Published in final edited form as:

Int J Comput Biol Drug Des. 2011 ; 4(4): 387–395. doi:10.1504/IJCBDD.2011.044445.

Using protein abundance to indicate underlying mRNA expression levels in *E. coli*: an SEM modelling approach

Jacqueline J. Harris*,

Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA

Christine W. Duarte, and

Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA

Michael C. Mossing

Departments of Chemistry and Biochemistry, University of Mississippi, Mississippi 38677, USA

Christine W. Duarte: cduarte@ms.soph.uab.edu; Michael C. Mossing: mmossing@olemiss.edu

Abstract

Do steady-state protein levels accurately predict mRNA levels? Based on the central dogma (DNA → RNA → protein) current protein levels are representative of mRNA present at an earlier time. However, most cellular mRNA – protein comparative studies try to relate steady-state protein levels to current mRNA levels in cells. Protein steady-states are more correctly related to protein production, protein degradation and other complex cellular conditions. Using Structural Equation Modelling (SEM) we relate linear protein measurements to latent mRNA in *E. coli*. This method can be used to find the optimal protein measurements that explain underlying mRNA expression, and better understand the proteomic and transcriptomic relationship in *E. coli* gene expression.

Keywords

systems biology; SEM; structural equation modeling; gene expression; proteomics transcriptomic relationship

1 Introduction

When a gene is turned ‘on’ several events take place before the effects of activation are realised. The gene is transcribed into RNA, translated into a polypeptide which then folds in three dimensions and assembles into a protein product. *In vivo* studies of β -galactosidase expression in *Escherichia coli* show an initial lag of about three minutes after the addition of artificial inducer, isopropyl- β -D-galactoside (IPTG) before the first β -gal molecules are detected (Boezi and Cowie, 1961; Kepes, 1969; Pardee and Prestidge, 1961, 1963). There is also a similar lag after the withdrawal of inducer. These lags represent the completion of steps in β -gal expression. These steps are ordered:

- transcription
- translation

- protein folding
- protein assembly.

Simply put, the carriage does not come before the horse, nor do they simultaneously exist. That being said, most mRNA–protein comparative studies try to relate current protein levels to current mRNA levels (Nie et al., 2006; Heijne et al., 2003), but the mRNA that gives rise to current protein levels clearly comes before.

Since gene expression is a dynamic process, it makes sense to monitor how a system changes over time. Steady-state differential protein measurements are just snapshots of a dynamic system, and may or may not capture the true complexity of the system. Some studies have found that differential steady-state protein measurements are biased by the cellular environment, and were only modestly correlated with mRNA (Nie et al., 2006). To get a more comprehensive description we need to take several snapshots over time as the system transitions from one state to the next. In this study we demonstrate that taking repeated protein measurements over time can reveal gene expression dynamics, and help us understand underlying mRNA responsible for observed protein abundance.

We have designed synthetic circuits in *E. coli* to isolate a genetic response and monitor the steps of gene expression. Using a two step protein quantification assay which resolves transcription and translation, circuit activity is monitored over time. Although mRNA is not measured directly we can use a latent variable analysis such as Structural Equation Modelling (SEM) to estimate underlying mRNA expression.

Structural Equation Modelling, also known as casual modelling, is a theoretical description of how an unobserved system functions, and yields predictions of observed outcomes (Hancock and Mueller, 2006). Each structural model is expected to encompass the components of causation which are:

- variable isolation
- variable association
- temporal priority.

Our synthetic genetic system allows us to construct a causal model that estimates its isolated mRNA expression based on observed protein abundance. Here SEM is used as a tool to analyse circuit response, specifically Latent Growth Modelling (LGM).

2 Results and discussion

Isolated Genetic Response

The design and construction of synthetic genetic circuits in *E. coli* has allowed us to control and monitor the dynamics of β -galactosidase gene expression in living cells (cf. experimental part). The circuits control elements originate from the lac operon and the immunity region of bacteriophage lambda (Figure 1), and allows rapid induction and repression of the circuit by addition or dilution of IPTG (isopropyl- β -D-thiogalactoside) respectively.

For unique characterisation of β -gal expression from our circuits we use a method that resolves transcription and translation of β -gal synthesis (cf. experimental part). In this method IPTG is added and then diluted giving short pulses of induction. Each pulse represents the initiation of mRNA molecules for that time. The lag in Expression represents the completion of the remaining steps of β -gal expression from that pulse of transcription.

The output from our circuits is units of β -gal, but we can model the relationship between enzyme expression and mRNA initiation during induction to estimate mRNA production.

In part one of the two part parallel method, IPTG is added and then diluted. Upon the addition of IPTG lac repressor (lacR) is inactivated, and the promoter is free to initiate transcription. After dilution of IPTG lacR re-binds the promoter and terminates transcription. However, any mRNA already present can continue on the path of expression resulting in a proportional linear yield of β -gal Figure 2(A), which is the total amount of β -gal molecules possible from a given amount of mRNA.

In part two of the method β -gal expression is directly assayed by the addition of Z-buffer and chloramphenicol (CAM) which stops protein synthesis (cf. experimental part). Only protein that has completed all steps of expression are expressed Figure 2(A). It is important to point out that even though mRNA is present; it is not yet expressed as β -gal. However, after waiting for mRNA to be completely expressed as β -gal, in part one, we know that mRNA is present, even if not full expressed as β -gal. It is because of this isolated transcription/translation response we can construct a structural model that estimates the underlying mRNA expression.

Modelling Protein Accumulation

As demonstrated in our experiments, measurements during linear synthesis of β -gal (after an initial lag and before protein approaches its steady-state) are indications of the underlying steady-state mRNA level. The IPTG dilution experiment (part 1 of our parallel experiment) (Figure 2(A)) is most representative of β -gal linear accumulation due to mRNA. To model this relationship we use LGM, which is a form of SEM. Latent growth modelling is applicable to repeated measures data, and focuses on how the data changes over time. Latent growth analysis has two levels. Level 1 is the measurement part of the model, which is the analysis of repeated measures and their trajectory or slope. Level 2 is the structural part of the model, and has units of the experiment. The latent variables are the intercept and slope. The loading for each latent intercept is 1, indicating a constant coefficient for the intercept in the model. The slope is the linear trajectory of the repeated measurements. The level 1 model for our experiments is expected to take the form of a line, and express linear accumulation of β -gal over time. The measurement model is shown below:

$$Y_{ij} = \pi_{0j} + \pi_{1j}t_{ij} + r_{ij} \quad (\text{measurement model}). \quad (1)$$

Y_{ij} represents a β -gal measure in experiment j at time i . π_{0j} is the intercept for β -gal trajectory in experiment j or the predicted β -gal level at baseline, where time = 0 is the baseline. π_{1j} is the slope of the trajectory in experiment j . t_{ij} is the time i corresponding to each measurement in experiment j . r_{ij} is the random error for unexplained deviation from the line in experiment j .

The structural model has two parameters which are,

$$\pi_{0j} = \beta_0 + u_{0j} \quad (2)$$

$$\pi_{1j} = \beta_1 + u_{1j}. \quad (3)$$

These parameters are modelled with a fixed component expressing the mean β_i across all experiments, and a random component u_{ij} representing each experiment's deviation from the

mean. Variance and covariance can be estimated for the random variables. In Figure 3 the path diagram for the latent growth model of β -gal expression includes latent intercept and slope variables. These variables have fixed loadings that convey information about the timing of measurements in each experiment. The model was fit in Mplus (Muthen and Muthen, 2011) using multiple simulated experiments similar to the actual experiments presented in this paper. Specifically, 1000 simulations were run under the model shown in Table 1 (see first column for simulated population). Figure 4 shows a random subset of simulated experiments used to fit the model.

Table 1 shows the results of the parameter estimates from the β -gal expression simulations. As shown in the table, the actual value for β -gal is set to 10 upon the initiation of expression (induction of mRNA) this is when IPTG is added at time 0. Our model estimates it to be 10.04 β -gal units with 95% confidence. During actual accumulation, β -gal increases on average 15 β -gal units per minute. Our model predicts the slope to be 15.01, which again contains the true value is within its 95% confidence interval.

Estimates of the variance components were consistent with their true values (Table 1). A test of model fit indicated that the linear model was a good fit for β -gal expression data. For a simulation with a 1000 observations $\chi^2(31) = 39.6$, $p = 0.14$, and $RMSEA = 0.017$.

3 Conclusions

An important goal in gene expression studies is to link transcriptomic and proteomic data to understand the dynamics of a genetic network. Based on the central dogma current protein levels are representative of mRNA present at an earlier time. Our synthetic circuits were constructed to isolate cause/effect variables of gene expression in *E. coli*. The circuits contain a control sequence that allows rapid induction and re-repression by adding and diluting an artificial inducer, IPTG. The induction patterns from the circuits were examined using a two part analysis that selectively arrest transcription or translation. After the addition of IPTG there is a lag before β -gal starts to accumulate. The duration of this lag is about three minutes, followed by a steady increase in β -gal accumulation. Both periods are representing the dynamics of protein production, but during the lag period β -gal accumulation is low. However, mRNA is proven to be present from the start of circuit induction by the two part analysis (cf. experimental part). This steady-state period (lag) is representative of the time it takes to complete all steps of β -gal gene expression. However, the increase in β -gal accumulation after the lag is the trajectory of β -gal reflected by the underlying level of mRNA.

It is common practice to take steady-state differential protein measurements and use fold change as a quantitative measure of an induced genetic system. However, system dynamics are omitted. This method allows us to look at the dynamic accumulation of proteins, and relate them to latent mRNA of the system. Our structural model relates the time during the genetic response when β -gal accumulation is linear and more representative of mRNA without the bias of conditional properties such as time to complete β -gal gene expressing steps. These important timing constraints can only be investigated during repeated measure analysis as a genetic system transitions from one state to the next. We feel that this study and others like it will give us a better understanding of the relationship between transcriptomics and proteomics of dynamic genetic networks.

4 Experimental part

Construction of genetic circuit

Simplified control circuits were constructed using the plasmid vector pAH125, which contains the γ replication origin of R6K, which requires the trans-acting Π protein (encoded by *pir*) for replication (Haldimann and Wanner, 2001). The vector also contains a *lacZ* gene downstream of the Multi-Cloning Site (MCS). Cloned in the MCS is a binding site for RNA polymerase (P_R promoter) is under dual control of the *lac* operator (*lacO*) and the lambda operator (λO). The *lacO* overlaps P_R promoter and represses transcription when bound by *lac* repressor. *Lac* repressor is produced by the *lacI* gene that is synthesised continuously from an unregulated promoter (Figure 1).

Strains

All strains are derivatives of *E. coli* K-12 from the Barry Wanner lab (Haldimann and Wanner, 2001). Strains selected can express the plasmid vector (pAH125) in single, medium and high copy. The single copy strain (BW25113) can integrate pAH125 into its chromosome via a temperature sensitive plasmid (pINT-ts).

Plasmid integration

BW25113 is unable to replicate pAH125 without the *pir* protein. Single copy integrants are generated by transforming BW25113 with pINT-ts. BW25113 strains containing pINT-ts, were grown at 30°C in SOC and made competent using CCMB method. Plasmid DNA was added to the competent cell culture and incubated at 37°C for 1 h and then at 42°C for 30 min, and then spread on selective agar plates and incubated at 37°C.

PCR verification of integrant copy number

Isolated colonies were picked and suspended in 20 μ L of water. 5 μ L of the cell suspension, 10pmol of each primer (CCAGGCATCACGGCAATATAC (P1), GGCTCTGGTCTGGTAGCAATG (P4), AACACTTAACGGCTGACATGG (P2), AACGAGTATCGAGATGGCACA (P3)) together, and 0.5 U of *goTaq* DNA polymerase were combined in 5x *goTag* buffer-2.5 $MgCl_2$ with deoxynucleoside triphosphates in a final volume of 25 μ L. PCR was carried out for 25 cycles with denaturing for 1 min at 95°C, annealing for 1 min at 60°C, and extension for 1 min at 72°C.

IPTG concentration determination for induction/repression

In order to induce and un-induce the circuits the working concentration of IPTG was determined by using a range of (IPTG) Through the completion of a comparative assay of induction efficiency using IPTG concentrations from 1 μ M to 2 mM, it was determined that induction begins at (IPTG) > 2 μ M and induction is maximum at (IPTG) $\geq$ 400 μ M. This information was applied to time-course induction experiments in order to induce and suppress β -galactosidase synthesis.

β -galactosidase two assay method

Using two different treatments that selectively arrest transcription or translation, we can measure the dynamics of gene expression in each circuit. Re-binding of *lacR* after induction stops transcription, and treatment with CAM stops protein synthesis. Comparison of these two treatments, help estimate change in mRNA production rate over time (Kepes, 1969). From the induced culture time points are taken for treatment 1, diluting the effects of IPTG, and allowing all mRNA transcribed to be translated, and treatment 2, direct assay (Figure 5). Treatment 1 is achieved by diluting induced cells 50-fold into pre-warmed M9 media at 30°C and waiting until all possible enzymes have been made. Time points are assayed using

the Miller method. Z-buffer is added to a final volume of 1ml to each time point. 40 μ L of chloroform, 20 μ L 1% SDS, and a final concentration of 50 μ g/mL chloramphenicol (CAM) is added. Then the samples are lysed using a multi-channel pipette. After the cell lysis 200 μ L of 4 mg/ml ONPG is added. After ONPG is added the time that it takes for yellow colour to appear and the absorbance of that yellow colour at OD420 are used to calculate Miller units of β -galactosidase. A Miller unit is defined as 1nMole of o-nitrophenol (ONP) produced per minute from hydrolysis of o-nitrophenol- β -D-galactoside (ONPG) (Miller, 1972, 1992). The Miller calculation also has a correction for cell density at OD600. the equation used to calculate Miller units is as follows:

$$\frac{(A_{420} - 1.75 \times A_{550}) \times 1000}{(A_{600} \times \text{time}(\text{min}) \times \text{volume of cells}(\text{ml}))} \quad \text{Miller unit equation.}$$

Acknowledgments

This research is supported in part by grants NSF # MCB- 0317026 and T32HL072757 from the National Heart, Lung, and Blood Institute.

References

- Boezi JA, Cowie DB. Kinetic studies of beta-galactosidase induction. *Biophys J*. 1961 November. 1:639. [PubMed: 13870521]
- Haldimann A, Wanner BL. Conditional-replication, integration, excision, and retrieval plasmid-host system for gene structure-function studies of bacteria. *J Bacteriol*. 2001; 183:6384. [PubMed: 11591683]
- Hancock, GR.; Mueller, RO. Structural Equation Modeling: A Second Course. Information Age Publishing, Inc; Greenwich: 2006.
- Heijne WHM, et al. Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach. *Biochim Pharmacology*. 2003; 65:857.
- Kepes A. Transcription and translation in the lactose operon of *Escherichia coli* studied by in vivo kinetics. *Prog Biophys Mol Biol*. 1969; 19(1):199–236. [PubMed: 4192354]
- Miller, JH. Experiments in Molecular Genetics. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, New York: 1972.
- Miller, JH. A Short Course in Bacterial Genetics. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, New York: 1992.
- Muthen, B.; Muthen, L. Mplus version 6.1, Program for Latent Modeling. Muthen & Muthen; 2011.
- Nie L, Wu G, Zhang W. Correlation between mRNA and protein abundance in *Desulfovibrio vulgaris*: a multiple regression to identify sources of variations. *Biochim Biophys Research Communications*. 2006; 336:603.
- Pardee AB, Prestidge LS. The initial kinetics of enzyme induction. *Biochim Biophys Acta*. 1961 April. 49:77. [PubMed: 13732743]
- Pardee AB, Prestidge LS. Inactivation of beta-galactosidase induction by ultraviolet light. *Biochim Biophys Acta*. 1963 April. 76:614. [PubMed: 14099619]

Biographies

Jacqueline J. Harris received her PhD in Chemistry from the University of Mississippi, Departments of Chemistry and Biochemistry, University, MS, USA in 2010. Currently, she is a postdoctoral fellow in the Department of Biostatistics, Section on Statistical Genetics at the University of Alabama at Birmingham. Her research interests include finding ways to relate genomic and proteomic data, systems biology, and GWAS analysis.

Christine W. Duarte received her PhD in Bioinformatics with a co-major in statistics at North Carolina State University in 2009. Currently, she is an Assistant Professor in the Department of Biostatistics, Section on Statistical Genetics at the University of Alabama at Birmingham. Her research interests include developing statistical methods for applying systems biology approaches to high dimensional genomic data analysis.

Michael C. Mossing received his PhD in Cellular and Molecular Biology from the University of Wisconsin in 1985. Currently, he is an Associate Professor in the Departments of Chemistry and Biochemistry at the University of Mississippi. His research interests include biophysical chemistry, DNA binding protein structure and function, combinatorial mutagenesis, protein engineering and expression, circular dichroism, fluorescence and NMR spectroscopy, thermodynamics and kinetics of protein folding and macromolecular recognition, the coupling of protein folding and assembly to DNA recognition.

Figure 1.

To make pAH125-laci-OPO-lacZ, PCR product laci-OPO was digested with *Pst*I and *Xba*I. The product of the digest is the laci-OPO sequence, which can be ligated into pAH125's MCS using *Pst*I and *Xba*I sticky ends (see online version for colours)

Figure 2.

(A) β -gal expression as a function of the time in contact with IPTG (cf. experimental part) and (B) chloramphenicol is used to stop β -gal expression (see online version for colours)

Figure 3.
Latent growth model of β -galactosidase expression

Figure 4.
Two randomly selected simulated β -gal trajectories (see online version for colours)

Figure 5.

Two-part assay is used to monitor the time course of β -gal activity of induced cell cultures (cf. experimental part)

Table 1Fixed and random parameter estimates from simulated data for β -gal expression model

	Simulated population (β -gal unit)	Estimate (β -gal unit)	SE	t	P-value
<i>Fixed parameters</i>					
Intercept	10	10.0402	0.021	487.181	<0.001
β -gal (slope)	15	15.0158	0.015	1013.194	<0.001
<i>Random parameters</i>					
<i>Variances</i>					
Intercept	0.20	0.2219	0.020	11.115	<0.001
β -gal (slope)	0.20	0.2167	0.010	22.063	<0.001
<i>Covariances</i>					
Intercept/ β -gal (slope)	0.040	0.0538	0.010	5.519	<0.001