

This un-edited manuscript has been accepted for publication in Biophysical Journal and is freely available on BioFast at <http://www.biophysj.org>. The final copyedited version of the paper may be found at <http://www.biophysj.org>.

A fluctuation method to quantify *in vivo* fluorescence data

Nitzan Rosenfeld

Depts. of Molecular Biology and Physics of Complex Systems
Weizmann Institute of Science
Rehovot 76100, Israel

Theodore J. Perkins

McGill Centre for Bioinformatics
McGill University, 3775 University Street
Montreal, Quebec H3A 2B4, Canada

Uri Alon

Depts. of Molecular Biology and Physics of Complex Systems
Weizmann Institute of Science
Rehovot 76100, Israel

Michael B. Elowitz

Division of Biology and Dept. of Applied Physics
Caltech, Pasadena CA 91125, USA

Peter S. Swain¹

Centre for Non-linear Dynamics
Dept. of Physiology, McGill University
3655 Promenade Sir William Osler
Montreal, Quebec H3G 1Y6, Canada

¹Corresponding author. Email: swain@cnd.mcgill.ca, Tel: +1 514 398 4360, Fax: +1 514 398 7452.

Abstract

Quantitative *in vivo* measurements are essential for developing a predictive understanding of cellular behavior. Here we present a technique that converts observed fluorescence intensities into numbers of molecules. By transiently expressing a fluorescently tagged protein and then following its dilution during growth and division, we observe asymmetric partitioning of fluorescence between daughter cells at each division. Such partition asymmetries are set by the actual numbers of proteins present, and thus provide a means to quantify fluorescence levels. We present a Bayesian algorithm that infers from such data both the fluorescence conversion factor and an estimate of the measurement error in the data. Our algorithm works for arbitrarily sized data sets and handles consistently any missing measurements. We verify the algorithm with extensive simulation and demonstrate its application to experimental data from *Escherichia coli*. Our technique should provide a quantitative internal calibration to systems biology studies of both synthetic and endogenous cellular networks.

Introduction

A goal of systems biology is to build a predictive, computational cellular model (1). A major challenge, however, is the lack of quantitative *in vivo* data for the many parameters required, such as protein concentrations and reaction rates (2). Well established techniques that address this issue directly do exist for the confocal microscope, such as fluorescence correlation spectroscopy (3, 4), image correlation spectroscopy (5), photon counting histogram analysis (6), and fluorescence intensity distribution analysis (7), but unambiguous results usually require considerable expertise. Another approach is to construct a model and extract parameters by fitting *in vivo* data (8, 9, 10). Nevertheless, experimental data, such as fluorescence levels of tagged proteins or immunoblots, is usually limited to unit-less ratios of expression levels that are only proportional to the actual protein concentrations. Not having direct measures can significantly hinder or complicate finding parameter values (9). In many studies (see (11, 12) for reviews), the linear relation between the concentrations of fluorescent proteins and their measured fluorescence intensities has been used to measure protein levels in living cells, though only in relative terms and not in absolute numbers. Here we present a fluctuation method for measuring and calculating the conversion factor between the amount of a fluorescent protein and the measured fluorescence level. We will denote this conversion factor by ν ; it is measured in fluorescence units per fluorescent protein (or, more generally, fluorescence units per fluorescent particle).

For a cell or a cellular compartment with a fluorescence intensity of y , the number of protein molecules, n , is given by y/ν . The measured fluorescence is, however, perturbed by measurement error. Assuming that this error is additive, each fluorescence measurement f actually satisfies

$$f = y + \epsilon = \nu n + \epsilon \quad (1)$$

with the magnitude of ϵ reflecting the size of the measurement error.

Analyzing the partitioning of proteins or other molecules in daughter cells upon cell division provides, in principal, a means to quantify fluorescence data (13, 14). We constructed a synthetic network in *Escherichia coli* that enables control of the expression levels of a reporter protein (13). The lambda phage protein, CI, was fused to yellow fluorescent protein (YFP) to make a fluorescent reporter CI-YFP. The reporter was placed on a plasmid under the control of the tetracycline promoter, pTet: a promoter tightly repressed by the tetracycline repressor, TetR. The tetR gene, itself, was chromosomally inserted into the bacterial genome where it is constitutively expressed. Consequently, only the presence of the inducer, anhydrotetracycline (aTc), which inhibits the DNA binding properties of TetR, allows fluorescent protein production. By washing out aTc, fluorescent gene expression is cut off. Fig. 1 shows time-lapse images of an *E. coli* microcolony. The colony originates from one cell taken from a population that was briefly induced by aTc and therefore contains a fixed amount of CI-YFP. No synthesis or significant degradation or photobleaching of CI-YFP takes place (see Fig. 1C), and its concentration only dilutes through microcolony growth (the average number of molecules per cell halving at each division). Analysis of such movies gives not only quantitative fluorescence levels, but also the lineage tree shown in Fig. 1B. We shall denote by f_{2i} and f_{2i+1} the fluorescence levels in the two daughter cells that originate from a mother cell with fluorescence f_i ; see Fig. 1D.

We present two methods to infer ν from the fluorescence data of such lineage trees. Assuming that fluorescent proteins are distributed with equal probability to either daughter cell at division, daughter cells will only have *on average* equal fluorescence levels. Our technique gathers information on ν by examining the deviation of actual daughter fluorescence levels from this average behavior. Such fluctuation analyses, although perhaps uncommon in molecular biology, are well established in neuroscience (15). For example, fluctuations in membrane current through a patch have long been used to infer the numbers of conducting ion channels in the patch (16). Although the mathematics of our analysis is different, we follow the same philosophy.

Method I: an approximate solution

Method I ignores the structure of the lineage tree and assumes no measurement error, i.e. $\epsilon = 0$ in Eq. 1. The data is collected into triads, each triad containing fluorescence from a mother cell and its two daughters. A triad is represented as (y_i, y_{2i}, y_{2i+1}) , where i denotes a mother cell and runs from 1 to L , say, and y_{2i} and y_{2i+1} are the fluorescence levels in the daughters.

Given a triad, we wish to infer the most probable value of ν . Using Bayes rule, the probability of ν given the data, \mathbf{y} , is

$$P(\nu|\mathbf{y}) \sim P(\mathbf{y}|\nu)P(\nu) \quad (2)$$

where $P(\mathbf{y}|\nu)$ is the likelihood of the data given a value of ν and $P(\nu)$ is the prior distribution for ν . We assume the prior distribution to be constant over a range of ν and zero elsewhere, so that, *a priori*, ν is equally likely to be found anywhere between a minimum (1, say) and a maximum (100, for example). As the fluorescent protein is neither synthesized nor degraded, the number of proteins in the parent cell, n_i , is equal to the sum of the numbers in the daughters, $n_i = n_{2i} + n_{2i+1}$. With no measurement error, conservation of proteins implies that the fluorescence of the mother cell must also equal the sum of the fluorescent values of the two daughters, $y_i = y_{2i} + y_{2i+1}$. In reality, measurement error causes this relationship to hold only approximately; method I ignores these errors.

Calculation of $P(\nu|\mathbf{y})$ involves evaluating the likelihood, $P(\mathbf{y}|\nu)$. Considering one triad, y_i, y_{2i} , and y_{2i+1} , denoted by the vector \mathbf{t} , the likelihood obeys

$$P(\mathbf{t}|\nu) = \int d\mathbf{n} \delta(\mathbf{t} - \nu\mathbf{n})P(\mathbf{n}) \quad (3)$$

where the n_i are approximated to be continuous, and the vector notation implies three delta functions, one for each member of the triad. The probability $P(\mathbf{n}) = P(n_i, n_{2i}, n_{2i+1})$ for the protein numbers can be factorized

$$P(\mathbf{n}) = P(n_{2i+1}|n_i, n_{2i})P(n_{2i}|n_i)P(n_i) \quad (4)$$

where $P(n_{2i+1}|n_i, n_{2i})$ is set by the constraint $n_{2i+1} = n_i - n_{2i}$, and $P(n_{2i}|n_i)$ is an even binomial distribution: each fluorescent molecule has the same chance of going to either daughter upon cell division.

An even binomial distribution can be approximated by a normal distribution with mean $n_i/2$ and standard deviation $\sqrt{n_i}/2$ (17). Using a delta function to enforce the

conservation of protein numbers, Eq. 4 then becomes

$$P(n_i, n_{2i}, n_{2i+1}) \sim \delta(n_{2i+1} - n_i + n_{2i}) \times \frac{e^{-\frac{(2n_{2i}-n_i)^2}{2n_i}}}{\sqrt{n_i}} \times \frac{1}{n_i^{\max} - n_i^{\min}} \quad (5)$$

where the prior, $P(n_i)$, is a uniform bounded distribution such that n_i lies anywhere between n_i^{\min} and n_i^{\max} *a priori*. Inserting Eq. 5 into Eq. 3, and using the relation $\delta(y - \nu n) = \delta(n - y/\nu)/\nu$ to carry out the integrations gives

$$P(\mathbf{t}|\nu) \sim \delta(y_{2i+1} - y_i + y_{2i}) \times \frac{e^{-\frac{(2y_{2i}-y_i)^2}{2\nu y_i}}}{\sqrt{\nu y_i}} \times \frac{1}{y_i^{\max} - y_i^{\min}} \quad (6)$$

Assuming independent measurements of each triad, $P(\mathbf{y}|\nu)$ for the full set of L triads is a product of terms like Eq. 6

$$P(\mathbf{y}|\nu) \sim \nu^{-\frac{L}{2}} \left(\prod_i y_i^{-\frac{1}{2}} \right) \exp \left[-\frac{1}{\nu} \sum_i \frac{(2y_{2i} - y_i)^2}{2y_i} \right] \quad (7)$$

assuming that the conservation constraints are satisfied. With a constant prior, $P(\nu)$, the posterior, see Eq. 2, has the same form as Eq. 7.

The most probable value of ν maximizes the posterior probability, and is found by differentiation. The maximum occurs at $\nu = \nu^*$, with

$$\nu^* = \left\langle \frac{[y_{2i} - y_{2i+1}]^2}{y_i} \right\rangle \quad (8)$$

where angled brackets denote an average over all L triads, and is equal to the average of the square of the difference in fluorescence of the two daughters divided by the value of the fluorescence in the mother cell. By evaluating the second derivative of Eq. 7 at $\nu = \nu^*$, the error in the inferred value, Eq. 8, is estimated as $\pm \nu^*/\sqrt{L}$.

Method I, exemplified by Eq. 8, is equivalent to the more *ad hoc* approach used previously (13, 14). Although it does ignore measurement error, it involves only a few simple computations.

Method II

We assume that each fluorescence measurement satisfies Eq. 1, where the measurement error term, ϵ_i , or equivalently $f_i - y_i$, has a normal distribution with zero mean and standard deviation σ . The size of σ sets the magnitude of the measurement error. The posterior for both ν and σ satisfies $P(\nu, \sigma|\mathbf{f}) \sim P(\mathbf{f}|\nu, \sigma)P(\nu, \sigma)$, or

$$P(\nu, \sigma|\mathbf{f}) \sim \int d\mathbf{y} P(\mathbf{f}|\mathbf{y}, \nu, \sigma)P(\mathbf{y}|\nu, \sigma)P(\sigma)P(\nu) \quad (9)$$

using the product rule of probability theory and noting that the prior for σ is independent of ν .

Including measurement error

The probability of the data \mathbf{f} given \mathbf{y} depends only on the measurement error, $P(\mathbf{f}|\mathbf{y}, \nu, \sigma) = P(\mathbf{f}|\mathbf{y}, \sigma)$. Using the normal distribution model for ϵ_i , we have

$$P(\mathbf{f}|\mathbf{y}, \sigma) \sim \frac{1}{\sigma^N} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (f_i - y_i)^2 \right] \quad (10)$$

assuming that the errors in each measurement are independent and that there are N measurements.

Including the tree

The second probability in Eq. 9, $P(\mathbf{y}|\nu, \sigma)$, is independent of the measurement error σ . From conservation of proteins, the y_i obey $P(y_{2i+1}|y_i, y_{2i}) = \delta(y_{2i+1} - y_i + y_{2i})$ for a mother cell and its daughters. Considering, for example, Fig. 1D, factorizing $P(\mathbf{y}|\nu)$ implies

$$\begin{aligned} P(\mathbf{y}|\nu) &= P(y_1) \times P(y_2|y_1, \nu) \delta(y_3 - y_1 + y_2) \times P(y_4|y_2, \nu) \delta(y_5 - y_2 + y_4) \\ &\quad \times P(y_6|y_3, \nu) \delta(y_7 - y_3 + y_6) \end{aligned} \quad (11)$$

where $P(y_1)$ is the prior for the fluorescence level in the first cell. For one mother-daughter pair, $P(y_{2i}|y_i, \nu)$ is given by the exponential term of Eq. 6, and so

$$\begin{aligned} P(\mathbf{y}|\nu) &\sim P(y_1) \times \delta(y_3 - y_1 + y_2) \delta(y_5 - y_2 + y_4) \delta(y_7 - y_3 + y_6) \\ &\quad \times \frac{\nu^{-\frac{3}{2}}}{\sqrt{y_1 y_2 y_3}} \exp \left[-\frac{1}{\nu} \left\{ \frac{(2y_2 - y_1)^2}{2y_1} + \frac{(2y_4 - y_2)^2}{2y_2} + \frac{(2y_6 - y_3)^2}{2y_3} \right\} \right] \end{aligned} \quad (12)$$

after re-arranging the delta functions.

The posterior distribution

Assuming constant (but bounded) priors for ν , y_1 , and σ , the posterior distribution satisfies

$$\begin{aligned} P(\nu, \sigma|\mathbf{f}) &\sim \int d\mathbf{y} \delta(y_3 - y_1 + y_2) \delta(y_5 - y_2 + y_4) \delta(y_7 - y_3 + y_6) \\ &\quad \times \frac{1}{\sigma^7} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^7 (f_i - y_i)^2 \right] \\ &\quad \times \frac{\nu^{-\frac{3}{2}}}{\sqrt{y_1 y_2 y_3}} \exp \left[-\frac{1}{\nu} \left\{ \frac{(2y_2 - y_1)^2}{2y_1} + \frac{(2y_4 - y_2)^2}{2y_2} + \frac{(2y_6 - y_3)^2}{2y_3} \right\} \right] \end{aligned} \quad (13)$$

for the tree of Fig. 1D (a more general expression is given in the appendix). Partly integrating Eq. 13, y_3 is replaced by $y_1 - y_2$, y_5 by $y_2 - y_4$, and y_7 by $y_1 - y_2 - y_6$.

It is instructive to consider only three cells, i.e. just one division event, and a *given* σ : all the integrals in the equivalent of Eq. 13 can be evaluated analytically, and the value of ν that maximizes this posterior is

$$\nu^* \simeq \frac{(f_2 - f_3)^2 - 2\sigma^2}{(2f_1 + f_2 + f_3)/3} \quad (14)$$

In the limit of $\sigma \rightarrow 0$, f_1 exactly equals $f_2 + f_3$, and Eq. 14 recovers Eq. 8. Notice that the best estimate for y_1 , the denominator of Eq. 14, is now a weighted sum of f_1 and $f_2 + f_3$, the latter being a second estimate of y_1 in a data set with measurement errors.

To evaluate Eq. 13 in general, we use the variable elimination method (18) to numerically find the integral for any given ν and σ . Let \mathbf{y}_s denote the set of independent \mathbf{y} variables — those not defined by the conservation of number constraints. As the integrand factorizes into a product of terms, one for each triad in the tree, the M -dimensional integral over \mathbf{y}_s transforms into a series of tractable 2-dimensional computations. Typically, we perform this calculation over a grid defined by *a priori* ranges of ν and σ , thus giving a two-dimensional posterior distribution (see Fig. 2A for an example). For the case of a normal distribution model of measurement error, however, we can derive an accurate estimate of σ (see Appendix):

$$\sigma^* \simeq \left[\frac{\min_{\mathbf{y}_s} \frac{1}{2} \sum_i (f_i - y_i)^2}{N - M} \right]^{\frac{1}{2}} \quad (15)$$

where the delta function constraints in Eq. 13 hold. There are N data points and M ($< N$) independent \mathbf{y} variables. We use this estimate with twenty steps of Golden section search to efficiently explore the *a priori* interval for ν . The results of Fig. 3 were generated with this method. For such simulated data, we compared the posteriors found with those generated using the true value of σ instead of σ^* ; their difference is negligible.

Extra and missing data

It is often possible to measure fluorescence levels in a cell several times before it divides, see Fig. 1B. The data is then stored in a matrix, rather than a vector, where f_{ij} is the j 'th measurement of fluorescence in cell i . Each measurement for cell i improves the estimate of y_i . The number of y_i variables does not change, and Eq. 10 just gains more terms. For example, with C measurements per cell

$$P(\mathbf{f}|\mathbf{y}, \sigma) \sim \frac{1}{\sigma^{NC}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^N \sum_{j=1}^C (f_{ij} - y_i)^2 \right] \quad (16)$$

and the general form of Eq. 13 (see appendix) remains essentially unchanged.

Cells do not all divide synchronously, Fig. 1B, and so for some we have more measurements than for others. Consequently, the C in Eq. 16 vary from cell to cell. Fig. 1B also shows that data is missing from parts of the lineage tree, particularly at the extremities. Usually these cells are obscured by surrounding cells, as large microcolonies no longer grow in a plane and ‘terrace’. Excluding the missing data corresponds to deleting the unnecessary terms (including those generated by the daughters of missing cells) in the general form of Eq. 13. These cases are automatically handled, our code generating as many terms as is appropriate for each cell and only when the corresponding data exist.

Experimental methods

Cultures of ‘ λ -cascade’ strains (13) were grown overnight in LB +15 μ g/mL kanamycin at 37°C from single colonies and diluted 1:100 in MSC media (M9 minimal medium + 0.6% succinate + 0.01% casamino acids + 0.15 μ g/ml biotin + 1.5 μ M thiamine). Cultures were grown to $OD_{600} \sim 0.1$ at 32°C and then induced by adding aTc to a concentration of 100 ng/mL for 3 minutes at ambient temperature, followed by 2 washes with MSC to remove aTc. Cells were allowed to grow and then diluted to give $\simeq 1$ cell per visual field when placed between a coverslip and 1.5% low melt MSC agarose. Growth of microcolonies was observed at 32°C using a Leica DMIRB/E automated fluorescence microscope at 100x magnification with a mercury light source and YFP filter cube Chroma #41028.

Custom software was used to control the microscope and related equipment (Ludl motorized stage and Hamamatsu Orca ER CCD camera), via ImagePro Plus and ScopePro packages (Media Cybernetics). Fluorescence background values were estimated from regions of the fluorescent images containing no cells. One background value was chosen for each movie (the minimum of the measured background levels for the first 10 to 20 frames). Cellular autofluorescence was measured in cells containing no YFP reporters and was low, having values similar to the change in background fluorescence from image to image. An autofluorescence value was therefore selected within this range that led to the most constant YFP signal for the entire microcolony and to constant YFP levels in cell divisions (so that the sum of YFP fluorescence in the daughters would equal that in the mother cell). Flatfield corrections were found to be negligible using an analytic correction method developed previously (19).

With this normalization, the total colony YFP remains constant (to within 5%) during the first 4 hours of growth. For later times, there is more variation (arising from crosstalk with the Cyan Fluorescent Protein designed to be induced at low CI-YFP levels in the lambda cascade strains (13)). The measurement error model incorporated in method II automatically attaches less weight to those cells whose fluorescence values are of order σ or lower, and the inferred ν is not significantly changed when data from the last generations of the lineage tree in Fig. 1B is ignored. Cellular YFP levels are given in Fig. 1C.

Software written in Matlab (MathWorks) identified and tracked cells from the phase contrast images and quantified their fluorescence levels using the segmented images to identify the appropriate pixels for each cell (13). Cellular fluorescence was measured by summing the fluorescence intensities of all pixels within a particular cell. Typical intervals between exposures were 9 minutes (for accurate tracking), but YFP fluorescence images were taken on alternate frames to reduce photobleaching. Images were usually acquired for around 8 hours, with colonies reaching 8 or 9 generations.

Results

The performance of the algorithm is determined by two competing effects: first, as the number of molecules increases binomial deviations become less significant relative to mean values, and so the numbers of molecules in daughter cells become effectively indistinguishable; second, low numbers of molecules have small signal to noise ratios and measurement error can potentially swamp binomial deviations.

Fig. 2A shows the two dimensional posterior inferred by method II from the data of Fig. 1. The most probable values of ν and σ are given by the posterior maximum. The distribution is orientated approximately parallel to the coordinate axes implying that the two parameters can be inferred independently, and that the errors in such inference will be given by the width of the distribution at half maximum along the appropriate axis. Summing over all values of σ gives the marginalized distribution for ν shown in Fig. 2B. For this data set, $\nu^* \simeq 15 \pm 4$ fluorescence units per fluorescent particle, implying that the predicted number of fluorescent proteins ranges from $\simeq 840$ in the initial cell to $\simeq 10$ for the generation 8 cells. Method I predicts $\nu^* \simeq 15 \pm 1$. The inferred value of ν is consistent (within a factor of two) over 4 separate experiments, where a microcolony was grown from a different initial cell (Fig. 2B). This level of accuracy is certainly high enough to provide parameters for cellular models, where only ballpark estimates are usually required (9), and could be improved by increasing the number of measurements per cell or the levels of intracellular fluorescent proteins (see below).

To illustrate the importance of the signal to noise ratio (f_i/σ for a cell with fluorescence f_i), we simulated data for an 8 generation tree. Starting from an initial number of proteins in the first cell, a microcolony was created by equally binomially partitioning protein into daughter cells. Each data point was multiplied by ν_0 , arbitrarily set to 25. Normally distributed samples with zero mean and standard deviation σ were added to each data point to include measurement errors. If, for example, three measurements per cell were desired, three different normal samples were added to the original data point to give three final data points.

Fig. 3A shows that Method II performs robustly: its accuracy increases steadily as the number of data points grows. Method I, which ignores measurement error, performs as well as method II only for those lineage trees whose cells all have a high signal to noise ratio. We simulated two types of data: one generated from an initial cell having 500 molecules (low signal to noise ratios) and the other generated from an initial cell containing 5000 molecules (high signal to noise). Although lineage trees with 7 generations were initially created, the inference algorithms were run on data sets with just the first 3 generations and then new data added generation by generation to explore the inclusion of additional layers of the lineage tree. In all cases, the true value of ν , ν_0 , was set to 25. We use a relative measure, $i_\epsilon = |\log_2(\nu/\nu_0)|$, to score an inferred value of ν . An inference score $i_\epsilon = 0$ is thus an exact inference, while $i_\epsilon = 1$ implies that the inferred value is either twice or half the true value.

Although large numbers of proteins increase the signal to noise ratio, too many, as mentioned earlier, can degrade the inference. Large numbers lead to very tight binomial distributions and so to potentially unmeasurable differences between the number of proteins in each daughter cell. For 5,000 molecules ($\simeq 3\mu M$), see Fig. 3B, and 50,000 molecules (not shown), accurate inference is not significantly affected, at least for simulated data: the inference score for the 50,000 case increases by approximately 20% for colonies of 6 or 7 generations. Such high numbers of molecules may be more frequent in eukaryotic cells.

Higher σ degrades inference, but this degradation is reduced by increasing the number of measurements per cell, see Fig. 3B. Six measurements gives accurate inference for σ as high as 200 (i.e. 8ν). For the same data, method II performed as well as or better than method I 78% of the time, increasing to 93% of the time when

$\sigma = 200$.

Discussion

The difference in fluorescence levels between two daughter cells after cell division is determined by the number of fluorescent proteins in the mother cell: the difference is on average larger if the number of proteins in the mother cell increases. We exploit this phenomenon to deduce *in vivo* numbers of molecules by following a cell that has transiently expressed fluorescent protein and recording the daughter cell fluorescence levels as the fluorescent protein is diluted out during growth. Our method assumes equal binomial partitioning of proteins at cell division, so that each protein has the same chance of going to either daughter cell.

We have introduced two algorithms to infer ν . Method I, exemplified by Eq. 8, is fast and easy to compute. It is reliable when the signal to noise is significantly greater than one. Method II, Eq. 13, although more computationally demanding, is valid for both high and low signal to noise ratios. It returns both the posterior probability for ν and the measurement error σ . Matlab (MathWorks) code for both methods is available on request.

Although we use a normal distribution model for measurement error, other distributions can be adopted providing they allow Eq. 13 to be factored into triad terms. For example, multiplicative measurement error with a corresponding log normal distribution only changes the middle term of Eq. 13 to

$$\prod_i (\sigma y_i)^{-1} \exp \left[-\frac{1}{2\sigma^2} \sum_i (\log f_i - \log y_i)^2 \right] \quad (17)$$

with the variable elimination method working as before.

The fluorescent protein of interest must only be transiently expressed, and then its expression fully repressed once fluorescence measurements are begun. The copy number of the protein's gene is therefore not important providing repression remains tight after the transient expression. Consequently, the fluorescent protein can be conveniently added on a plasmid. Given the value of ν of YFP, other fluorescent proteins can be quantified by comparing their expression to that of YFP in a bacterial strain that expresses both proteins from identical promoters (13).

Our analysis assumes even partitioning of fluorescent proteins into daughter cells, i.e. a protein is as equally likely to go to one daughter as to the other. This assumption is reasonable for a repressor protein: non-specific DNA binding presumably causes most to be carried from mother to daughters by (evenly partitioning) chromosomes. For cytosolic proteins in cells that divide asymmetrically, for example by budding rather than by fission, the situation may be more complicated. Our algorithm could be adapted to include such asymmetries, where, for example, the probability of a protein going to a particular daughter cell could be proportional to the volume of the daughter cell. Alternatively, it may be possible to pre-process the data, restricting the analysis to daughter cells of equal size where division events are, presumably, even. In our movies, over 85% of division events generated daughters with a 5% difference in cell volume or less.

A protein that exists in several different multimer forms can cause additional difficulties. If the distribution of the protein between multimers changes at each

measurement and particularly at each cell division, not only must our algorithm be extended but many more measurements will be needed to gain reliable statistics. The fluorescent protein considered here, CI-YFP, dimerizes. Nevertheless, given that its dissociation constant is approximately 10 nM (20, 21), we assume that it only exists in dimer form, and so only need to halve the inferred value of ν to find the proportionality constant per YFP molecule.

By calibrating fluorescence measurements, our method allows parameters fit to network output to be expressed in absolute rather than relative units, and so enables information from different experiments to be easily combined into a larger, predictive framework. The technique can be applied in parallel to measurements of network function (13) and properties (14), and potentially to eukaryotic cells.

Acknowledgments

We thank Derek Bowie, Robert Sidney Cox III, Jon Young, and the anonymous referees for useful comments. M.B.E. acknowledges a CASI award from the Burroughs Wellcome Fund and the Searle Scholars Program. U.A. and M.B.E. are supported by the Human Frontiers Science Program. P.S.S. is supported by N.S.E.R.C. (Canada) and by a Tier II Canada Research Chair.

Appendix

General expression for the posterior probability

For a complete lineage tree with N cells, Eq. 13 becomes

$$P(\nu, \sigma | \mathbf{f}) \sim \int d\mathbf{y} \sum_{i=1}^{N-M} \delta(y_{2i} + y_{2i+1} - y_i) \times \frac{1}{\sigma^N} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (f_i - y_i)^2 \right] \\ \times \frac{\nu^{-\frac{N-M}{2}}}{\prod_{i=1}^{N-M} y_i^{\frac{1}{2}}} \exp \left[-\frac{1}{\nu} \sum_{i=1}^{N-M} \frac{(2y_{2i} - y_i)^2}{2y_i} \right] \quad (18)$$

with $N - M = (N - 1)/2$.

Derivation of an estimate for the measurement error

The posterior for σ satisfies

$$P(\sigma | \mathbf{f}) = \int d\nu P(\sigma, \nu | \mathbf{f}) \quad (19)$$

where $P(\sigma, \nu | \mathbf{f})$ is given by Eq. 18. After integrating out the delta functions in Eq. 18, the number of y variables in the integral drops to M because of the conservation of numbers constraints. We denote this set of M variables by $\mathbf{y}_s = \{y_1, y_2, \dots, y_{N-M}\}$.

To estimate the remaining integral in Eq. 18 (and so evaluate Eq. 19), the exponent of the middle term of Eq. 18 can be re-arranged into a quadratic form in the \mathbf{y}_s , i.e. as $-\frac{1}{\sigma^2} \left(\frac{1}{2} \mathbf{y}_s^T \mathbf{A} \mathbf{y}_s + \mathbf{b}^T \mathbf{y}_s + c \right)$, where \mathbf{A} is a symmetric $M \times M$ matrix and \mathbf{b} is a $M \times 1$ vector. This quadratic form can be diagonalized through the transformation

$$\tilde{\mathbf{y}} = R(\mathbf{y}_s + A^{-1}\mathbf{b})/\sigma \quad (20)$$

with R the matrix of eigenvectors of A . The exponent then becomes $-\left[\frac{1}{2}\tilde{\mathbf{y}}^T\Lambda\tilde{\mathbf{y}} + \frac{\tilde{c}}{\sigma^2}\right]$, where Λ is the diagonal matrix of eigenvalues λ_i of A and $\tilde{c} = c - \frac{1}{2}\mathbf{b}^T A^{-1}\mathbf{b}$.

Once diagonalized, the middle term of Eq. 18 can be written as a product of normal distributions, one for each eigenvalue λ_i , which we can integrate. Defining $\mathcal{N}(x|0, \sigma) = \exp[-x^2/(2\sigma^2)]/(\sqrt{2\pi}\sigma)$, the middle term satisfies

$$\frac{1}{\sigma^N} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (f_i - y_i)^2\right] = \sigma^{-N} e^{-\frac{\tilde{c}}{\sigma^2}} \prod_i \mathcal{N}\left(\tilde{y}_i|0, \lambda_i^{-\frac{1}{2}}\right) \quad (21)$$

remembering Eq. 20.

Integrating out the delta functions in Eq. 18 leads to its last term becoming a function of \mathbf{y}_s rather than \mathbf{y} :

$$\frac{\nu^{-\frac{N-M}{2}}}{\prod_{i=1}^{N-M} y_i^{\frac{1}{2}}} \exp\left[-\frac{1}{\nu} \sum_{i=1}^{N-M} \frac{(2y_{2i} - y_i)^2}{2y_i}\right] = \frac{\nu^{-\frac{N-M}{2}}}{\alpha(\mathbf{y}_s)^{\frac{1}{2}}} e^{-\frac{\beta(\mathbf{y}_s)}{\nu}} \quad (22)$$

where $\alpha(\mathbf{y}_s)$ and $\beta(\mathbf{y}_s)$ are algebraic functions of \mathbf{y}_s .

The posterior distribution for σ , from Eq. 19, is thus

$$P(\sigma|\mathbf{f}) \sim \sigma^{-(N-M)} e^{-\frac{\tilde{c}}{\sigma^2}} \int d\tilde{\mathbf{y}} \prod_i \mathcal{N}\left(\tilde{y}_i|0, \lambda_i^{-\frac{1}{2}}\right) \alpha(\mathbf{y}_s)^{-\frac{1}{2}} \int_0^\infty d\nu \nu^{-\frac{N-M}{2}} e^{-\frac{\beta(\mathbf{y}_s)}{\nu}} \quad (23)$$

from Eqs. 21 and 22. Evaluating the integral over ν gives

$$P(\sigma|\mathbf{f}) \sim \sigma^{-(N-M)} e^{-\frac{\tilde{c}}{\sigma^2}} \int d\tilde{\mathbf{y}} \prod_i \mathcal{N}\left(\tilde{y}_i|0, \lambda_i^{-\frac{1}{2}}\right) \alpha(\mathbf{y}_s)^{-\frac{1}{2}} \beta(\mathbf{y}_s)^{-\frac{N-5}{4}} \quad (24)$$

where

$$\mathbf{y}_s = \sigma R^{-1}\tilde{\mathbf{y}} - A^{-1}\mathbf{b} \quad (25)$$

from Eq. 20.

Most of the contribution to the integral in Eq. 24 will come from the maxima of the normal distributions at $\tilde{y}_i = 0$. At these values, Eq. 25 shows that \mathbf{y}_s will have little σ dependence and consequently that the entire integral can be approximated as being independent of σ . The term before the integral dominates, and maximizing this term with respect to σ gives the estimate in Eq. 15. For 5000 simulated data sets with randomly assigned measurement errors of σ_0 , the mean of $\log_2(\sigma^*/\sigma_0)$ was $\simeq -0.009$, a negligible difference for accurately inferring ν .

References

1. Ideker, T., T. Galitski, and L. Hood. 2001. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2:343–372.
2. Gilman A. and A. P. Arkin. 2002. Genetic ‘code’: representations and dynamical models of genetic components and networks. *Annu. Rev. Genomics Hum. Genet.* 3:341–369.
3. Elson, E. L. and D. Magde. 1974. Fluorescence correlation spectroscopy I. Conceptual basis and theory. *Biopolymers.* 13:1–27.

4. Magde, D., E. L. Elson, and W. W. Webb. 1974. Fluorescence correlation spectroscopy II. An experimental realization. *Biopolymers*. 13:29–61.
5. Wiseman, P. W., J. A. Squier, M. H. Ellisman, and K. R. Wilson. 2000. Two-photon image correlation spectroscopy and image cross-correlation spectroscopy. *J. Microsc.* 200:14–25.
6. Chen, Y., J. D. Muller, P. T. So, and E. Gratton. 1999. The photon counting histogram in fluorescence fluctuation spectroscopy. *Biophys. J.* 77:553–567.
7. Kask, P., K. Palo, D. Ullmann, and K. Gall. 1999. Fluorescence-intensity distribution analysis and its application in biomolecular detection technology. *Proc. Natl. Acad. Sci. U. S. A.* 96:13756–13761.
8. Ronen, M., R. Rosenberg, B. I. Shraiman, and U. Alon. 2002. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. U. S. A.* 99:10555–10560.
9. Brown K. S. and J. P. Sethna. 2003. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E.* 68:021904.
10. Jaeger, J., S. Surkova, M. Blagov, H. Janssens, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz. 2004. Dynamic control of positional information in the early Drosophila embryo. *Nature*. 430:368–371.
11. Kaern, M., T. C. Elston, W. J. Blake, and J. J. Collins. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6:451–464.
12. Sprinzak D. and M. B. Elowitz. 2005. Reconstruction of genetic circuits. *Nature*. 438:443–448.
13. Rosenfeld N., J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. 2005. Gene regulation at the single-cell level. *Science*. 307:1962–1965.
14. Golding, I., J. Paulsson, S. M. Zawilski, and E. C. Cox. 2005. Real-time kinetics of gene activity in individual bacteria. *Cell*. 123:1025–1036.
15. Traynelis, S. F. and F. Jaramillo. 1998. Getting the most out of noise in the central nervous system. *Trends Neurosci.* 21:137–145.
16. Neher, E. and C. F. Stevens. 1977. Conductance fluctuations and ionic pores in membranes. *Annu. Rev. Biophys. Bioeng.* 6:345–381.
17. Mathews, J. and R. L. Walker. 1970. *Mathematical methods of physics*. Addison Wesley. New York, New York.
18. Zhang, N. L. and D. Poole. 1996. Exploiting causal independence in Bayesian network inference. *J. Artif. Intell.* 5:301–328.
19. Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain. 2002. Stochastic gene expression in a single cell. *Science*. 297:1183–1186.

20. Johnson, A. D., C. O. Pabo, and R. T. Sauer. 1980. Bacteriophage lambda repressor and cro protein: interactions with operator DNA. *Methods Enzymol.* 65:839–856.
21. Koblan, K. S. and G. K. Ackers. 1991. Energetics of subunit dimerization in bacteriophage lambda cI repressor: linkage to protons, temperature, and KCl. *Biochemistry.* 30:7817–7821.

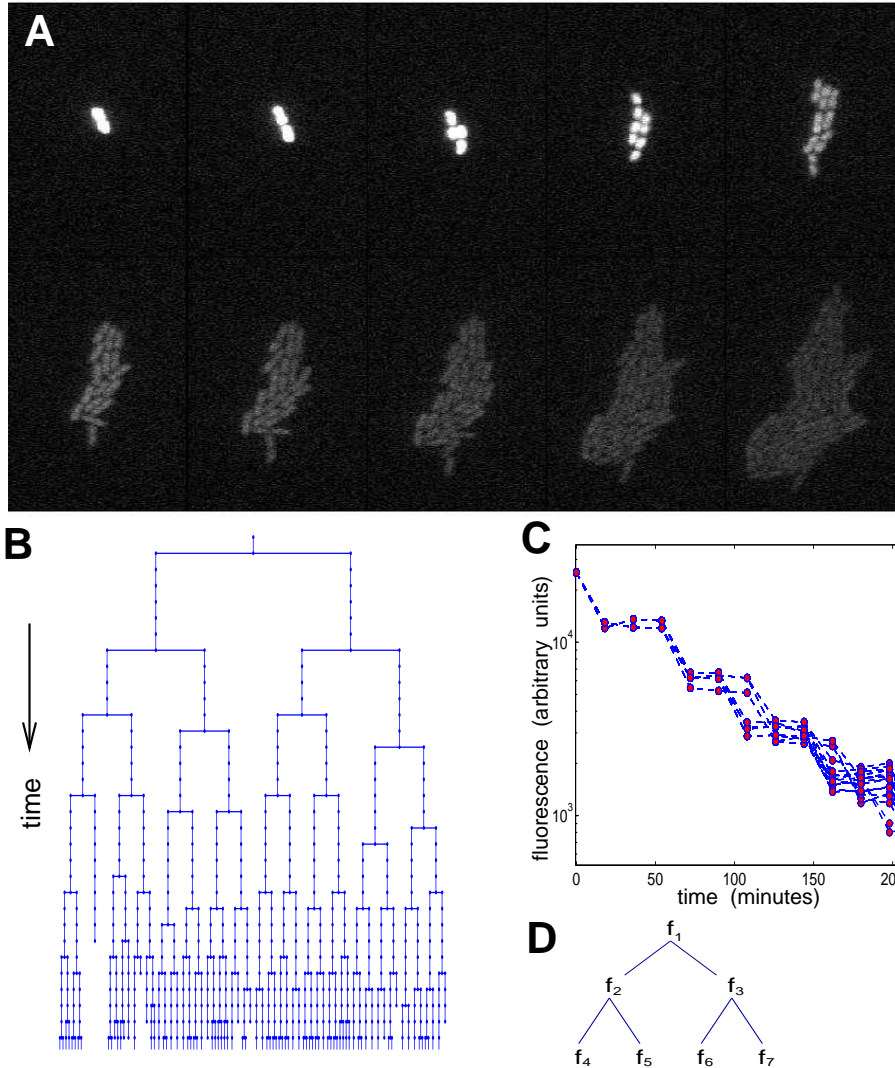


Figure 1: Partitioning of a fluorescent protein during microcolony growth. **A** Snapshots from a typical dilution experiment. Images are taken using yellow fluorescence filters, and the time between the frames shown is approximately 36 minutes. **B** The lineage tree extracted from the same movie. Time increases downwards as more and more divisions occur. Measurements are marked by a dot and were taken approximately every 9 minutes (giving a total of just over 700). **C** Cellular fluorescence only decreases significantly at cell division. For this dataset, fluorescence measurements were taken every second frame, i.e. every 18 minutes. The first cell divides after approximately 20 minutes, and its daughters in turn both divide at 50 minutes. Notice the different fluorescence values in each daughter cell become more apparent at later divisions. For clarity, only the initial part of the movie is shown. **D** Schematic of a lineage tree generated from an initial cell with fluorescence value f_1 . Second generation cells have fluorescence f_2 or f_3 , while third generation cells have f_4 , f_5 , f_6 , or f_7 .

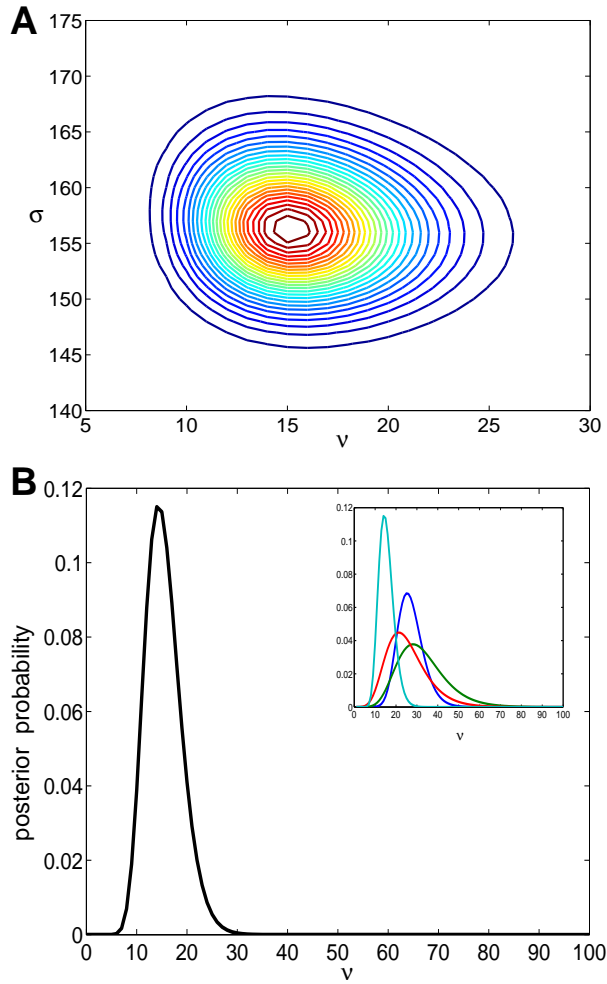


Figure 2: Results of method II applied to the data of Fig. 1. **A** A contour plot of the posterior probability of σ and ν . The distribution has a single peak with a most probable value of $\nu \simeq 15 \pm 4$ fluorescence units per fluorescent particle and of $\sigma \simeq 156 \pm 5$ fluorescence units. **B** The posterior marginalized over σ . The most probable value of ν is given at the maximum and the error in this estimate by the peak width at half maximum. The inset shows the marginalized posteriors for the data of Fig. 1 (left curve) and for three other data sets taken on the same microscope on the same day. Inferred values of ν are consistent.

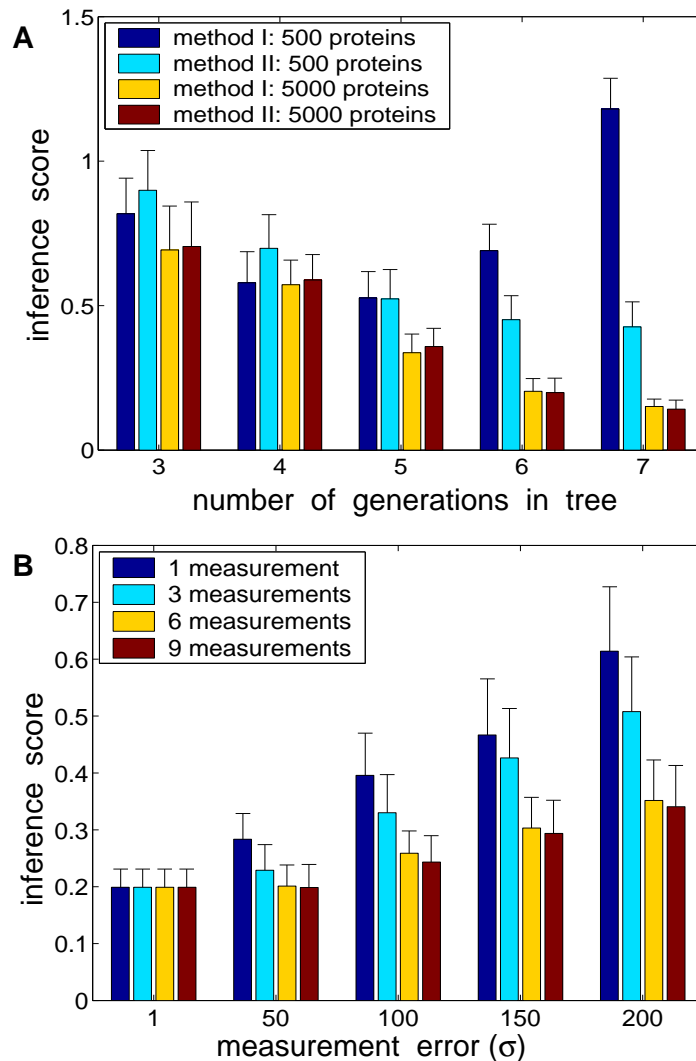


Figure 3: Evaluation of the inference methods for simulated data using the inference score $i_\epsilon = |\log_2(\nu/\nu_0)|$ (a perfect fit has an inference score of zero), with simulated $\nu_0 = 25$. **A**. Inference improves with more data, but is sensitive to signal to noise ratios; method II is robust. Twenty different data sets, each of 7 generations, were created from an initial cell having either 500 (first two bars) or 5000 fluorescent proteins (last two bars). For generation numbers below 7, the appropriate lower part of the data tree was discarded. Measurement error was added with $\sigma = 150$, and 3 measurements were taken per cell. **B**. The performance of method II as the size of the measurement error grows; inference improves by increasing the number of measurements per cell. Notice the new y -axis scale. Twenty different data sets with 500 proteins in the initial cell were fit, and the results averaged to generate each bar. For both figures, we used the analytical estimate for σ , Eq. 15, and took the maximum of the posterior as the best estimate for ν . Error bars are standard errors.