Research Article

Inferring gene expression dynamics from reporter protein levels

David Brown[1] *and C. Phoebe Lostroh*[2]

[1] Department of Mathematics and Computer Science, Colorado College, Colorado Springs, CO, USA
[2] Department of Biology, Colorado College, Colorado Springs, CO, USA

We present a mathematical method for inferring the dynamics of gene expression from time series of reporter protein assays and cell populations. We show that estimating temporal expression dynamics from direct visual inspection of reporter protein data is unreliable when the half-life of the protein is comparable to the time scale of the expression dynamics. Our method is simple and general because it is designed only to reconstruct the pattern of protein synthesis, without assuming any specific regulatory mechanisms. It can be applied to a wide range of cell types, patterns of expression, and reporter systems, and is implemented in publicly available spreadsheets. We show that our method is robust to a several possible types of error, and argue that uncertainty about the decay kinetics of reporter proteins is the limiting factor in reconstructing the temporal pattern of gene expression dynamics from reporter protein assays. With improved estimates of reporter protein decay rates, our approach could allow for detailed reconstruction of gene expression dynamics from commonly used reporter protein systems.

Received 18 June 2008
Revised 30 August 2008
Accepted 15 September 2008

Keywords: β-galactosidase · Gene expression · Mathematical model · Promoter activity · Reporter proteins

1 Introduction

Assays of reporter proteins such as *β*-galactosidase, luciferase, chloramphenicol acetyl transferase (CAT), and green fluorescent protein (GFP) are among the most widely used methods for investigating the dynamics of gene expression and protein synthesis [1]. In this approach a gene fusion is created in which the coding region for the reporter protein is placed under the control of the promoter of a gene of interest. The synthesis of the reporter protein reflects the target gene's promoter activity for transcriptional fusions. Alternatively, a translational gene fusion can be created, in which the reporter gene is cloned in frame with the first few codons of the target gene, so that the reporter protein levels reflect a combination of promoter activity and post-transcriptional processes. In a typical experiment, the reporter protein concentration in a population of cells is assayed at a series of time points, and the concentration of reporter protein per cell is plotted *versus* time (or *vs.* population size) to indicate the dynamic pattern of gene expression.

The use of reporter gene fusions to study gene expression is not without its issues. For example, fusions can affect mRNA stability and transcription or translation rates in unintended ways [2]. Other methods, including quantitative real-time PCR and Northern blot, offer more direct ways of measuring gene expression dynamics. However, reporter gene fusions are likely to remain a popular method for studying gene expression for several reasons. First, it is a relatively inexpensive technology that is well established in many laboratories. Second, results from transcriptional and translational fusions can be combined to shed light on each of the steps involved in gene expression. Finally, reporter gene fusions have a wide variety of other experimental uses [1, 3], so investing time and resources to create them is convenient for many experimenters.

Correspondence: Dr. David Brown, Department of Mathematics and Computer Science, 14 E. Cache la Poudre St., Colorado Springs, CO 80903, USA
E-mail: dbrown@coloradocollege.edu
Fax: + 1 719 389 6841

Abbreviations: CAT, chloramphenicol acetyl transferase

In this paper, we address another subtlety regarding the interpretation of reporter protein time series data. The reporter protein level at a particular time does not simply reflect the level of gene expression at that instant, but rather the cumulative effect of protein synthesis, degradation, and dilution (through population change) up to that time. As a result, the underlying temporal pattern of gene expression is not always obvious from the pattern of reporter protein accumulation. The more stable the reporter protein is *in vivo*, the more history is contained in each data point, and the more difficult it is to estimate the gene expression dynamics from visual inspection of graphs displaying reporter protein levels over time. Whatever challenges exist with using reporter gene fusions to monitor gene expression dynamics, it is clear that the quantity of interest is not the amount of reporter protein that has accumulated by some time point, but rather the rate of synthesis of the reporter protein at that time.

We present two simple mathematical models of reporter protein synthesis that allow one to infer the underlying pattern of gene expression from a time series of reporter protein levels and cell population size. One model is specific to GFP, which is synthesized in an inactive form and must undergo conformational changes to become fluorescent. The other model is appropriate for reporter proteins that do not undergo extensive modifications (*i.e.*, β-galactosidase, luciferase, and CAT). The models provide a method for inferring the relative rate of reporter protein synthesis *per* cell from time series data on reporter protein levels and cell populations, provided that an estimate of the protein's decay rate is available. Our approach can be thought of as a model-based transformation of data, which converts the time series for cell population and reporter protein assay into a time series for reporter protein synthesis.

We are not attempting to model the mechanisms driving changes in gene expression, only to reconstruct the time course of reporter protein synthesis. Because this reconstruction is based only on the balance of protein synthesis, dilution, and decay, our method is both simple and general. This rate of reporter protein synthesis reflects the relative promoter activity (in the case of a transcriptional fusion) or the relative rate of target protein synthesis (in the case of a translational fusion). In our approach, we do not need to distinguish between these two cases, since it only affects how the data is interpreted. Thus, we follow the convention of referring to both promoter activity and translational control as "gene expression". Throughout this paper it should be remembered that we are only inferring the rate of reporter protein synthesis; the interpretation of this in terms of the mechanisms of gene expression will depend on the experiment.

As far as we are aware, this represents the first published general method for inferring gene expression dynamics from reporter protein data. Subramanian and Srienc [4] developed a mathematical model of GFP synthesis to aid in the analysis of gene expression in mammalian cells. However, they assumed constant promoter activity and an exponentially growing cell population; they did not provide a general framework for determining gene expression dynamics from protein assays. Their model was modified and further analyzed by Leveau and Lindow [5], who continued to assume constant promoter levels and used steady-state protein levels to get a single estimate of promoter activity under each experimental condition. In contrast, our approach allows one to infer the dynamics of gene expression, providing valuable insights into regulatory events. Moreover, our approach is generic, making no specific assumptions about the type of cells, regulatory mechanisms, or other experimental conditions. Bhasi *et al*. [6] developed a mathematical model for estimating transcription and translation rates from simultaneous time series of mRNA and protein concentrations in a constant cell population. Our approach differs from theirs in the type of data to which it can be applied: we assume that mRNA levels are not known and that the cell population may be changing. In addition, while the underlying models of protein synthesis are similar, Bhasi *et al*. [6] use a complex data-smoothing algorithm requiring specialized software, while we have opted for a simpler approach that can be implemented in a spreadsheet.

Finally, our method is designed for data from population experiments, in which the reporter protein is synthesized by many cells. This is a common experimental technique that provides a first approximation to gene expression dynamics. The observed dynamics are typically quite smooth, and can be captured by deterministic models like ours. In contrast, gene expression studies using reporter fusions in single cells have revealed a more complex picture, in which gene expression is stochastic, with bursts of transcription and possible coordination between transcriptional and translational regulation [7, 8]. Clarifying the relationship between events at the single cell level and the population level requires more complex approaches than we present here. However, a necessary first step is to reconstruct the rate of reporter protein synthesis by the population, which our models accomplish.

2 Materials and methods

2.1 Model structure and assumptions

The general model consists of the following balance equation:

[rate of change of reporter protein concentration *per* cell] = [synthesis] – [degradation] – [dilution].

This equation makes no specific assumptions about the processes involved; it is an accounting of the factors that contribute to changes in protein concentration. Indeed, it is simply a statement of the conservation of matter in the form of a dynamic equation. We are interested in determining the relative dynamics of gene expression, rather than absolute rates of transcription or translation. Thus, relative measures of reporter protein concentration (*i.e.*, assay values) and cell population (OD or CFU) can be used in the balance equation, provided that these relative measures are proportional to the true values.

Of the four terms in the balance equation, the change in protein concentration and the dilution rate can be determined from typical time series data from a reporter gene fusion study. If we make a suitable assumption about the degradation rate of the protein, the balance equation can be solved for the rate of protein synthesis as a function of time; this quantity indicates the relative dynamics of gene expression, defined by the nature of the reporter gene fusion.

Let $B(t)$ be the value of the reporter assay per cell and $V(t)$ be the relative cell population at time t. Let $m(t)$ be the specific degradation rate of the reporter protein, and $p(t)$ be the rate of reporter protein synthesis. For transcriptional fusions, $p(t)$ will be proportional to the target gene's promoter activity. For translational fusions, it will be proportional to the rate of synthesis of the target protein, reflecting the net rate of transcription and translation. In both cases, we will refer to $p(t)$ throughout this paper as the level of gene expression. The dynamics of $p(t)$ over time are what we want to infer from the data for $B(t)$ and $V(t)$ and the assumed or measured values of $m(t)$.

The rate of protein degradation is the only term about which specific assumption must be made in order to complete the model. For generality, we represent the *in vivo* degradation of the protein in the form of first-order kinetics (*i.e.*, proportional to the amount of protein), with a decay rate $m(t)$ that is allowed to vary with time. In principle, the variable decay rate allows this term to reproduce any kind of decay kinetics. Yildirim and Mackey [9] obtained good agreement with experimental data in a detailed model of the *lac* operon in *Escherichia coli*, assuming first-order decay of β-galactosidase with a constant half-life of 14 h (as measured by Mandelstam [10]). However, Leveau and Lindow [5] assumed that GFP degradation was limited by the availability of proteolytic proteins, yielding Michaelis–Menten kinetics. We discuss the effect of Michaelis–Menten proteolysis below.

As the cell population grows or declines, the concentration of reporter protein per cell volume changes even if no new production or degradation occurs. This dilution process occurs at the same specific rate as the specific growth rate of the population,

$$\frac{1}{V}\frac{\mathrm{d}V}{\mathrm{d}t}.$$

For example, if the population at some time is growing at 10%/h, then the protein concentration *per* cell, in the absence of any new synthesis or degradation, is decreasing at 10%/h.

The balance equation can now be expressed in the form of a differential equation:

$$\frac{\mathrm{d}B}{\mathrm{d}t} = p(t) - \left(m(t) + \frac{1}{V}\frac{\mathrm{d}V}{\mathrm{d}t}\right)B. \tag{1}$$

This equation can be integrated to obtain the protein levels in terms of the gene expression and population growth (see appendix). This reveals that the protein level at time t is a weighted sum of the history of gene expression, with the "memory" of the system decaying exponentially at a rate given by the sum of degradation and dilution terms.

More importantly, the balance equation can be rearranged to solve for the gene expression level at time t:

$$p(t) = \frac{\mathrm{d}B}{\mathrm{d}t} + \left(m(t) + \frac{1}{V}\frac{\mathrm{d}V}{\mathrm{d}t}\right)B. \tag{2}$$

Time series data can be used to estimate the rates of change of B (reporter protein assay) and V (cell population), yielding an approximation to the dynamics of gene expression.

In particular, assume that cell population and protein assay values are known at times t_0, $t_1,\ldots t_i,\ldots t_n$ (these sampling times do not need to be evenly spaced). Let $\Delta t_i = t_{i+1} - t_i$, the size of the i^{th} time step. The instantaneous rate of change of B at time t_i can be approximated using the following discretization of the derivative:

$$\begin{aligned}\frac{\mathrm{d}B(t_i)}{\mathrm{d}t} \approx & -B(t_{i-1})\left(\frac{\Delta t_i}{\Delta t_{i-1}\left(\Delta t_{i-1} + \Delta t_i\right)}\right) \\ & + B(t_i)\left(\frac{\Delta t_i - \Delta t_{i-1}}{\Delta t_{i-1}\Delta t_i}\right) + B(t_{i+1})\left(\frac{\Delta t_{i-1}}{\Delta t_i\left(\Delta t_{i-1} + \Delta t_i\right)}\right)\end{aligned} \tag{3}$$

This is a generalization of the standard centered difference formula approximation to the derivative, valid for unequal time steps. It is a second order approximation, with the error scaling as the square of the time step (reducing the time step by half will reduce the error to one-fourth its previous size). The derivative of V can be approximated in an analogous way. Plugging these discretizations into the balance Eqn. (2) yields a formula for approximating the protein synthesis dynamics. We have implemented this formula in a spreadsheet, available online at http://faculty1.coloradocollege.edu/~dbrown/genedynamics.html.

It is important to keep in mind that there are two possible sources of error introduced by the model: the uncertainty about the protein decay rate, and the use of a discrete difference to estimate an instantaneous rate of change. We explore the consequences of this in our simulations below.

2.2 GFP model

Thus far, we have assumed that gene expression results in the production of a detectable form of the reporter protein. However, GFP is produced in an inactive form, which must be converted to an active form before it can be detected via fluorescence. Following Subramanian and Srienc [4] and Leveau and Lindow [5], we assume that this conversion obeys first order kinetics. This yields the following model for the concentrations of nonfluorescent ($N(t)$) and fluorescent ($F(t)$) protein:

$$\frac{dN}{dt} = p(t) - \left(c + m(t) + \frac{1}{V}\frac{dV}{dt}\right)N \qquad (4)$$
$$\frac{dF}{dt} = cN - \left(m(t) + \frac{1}{V}\frac{dV}{dt}\right)F$$

where c is the rate of conversion from nonfluorescent-to-fluorescent protein. We assume that both forms of the protein have the same decay rate m (this can easily be modified if data warrant it). These equations can be combined to obtain a gene expression reconstruction formula for the GFP model:

$$p(t) = \frac{1}{c}\left[\frac{d^2F}{dt^2} + \left(c + 2m(t) + \frac{2}{V}\frac{dV}{dt}\right)\frac{dF}{dt} + \left(cm(t) + m^2(t) + \frac{dm}{dt} + \frac{c + 2m(t)}{V}\frac{dV}{dt} + \frac{1}{V}\frac{d^2V}{dt^2}\right)F\right]. \qquad (5)$$

In the limit of very fast conversion ($c \to \infty$), this equation converges to equation 2. The presence in the GFP formula of second derivatives implies that the underlying gene expression depends in part on the concavity of the fluorescence and population time series. Thus, the existence of an unobservable state of the reporter protein complicates the relationship between gene expression and assay values in a nonintuitive way.

The GFP equation can be discretized to yield a formula for estimating $p(t)$ from time series data for F (GFP fluorescence per cell) and V (cell population). We have done so using the same discretization formula for the first derivatives, and the following second order approximation for the second derivative:

$$\frac{d^2F(t_i)}{dt^2} \approx 2F(t_{i-1})\left(\frac{1}{\Delta t_{i-1}(\Delta t_{i-1} + \Delta t_i)}\right) - 2F(t_i)\left(\frac{1}{\Delta t_{i-1}\Delta t_i}\right) + 2F(t_{i+1})\left(\frac{1}{\Delta t_i(\Delta t_{i-1} + \Delta t_i)}\right) \qquad (6)$$

The resulting formula for inferring gene expression is more complex than the general model, but is still easily implemented in a spreadsheet (also available online).

The discrete approximations that we have used are not the only way to estimate the derivatives from time series. For example, one could use cubic spline interpolation or nonparametric regression to produce smooth functions for $B(t)$ and $V(t)$. An approach like this was used by Bhasi *et al.* [6]. However, a more complex approach requires specialized software that experimenters may not have readily available. Given the inherent limitations to precision in the type of data we are considering (see below), we expect that the simple method will be sufficiently accurate and is more likely to be widely used than a more complex approach.

One advantage of approaches using interpolation and regression is that they smooth out some of the noise inherent in the data, allowing the stronger trend to show through. In our approach, any noise or error in the data is propagated directly to the reconstruction formulas. To reduce the sensitivity of our method to noise in the data, we have incorporated a smoothing step that can be applied to the reporter protein and cell population time series before the reconstruction of the protein synthesis is carried out. Many methods exist for smoothing time series, each with advantages and disadvantages. In general, the more aggressive one is in attempting to smooth out the noise, the more one risks losing the real signal. In choosing our smoothing method, we have opted for one that is simple, appropriate for short time series, and can be adjusted by the user depending on how much smoothing they want to carry out.

Our smoothing procedure is a type of moving average: a smoothed time series is produced by replacing each data point by a weighted average of itself and some neighboring data points. Specifically,

we replace each point by a weighted average of itself and its nearest neighbor on either side. The weights are determined by a smoothing parameter that the user sets, and by the distances between time points (so that the nearer neighbor carries more weight). The formula for producing the smoothed protein assay time series is:

$$\tilde{B}(t_i) = (1-\omega)B(t_i) + \omega\frac{\Delta t_i}{\Delta t_i + \Delta t_{i-1}}B(t_{i-1}) + \omega\frac{\Delta t_{i-1}}{\Delta t_i + \Delta t_{i-1}}B(t_{i+1}) \quad (7)$$

The formula for smoothing the cell population count is analogous. The smoothing parameter, ω, can be set between 0 and 1. The case $\omega = 0$ corresponds to no smoothing; the time series is unchanged. The case $\omega = 1$ corresponds to the most aggressive smoothing; each data point is replaced by a point on the line that connects the previous and following data points. We explore the effects of smoothing on protein synthesis reconstruction from noisy datasets below.

2.3 Some assumptions

In both models, we have assumed that the rate of reporter protein synthesis at time t is proportional to the level of gene expression at that time. However, in transcriptional fusions there is a time lag between promoter activity and the completion of protein synthesis. This time lag is short in most prokaryotic systems because of the coupling of transcription and translation and because post-transcriptional regulation of mRNA is a rare phenomenon. If this time lag is not negligible relative to the timescale of the experiment, it can be incorporated in either version of the model by replacing $p(t)$ by $p(t-\tau)$, where τ is the length of the time lag. This might be necessary in analyzing some reporters in eukaryotic systems, or in short timescale experiments in prokaryotes.

Our method relies on several other standard assumptions about the data. We assume that the reporter assay is proportional to the protein concentration over the entire range of concentrations in the data. We assume that the value of the reporter assay depends only on the protein concentration, although GFP fluorescence may also depend on the activity state of the cell [11]. For transcriptional fusions, we assume that the amount of reporter protein produced *per* mRNA is constant over the course of a particular experiment, so that the rate of protein synthesis is proportional to the promoter activity. This is an assumption usually made by biologists using transcriptional fusions. The production of β-galactosidase *per lacZ* mRNA is known to vary with temperature and cloning vector [12], so comparisons of activity across experiments must be done with care. We assume that the total cell volume is proportional to the population measurement, but turbidity measurements are known to deviate from proportionality to cell population at high OD [13]. When any of these assumptions is not valid for a particular system, it must be corrected either through a transformation of the data or modification of the model. In the next section, we examine the sensitivity of the model to violations of the assumption that the observed cell population and reporter protein concentration are proportional to their true values.

3 Results

3.1 Case study: rpoS

Before presenting a thorough analysis of the method, we illustrate its use by applying the general model to a classic example, *rpoS* expression in *E. coli*. *rpoS* is the structural gene for σ^S, the σ factor necessary for induction of many stationary phase specific genes. Lange *et al*. [14] studied the activity of the major promoter (*rpoS*p_1) using an *rpoS*::*lacZ* transcriptional reporter fusion. They observed that, in wild-type cells growing in rich medium, β-galactosidase activity *per* cell in an *rpoS*::*lacZ* reporter strain increased dramatically when the growth curve indicated the population was entering stationary phase ([14], Fig. 3B). From these data, they concluded that *rpoS*p_1 activity is induced "during entry into stationary phase".

When we apply our method to these data, we can gain additional information about the pattern of *rpoS*p_1 activity. The rate of β-galactosidase decay in *E. coli* has been estimated as 0.01–05/h [10, 15]. Accordingly, we include reconstructions of gene expression dynamics using protein half-lives of 5, 10, and 20 h. (In the figures we indicate the half-life of the reporter protein,

$$h = \frac{\ln 2}{m},$$

rather than the decay rate for ease of interpretation.) It is clear that the promoter activity increased throughout the exponential growth phase, peaked prior to entry into stationary phase, then declined dramatically (Fig. 1a). These conclusions hold over a wide range of values for the half-life of β-galactosidase, suggesting that they are quite robust. Moreover, the result holds after smoothing the data (for example, with $\omega = 0.5$; Fig. 1b). From visual in-

Biotechnol. J. 2008, 3, 1437–1448

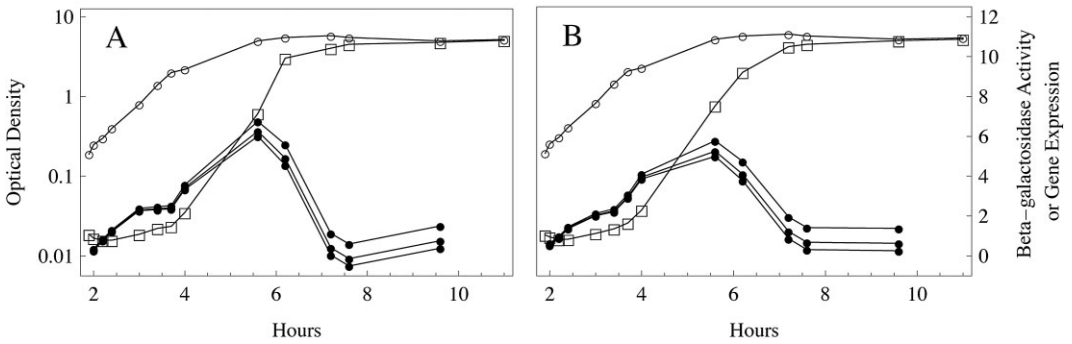

Figure 1. Inferred promoter activity for *rpoS*, from data reported in Lange *et al*. [14]. We used visual inspection of published figures to plot the β-galactosidase (open squares) and cell population (OD578, open circles) over time. (a) We used our general model to calculate $p(t)$, the activity of the promoter driving *rpoS*::*lacZ* activity, using a β-galactosidase half-life of $h = 5$, 10, or 20 h (solid circles, highest curve to lowest, respectively). (b) We repeated the procedure after smoothing the data using the smoothing parameter value $\omega = 0.5$.

spection of the original β-galactosidase time series, it was not obvious that *rpoS* expression increased throughout the exponential growth phase, nor that it shut down following entry into stationary phase. This suggests that model-based reconstruction of gene expression dynamics has the potential to refine our understanding of many expression profiles.

3.2 Patterns of activity

We next use the model to explore the relationship between simple patterns of underlying gene expression $p(t)$ and the reporter protein dynamics. For simplicity, we assume a logistically growing cell population

$$\left(\frac{dV}{dt} = rV\left(1-\frac{V}{K}\right)\right),$$

and a constant rate of protein decay. Figure 2 displays the patterns of gene expression that would correspond to various simple patterns of reporter protein concentrations and different protein decay rates.

Several phenomena emerge in Fig. 2. First, when population growth and protein decay occur over the same time scale as the gene expression dynamics, the relationship between the underlying pattern of gene expression and protein levels may not be intuitively obvious. In general, gene expression is closely related to both the slope of the protein curve and the protein level itself, but the relationship is complicated by the changing cell population. Second, superficially similar protein data can arise from quite different patterns of gene expression. For example, steadily increasing protein levels may reflect gene expression that is constant, linearly increasing, or saturating (Figs. 2A–C).

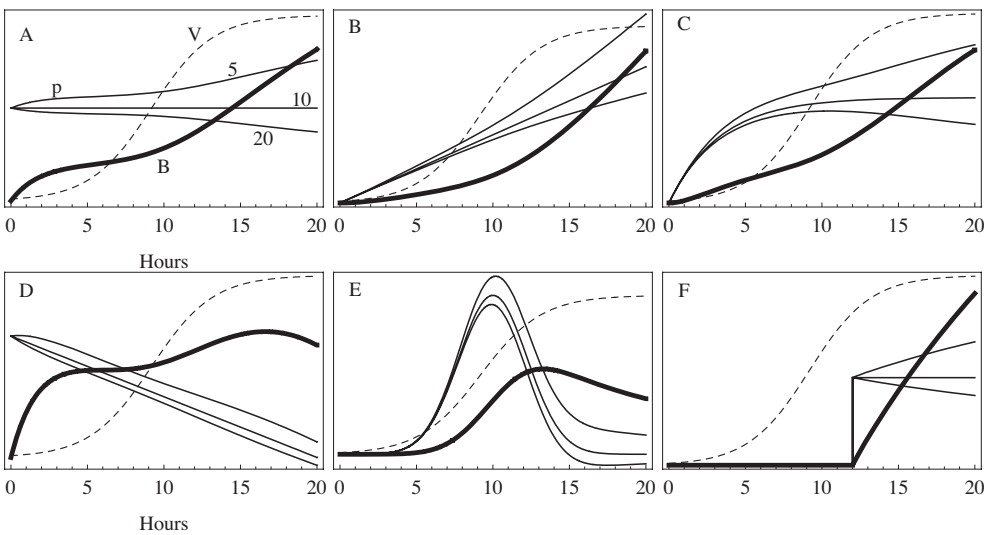

Figure 2. Simulated patterns of gene expression (p) and reporter protein levels (B) in a logistically growing cell population (V). In each of the six cases, the reporter protein concentration corresponds to several different gene expression dynamics using protein half-lives of $h = 5$, 10, and 20 h. Gene expression levels are always highest for the shortest half-life. The scale of the y-axis is relative, and scaled differently for each simulation.

Third, the general patterns are robust over a wide range of protein half-lives (5–20 h in these simulations). However, the specific gene expression levels corresponding to a given reporter protein pattern do depend on the protein half-life, with the differences increasing over time. This suggests that accurate determination of detailed gene expression patterns from reporter protein assays will, in some circumstances, require estimates of the protein's decay rate in the organism and experimental conditions of interest.

We can gain some insight into this by considering two limiting cases of the general model. Ignoring dilution due to population growth, suppose that the reporter half-life is much longer than the timescale of the gene expression dynamics. In this case, we have approximately $m(t) = 0$,

$$\text{so } p(t) \approx \frac{dB}{dt}.$$

Thus, the relative level of gene expression is given by the slope of the protein assay curve, and a precise estimate of the decay rate is not needed. At the other extreme, suppose that the protein is very short-lived relative to the gene expression dynamics. In this case, the accumulation of the protein will track changes in gene expression almost instantaneously, so the relative level of gene expression is given approximately by $B(t)$. In this case also, precise knowledge of the decay rate is not needed. The challenge occurs when the half-life of the reporter protein is comparable to the timescale of gene expression dynamics, a situation that often occurs in experimental systems and that we model in our simulations. In this regime, our simulations suggest that knowing the decay rate within a factor of two will usually yield reconstructions that are qualitatively correct, but that quantitative precision will be limited by inaccuracy in the assumed decay rate.

Reporter protein decay rates can be determined by a variety of methods. The standard methods in eukaryotic cells are pulse-chase, in which the protein of interest is labeled with radioactivity for a short pulse and then its persistence is measured over time, or cylclohexamide poisoning, in which all protein synthesis is blocked in order to follow the decay of proteins after the drug prevents ribosome function (recently reviewed in ref. [16]). Each of these methods typically involves quantitative immunoprecipitaiton or immunoblotting to observe the specific protein of interest. The cyclohexamide poisoning methods was recently used in combination with a genomic library of *Saccharomyces cerevisiae*, to determine the half-lives of most yeast proteins [17]. In bacteria, pulse-chase studies are also used see (ref. [18]).

Our model provides an alternative method for determining the reporter protein decay rate. If the gene is silenced beginning at some time ($p(t) = 0$), the balance equation can be solved to provide an estimate of the (time-dependent) protein decay rate in terms of the cell population and protein assay data:

$$m(t) = -\left(\frac{1}{B}\frac{dB}{dt} + \frac{1}{V}\frac{dV}{dt}\right). \quad (8)$$

Here, B (protein assay) and V (cell population) are directly available from the experimental data, and the derivatives can be estimated from the data using the method described above. This generalizes the common approach of determining the protein half-life by monitoring the reporter activity levels following downregulation of the gene [15]; in our formulation, the cell population does not need to be constant. If the gene is not completely silenced, this estimate provides a lower bound on the protein decay rate, provided that transcription and translation of the protein of interest are known to be tightly coupled. The rhamnose and arabinose-inducible promoters in *E. coli* are examples of such tightly regulated promoters [19].

Thus far, we have assumed first-order kinetics for the degradation of the reporter protein *in vivo*. In some cases, however, the degradation may be limited by the availability of proteolytic enzymes, yielding Michaelis–Menten type kinetics. In principle, this can be captured in our model framework by varying the protein decay rate over time. However, it is more natural to include a different form of the decay kinetics. We modified our general model to incorporate Michaelis–Menten decay kinetics, and simulated the same patterns of gene dynamics as above. Over a wide range of saturation parameters, we found that the relationship between reporter protein accumulation and gene expression is very similar to the patterns using first-order kinetics. This suggests that in many cases the use of first-order kinetics will be suitably accurate. However, improved understanding of reporter protein decay is probably a key factor required for detailed analysis of gene expression dynamics; more complex forms of decay kinetics can be incorporated in our general framework when warranted.

The GFP model can be used to simulate observed fluorescence corresponding to the same patterns of gene expression (Fig. 3). In these simulations, we have used parameter values of $c = 1.54$/h (rate of conversion to fluorescent form) and $m = 1.08$/h (decay rate), corresponding to the unstable engineered GFP variant LVA [20, 21]. Despite the complication introduced by the nonfluorescent stage, the assay dynamics more closely

Biotechnol. J. 2008, 3, 1437–1448

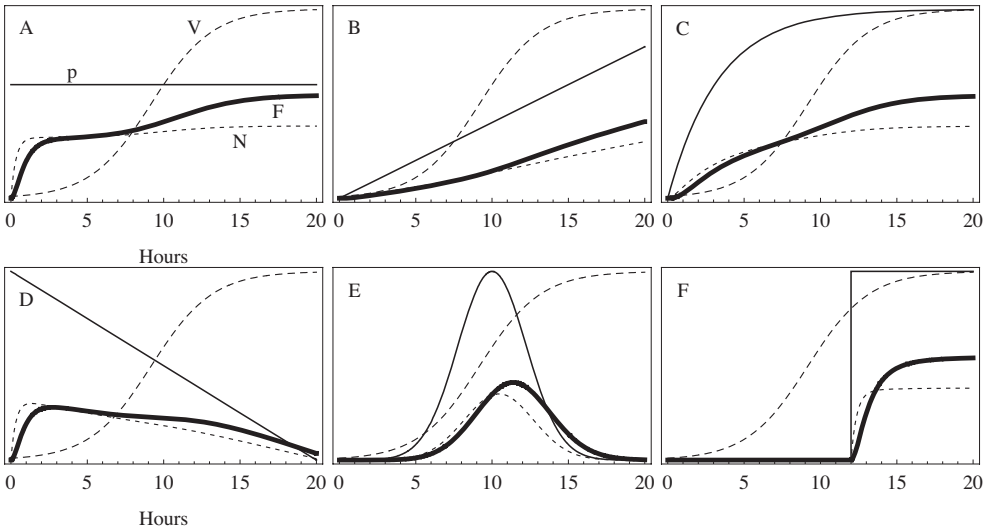

Figure 3. Simulated patterns of gene expression (p) and GFP fluorescence (F) in a logistically growing cell population (V). The nonfluorescent protein (N) is also plotted. In each of the six cases, the gene expression was specified, and the resulting protein levels were simulated using the GFP model.

match the underlying gene expression, due to the rapid turnover of GFP. However, the relationship between gene expression and assay values will be obscured for dynamics that occur on a time scale comparable to GFP turnover (roughly 1 h). Thus, even for unstable reporter proteins, it is important to have a quantitative method for inferring patterns of gene expression.

3.3 Inferring gene expression from protein assays

In order to evaluate the ability of the inference procedure to reconstruct gene expression dynamics, we applied it to simulated data. We chose simple patterns of gene expression $p(t)$, then used our model to simulate the resulting protein dynamics in a logistically growing population. We sampled the resulting protein and population functions at discrete times to generate simulated datasets and used the formulas above to reconstruct the underlying gene expression. If the reconstruction method is accurate, the inferred time series of gene expression should match the originally specified function $p(t)$. A similar approach of using simulated data to evaluate the performance of a reconstruction algorithm was used by Bhasi *et al.* [6].

As Fig. 4 shows, our method yields accurate reconstruction of the simulated gene expression dynamics. Importantly, the method distinguishes between different activity patterns that correspond to superficially similar protein patterns (Figs. 4A, B). We obtained similar results using the discretized GFP model to reconstruct gene expression dynamics (Fig. 5).

Using discrete data points to approximate the instantaneous rates of change introduces a discretization error. This error can be minimized by reducing the time interval between successive measurements. Simulations using a peaked pattern of gene expression (as in Fig. 2E) indicated that increasing the sampling frequency from 4 to 2 h dramatically improved the accuracy of the reconstruction, while a further increase to 1 h yielded only a moderate improvement. In general, the appropriate time between data points will depend on the time scales of expression dynamics and population change.

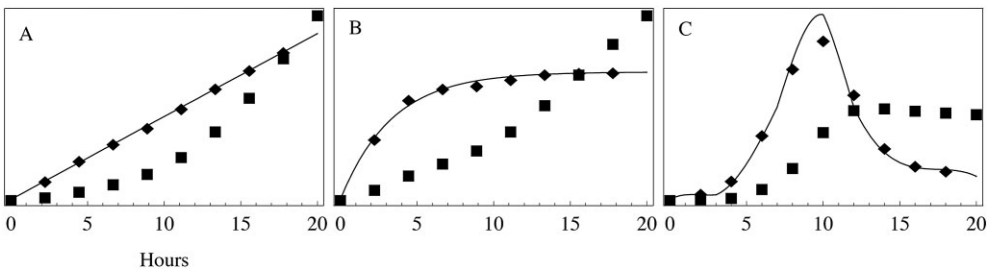

Figure 4. Reconstructing gene expression from simulated data using the general model. We ran the model with several specified gene expression patterns $p(t)$ (solid curves), and a cell population growing logistically as in Fig. 2. Sampling the resulting protein levels at discrete times yielded simulated data (squares). This was used, together with the population data (not shown) to reconstruct the gene expression (diamonds).

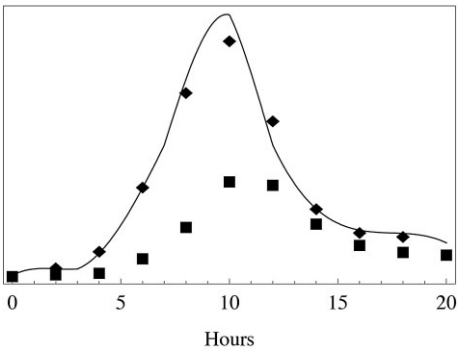

Figure 5. Reconstructing gene expression from simulated data using the GFP model. We ran the model with a specified gene expression pattern $p(t)$ (solid curve), and a cell population growing logistically as in Fig. 2. Sampling the resulting protein levels at discrete times yielded simulated data (squares). This was used, together with the population data (not shown) to reconstruct the gene expression (diamonds).

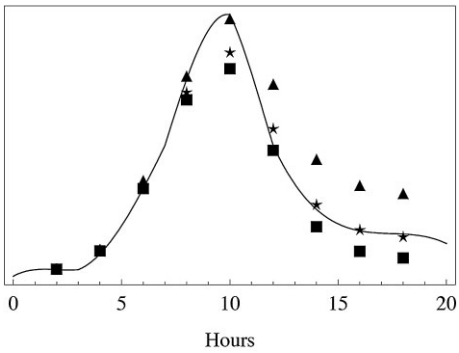

Figure 6. Effect of decay rate error on gene expression reconstruction. Simulated data (not shown) were produced from the specified gene expression $p(t)$ (solid curve), as in Fig. 4, using a protein half-life of $h = 10$ h. The simulated data were then used to reconstruct the activity, using half-lives of $h = 5$ (triangles), 10 (stars), or 20 (squares) hours.

The main source of possible errors in applying our model is the use of incorrect decay rates for the reporter protein. To investigate how serious a problem this may be, we repeated the exercise of reconstructing a known activity pattern from a simulated dataset. This time, we generated the data using a protein half-life of 10 h, then carried out the activity inference procedure using half-lives of 5, 10, and 20 h. As Fig. 6 shows, the reconstructed pattern is robust to this type of error, at least for this activity profile. The magnitude of the error grows over time, indicating that greater effort should be made to measure the decay rate when experiments last a period of time comparable to the half-life of the reporter protein. Our method allows for a decay rate that changes over time; in systems where time-dependent estimates of decay rates are available, including this feature will improve the accuracy of the reconstruction. However, our simulations indicate that variation in the decay rate by a factor of two does not affect the most important features of the gene expression reconstruction. This suggests that a time-averaged single value of the decay rate may provide a suitable approximation in many cases.

Our models are based on the assumption that the population data and reporter protein assay are proportional to the true population and protein concentrations over the entire range of the data. However, OD measurements are known to deviate from proportionality, as objects at higher density are underdetected due to shadowing [13]. To assess the sensitivity of our model to violations of the proportionality assumptions, we carried out simulated reconstructions of gene expression in which the "observed" cell population was not proportional to the true value, deviating by a maximum of 17% at $OD_{600} = 1$ [13]. In general, the reconstructed gene expression was very robust to this form of error. The difference between the true and reconstructed gene expression levels was less than 5% for most of the patterns investigated, and peaked during the most rapid population growth (mid-exponential phase). Of course, the error may be larger for high population densities in which the departure from proportionality is more severe.

We carried out a similar analysis assuming that optical reporter protein assays may be subject to nonlinearities at high concentrations. Here, the percentage error in the reconstructed gene expression was slightly larger than the percentage difference between the true and "observed" protein concentrations. As would be expected, the error was largest when protein concentrations were highest. For both cell population and reporter protein concentration, violating the assumption of proportionality between observed and actual values usually has a modest quantitative effect on the reconstructed gene expression, and a negligible effect on the qualitative dynamics.

Finally, we tested our reconstruction method's sensitivity to experimental noise, and the effect of the smoothing procedure. Specifically, we simulated measurement errors in the population counts and protein assay by assuming a multiplicative noise factor with a SD of 10%. (Error bars in published reporter protein time series suggest that experimental noise may be much smaller than this.) Each data point was multiplied by a factor $1 + X_i$, where X_i are independent, normally distributed random variables with mean 0 and SD 0.1. Each simulated noisy dataset was then used to reconstruct the gene expression dynamics.

Biotechnol. J. 2008, 3, 1437–1448

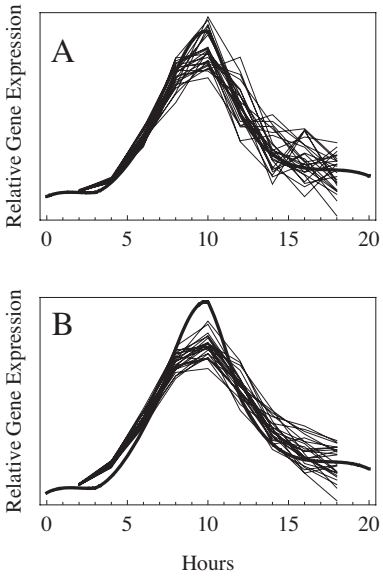

Figure 7. Effect of noisy data and smoothing on gene expression reconstruction. Simulated data were produced from the specified gene expression $p(t)$ (heavy curve), in a logistically growing population. Twenty sets of data (not shown) were produced by including 10% multiplicative noise in each protein assay and population data point. Each noisy dataset was used to reconstruct the gene expression dynamics (lighter curves). (a) Reconstructions based on the noisy simulated data. (b) Reconstructions based on the noisy data after smoothing with parameter value $\omega = 0.5$.

As Fig. 7a illustrates, the unsmoothed reconstruction method was generally robust to substantial measurement error. The 20 simulations depicted all follow the same general trend, and most agree in important details like the timing of the onset and peak of gene expression. However, some simulations yielded large swings in the inferred protein synthesis rates due to the noise. In Fig. 7b, we see that a moderate amount of data smoothing (using $\omega = 0.5$ for both the population and protein assay time series) removed these spurious dynamics and lowered the variability in the reconstructions. Smoothing comes at a cost; however; in this case the rapid change in protein levels was partially smoothed away, so that the peak synthesis rate was underestimated by the reconstructions. Since there is no *a priori* method for choosing the amount of smoothing, one should experiment with different values of the smoothing parameter, and use the smallest one that seems to produce an acceptable signal to noise ratio based on visual inspection of the smoothed time series.

In this example, the variability in the inferred gene expression grows over time. This is not due to the accumulation of error; our method uses a local approximation to the derivative at each point, so each data point only influences the current and immediately preceding and following synthesis rates. Rather, the increasing error over time is due to our assumption of multiplicative noise and the fact that the population and protein assay levels increase over time in this example. The variability in the reconstructed gene expression at a given time simply reflects the magnitude of the experimental errors near that time.

4 Discussion

We have presented a pair of simple mathematical models that relate the dynamics of gene expression to the accumulation of reporter proteins. The models provide a straightforward way of reconstructing the approximate dynamics of gene expression from time series data for the reporter protein concentration and cell population. While the models rely on several assumptions, they are ones commonly used in interpreting reporter protein data. Moreover, our simulations indicate that the method is robust to moderate nonlinearities in the protein and population measurements, to uncertainty about the protein decay kinetics, and to realistic levels of experimental noise.

The models are simple and general because they are intended only to infer an unobserved process (gene expression) from an observed one (reporter protein accumulation). They make no assumptions about the mechanisms regulating gene expression, allowing the dynamics to take on any pattern. The result is a simple algorithm, implemented in a spreadsheet, that can be used to infer the dynamics of promoter activity or protein synthesis from standard time series data. Nevertheless, the approach is sufficiently mechanistic to provide a more detailed analysis of gene expression than approaches based on linear regression [22, 23], and provides a much clearer picture of the dynamics than visual inspection of reporter protein data alone.

In the scenarios that we simulated, knowing the protein half-life within a factor of two was usually sufficient for recovering the general pattern and approximate timing of gene expression. However, details may only be reliably captured with more precise knowledge of the decay rate, especially if the reporter protein half-life is comparable to the timescale of the gene expression dynamics. While the general patterns are robust to variation in the decay rate or proteolytic saturation, uncertainty about the lifespan of reporter proteins *in vivo* may be an important limitation to precise inference about the dynamics of gene expression from reporter fusions. To date, there does not appear to

have been a major effort to determine these kinetics in a wide range of cell types and experimental conditions.

In principle, the reconstructed reporter protein synthesis produced by our method could be used as input to a model for predicting the dynamics of the functional protein or RNA of interest. For example, suppose our method were applied to data from a transcriptional fusion. The output would consist of the time course of the (relative) production of the target gene's mRNA. This could then be incorporated into a more complex model for simulating the target protein's translation, folding, dimerization, decay, *etc*. The result would be a prediction of the amount of functional target protein over time. We have not pursued this in the current study because of the variety of possible scenarios and the number of parameter values that would be required. However, it may be a promising avenue for future research, and it indicates the extent to which mathematical models can be used to extend the type of predictions made from gene fusion data.

More direct methods of measuring gene expression have been developed, including microarrays and quantitative reverse transcription PCR. However, reporter gene fusions are likely to remain a popular method for inferring expression dynamics because of their relatively low cost and use in other applications [1, 24]. Model-based inference of gene expression dynamics from reporter fusions has the potential to increase the information provided by this important technology.

David Brown was supported by NSF grant DEB 0120169. David Brown and Phoebe Lostroh were supported by MacArthur Professorships from Colorado College.

The authors have declared no conflict of interest.

Appendix

The balance equation in the general model can be solved using the method of integrating factors to obtain:

$$B(t) = B(0)e^{-\int_0^t m(s)+\frac{1}{V(s)}V'(s)\mathrm{d}s} + \int_0^t e^{-\int_w^t m(s)+\frac{1}{V(s)}V'(s)\mathrm{d}s} p(w)\mathrm{d}w.$$

In the simplest case where the protein decay rate and cell population specific growth rate are both constants

$$(m(t) = m \text{ and } \frac{1}{V}\frac{\mathrm{d}V}{\mathrm{d}t} = r),$$

this simplifies to:

$$B(t) = B(0)e^{-(m+r)t} + \int_0^t e^{-(m+r)(t-w)} p(w)\mathrm{d}w.$$

Thus, we see that the protein level at time t is a weighted integral of the history of gene expression, with the "memory" of the system decaying exponentially at rate $m + r$. In the more general case, the same idea holds, but the weighting function is more complicated.

5 References

[1] Naylor, L. H., Reporter gene technology: The future looks bright. *Biochem. Pharmacol.* 1999, *58*, 749–757.
[2] Pessi, G., Blumer, C., Haas, D., lacZ fusions report gene expression, don't they? *Microbiology (Reading, England)* 2001, *147*, 1993–1995.
[3] Silhavy, T. J., Gene fusions. *J. Bacteriol.* 2000, *182*, 5935–5938.
[4] Subramanian, S., Srienc, F., Quantitative analysis of transient gene expression in mammalian cells using the green fluorescent protein. *J. Biotechnol.* 1996, *49*, 137–151.
[5] Leveau, J. H., Lindow, S. E., Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *J. Bacteriol.* 2001, *183*, 6752–6762.
[6] Bhasi, K., Forrest, A., Ramanathan, M., SPLINDID: A semiparametric, model-based method for obtaining transcription rates and gene regulation parameters from genomic and proteomic expression profiles. *Bioinformatics (Oxford, England)* 2005, *21*, 3873–3879.
[7] Levsky, J. M., Singer, R. H., Gene expression and the myth of the average cell. *Trends Cell Biol*. 2003, *13*, 4–6.
[8] Golding, I., Paulsson, J., Zawilski, S. M., Cox, E. C., Real-time kinetics of gene activity in individual bacteria. *Cell* 2005, *123*, 1025–1036.
[9] Yildirim, N., Mackey, M. C., Feedback regulation in the lactose operon: A mathematical modeling study and comparison with experimental data. *Biophys. J*. 2003, *84*, 2841–2851.
[10] Mandelstam, J., Turnover of protein in starved bacteria and its relationship to the induced synthesis of enzyme. *Nature* 1957, *179*, 1179–1181.
[11] Hakkila, K., Maksimow, M., Karp, M., Virta, M., Reporter genes lucFF, luxCDABE, gfp, and dsred have different characteristics in whole-cell bacterial sensors. *Analyt. Biochem.* 2002, *301*, 235–242.
[12] Liang, S. T., Dennis, P. P., Bremer, H., Expression of lacZ from the promoter of the Escherichia coli spc operon cloned into vectors carrying the W205 trp-lac fusion. *J. Bacteriol.* 1998, *180*, 6090–6100.
[13] Bipatnath, M., Dennis, P. P., Bremer, H., Initiation and velocity of chromosome replication in Escherichia coli B/r and K-12. *J. Bacteriol* 1998, *180*, 265–273.
[14] Lange, R., Fischer, D., Hengge-Aronis, R., Identification of transcriptional start sites and the role of ppGpp in the expression of rpoS, the structural gene for the sigma S subunit of RNA polymerase in *Escherichia coli*. *J. Bacteriol.* 1995, *177*, 4676–4680.
[15] Bachmair, A., Finley, D., Varshavsky, A., In vivo half-life of a protein is a function of its amino-terminal residue. *Science (New York, NY)* 1986, *234*, 179–186.

[16] Zhou, P., Determining protein half-lives. *Methods Mol. Biol. (Clifton, NJ)* 2004, *284*, 67–77.
[17] Belle, A., Tanay, A., Bitincka, L., Shamir, R., O'Shea, E. K., Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 13004–13009.
[18] Grunenfelder, B., Rummel, G., Vohradsky, J., Roder, D. *et al.*, Proteomic analysis of the bacterial cell cycle. *Proc. Natl. Acad. Sci. USA* 2001, *98*, 4681–4686.
[19] Haldimann, A., Daniels, L. L., Wanner, B. L., Use of new methods for construction of tightly regulated arabinose and rhamnose promoter fusions in studies of the Escherichia coli phosphate regulon. *J. Bacteriol.* 1998, *180*, 1277–1286.
[20] Cormack, B. P., Valdivia, R. H., Falkow, S., FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* 1996, *173*, 33–38.
[21] Andersen, J. B., Sternberg, C., Poulsen, L. K., Bjorn, S. P. *et al.*, New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl. Environ. Microbiol.* 1998, *64*, 2240–2246.
[22] Bruenn, J., Hollingsworth, H., A mutant of *Escherichia coli* with a new, highly efficient promoter for the lactose operon. *Proc. Natl. Acad. Sci. USA* 1973, *70*, 3693–3697.
[23] Mecsas, J., Rouviere, P. E., Erickson, J. W., Donohue, T. J., Gross, C. A., The activity of sigma E, an *Escherichia coli* heat-inducible sigma-factor, is modulated by expression of outer membrane proteins. *Genes Dev*. 1993, *7*, 2618–2628.
[24] Beckwith, J., The all purpose gene fusion. *Methods Enzymol.* 2000, *326*, 3–7.