Computational design of RNA-based oscillatory circuits

J. Binysh
* *University of Warwick, Complexity Department*

***Abstract*—genetic circuitry, RNA's offers an attractive alternative to more traditional methods, which typically involve using proteins to regulate DNA transcription. In comparison to proteins, it is relatively straightfor**

I. Introduction

The process of gene expression can be summarised as follows: DNA is read, and a copy of it is made, in the form of an RNA molecule (this is called *transcription*). This RNA molecule (known as messenger RNA, or mRNA) makes its way to a piece of cellular machinery called the Ribosome, which reads it, and makes a protein - which protein is made depends on the DNA sequence originally read (*translation*).

The path from genetic transcription to protein expression is naturally regulated in many ways [**?**]. This regulation allows the cell to control protein expression, and so cell behaviour, in response to various environmental cues. The natural cell machinery which performs it takes the form a genetic circuits - networks of interacting gene expression regulators. This genetic circuitry offers rich possibilities for modification, and an important goal within synthetic biology is to understand and manipulate it.

As well as acting as the intermediate between DNA and protein, RNA molecules play direct and important roles in regulating gene expression [1]. For the synthetic biologist looking to engineer regulation of genetic circuitry, RNA offers an attractive alternative to more traditional methods, which typically involve using proteins to regulate DNA transcription. In comparison to proteins, it is relatively straightforward to predict the structure and function of an RNA from its sequence using physiochemical models. Recently, this has been exploited to computationally design synthetic sRNA's - small RNA's which do not code for a protein, but rather have some direct regulatory function - with regulatory behaviour that can be predicted [2] [3].

This report will focus on one such sRNA system, introduced in [3]. It will extend existing understanding of it beyond the qualitative by proposing a quantitative model of gene expression, in the form of a set of ODE's, and fitting this model to available time series data to estimate its unknown parameters.

The report is structured as follows. In the remainder of this section we review the sRNA regulatory system we will consider, and discuss recent single cell fluorescence experiments performed on this system. In section II we propose a set of ODE's to model the system. In sect III and estimate its unknown parameters by fitting to time series data. Finally, in section IV, we conclude, and suggest directions for further work.

A. The sRNA regulatory system

In bacteria, one mechanism by which gene expression is regulated is as follows [4]: In order for a bacterial mRNA to be translated into a protein, the Ribosome must initially bind to the mRNA (Fig. 2). This occurs at the Ribosome Binding Site (RBS) [5], a specific nucleotide sequence found on the mRNA. In an mRNA there is an untranslated region of nucleotides at the 5' end of the molecule (the UTR), upstream of the RBS. Translation may be self repressed by this 'tail' of the mRNA folding over and binding across the RBS, forming a stem loop in the mRNA and preventing the Ribosome from binding (Fig. 2). This self repression may be released with an sRNA which binds to the same region on the mRNA - the new conformation of the sRNA:mRNA complex uncovers the RBS, allowing the Ribosome to bind. In summary, the presence of the sRNA positively regulates gene expresssion.

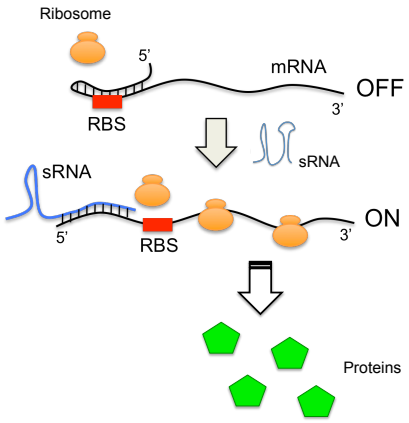


Fig. 2: A mechanism by which sRNA's can regulate gene expression. Initially, the 5' UTR of the mRNA is folded over the RBS, forming a loop and blocking Ribosome binding. The sRNA binds to this mRNA, causing a conformational change which uncovers the RBS, and allows translation to occur.Image reproduced from [3]

[3] proposed a computational methodology to design general genetic circuits based on RNA interactions, and as a case study of the methodology chose to design a synthetic sRNA-

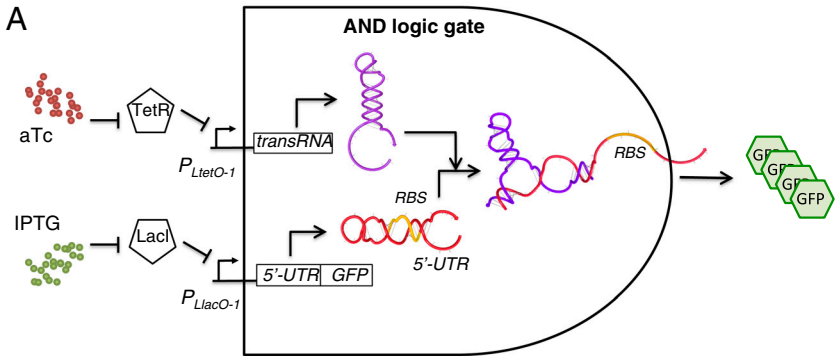


Fig. 1: A logical AND gate formed from a self repressed mRNA, and an sRNA which uncovers its RBS. In this system, transcription of the sRNA (transRNA) and mRNA (5'-UTR,GFP) are controlled by two promoter regions, $P_{\mathrm{LtetO}-1}$ and $P_{\mathrm{LlacO}-1}$. These are disabled by the presence of two chemical repressors, TetR and LacI, found naturally in the strain of *E. coli* discussed. These chemical repressors are themselves disabled by two chemicals, aTC and IPTG. In the notation of the diagram, a barred line indicates repression, and an arrowed line indicates production. We see a 'double negative' in aTc repressing TetR, which itself represses transcription of the sRNA (likewise for IPTG and the mRNA). Thus presence of the sRNA and mRNA are controlled by the presence of aTc and IPTG, which can be experimentally introduced to the cell. Image reproduced from [3].

mRNA pair capable of acting in the manner described above. The algorithm assumed an interaction scheme between the RNA's as shown in Fig. A.3. The two RNA's, originally in their own individually folded states, would initially interact via a small 'toehold' sequence of unpaired nucleotides to form an unstable transition state. This intermediate complex would then rapidly form a final, stable complex with the desired conformation. By suggesting sRNA and mRNA sequences which optimised this energy landscape, [3] suggested several devices which would form a stable hybrid with the RBS free, and experimentally validated their function in *E. coli*, using an mRNA which codes for GFP for experimental ease (note the algorithm only optimises the 5' UTR of the mRNA, so the actual protein being coded for is unimportant).

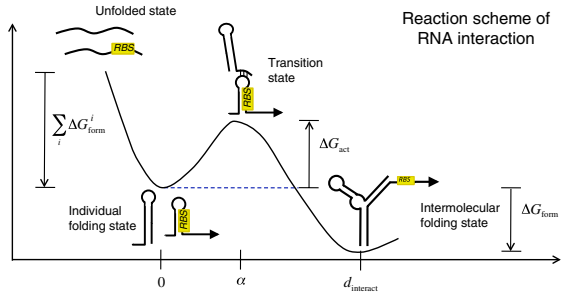


Fig. 3: A network with an intuitively clear community structure, which is captured by the partition chosen, shown in gray. Image reproduced from [3]

Further, by placing the concentrations of the sRNA and mRNA under the control of tuneable promoters, [3] constructs a logical AND gate from one of the proposed devices (RAJ11) *in vivo* (Fig. 1). In this system, transcription of the designed sRNA and mRNA are placed under the control of promoter regions, $P_{\mathrm{LtetO}-1}$ and $P_{\mathrm{LlacO}-1}$ [6]. These are in turn controlled by two transcriptional repressors, TetR and LacI, which are naturally present in the strain of E. coli considered. These repressors disable the promoter regions, and so by default transcription of the RNA's is turned off, and no protein is produced. These repressors can themselves be disabled by the presence of two chemicals, aTC and IPTG, which can be introduced externally into the cell (Fig. 1). So transcription of the two RNA's is indirectly controlled by the presence of two chemicals - if neither is present, sRNA and mRNA transcription is repressed, and no protein is produced. If only one is present, the AND gate remains off, either because there is no mRNA to be translated into protein, or because the mRNA is self repressed. But when both are present, the conformational change discussed above occurs, and protein is produced.

promoter?

Although a qualitative understanding of this system exists [3], it is of interest to attempt a quantitative understanding of the genetic circuit involved. Such an understanding would allow, for example, tailoring of the system in reponse to design requirements, by altering the values of the important parameters of the model. By changing which sRNA-mRNA device is used in the system, it would also allow exploration of the relationship between the thermodynamic properties of

each device, and the model's rate constants.

B. Single Cell Fluorescence Data

Recent experiments have used timelapse microscopy to observe the fluoresence of bacteria which contain the above device, as they are periodically forced with a varying aTc or IPTG concentration [7].

The bacteria are grown a single layer thick in rows of chambers. A medium constantly flows through these chambers, allowing normal feeding of the bacteria, and the introduction of aTc or IPTG. The chambers are monitored with software which traces the position of each cell over time, allowing timeseries of individual cell fluorescences to be recorded.

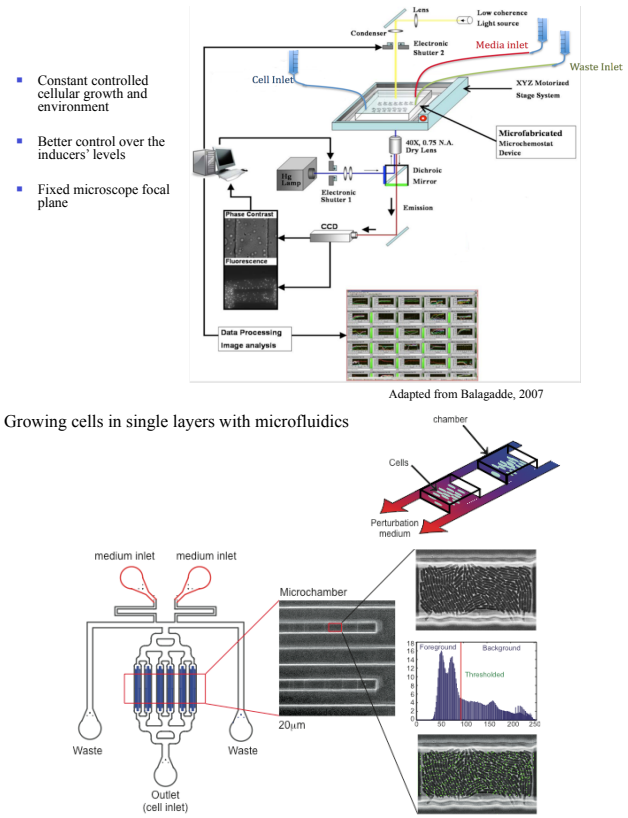


Fig. 4: A schematic of an experimental setup which allows single cell fluorescences to be recorded over time. Bacteria are grown in chambers, with a constant supply of medium flowing through them from the media inlet. This inlet also allows the introduction of aTc IPTG. Software then traces these cells, allowing fluorescence timeseries to be built. Image reproduced from [7].

The data we will consider consists of three sets of individual cell timeseries, labelled *13_9*, *14_7* and *14_9*. They correspond to three different experimental runs of the above apparatus, for which IPTG concentration was held constant, at a value assumed large enough to saturate the cell's response, and aTc concentrations were varied periodically. Section A shows the full datasets, with their forcing functions.

II. Model Derivation

We propose a simple model, consisting of a set of ODE's with mass action kinetics [8], to describe the system. Tables I, II give complete descriptions of the parameters and state variables.

$$\frac{ds}{dt} = \frac{N\alpha_T}{f_T}y(t) - (\mu+\delta_s)s - k_{\text{on}}sm + k_{\text{off}}s:m \quad (1)$$

$$\frac{dm}{dt} = \frac{N\alpha_L}{f_L}x(t) - (\mu+\delta_m)m - k_{\text{on}}sm + k_{\text{off}}s:m \quad (2)$$

$$\frac{ds:m}{dt} = k_{\text{on}}sm - (k_{\text{off}}+k_{\text{hyb}})s:m - (\mu+\delta_{sm})s:m \quad (3)$$

$$\frac{dc}{dt} = k_{\text{hyb}}s:m - (\mu+\delta_c)c \quad (4)$$

$$\frac{dp}{dt} = \beta m + f_s\beta c - (\gamma+\mu+\delta_g)p - \frac{v_z p}{K_z+p+g} \quad (5)$$

$$\frac{dg}{dt} = \gamma p - (\mu+\delta_g)g - \frac{v_z g}{K_z+p+g} \quad (6)$$

$$z = z_0 + \frac{g}{\Theta} \quad (7)$$

Based on the reaction mechanism in Fig. A.3, the hybridization of the sRNA and mRNA first into an unstable complex, then a stable one, is modelled in eqs.1 - 4. The initial binding is modelled as a reversible reaction with forward and backward rates k_{on} and k_{off}.

$$\text{sRNA} + \text{mRNA} \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} \text{sRNA}:\text{mRNA}_{\text{unstable}}$$

after which the stabilization is modelled as an irreversible reaction with rate k_{hyb}.

$$\text{sRNA}:\text{mRNA}_{\text{unstable}} \xrightarrow{k_{hyb}} \text{sRNA}:\text{mRNA}_{\text{stable}}$$

In addition, these complexes are given degradation rates, δ_s, δ_m, δ_{sm}, δ_c, and dilution of chemical concentrations due to cell growth are modelled with a dilution rate μ.

Control of the system by aTc and IPTG is modelled by $y(t)$ and $x(t)$ in eqs.1-2. $y(t)$ models the response of the sRNA transcription rate to a time varying aTc concentration - it is normalised to lie between 1 and f_T, and is typically sigmoid in response to aTc concentration [3]. Thus the forcing may vary between $N\frac{1}{f_T}$ and $N\frac{\alpha_T}{f}$, where α_T is the maximal transcription rate of the $P_{\text{LtetO}-1}$ promoter. When engineering the system, many copies of the $P_{\text{LtetO}-1}$ promoter and section of DNA coding for sRNA may be placed in the plasmid DNA - this is modelled by the copy number, N. Identical considerations hold for $x(t)$ and IPTG concentrations.

We explicitly model translation as a simple one step process in eqs. 5-7. There is a small rate of translation of the self repressed mRNA [3], which is modelled at rate β, and a larger one for translation of the stable complex, βf_s. Here f_s represents the fractional change in translation rate between the repressed mRNA and the unrepressed complex.

$$\text{sRNA} : \text{mRNA}_{\text{Stable}} \xrightarrow{\beta f_s} \text{GFP}_{\text{Immature}}$$

$$\text{mRNA} \xrightarrow{\beta} \text{GFP}_{\text{Immature}}$$

Initially, the translated GFP is in an immature state, and will not fluoresce. To account for this, we include a maturation rate, γ

$$\text{GFP}_{\text{Immature}} \xrightarrow{\gamma} \text{GFP}_{\text{Mature}}$$

Degradation of the immature and mature GFP is modelled in two ways. Firstly a generic degradation rate δ_g, assumed identical for the mature and immature species, with the dilution rate μ shared by all species.

para about clpX

Finally, eq. 7 simply represents calibration of mature GFP levels to experimentally observed fluorescence.

TABLE I: Model Parameters (those to be estimated shown in bold)

Parameter	Units	Definition
N		Number of copies of promoter existing on plasmid DNA
z_0	AFU	Baseline experimental fluorescence
α_L	nM/min	Maximal transcription rate of $P_{\text{LlacO}-1}$ promoter
α_T	nM/min	Maximal transcription rate of $P_{\text{LtetO}-1}$ promoter
f_L		Unitless ratio between repressed and unrepressed $P_{\text{LlacO}-1}$ transcription rate
f_T		Unitless ratio between repressed and unrepressed $P_{\text{LtetO}-1}$ transcription rate
δ_g	/min	GFP degradation rate
γ	/min	GFP maturation rate
v_z	nM/min	degradation constant of clpx
K_z	nM/min	Dissociation constant of clpx
$\boldsymbol{\Theta}$	nM/AFU	Ratio between GFP concentration and observed fluoresence
$\boldsymbol{\mu}$	/min	Dilution rate
$\boldsymbol{\delta_m}$	/min	mRNA degredation rate
$\boldsymbol{\delta_s}$	/min	sRNA degredation rate
δ_{sm}	/min	Unstable sRNA:mRNA degradation rate
δ_c	/min	Stable sRNA:mRNA degradation rate
$\boldsymbol{k_{on}}$	/min	sRNA:mRNA binding rate
$\boldsymbol{k_{off}}$	/min	sRNA:mRNA unbinding rate
$\boldsymbol{k_{hyb}}$	/min	sRNA:mRNA hybridization rate
$\boldsymbol{\beta}$	/min	Baseline translation rate of repressed mRNA
$\boldsymbol{f_s}$		Ratio of repressed mRNA to unrepressed complex translation rate.

III. Parameter Estimation

Our next goal is to estimate the unknown parameters of this model, given the available fluorescence time series data, by fitting predicted time series from the model to the data. Typically, this is done by minimising the least squares error between model prediction and the experimental data [9]–[11]. Suppose we have some ODE model of our system

TABLE II: State Variables

State variable	Units	Definition
s	nM	sRNA concentration
m	nM	mRNA concentration
$s:m$	nM	Unstable sRNA:mRNA complex concentration
c	nM	Stable sRNA:mRNA complex concentration
p	nM	Immature GFP concentration
g	nm	Mature GFP concentration
z	AFU	Observed fluorescence
$y(t)$		Unitless aTc forcing function
$x(t)$		Unitless IPTG forcing function

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}, t) \tag{8}$$

where $\mathbf{y}$ is our state vector, $\boldsymbol{\theta}$ is a vector of model parameters, and t is time. The model may be integrated numerically, giving a prediction $\mathbf{y}(t, \boldsymbol{\theta})$. An error between the model prediction and an experimental time series is defined as

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (\mathbf{y}_{\text{exp}}(t_i) - \mathbf{y}(t_i, \boldsymbol{\theta}))^2 \tag{9}$$

where the experimental timeseries, $\mathbf{y}_{\text{exp}}(t_i)$ is evaluated at timepoints t_i, $i = 1 \ldots N$. This error function defines a landscape in $\boldsymbol{\theta}$ space, and we seek to minimise it by varying $\boldsymbol{\theta}$. In our case, we do not have experimental data on the full state vector, but only one component of it - the observed fluorescence, $z(t)$. In addition, rather than a single experimental run, we have many, corresponding to a timeseries from each cell. We incorporate this by fitting to the experimental mean of the data, and only minimising over the observed component. Our minimisation problem is thus

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} (z_{\text{exp,mean}}(t_i) - z(t_i, \boldsymbol{\theta}))^2 \tag{10}$$

The next step is performing the minimisation. In general, the landscape defined by the error function is multimodal, and may be very rugged. [1] A local optimisation algorithm will often get stuck in local minima. To try and surmount this problem, [10] suggests the use of a global optimisation algorithm, and in particular recommends several Evolutionary Algorithms, of which we choose one, the CMA-ES [12], [13].

In order to reduce the dimensionality of our search space, we can perform a literature search for existing values of some of our parameters, simplify our model to remove others, and place bounds on those that remain. Section C contains a list of parameter values found in the literature, where available, and their reference, as well as initial bounds placed on parameters

[1] this is true even if ODE model is linear in its parameters, as is almost the case for us - though the ODE model is linear, the resulting solutions are in general not. A counterexample is the harmonic oscillator.

not found in the literature. To further reduce the search space, we assume that δ_m, δ_{sm} and δ_c all take similar values, and set them equal. After this is done, we are left with a 9 dimensional search space, bounded by a hypercube (parameters to be estimated are shown in bold in I).

A. Initial Parameter Estimates

We begin by fitting two of the datasets, *13_9* and *14_7*, by choosing 200 points, uniformly distributed over our initial parameter bounds, and running the CMA-ES starting from them. Results are shown in Figs. 5, B.1. Fig. 5 demonstrates that the model is capable of quantitatively capturing the data - however, the fitting also suggests that some parameters are not tightly constrained, taking values right across the initial bounding range specified. The correlation matrices in Fig. 5 also indicate that there are spaces of parameters within which the fitness function remains approximately constant - for example, in both datasets there exists of positive correlation between values of μ and β. Referring to eq. 5, this makes heuristic sense - the two parameters may have similar effects on model predictions, and may be able to co-vary in such a way as to leave the model prediction unchanged. The results suggest a fitness landscape relatively flat to perturbations in certain combinations of parameters, and indicate we may have difficulty obtaining unique estimates of the model parameters.

We can test the predictive ability of our model by cross validating - taking the parameter values found in the fitting of one dataset, and using them to give model predictions for another. We thus take the parameter values giving the best fit for the *13_9* dataset and use them to predict the *14_7* dataset, and vice versa. Results are shown in Fig.6.

Though the predictions are reasonable, they are substantially worse than the predictions for the data they were trained on. The first reason for this may be experiment to experiment variability of parameter values [11]. The second may be that the model is overfitting.

Taken together, these results suggest that some parameters may be inestimable, and that the data may be better described by a simplified model.

B. Parameter Estimability

There are two main reasons why a parameter may not be estimable [14]–[16]: Model predictions may be insensitive to the value of a particular parameter, or the effects of varying one parameter on model predictions may be highly correlated with the effects of varying several others.

These problems may stem from structural inadequacies in the model (often termed *structural identifiability*), in which two different parameter sets can give identical model predictions [17], [18]. If this is the case, no amount of experimental data will allow us to estimate parameters, and we must consider reformulating the model. [2]

Problems may also arise for more practical reasons (*practical identifiability* [14]). For example, it is possible that in the experimental regime we operate in, parameter effects may be weak, or highly correlated, but in general this is not true. [3] In this case, parameter estimates be may improved by taking data in more varied experimental conditions, and attempting to observe as many components of the model output as possible.

We may begin to investigate these issues in our model by performing a local sensitivity analysis about one of the solutions found in our initial parameter estimation. We numerically estimate the sensitivity matrix, S:

$$S_{ij} = \hat{\theta}_j \frac{\partial z}{\partial \theta_j}\Big|_{t_i} \tag{11}$$

where S_{ij} is the derivative of the observed fluorescence, z, with parameter θ_j, evaluated at timepoint t_i. This matrix contains information about how sensitive z is to perturbations in parameter values, and at what times it is most sensitive. $\hat{\theta}_j$ is the value of the parameter that the derivative is being evaluated at. It is included to set the scale that parameters may vary at, to ensure that apparently small sensitivity values do not result from a poor choice of units.

It can be shown that if sensitivity co-efficients are linearly dependent over the range of observation values, the associated parameters cannot be simultaneously estimated [16]. Related to this fact, a number of measures have been proposed to asses parameter estimability from the sensitivity matrix [14], the simplest of which is to simply plot the sensitivity co-effecients as a function of time, and visually check for obvious linear relations between the curves.

Fig. 7 shows plots of each column of S, evaluated at the parameter set giving the lowest error in the *13_9* dataset - these show the evolution of the sensitivity matrix over time.

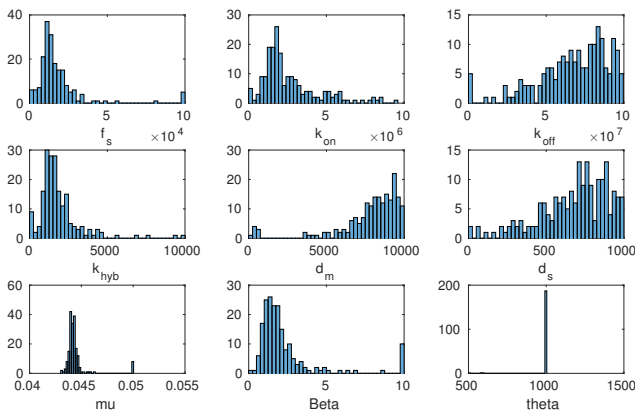
Fig. 7a shows them unscaled, Fig. 7b shows them scaled by the norm of each column of S, so that their shapes may be more easily compared. We see that many of the parameters give sensitivity curves of similar shapes - the effects of perturbing any one of these parameters all look similar in terms of model output.

This may help to explain why some of our initial parameter estimates are very loose - this result suggests there is a family of parameters all of which, in terms of the model output we have available, cannot be resolved. As such, we should view estimates of these parameters with extreme caution. Note that the sensitivity curve that looks least similar to the others in Fig. 7b - μ - corresponds to a relatively tightly estimated value in Fig. 5.

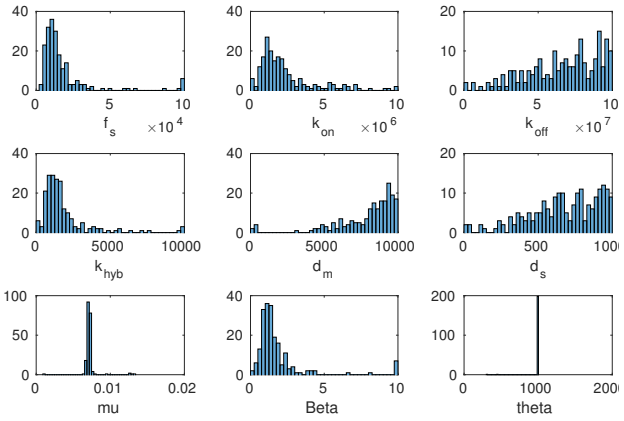
We suggest that part of the issue is the step function forcing used in model predictions, and the timescales on which the system responds to it. We hypothesize that the system may have two timescales in it - a fast timescale in which eqs. 1 - 4, representing the hybridization of the sRNA and mRNA into a stable complex, equilibrate in response to external forcing,

[2] A simple example of a structurally non - identifiable model is $y = \beta_1\beta_2 x$, where we are given data (x, y), and asked to estimate parameters β_1 and β_2 - we can see that, in principle, only the product can ever be estimated, a problem no amount of data can fix.

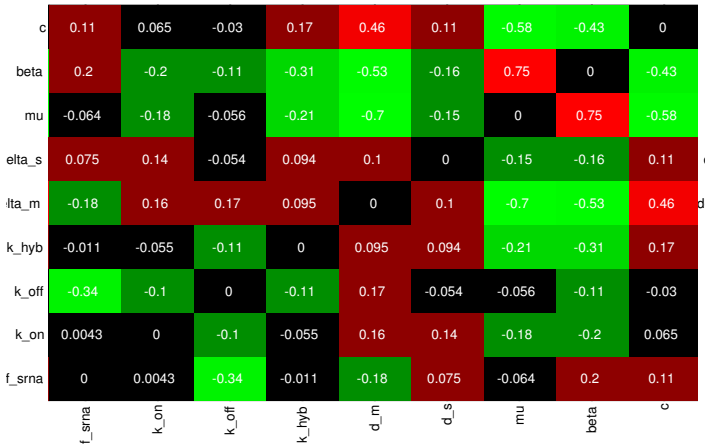
[3] [16] gives an example in which a parameter only effects model predictions after several hours, though others will affect it at all timescales - in this example, if we only took data for a few minutes, the parameter would be inestimable, but in principle it is not.



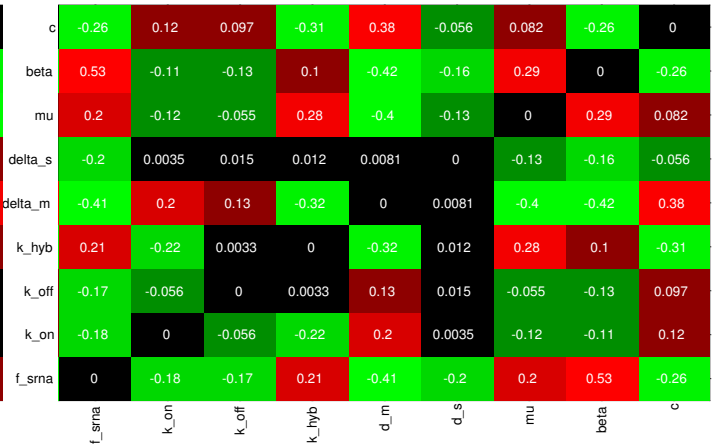
(a) Histogram of estimated parameter values, found from 200 runs of the CMA-ES algorithm. Fitted to the *13_9* dataset.



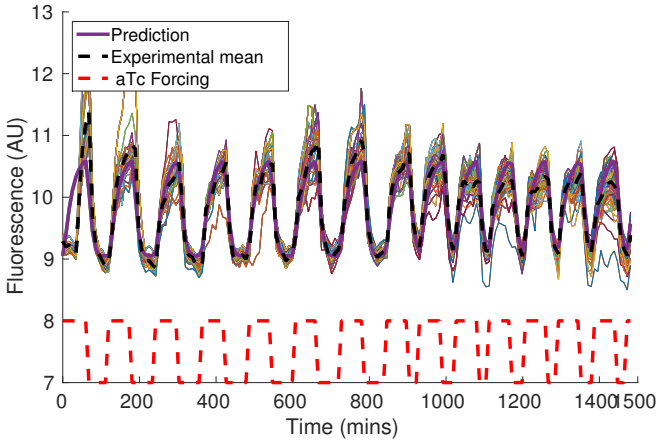
(b) Histogram of estimated parameter values, found from 200 runs of the CMA-ES algorithm. Fitted to the *14_7* dataset.



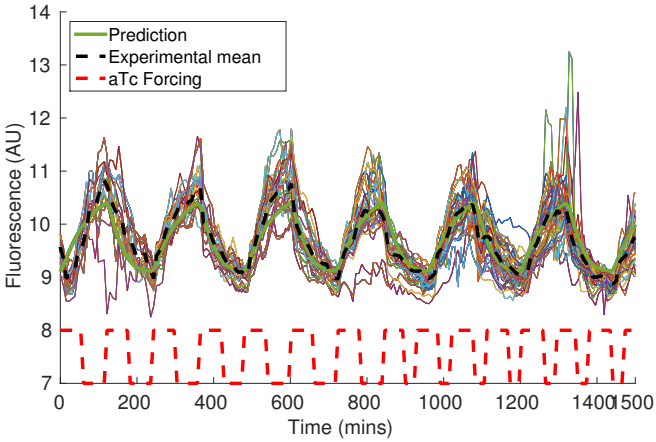
(c) Correlation matrix of estimated parameter sets, from the *13_9* dataset.



(d) Correlation matrix of estimated parameter sets, from the *14_7* dataset.

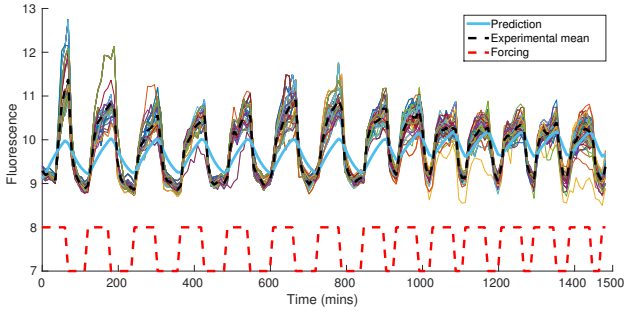


(e) Model prediction, using the parameter set with the smallest error value of the initial 200 found, for the *13_9* dataset. A

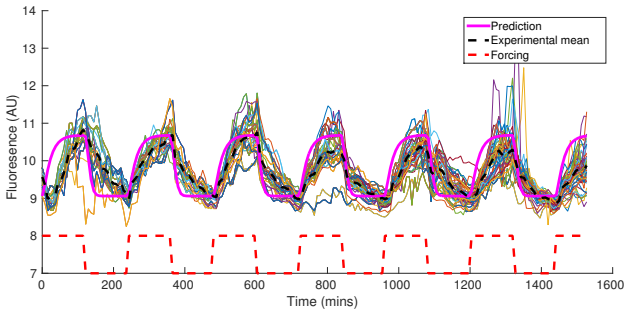


(f) Model prediction, using the parameter set with the smallest error value of the initial 200 found, for the *14_7* dataset.

Fig. 5: Parameter estimates, inter-parameter correlation values, and model predictions for the *13_9* and *14_7* datasets. Note that in the model predictions, aTc forcing is shown - IPTG concentration is constant at a level which saturates the cell's response. The forcing curve's height is schematic - aTc concentration is switched between off, and a level which saturates the cell's response

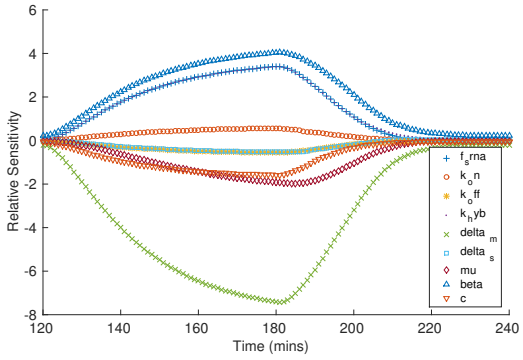


(a) *14_7* prediction *13_9* data

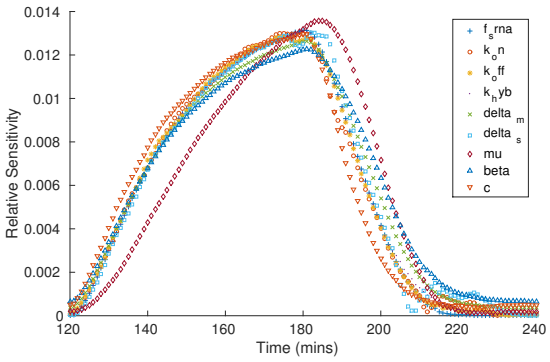


(b) *13_9* prediction *14_7* data

Fig. 6: Cross validating data by taking parameter estimates from one dataset, and using them to predict another. Here we take the best parameter set fitted on the *13_9*, and fit it against the *14_7* dataset, and vice versa



(a) Unscaled



(b) Scaled

Fig. 7: Sensitivity co-efficients S_{ij} evaluated about a set of estimated parameters from the *13_9* dataset. 7a shows them unscaled, 7b shows them scaled by the norm of each column of S.

and a slower timescale, in eqs. 5, 6, in which measured fluorescence changes in response to the forcing. If this is the case, then it may be that our experimental data can only probe the system via the fixed point of eqs. 1 - 4 , and that the parameters contained within those equations can only work to set the particular fixed point values of p and g that the system tends toward. If this were the case, the particular shape of the models response curve would be primarily determined by the parameters appearing in eqs. 5, 6, 7 - namely μ, β, f_s and Θ. If all the information contained in the parameters in eqs.1 - 4 can only be expressed in the form of the models fixed point, it would not be surprising if there were large spaces of parameters, all giving the same model fixed point, which were indistinguishable.

Fig. 8 shows model output for all state variables, using the parameter values giving the lowest error on the *13_9* dataset, and normalised to lie on the same scale.

It demonstrates that, at least in some of the parameter sets initially estimated, this difference in time scales exists. The fixed points of eqs. 1 - 4 affects eqs. 5, 6 through the $\beta m + f_s \beta c$ term - we would expect that, if the fixed point value of this term is what is relevant in determining model response, there would be some consistency in its value across the parameter sets found. Fig. 9 shows a scatterplot of model fixed point values against error function value, and demonstrates that, though individual parameter values can be spread across very large ranges, they are correlated in such a way as to give similar model fixed points.

In Fig. 10, we also see a strong positive correlation between the fixed point values and the values of μ ($R = 0.9854$ for the visually tightly clustered data). Taken together with the relatively tightly constrained values of μ found in Fig. 5, these results suggest that in order to minimize error, the algorithm is effectively trying to vary $\beta m + f_s \beta c$ and μ, and finding a 'trench' of highly correlated values, which is shown in Fig. 10.

These results further suggest that, while we are trying to minimise model error over our initial high dimensional space, we are effectively working in a lower dimensional space, in which one dimension is the value of the models $\beta m + f_s \beta c$.

- explore degeneracies in fixed point equations
- discuss estimatbility criteria- uniqeuess and sensitivity
- put in some sensitivity analysis results
- discuss model fixed points
- discuss systematic paramater shift.

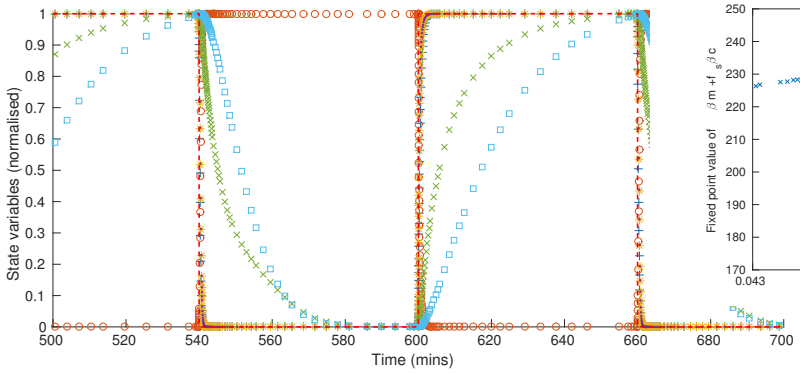


Fig. 8

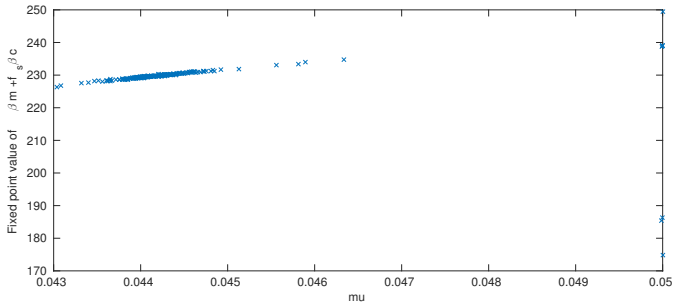


Fig. 10

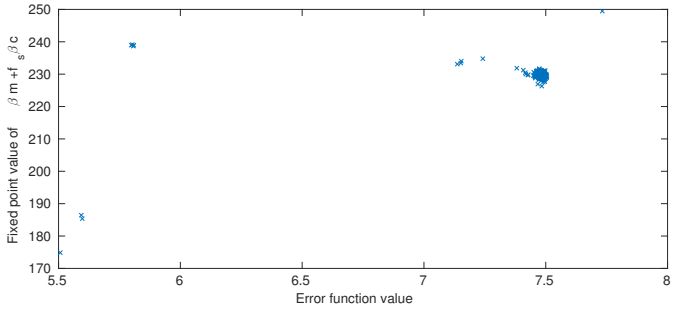


Fig. 9

- discuss cross - validation and predicitive ability of parameter sets found.
- suggest two times, work out fixed point for all the solutions found, look at structure of soln,
- re run from starting pararms.

IV. Conclusions and Further Work

- discuss evidence for model being very over parametrized, and suggest a simpler model may work equally well - suggest biologically interesting parameters may be unobservable.
- discuss general methodological flaws - throwing marbles in the air, local minima, is MLE even any good? say data shows even the EA doesn't avoid local minima. The space is so unconstrained. the algorithm can't be expected to just fix all these problems.
- suggest additional forcing time data may not help due to two timing nature,
- suggest there MAY be other ways to fit the data, but the wide range of parameter bounds mean we just don't know - the landscape is complicated and huge, and there may be minima widely separated in the parameter space. whose to say which is right?
- Further work - on experimnetal side, propoer lit review to rightly constrain some values, fix more. More state variables! should easily be able to fix mu, theta, maybe others too? other forcing functions?
- a proper bloody methodology - gaussian processes.
- mention this is how sys bio papers do it, and we were given some earlier fitting work which we think is now crap.
- suggest a general structural analysis be carried out.

some stuff

References

[1] F. J. Isaacs, D. J. Dwyer, and J. J. Collins, "RNA synthetic biology." *Nature biotechnology*, vol. 24, no. 5, pp. 545–554, 2006.

[2] G. Rodrigo, T. E. Landrain, S. Shen, and A. Jaramillo, "A new frontier in synthetic biology: Automated design of small RNA devices in bacteria," pp. 529–536, 2013.

[3] G. Rodrigo, T. E. Landrain, and A. Jaramillo, "De novo automated design of small RNA circuits for engineering synthetic riboregulation in living cells," *Proceedings of the National Academy of Sciences*, vol. 109, no. 38, pp. 15 271–15 276, 2012.

[4] T. Soper, P. Mandin, N. Majdalani, S. Gottesman, and S. a. Woodson, "Positive regulation by small RNAs and the role of Hfq." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 21, pp. 9602–9607, 2010.

[5] J. Shine and L. Dalgarno, "Identical 3'-terminal octanucleotide sequence in 18S ribosomal ribonucleic acid from different eukaryotes. A proposed role for this sequence in the recognition of terminator codons." *The Biochemical journal*, vol. 141, no. 3, pp. 609–615, 1974.

[6] R. Lutz and H. Bujard, "Independent and tight regulation of transcriptional units in escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements," *Nucleic Acids Research*, vol. 25, no. 6, pp. 1203–1210, 1997.

[7] A. Jaramillo, "Predicitve Modelling of Riboregulatory Circuits to Re-engineer Living Cells." [Online]. Available: http://www2.warwick.ac.uk/fac/sci/wcpm/seminars/wcpm_seminar_presentation_alfonso_jaramillo.pdf

[8] Uri Alon, *An Introduction to Systems Biology*, 1st ed.

[9] D. Brewer, M. Barenco, R. Callard, M. Hubank, and J. Stark, "Fitting ordinary differential equations to short time course data." *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 366, no. 1865, pp. 519–544, 2008.

[10] E. Algorithms, E. Algorithms, C. G. Moles, C. G. Moles, P. Mendes, P. Mendes, J. R. Banga, and J. R. Banga, "Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods," *Genome Research*, pp. 2467–2474, 2003.

[11] C. Y. Hu, J. Varner, and J. B. Lucks, "Generating effective models and parameters for RNA genetic circuits," *ACS Synthetic Biology*, p. 150605124221004, 2015. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/acssynbio.5b00077

[12] N. Hansen, "The CMA evolution strategy: A comparing review," *Studies in Fuzziness and Soft Computing*, vol. 192, no. 2006, pp. 75–102, 2006.

[13] ——, "The CMA evolution strategy: A tutorial," *Vu le*, pp. 1–34, 2011. [Online]. Available: http://www.lri.fr/~hansen/cmatutorial110628.pdf

[14] K. a. P. Mclean and K. B. Mcauley, "Mathematical modelling of chemical processes-obtaining the best model predictions and parameter estimates using identifiability and estimability procedures," *Canadian Journal of Chemical Engineering*, vol. 90, no. 2, pp. 351–366, 2012.

[15] K. Z. Yao, B. M. Shaw, B. Kou, K. B. McAuley, and D. W. Bacon, "Modeling Ethylene/Butene Copolymerization with Multisite Catalysts: Parameter Estimability and Experimental Design," *Polymer Reaction Engineering*, vol. 11, no. 3, pp. 563–588, 2003.

[16] J. Beck, *Parameter Estimation in Engineering and Science*, 1st ed. Wiley, 1977.

[17] J. E. Jiménez-Hornero, I. M. Santos-Dueñas, and I. García-García, "Structural identifiability of a model for the acetic acid fermentation process," *Mathematical Biosciences*, vol. 216, no. 2, pp. 154–162, 2008.

[18] M. Grewal and K. Glover, "Identifiability of linear and nonlinear dynamical systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 6, pp. 833–837, 1976.

[19] J. B. Andersen, C. Sternberg, L. K. Poulsen, S. P. Bjø rn, M. Givskov, and S. r. Molin, "New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria," *Applied and Environmental Microbiology*, vol. 64, no. 6, pp. 2240–2246, 1998.

[20] R. Iizuka, M. Yamagishi-Shirasaki, and T. Funatsu, "Kinetic study of de novo chromophore maturation of fluorescent proteins," *Analytical Biochemistry*, vol. 414, no. 2, pp. 173–178, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.ab.2011.03.036

[21] G. L. Hersch, T. a. Baker, and R. T. Sauer, "SspB delivery of substrates for ClpXP proteolysis probed by the design of improved degradation tags." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 33, pp. 12 136–12 141, 2004.

APPENDIX A
INITIAL EXPERIMENTAL DATA

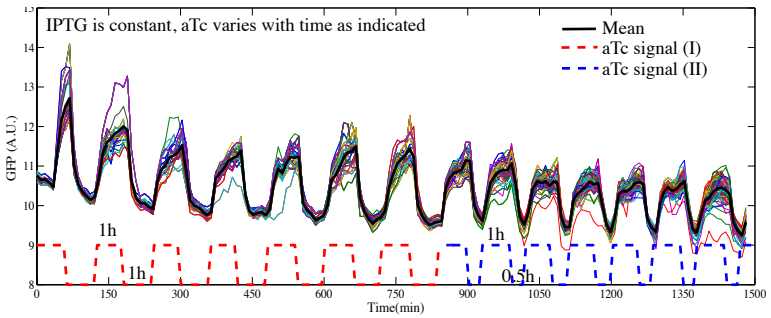


Fig. A.1: The *13_9* dataset, with aTc forcing shown. IPTG concentration is constant at a level which saturates the cell's response. Note the forcing curve's height is schematic - aTc concentation is switched between off, and a level which saturates the cell's response

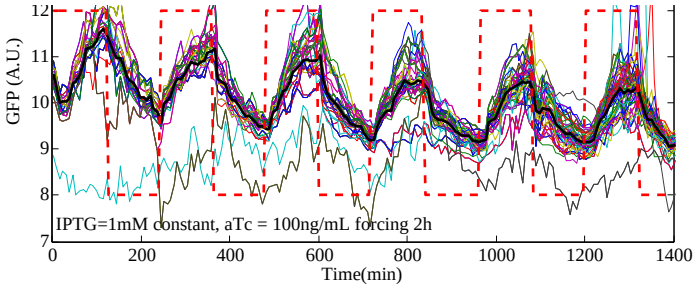


Fig. A.2: The *14_7* dataset, with aTc forcing shown. IPTG concentration is constant at a level which saturates the cell's response. Note the forcing curve's height is schematic - aTc concentation is switched between off, and a level which saturates the cell's response

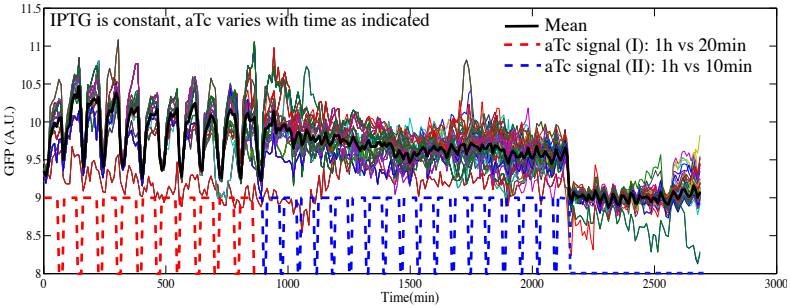
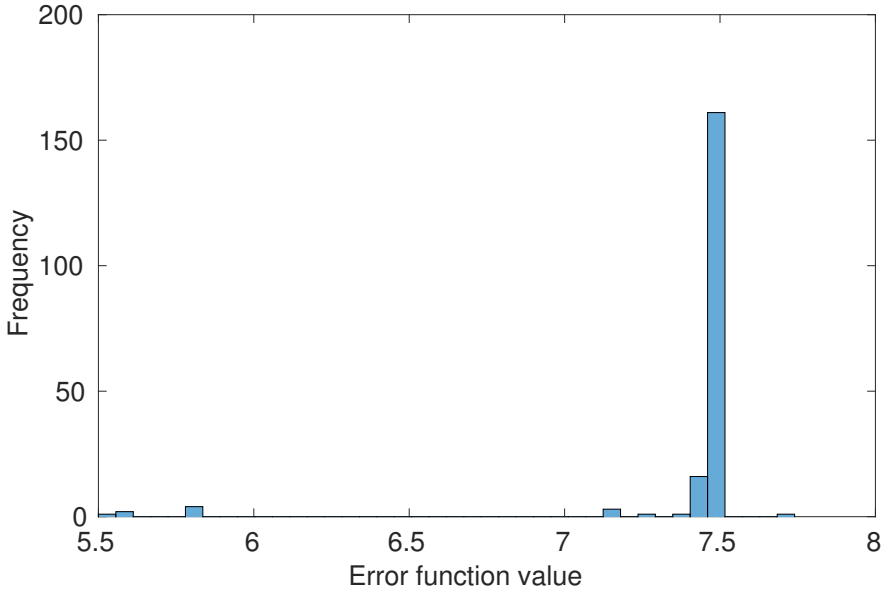
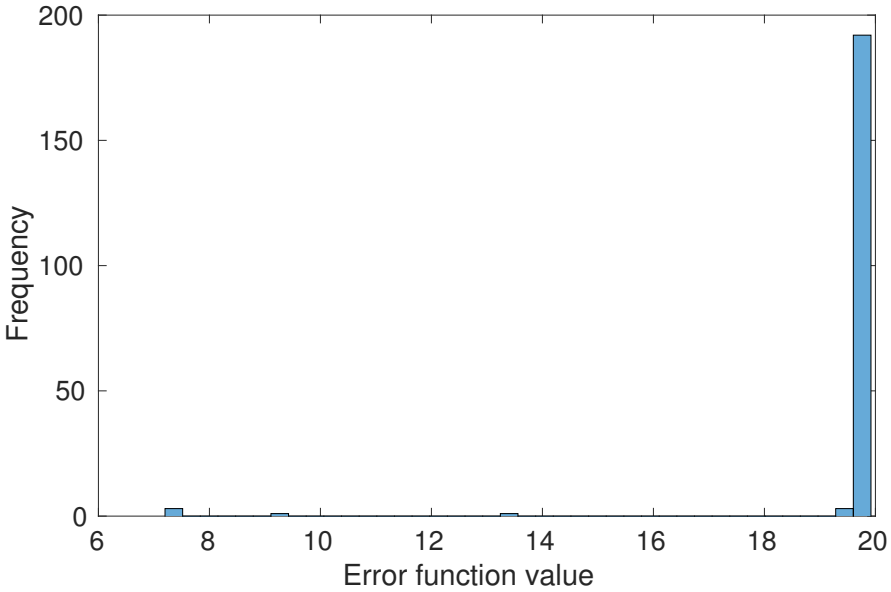


Fig. A.3: The *14_9* dataset, with aTc forcing shown. IPTG concentration is constant at a level which saturates the cell's response. Note the forcing curve's height is schematic - aTc concentation is switched between off, and a level which saturates the cell's response

APPENDIX B
HISTOGRAM OF INITIAL FIT ERROR VALUES



(a) Error values found from 200 initial parameter estimates, *13_9 dataset*.



(b) Error values found from 200 initial parameter estimates, *14_7 dataset*

Fig. B.1

APPENDIX C
PARAMETER LITERATURE REVIEW

TABLE III: Literature references, or initial rough bounds, on parameter values, with those to be estimated shown in bold

Parameter	**Value**	**Definition**	**Reference**	**Initial Bounds**
N	300	Number of copies of promoter existing on plasmid DNA	Experimentally set	
z_0	9 AFU	Baseline experimental fluorescence	Experimentally determined	
α_L	11 nM/min	Maximal transcription rate of $P_{\mathrm{LlacO-1}}$ promoter	[6]	
α_T	11 nM/min	Maximal transcription rate of $P_{\mathrm{LtetO-1}}$ promoter	[6]	
f_L	620	Unitless ratio between repressed and unrepressed $P_{\mathrm{LlacO-1}}$ transcription rate	[6]	
f_T	2535	Unitless ratio between repressed and unrepressed $P_{\mathrm{LtetO-1}}$ transcription rate	[6]	
δ_g	0.0005 /min	GFP degradation rate	[19]	
γ	0.132 /min	GFP maturation rate	[20]	
v_z	100 nM/min	degradation constant of clpx	[21]	
K_z	75 nM/min	Dissociation constant of clpx	[21]	
$\boldsymbol{\Theta}$	nM/AFU	Ratio between GFP concentration and observed fluorescence		300 - 1000
$\boldsymbol{\mu}$	/min	Dilution rate		0.001-0.05
$\boldsymbol{\delta_m}$	/min	mRNA degredation rate		1 - 10^5
$\boldsymbol{\delta_s}$	/min	sRNA degredation rate		1 - 10^3
δ_{sm}	/min	Unstable sRNA:mRNA degradation rate		Set to δ_m
δ_c	/min	Stable sRNA:mRNA degradation rate		Set to δ_m
$\boldsymbol{k_{on}}$	/min	sRNA:mRNA binding rate		100 - 10^7
$\boldsymbol{k_{off}}$	/min	sRNA:mRNA unbinding rate		1 -10^8
$\boldsymbol{k_{hyb}}$	/min	sRNA:mRNA hybridization rate		1 - 10^4
$\boldsymbol{\beta}$	/min	Baseline translation rate of repressed mRNA		0.0001 - 10
$\boldsymbol{f_s}$		Ratio of repressed mRNA to unrepressed complex translation rate.		0.1 - 10^4