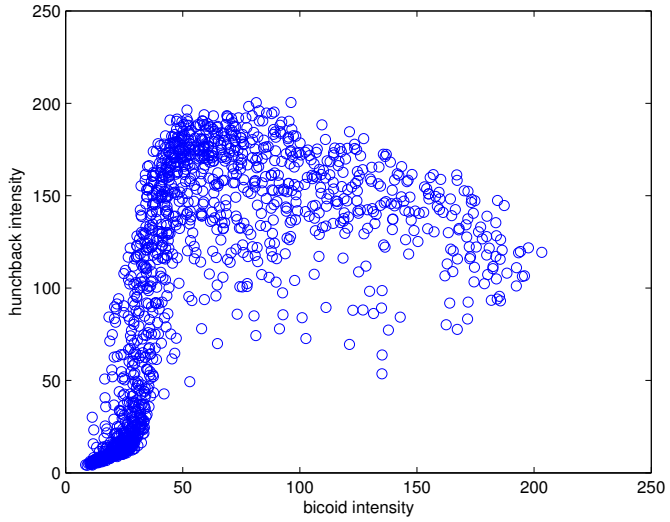


Fig. 35. Bicoid intensity vs. Hunchback intensity for the 1606 nuclei in embryo hx21 with A-P coordinate between 10% and 70% of the egg length.

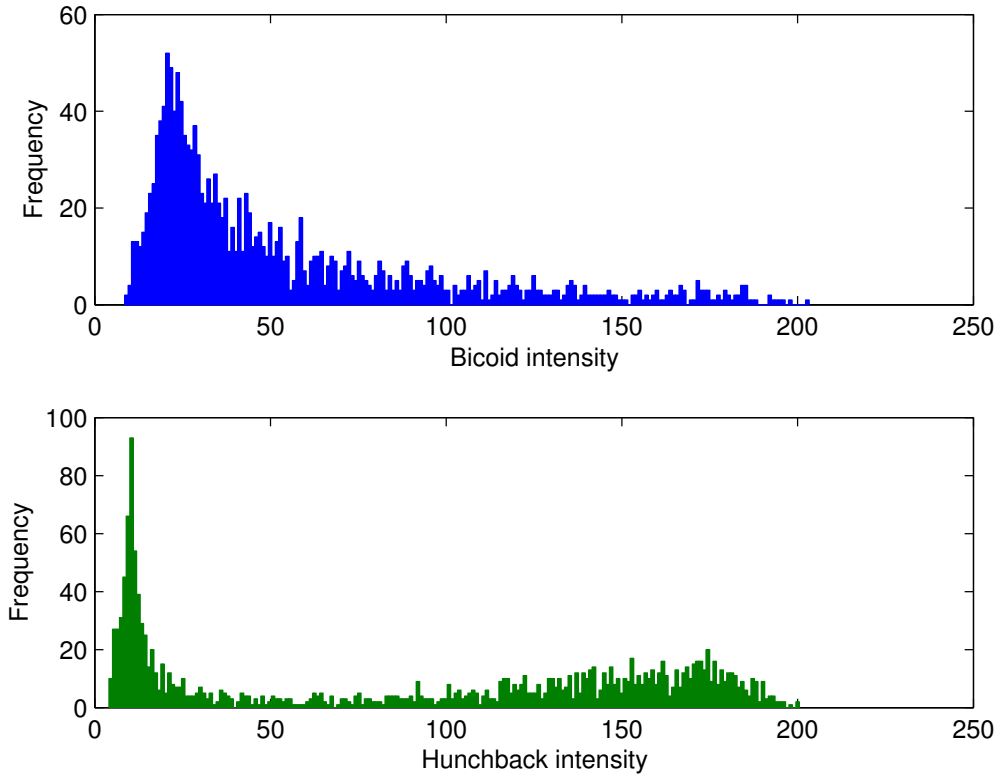


Fig. 36. Histograms of Bicoid intensity (top, blue) and Hunchback intensity (bottom, green) for the 1606 nuclei in embryo hx21 with A-P coordinate between 10% and 70% of the egg length.

Experimental Tests of Gene Expression Noise Within Single Cells

Single Cell-Tracking Experiments. Method: Ecoli cells MC4100egfp/ MC4100egfplaci were first grown in Luria Broth medium for ~5 h and then centrifuged at 8,000 rpm. The pellet was resuspended in PBS (pH7.4) at high dilution.

Holes of diameter 1 cm were bored into circular 35 mm petri dishes. 1 ml of LB agar (1.5%) was set in the 35mm petri dish after temporarily sealing its hole with a glass coverslip fixed with insulation tape. The coverslip was removed after the agar had set and the 20ul of the dilute bacterial solution was dropped on the agar at the petri dish rear. The petri dish was kept in a 37 degree incubator for 5 minutes and then the rear spotted with a glass coverslip.

Brightfield and fluorescence images were taken using a 100X objective with Kohler illumination – the sample maintained throughout at 30 degrees. Liquid LB with the right antibiotics were supplied to the petri dish at regular intervals to prevent the agar from shrinking and to allow for oxygen and nutrients to be replenished.

To calculate fluorescence per cell, line scans along the cell center of each cell was done and the sum of the intensity was treated as the fluorescence per cell – assuming that the cell width is similar for all cells. Each cell cycle fluorescence trace was fit with an exponential growth function to find the approximate gene expression rates. Cell division asymmetry was characterized from the difference in the protein content in two daughter cells just after division. $f_{\text{cell division}} = I_{\text{daughter}}/I_{\text{mother}}$.

Unregulated Images of Cells and Fluorescence Time Traces

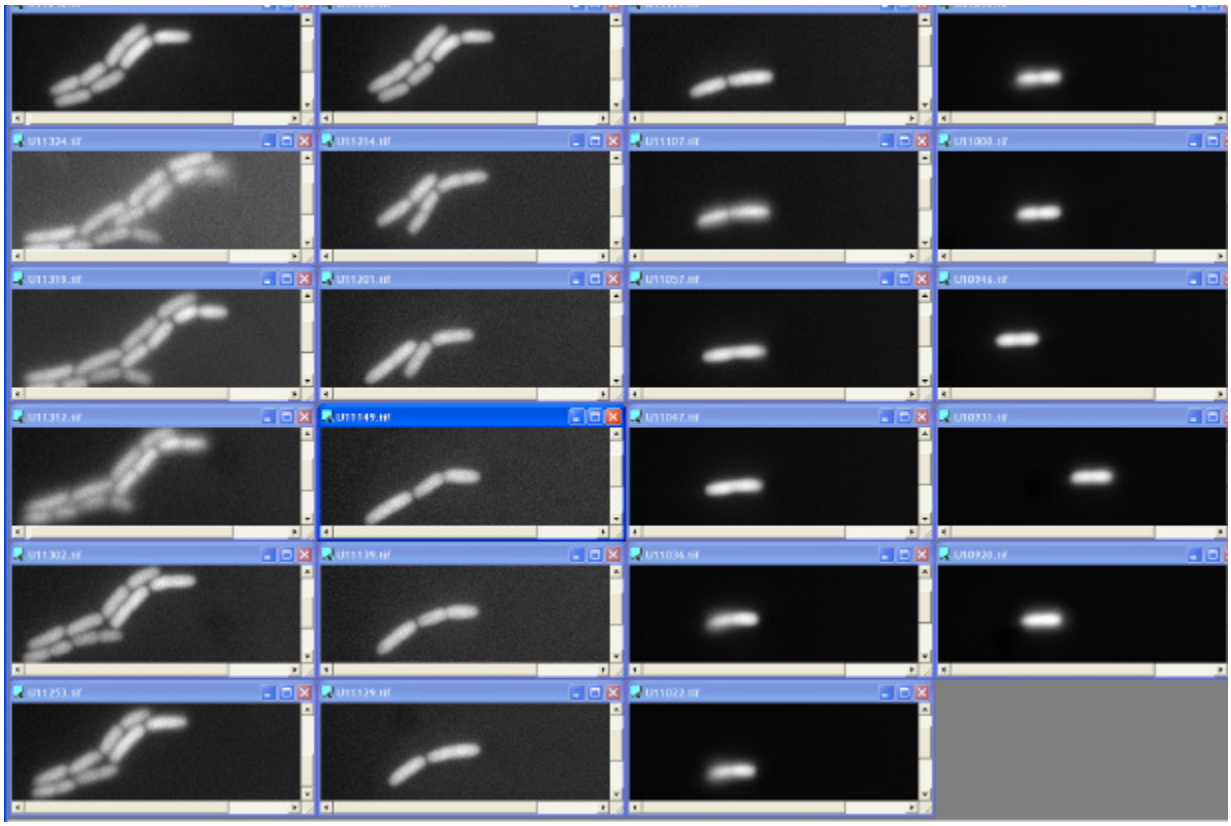


Fig. 37. Images of MC4100egfp colony through its growth from the single cell to a 12celled colony. The time points should be followed from the left bottom to right top.

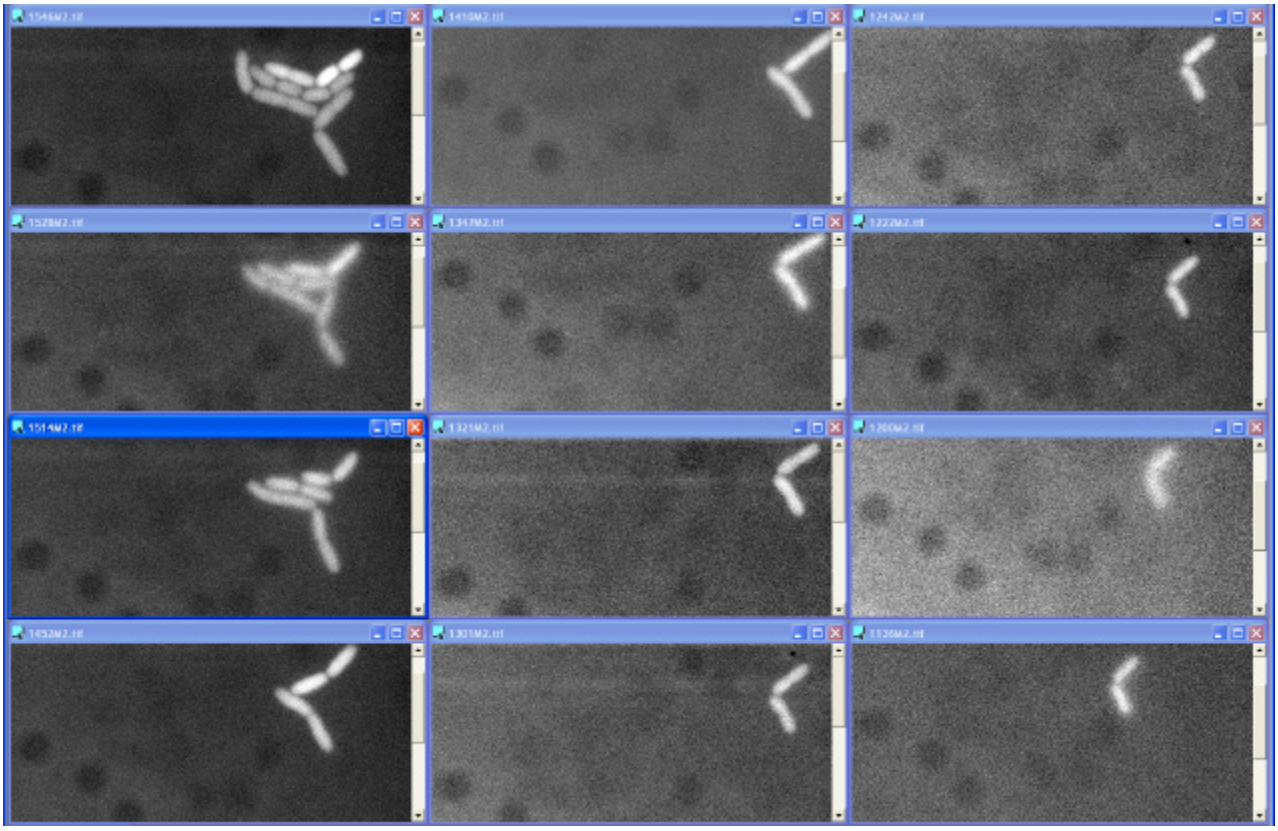


Fig. 38. Images of MC4100egfp colony M2 through its growth from the single cell to a 12celled colony. The time points should be followed from the left bottom to right top.

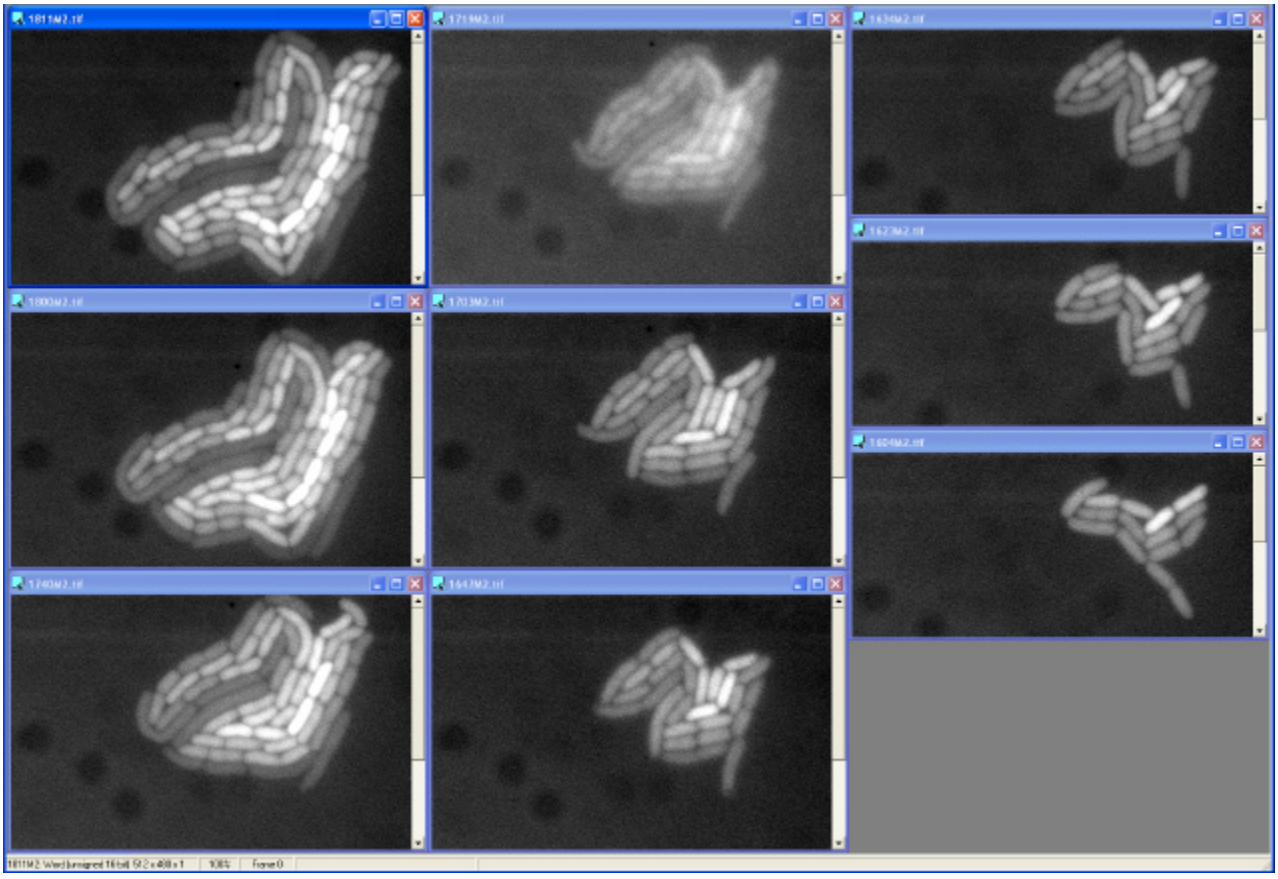


Fig. 39. Images of MC4100egfp colony M2 through its growth from a 14celled colony. The time points should be followed from the left bottom to right top.

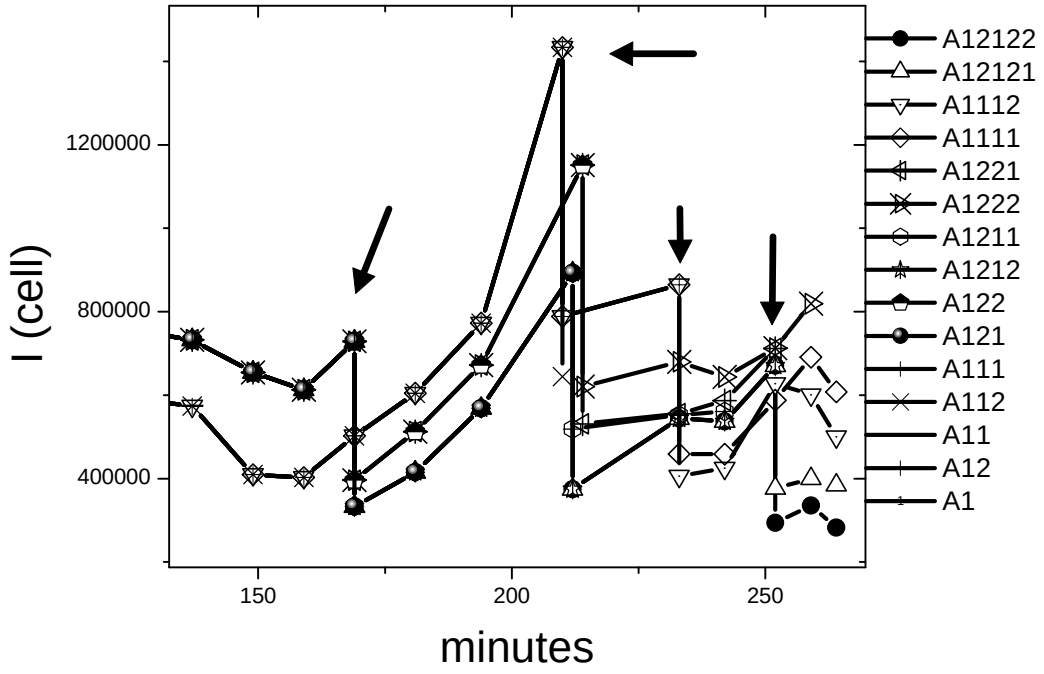


Fig. 40. Fluorescence time traces of 15 different unregulated cells from colony M1 . Arrows mark out division events.

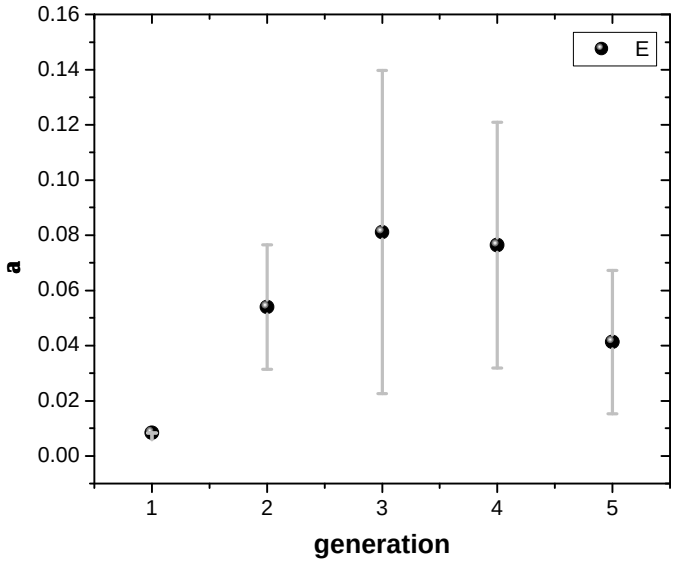


Fig. 41. Dependence of gene expression rate a on the generation number of the cell. Shows similar trend as assumed in the simulations.

Autoregulated Time Traces

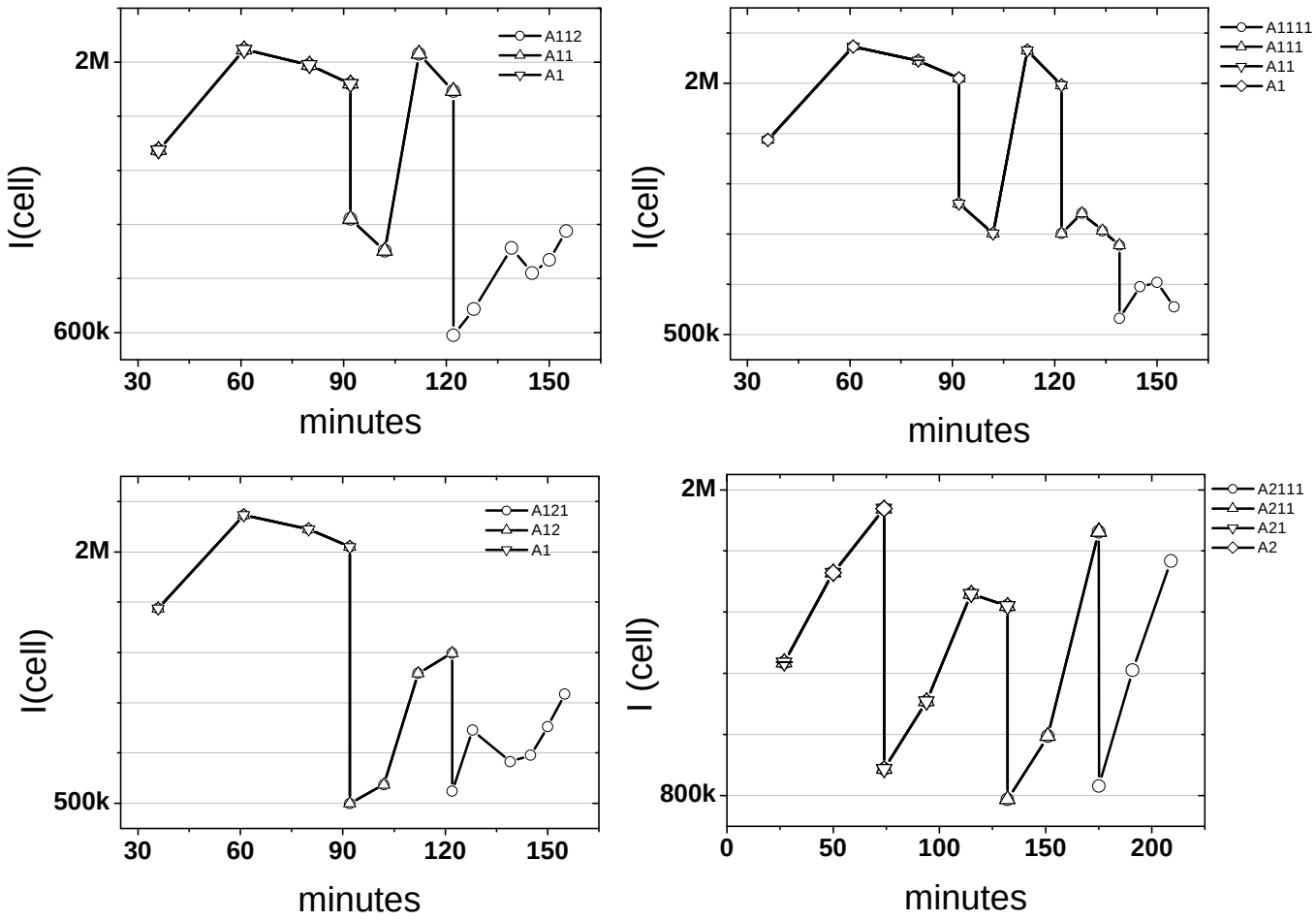


Fig. 42. Fluorescence time traces for regulated cells followed through two or three cell divisions. The nature of the rise of fluorescence before cell division is distinctly different from that observed for unregulated cells

Normalization of Fluorescence Signal With Cell Size

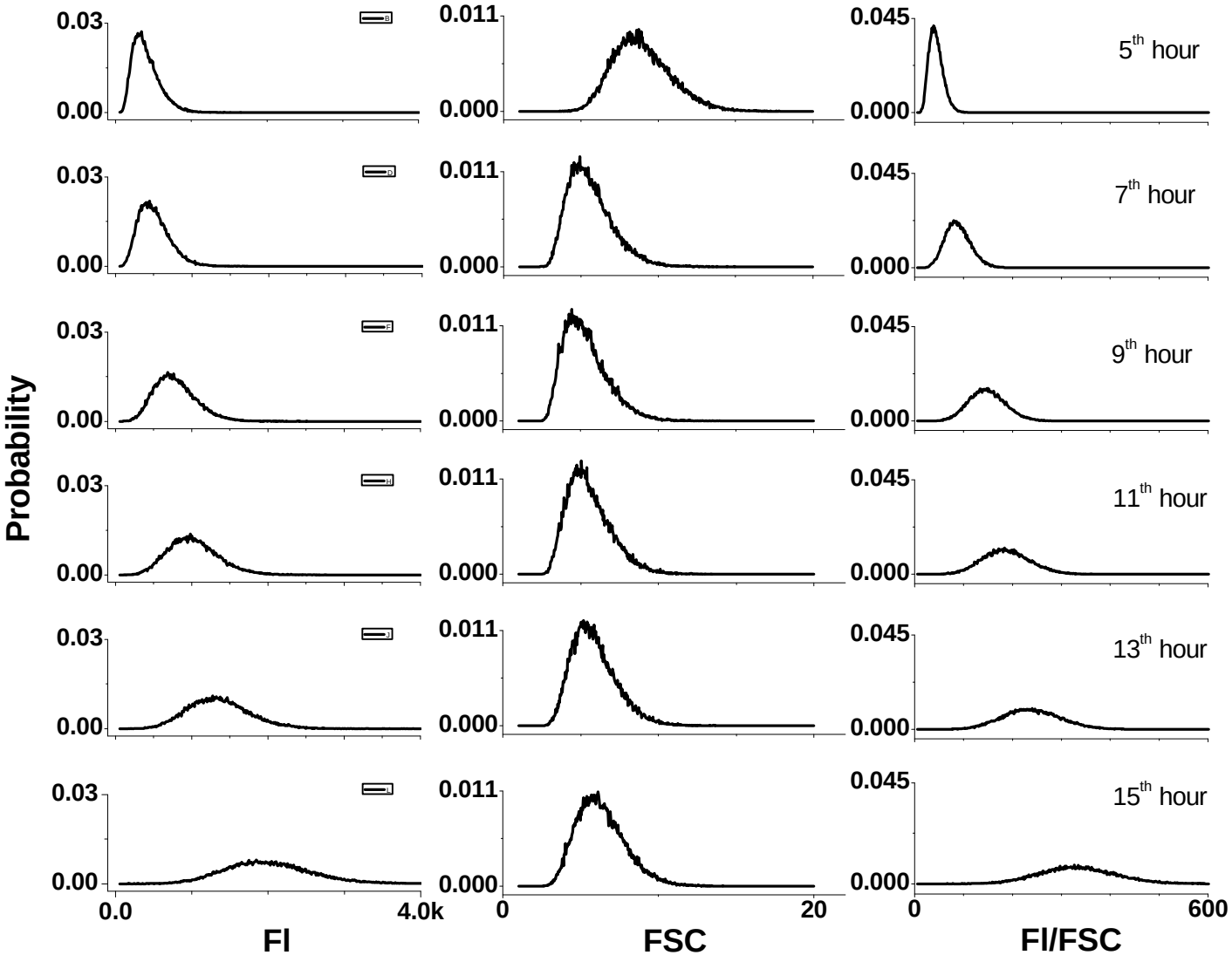


Fig. 43. Unregulated cells : From FACS data, distributions of Fluorescence (Fl :- column 1), Forward scatter (FSC:- column 2) and fluorescence normalized with the corresponding FSC (Fl/FSC :- column 3) for 5th to 15th hour (row 1 to 6). The distributions fitted to lognormal or gaussian to bring out the difference in their symmetries are plotted in the main text Fig. 4.

Bimodality

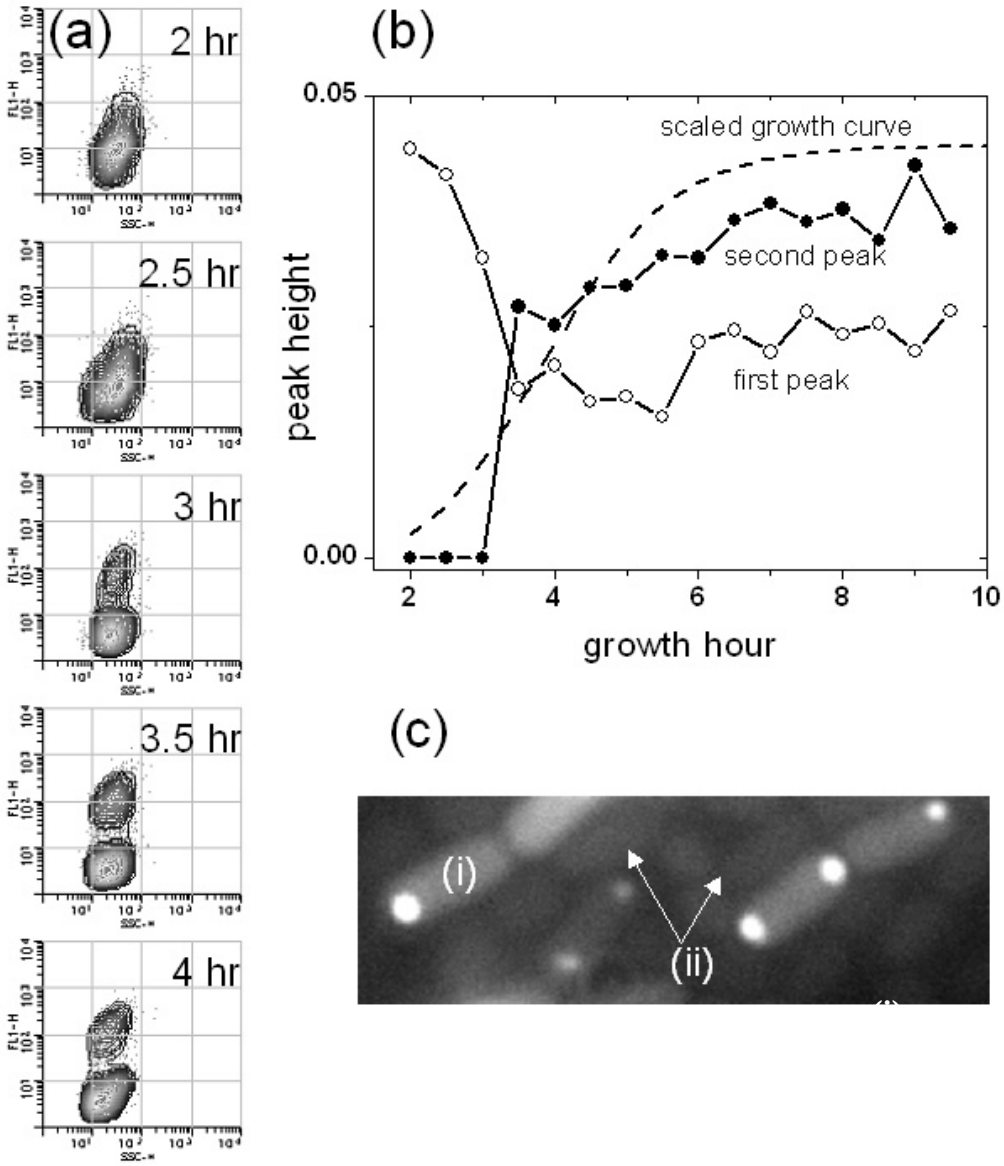


Fig. 44. (a) Scatter plot of fluorescence (Fl) vs. side scatter (SSC) for the regulated cells for hours 2 through 4. (b) Plot of peak heights (first peak and second peak) as a function of time. Response resembles that of a positive feedback switch as in Fig. 6. of the main text. (c) Image of the cells with (i) higher intensity having tight spots and (ii) lower intensities having very low but uniform haze.

References

1. Kierzek, A. M., Zaim, J., & Zielenkiewicz, P. (2001) *J. Biol. Chem.* **276**, 8165–8172.

2. Banerjee, B., Balasubramanian, S., Ananthakrishna, G., Ramakrishnan, T. V., & Shivashankar, G. V. (2004) *Biophys. J.* **86**, 3052–3059.

3. Swain, P., Elowitz, M., & Siggia, E. D. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 12795–12800.

4. Gillespie, D. T. (1977) *J. Phys. Chem.* **81**, 2340–2361.

5. Puchalka, J. & Kierzek, A. M. (2004) *Biophys. J.* **86**, 1357–1372.

6. Rossi, F. M. V., Kringstein, A. M., Spicher, A., Guicherit, O. M., & Blau, H. M. (2000) *Mol. Cell* **6**, 723–728.

7. Vilar, J. M. & Leibler, S. (2003) *J. Mol. Biol.* **331**, 981–989.

8. Rippe, K. (2001) *Trends Biochem. Sci.* **26**, 733–40.