Comprehensive biophysical model of the *lac* operon

Kate Patterson
Department of Mathematics,
The University of Auckland,
Auckland, New Zealand

Konstantin Mischaikow
Department of Mathematics and BioMaPS Institute,
Rutgers, The State University of New Jersey,
Piscataway, NJ, USA

Tomas Gedeon[1]
Department of Mathematics and Center for Computational Biology,
Montana State University,
Bozeman, MT, USA

[1] Department of Mathematics and Center for Computational Biology, Montana State University, Bozeman, MT 59715, USA

Abstract

We introduce and analyze a detailed biophysical model of the control of the *lac* operon. The model connects two spatial scales: the biophysical parameters of molecular interactions at a molecular scale and the resulting expression level of *lac* genes measured on the cellular level. We parameterize the model to the extent possible and find parameter bounds for 19 other parameters. We optimize a least square fit between the model predicted and experimentally determined repression values. We find that the standard model, based on the four basic mechanisms involved in the *lac* operon regulation, is not able to match the repression data. However, a larger model which includes an additional six biologically suggested mechanisms fits data well. We find several local minima which provide a very good fit to data, but require that the bending energy of a short DNA loop is smaller than predicted from tethered particle experiments. We conclude with a study of robustness of our fit in the parameter space.

Key words: gene regulation; lac operon; statistical mechanics; DNA looping

Introduction

The primary goal of this paper is to create a mathematical model of the control mechanism of the *lac* operon. We are faced with three challenges that are typical of many modeling efforts in the life sciences: (1) we need an understanding of the underlying biochemistry and biophysics, (2) we need to determine relevant parameter values, and (3) we need to be able to compare the predictions of the model against appropriate experimental results. In most situations none of the challenges is completely overcome. Nevertheless, and this is one of the consequences of this work, building a model that is as coherent as possible based on partial results can shed insight into which of the three challenges is most pressing for furthering our understanding.

The *lac* operon in *E. coli* controls the expression of proteins β-galactosidase, permease and transacetylase, which are responsible for lactose uptake and digestion. It is one of the most extensively studied gene regulatory systems, both experimentally and theoretically, and thus there is a relatively good understanding of the main biochemical and biophysical components of the *lac* operon control. Hence with regard to challenges (1) and (3) this an ideal system to study. The *lac* genes are expressed only when the *lac* repressor is not bound to its primary site, O1; its ability to bind DNA is impaired when an inducer, which signals presence of lactose or a lactose analog, binds the repressor. The up-regulation of the *lac* expression is achieved by cooperative binding of the CAP protein and RNA polymerase (RNAP). Another well known cooperative feature is that a repressor bound to O1 can simultaneously bind an additional site (either O2 or O3), which causes looping of the DNA and more durable repression. In addition to these well understood control parts, there are other control features, discussed more extensively in the Model section, which have been suggested to contribute to the function of the *lac* operon. This information allows us to construct a comprehensive mathematical model. Furthermore, Oehler et al. (1) have experimentally obtained repression data for various knockouts and permuted binding sites and thus we have an extensive experimental data set against which we can test our model.

With regard to challenge (2) many of the parameters of interest have not been experimentally measured. To address this issue we adopt, in the Parameter selection and optimization section, the following strategy. We begin by noting that there are a handful of parameters which can be computed directly from the existing experimental literature. For the rest of the parameters we begin by determining reasonable bounds. We then perform a constrained optimization to fit these remaining parameters by minimizing the distance between the repression levels predicted by the model and the repression data of (1).

At this point we are in a position to interrogate the model as is done in the Results. In particular we show that the comprehensive model fits the data better than a smaller standard model, which takes into account only the main *lac* control mechanisms mentioned above: binding of the repressor and its looping, impairment of this binding by the inducer and CAP induced up-regulation. The optimization procedure finds several local optima which produce a very similar fit. The question of how we interpret and understand such results is closely related to the question of robustness of model predictions to change in parameters. We view all of these optima as informative solutions and try to find common features that all of them share. The majority of parameter sets that differ from the local optima in each component by one or two standard deviation still produce good fits, suggesting robustness of the fit in the parameter space. However, the parameters cannot be arbitrary. We find that all parameter sets at local minima share the following property: the looping energies for at least one of the short DNA loops are predicted to be lower than the values reported from *in vitro* experiments.

Model

We begin by reviewing the main control mechanisms for *lac* operon. The segment of DNA containing both the *lac* repressor coding region and the *lac* operator is indicated in Fig. 1. Starting from the left are the RNAP binding site, P_i, followed by the gene lacI, which codes for *lac* repressor. The *lac* repressor is constitutively expressed. Next is the *lac* promoter region, P_{lac}, which is immediately followed by the *lac* genes lacZ, lacY and lacA. It is the combinatorial control of the *lac* genes expression at the P_{lac} region that is a subject of this paper.

Figure 1 about here

We highlight the following biochemical processes that we explicitly incorporate into our mathematical model. Note that the first four are related to the standard model for *lac* operon (2).

1. The *lac* genes are expressed when the O1 operator is not bound by a repressor and RNAP binds the promoter.

2. The *lac* repressor's ability to bind the DNA is impaired when an inducer such as allolactose (but also allolactose analogs IPTG or TMG) binds the repressor.

3. A protein complex called catabolite activator protein (CAP) up-regulates gene expression when it binds to C1 on the promoter region of the DNA.

4. The *lac* repressor binds any of the operator sites, O1, O2, or O3, with decreasing affinity. Furthermore, being a tetramer with two DNA binding domains, a single repressor can bind any two operators simultaneously, looping the DNA.

5. An additional RNAP binding site P2, weaker than P1, leads to transcription as long as the O1 operator is free from repressor (3).

6. It has been suggested that when CAP is bound to C1, the repressor may bind more favorably to an operator O3* than to O3 (4, 5), where the O3* site is overlapping the O3 operator in all but five DNA base pairs.

7. CAP assisted looping is a cooperative mechanism which reduces the energy of forming a DNA loop between the O3 or O3* operator and another operator when CAP is bound to C1. For example, the energy for looping O1 to O3 is lower when CAP is bound to C1 (5–7).

8. C1-O3 steric interference is a negative physical interaction between a CAP bound at C1 and a repressor bound at O3 (4, 5, 8).

9. It is not known if an inducer-impaired binding domain of a repressor is unable to bind the DNA, or whether this ability is only impaired. In our model we allow such binding, although the affinity is reduced in comparison to an unimpaired binding domain.

10. Finally, since in the experiments of Oehler et al. (1, 9) (described in detail below) the deleted binding sites were not physically removed, but only mutated, we also allow a repressor to bind these deleted sites, albeit at a reduced affinity.

A more detailed description of these processes is presented in the Supplement, Section S.1.

The standard model for gene regulation focuses on the combinatorial binding of the regulatory proteins to the DNA (10). The binding and unbinding of the regulatory factors is governed by a stochastic dynamical process which evolves on a much shorter time scale than the changes in the concentration of these factors. Using this difference in time scale we assume that the concentrations are fixed and that the binding and unbinding reactions are at statistical equilibrium. We assume a finite collection $\mathcal{S}$ of *states* and the model makes use of the equilibrium probability of occurrence of each state. This approach is a broadly accepted quantitative framework for modeling transcription (11) and has been experimentally validated on a variety of genes (12–16). A state s of the operon is a particular configuration of transcription factors (or their absence) on the DNA, that is admissible in view of possibly overlapping binding sites. The probability of the occurrence of a particular state s from the set of admissible states $\mathcal{S}$ is

$$\mathbb{P}_s := \frac{K_s[RNAP]^{\alpha_s}[CAP]^{\alpha_s^1}[R]^{\alpha_s^2}[R']^{\alpha_s^3}}{Z} \tag{1}$$

where $K_s = \exp(-E(s)/RT)$ is the equilibrium constant and Z is the partition function

$$Z = \sum_{s\in\mathcal{S}} K_s[RNAP]^{\alpha_s}[CAP]^{\alpha_s^1}[R]^{\alpha_s^2}[R']^{\alpha_s^3}. \tag{2}$$

In these formulas $E(s)$ denotes the change in Gibbs free energy between the empty state $s_\emptyset$, where nothing is bound to the DNA, and the state $s \in \mathcal{S}$ under the normalization $E(s_\emptyset) = 0$. The exponents $\alpha_s,\ \alpha_s^1, \alpha_s^2$, and α_s^3 represent the number of bound molecules of RNAP, CAP, the unimpaired *lac* repressor R, and the repressor with one free binding domain R', in state s, respectively. As is standard, RT denotes the universal gas constant times the temperature (17). Note that the universal gas constant will only appear as a part of a product with the temperature T, so all other occurrences of R in this paper refer to the repressor. Observe that not all combinatorial configurations are in $\mathcal{S}$. When, for example, RNAP binds the P1 region, it also overlaps a portion of the P2 region, and vice versa. Therefore two RNAP molecules cannot simultaneously bind P1 and P2 and therefore $\mathcal{S}$ does not contain a state with both promoter regions occupied by their own RNAP molecule.

Since the only states compatible with transcription are those in which the promoter is bound by RNAP, we assign to each RNAP-bound state s a rate of a transcription k_s from that state. Then the overall rate of transcription for the *lac* promotor is given by

$$f := \sum_s k_s\mathbb{P}_s = \frac{1}{Z}\sum_s k_sK_s[RNAP]^{\alpha_s}[CAP]^{\alpha_s^1}[R]^{\alpha_s^2}[R']^{\alpha_s^3}. \tag{3}$$

To evaluate this expression we need to know the set of states $\mathcal{S}$, and for each individual state s the energy K_s and the rate k_s, along with the concentrations of RNAP, CAP, R and R'. This is discussed in greater detail in the next section.

We do not know of any direct experimental measurements of transcription rates of the *lac* promotor and thus we test our model against data obtained by Oehler et al. (1) (Supplemental

Table S.6). Their work reports the value of *repression* for a series of mutants of *lac* operon. Repression is the ratio of β-galactosidase activity recorded for cells exposed to 1 mM concentration of inducer IPTG over the activity at 0 mM IPTG,

$$\mathcal{R}_p = \frac{\beta\text{-galactosidase}(1\text{mM})}{\beta\text{-galactosidase}(0\text{mM})}. \tag{4}$$

IPTG binds to the repressor R which affects the concentrations $[R]$ and $[R']$ and thus the rate of transcription f can viewed as a function of IPTG. In order to compute the repression in our model we assume that β-galactosidase activity is proportional to the β-galactosidase protein concentration in the cell, and that neither the translation from mRNA to protein, nor the degradation of protein or mRNA are affected by IPTG. Then β-galactosidase activity is proportional to the transcription of mRNA, thus

$$\mathcal{R}_p = \frac{\beta\text{-galactosidase}(1\text{mM})}{\beta\text{-galactosidase}(0\text{mM})} = \frac{f(1\text{mM})}{f(0\text{mM})}. \tag{5}$$

We use this equation to model *lac* operon repression.

Parameter selection and optimization

In order to evaluate expression (3), which is necessary for the repression calculation, equation (5), we must determine equilibrium constants K_s and transcription rates k_s for all equilibrium states s along with the concentrations of RNAP, CAP, R and R'. The determination of these values represents a significant portion of the work reported in this paper.

Our model has 648 states each of which can have a different equilibrium constant K_s and thus different free energy $E(s)$. However, these energies can be computed from free energies associated to a much smaller set of *elementary states*, corresponding to the binding a single element to its binding site, as well as cooperative energies between elementary states. In this approach, the free energy of a state $E(s_{ab})$ in which both elements A and B are bound to the DNA at the same time need not be the sum of the free energies of binding protein A and B separately. This energy difference is referred to as *cooperativity* and defined by $\Delta G_{AB} := E(s_{ab}) - (\Delta G_A + \Delta G_B)$, where $\Delta G_A = E(s_a)$ and $\Delta G_B = E(s_b)$ are free energies of binding protein A and B individually to the DNA. Cooperativity can be positive or negative. In the *lac* operon, the cooperativity is known to exist only between pairs of elementary states, even though in theory there can be higher order cooperative effects between more than two states.

In Table 1 we list the parameter values that we are able to determine from the existing literature. K_{P1} and K_{P2} are the binding constants of RNAP binding to their binding sites P1 and P2 respectively, while [CAP] and [RNAP] denote concentrations of CAP and RNAP in a typical *E. coli*. The constant K_{P1C1} captures the cooperative lowering of the binding energy when CAP is bound to C1 and RNAP is bound to P1. RT is the universal gas constant time temperature. The different values of the last two constants, k_f and k_{fC1} reflect the fact that the cooperativity between CAP bound to C1 and RNAP to P1 not only affects the binding energy, but also increases the transcription initiation rate. Therefore we use the transcription rate k_{fC1} for all states that include CAP bound to C1 and RNAP to P1, and the rate k_f for all other states. The computation of these values as well as our sources can be found in the Supplement, Section S.2.

Table 1 about here.

The remaining parameter values presented in Table 2, are computed by our optimization procedure detailed below. Table 2.A provides information about the free energy associated with elementary states, Table 2.B about the looping energies (explained below), and Table 2.C about the cooperativity effects.

To initiate the optimization procedure we must first determine bounds on all parameters in Table 2. To obtain bounds on free energies associated with the elementary states (Table 2.A) we used the work of Horton et al. (18) which reports the contribution of each base pair in the O1 operator to the overall binding energy between the O1 operator and a repressor. We use the sum of the reported mean contributions of each base pair as our estimate of *specific* binding energies ΔG_{*}^{spec} for all operators, based on their individual DNA sequences. Summing the error bars reported in Horton et al. (18) across all base pairs we arrive at upper and lower bounds of overall specific binding energies. As an example, our bound on ΔG_{O1} has the form $\Delta G_{O1} := \Delta G_{ns} + \Delta G_{O1}^{spec}$ where

$$\Delta G_{O1}^{spec} \in [-3.615 - 4.125, -3.615 + 2.125] = [-7.740, -1.490] \text{ kcal/mol.}$$

In addition, we take into account a non-specific, basal binding energy, ΔG_{ns}, between a repressor and the DNA. This energy does not depend on the DNA sequence of a particular binding site. Total binding energy between the repressor and a particular operator is then a sum of ΔG_{ns} and the specific energy of that operator. We selected the bounds for ΔG_{ns} to be $[-7, -13.1]$ kcal/mol, a range taken around the -9.7 kcal/mol cited in Horton et al. (18). We optimize separately the specific binding energies and the basal binding energy ΔG_{ns}.

The DNA loops form when a repressor is simultaneously bound to two operators. The energy corresponding to the equilibrium constant K_s (in equation (3)) for this state will have an additional looping energy ΔG_{Oij}, where O_{ij} represents the loop between operator i and operator j and $i, j \in 1, 2, 3, 3^{*}$. The looping energy thus enters the equation in the same way as the cooperative energy described above. Since the formation of the loops requires energy, we set the lower bound on all looping energies, listed in Table 2.B, to zero. The loops formed by binding simultaneously to O1 and O3 and O1 and O3* are much shorter than the loops formed between the O1, O2 and the O2, O3 sites. Since longer loops are easier to form, the upper bound for looping energy for longer loops should be lower than those for shorter loops. The highest reported energy for a short loop (80-250 base pairs (19)) is 17 kcal/mol, see Table 3. Therefore we set the upper bound on energy for the short loops (ΔG_{O13} and ΔG_{O13*}) to 17 kcal/mol. On the other hand, there are 401 base pairs in the O1-O2 loop, and the length between O2 and O3 is even longer. Han et al. (20) measure the looping energy associated with loops of length 300-310 base pairs and find the energy to be 10-11.5 kcal/mol. We therefore we set the upper bound on the energy for the longer O1-O2 and O2-O3 loops to 13 kcal/mol.

Finally, we discuss bounds on cooperativity energies listed in Table 2.C. The energy ΔG_{C1loop} represents looping assistance by CAP bound to C1 (6) which helps formation of O1-O3, O2-O3 and O2-O3* loops and therefore is negative. We bound ΔG_{C1loop} between zero and -7 kcal/mol, a range taken around the value -3.1 kcal/mol suggested by (5). On the other hand ΔG_{C1O3} represents steric interference between CAP and repressor bound to O3 and is therefore positive, bounded between zero and 10. The last entry in Table 2.C is the reduction in binding energy between the repressor and any of operators when the repressor is bound by the inducer. ΔG^{I} was bounded between 5 and 11 kcal/mol with the constraint that $\Delta G_{ns} + \Delta G^{I} < 3$. These bounds are discussed further in the Supplement Section S.2.8.2. Table S.7 contains exact bounds.

Table 2 about here.

Having obtained upper and lower bounds for the parameters in Table 2, we search within this bounded 19 dimensional hypercube B for the parameter values which produce the best fit to the data presented in (1). Oehler et al. (1) construct a variety of *lac* operons by mutating the repressor binding sites O1, O2, O3, or interchanging their positions on the DNA. For these modified *lac* operons, they report repression in the presence of the repressor at approximately 5 times and 90 times that of the wild type concentration (for the reported values see Table S.6 in the Supplemental Material). Since the repression values range from 1.3 to 8100 and we wish to weigh a two-fold change of the first value (say, from 1.3 to 2.6) in the same way than a two-fold change of the second value (from 8100 to 16200), we evaluate the least square fit in the log space

$$\xi = \sum_{m \in \mathcal{M}} \sqrt{\left(\log(\mathcal{R}_{D_1}(m)) - \log(\mathcal{R}_{P_1}(m))\right)^2 + \left(\log(\mathcal{R}_{D_2}(m)) - \log(\mathcal{R}_{P_2}(m))\right)^2}, \qquad (6)$$

where $\mathcal{M}$ is the set of all mutants; $\mathcal{R}_{D_1}(m)$ and $\mathcal{R}_{D_2}(m)$ are the repression values from Oehler et al. (1) for mutant m at $5\times$WT and $90\times$WT repressor concentrations respectively and $\mathcal{R}_{P_1}(m)$ and $\mathcal{R}_{P_2}(m)$ are the model predicted repression values at the same concentrations (1). We do not include the Oehler et al. (1) measurements for which there is only a lower bound measurements in the cost function, but these inequalities are satisfied in all our optimized models.

Denoting the 19 parameter values in Table 2 by the vector $\vec{x}$ we use the MATLAB function `fmincon` with the `active-set` algorithm to perform a constrained minimization of ξ over $\vec{x}$. In an attempt to explore the possible multiple local minima of $\vec{x}$ we initialize the minimization at 20 different initial parameter vectors. In this set we include the vector containing the lowest admissible value in each entry of $\vec{x}$ as well as the vector with the highest allowable value in each entry of $\vec{x}$. These represent two of the corners of B. We generate an additional eight random vectors on the boundary of B, where each entry of $\vec{x}$ is either an upper or a lower bound for the particular entry, and 10 random points from anywhere in the interior of B. The minimization is stopped when the magnitude of the directional derivative in the search direction is less than 2×10^{-6} and the value of $\vec{x}$ is not more than 1×10^{-6} beyond the constraint. We discuss this optimization step further in the Results.

Results

As indicated previously, we test our model against repression data for a series of *lac* operon mutants obtained by Oehler et al. (1). They constructed *lac* operon mutants by mutating the repressor binding sites O1, O2, O3, or interchanging their positions on the DNA, and measured repression in the presence of the repressor at approximately 5 times and 90 times that of the wild type concentration (for the reported values see Table S.6 in the Supplemental Material). In order to show the fit of our model to their data, we represent each mutant of the *lac* operon as a 3-tuple. The first slot represents the *position* on the DNA of the wild type (WT) O1 operator, the second slot represents the position on the DNA of the WT O2 operator, and the third slot represents the position on the DNA of the WT O3 operator. The numbers of the 3-tuple represent the DNA *sequence* present in that slot: one for O1, two for O2 and three for O3. Using this notation, (1,2,3) is the WT *lac* operon, and (2,2,0) is a mutant *lac* operon with the DNA sequence for O2 in the position of the

WT O1 operator as well as in the position of the WT O2 operator, while the zero indicates that the O3 operator has been 'deleted' (i.e mutated). The DNA sequences for the WT and deleted binding sites can be found in the Supplemental Material.

In Fig. 2(a) we show predicted repression level curves as a function of the repressor concentration for each mutant along with the repression data from Oehler et al. (1). The repression curves were obtained from our model based on the biochemical processes 1-10, using the optimized parameters shown in Table 2.

Figures 2a and 2b about here.

Observe that the only significant differences between our model and the reported data involves the mutant $(1, 0, 1)$ at both repressor levels and the mutants $(1, 0, 3)$ and $(1, 2, 3)$ at 90 times the WT repressor levels. However, these are precisely the data points points at which (1) report that their experimental procedure could only determine a lower bound on the actual repression value. Thus it is possible that our model predicts the correct repression value.

As is indicated in the prior section, we performed an optimization procedure on 20 initial conditions for the parameter vector $\vec{x}$. Eleven initial conditions lead to local minima with the values $0.51 \leq \xi \leq 0.56$, three of which are within 0.0005 of the minimal value $\xi = 0.5115$. The remaining nine initial values lead to a minima with values of $1.28 \leq \xi \leq 4.6$. These results suggest that the $\vec{x}$ landscape is relatively flat and the best solution depends only weakly on the initial condition of the minimization. We return to this issue below. The best model with the value $\xi = 0.5115$ is presented in Fig. 2a where we plot the repressor concentration in molar along the x-axis and the repression value along the y-axis. The minimizing values of the parameters are in Table 2.

A reasonable criticism of our strategy is that we used (1) to optimize the choice of parameter values and thus the strong agreement is due to over fitting of the data. To provide at least a partial rebuttal to this we considered a small model. As is indicated in Model section the standard description of *lac* operon focusses on the biochemical processes 1-4. Thus we repeated the process of optimizing the parameter values but only used the terms in equation (3) associated with 1-4. The results are presented in Fig. 2(b). Clearly the fit is much poorer. We conclude from this that the standard model is insufficient to explain the data.

Another reason for confidence in our model has to do with the internal consistency of the optimized parameter values. It has been observed experimentally (9) that repressor binds O1 about 10 times stronger than O2, and a repressor binds O1 about 300 times stronger than O3 which is in agreement with the values for ΔG_{O1}, ΔG_{O2}, and ΔG_{O3} reported in Table 2. Moreover, almost all of the free energy values are strictly within the error bound determined from the data, see Table S.7. The exceptions are the lower bound of O1 and O2 and the upper bound of O1^{-}a, O3^{*-} and O3* which we had to adjust so that the locally optimal point ended up in the interior of the search domain.

Figure 3 about here.

As a final internal test of our model we consider how robust our model is to small perturbations in the parameter values. Recall that our minimization procedure produced eleven vectors of parameter values for which $0.51 < \xi < 0.56$. We compute the standard deviation in each component of these vectors and then generate a set of 847 new vectors by changing one entry of each of the original eleven vectors by zero, one or two standard deviations. The fit ξ is computed for each of these vectors and the lowest 828 are presented in increasing order in Fig. 3 (the x-axis represents the vector identity (1-828) and the y-axis is the value of ξ.) The vectors not shown have a cost between $8 < \xi < 26$; we exclude these vectors to improve presentation of the data. Note that there is a sharp jump in the cost function at the value 1.31. This implies that there are vectors within one or two standard deviation of the optimal solutions which yield a bad fit, however, the vast majority (694 of 847 to be precise) have cost functions below this value. We present in the inset figure a solution that has a fit of the cost $\xi = 1.31$, which represents the worst of the good fit solutions. At first glance, the repression curves in Fig. 3 look like a reasonable fit, and the upper set of curves are. However, this set of curves results from decreasing the ΔG_{O13} looping parameter by two standard deviations, a change which is most noticeable in the (3,0,1), and (0,0,1) curves, each of which is well above the measured values. In summary, it appears that the optimal solutions do not lie in deep isolated wells of the cost function, and that the fit remains very good for a rather broad collection of parameters.

Based on the arguments presented above we now adopt the perspective that our full model successfully predicts the expression data based on the biophysics of the interactions and use it to investigate energies associated with DNA looping. Determining the energy required to loop the DNA between the O1 and O3 operators is an ongoing problem that is being addressed via a variety of approaches. Based on the assumptions that as the energy needed for loop formation decreases, the probability of loop formation increases, and an increase in loop formation correlates to an increase in repression, (21–27) measured (*in vivo*) how the distance between operators affect repression. Taking a different approach, (20, 28–30) performed tethered particle experiments (*in vitro*) where a segment of DNA containing two operators is tethered to a flat plane on one end and a bead is attached to the other. Finally, (5, 20, 30–33) have created and used mathematical models based on the structure of the DNA to predict the energy of looping. The results of these investigations for the short loop of length $\sim$ 90 base pairs, which is comparable to DNA loop between O1 and O3, are summarized in Table 3.

Table 3 about here.

In Fig. 4 and Table 4 we present looping energy predictions from our model at eleven best values of parameters, that correspond to $\xi \leq 0.56$. The first value, ΔG_{O13}, is the energy necessary to form the short DNA loop between O1 and O3 when the repressor binds these two operators. The second value ΔG_{O13*} represents the energy necessary to form the short DNA loop between O1 and O3*, where O3* is a shifted operator position for binding of the repressor in the presence of CAP bound to C1. While the predicted energies vary between the solutions, we can see that the typical values for ΔG_{O13} fall between 7.6-7.7 kcal/mol with the mean at 9.08 kcal/mol. While the values for ΔG_{O13*} have larger variance, both short looping values are frequently lower than the bounds predicted by tethered particle experiments and models (See Table 4).

Figure 4 and Table 4 about here.

The variability of the predicted energies ΔG_{O13} and ΔG_{O13*} brings up a question if these loop energies are sufficiently constrained by the repression data, which our model uses to constrain the parameters. We performed an optimization procedure leading to the values in Table 4 within very generous bounds - both ΔG_{O13} and ΔG_{O13*} were only constrained between 0 and 17 kcal/mol, while the other longer loops are constrained between 0 and 13 kcal/mol. To verify the robustness of our conclusion that the short loop energies ΔG_{O13} and ΔG_{O13*} are lower than previously reported, we optimize the model with progressively higher lower bound on short loop energies. If these energies are not constrained by our model, then we should not be able to find solutions that match data as well as our best solutions with $\xi = 0.5115$.

Indeed, when we constrain the looping energies ΔG_{O13} and ΔG_{O13*} between 11 to 17 kcal/mol, we still find a very good fit, since our best solution had $\xi = 0.6060$. However, there were only two minima with $\xi \leq 0.63$, while there were eleven such values with broadly constrained looping energy values. Furthermore, these optimization runs predict that either ΔG_{O13} or ΔG_{O13*} or both are at the lower bound of 11 kcal/mol in the optimization domain. This suggests that these energies would be decrease further if the lower bound was not enforced. When we constrain the looping energy even higher, with the lower bounds set to $\Delta G_{O13} = 16$ kcal/mol and $\Delta G_{O13*} = 14.7$ kcal/mol which are the lower bounds for these energies predicted by a recent looping model (5), the best fit from 20 initial conditions has value $\xi = 1.2778$. We show the repression curves and parameter values for the lowest minima associated with these two datasets in the supplement, Fig. S.1 and Tables S.10 and S.11.

These simulations confirm that the repression data do constrain the looping energies ΔG_{O13} and ΔG_{O13*}. This increases our confidence in the conclusion that both ΔG_{O13} and ΔG_{O13*} are in the range $6 - 13$ kcal/mol which agrees with the lowest estimate in Table 3 by Czapla et al. (34). Their DNA looping model is unique in that it includes the nucleoid protein HU. HU, found in *E. coli*, binds the DNA non-specifically, causing a sharp bend in the DNA. By incorporating HU into their model, the energy required to form short loop DNA configurations is lowered by the presence of the kink resulting from bound HU. Our results indirectly support conclusions from their model.

Discussion

We have developed a comprehensive biophysical model of *lac* operon based on all the accepted, as well as some suggested mechanisms of control of *lac* expression. We have parameterized the model by experimental biophysical parameters to the extent possible and found experimental bounds on the remaining parameters. This results in a compact space of potential parameter values in which we optimize the fit to the existing repression data. We find that even after optimization a standard model, which takes into account binding of the repressor and its looping, impairment of this binding by the inducer and CAP induced up-regulation, does not fit the repression data of Oehler et. al. (1). On the other hand, we find an excellent fit for a comprehensive model. When the lower bound constraint on looping energy for short DNA loops is set very low we find many different optimal solutions with the same quality of fit. On the other hand, when we set the lower bound on short DNA loop energy to levels estimated through tethered particle experiments (20), the optimization procedure finds fewer local minima producing a good fit. This suggests that enforcing a high lower bound of short loop energy produces a less robust fit. Our conclusions are compatible with suggestions, that the DNA looping is assisted by a non-specific binding protein, such as HU, which binds non-specifically to the DNA and lowers the energy of looping (34) by introducing a kink into the DNA. Such a protein may reconcile the discrepancy between the *ex* *vivo* tethered particle experimental data and the looping energy predictions which are determined by fitting models, like ours, to *in vivo* data.

Our results are robust in the parameter space. Starting from 20 initial conditions we have found 11 minimizers with essentially the same value of the cost function i.e. with the same quality of the fit. We have explored the neighborhood of these points and found that the majority of these points still provide a satisfactory fit. This may be interpreted as robustness of the model with respect to the parameters, or, alternatively, as a relative lack of constraint on the parameters by the experimental data.

We often see a radically different behavior of the model repression curves at the values of the repressor that were not yet interrogated by the experiments. This suggest that such experiments would provide further constraints on the parameters in the system, which in turn would lead to deeper understanding of the relative contributions of different components of *lac* operon to its function.

Acknowledgement

Research of T.G. was partially supported by National Science Foundation grant DMS-081878 Research of K.M. was partially supported by National Science Foundation grants DMS-0915019, DMS-1125174, and CBI-0835621 and by contracts from Defense Advanced Research Projects Agency and Air Force Office of Scientific Research. Research of K.P. was partially supported by National Science Foundation grant DMS-081878 and National Institutes of Health grant R01-DE19245.

Supporting citations

References (35–45) appear in the Supporting Material.

References

1. Oehler, S., M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill, 1994. Quality and position of the three *lac* operators of *E. coli* define efficiency of repression. *EMBO J* 13:3348–55.

2. Müller-Hill, B., 1996. The *lac* Operon: A Short History of a Genetic Paradigm. de Gruyter.

3. Malan, T., and W. McClure, 1984. Dual promoter control of the *Escherichia coli* lactose operon. *Cell* 39:173–80.

4. Fried, M., and J. Hudson, 1996. DNA looping and *lac* repressor-CAP interaction. *Science* 274:1930–1.

5. Swigon, D., and W. Olson, 2008. Mesoscale modeling of multi-protein-DNA assemblies: The role of the catabolic activator protein in Lac-repressor-mediated looping. *Int J Non Linear Mech* 43:1082–93.

6. Lawson, C., D. Swigon, K. Murakami, S. Darst, H. Berman, and R. Ebright, 2004. Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.* 14:10–20.

7. Kuhlman, T., Z. Zhang, M. Saier, Jr, and T. Hwa, 2007. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 104:6043–8.

8. Hudson, J., and M. Fried, 1990. Co-operative Interactions between the catabolite gene activator protein and the *lac* repressor at the lactose promoter. *J. Mol. Biol.* 214:381–96.

9. Oehler, S., E. Eismann, H. Krämer, and B. Müller-Hill, 1990. The three operators of the *lac* operon cooperate in repression. *EMBO J* 9:973–9.

10. Ptashne, M., and A. Gann, 2002. Genes and Signals. Cold Spring Harbor Laboratory Press.

11. Bintu, L., N. Buchler, H. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, 2002. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* 15:116–24.

12. Ackers, G., A. Johnson, and M. Shea, 1982. Qualitative Model for Gene Regulation by λ Phage Repressor. *Proc. Natl. Acad. Sci. USA* 79:1129–33.

13. Gedeon, T., K. Mischaikow, K. Patterson, and E. Traldi, 2008. Binding cooperativity in phage lambda is not sufficient to produce an effective switch. *Biophys. J.* 94:3384–92.

14. Santillán, M., and M. Mackey, 2004. Why the Lysogenic State of Phage λ Is So Stable: A Mathematical Modeling Approach. *Biophys. J.* 86:75–84.

15. Santillán, M., and M. Mackey, 2004. Influence of Catabolite Repression and Inducer Exclusion on the Bistable Behavior of the *lac* Operon. *Biophys. J.* 86:1282–92.

16. Shea, M., and G. Ackers, 1985. The O_R control system of bacteriophage lambda : A physical-chemical model for gene regulation. *J. Mol. Biol.* 181:211–30.

17. Hill, T., 1960. Introduction to Statistical Thermodynamics. Addison Wesley.

18. Horton, N., M. Lewis, and P. Lu, 1997. *Escherichia coli lac* Repressor-*lac* Operator Interaction and the Influence of Allosteric Effectors. *J. Mol. Biol.* 265:1–7.

19. Travers, A., 2006. DNA Topology: Dynamic DNA Looping. *Current biology* 16:R838–40.

20. Han, L., H. Garcia, S. Blumberg, K. Towles, J. Beausang, P. Nelson, and R. Phillips, 2009. Concentration and Length Dependence of DNA looping in Transcriptional Regulation. *PLoS ONE* 4:e5621.

21. Bellomy, G., M. Mossing, and M. Record, 1988. Physical properties of DNA *in vivo* as probed by the length dependence of the *lac* operator looping process. *Biochemistry* 27:3900–6.

22. Becker, N., J. Kahn, and L. Maher III, 2005. Bacterial Repression Loops Require Enhanced DNA Flexibility. *J. Mol. Biol.* 349:716–730.

23. Becker, N., J. Kahn, and L. Maher III, 2007. Effects of nucleoid proteins on DNA repression loop formation in *Escherichia coli*. *Nucleic Acids Res.* 35:3988–4000.

24. Becker, N., J. Kahn, and L. Maher III, 2008. Eukaryotic HMGB proteins as replacements for HU in *E. coli* repression loop formation. *Nucleic Acids Res.* 36:4009–21.

25. Müller, J., S. Oehler, and B. Müller-Hill, 1996. Repression of *lac* Promoter as a Function of Distance, Phase and Quality of an Auxiliary *lac* Operator. *J. Mol. Biol.* 257:21–9.

26. Law, S. M., G. R. Bellomy, P. J. Schlax, and M. T. Record, 1993. In Vivo Thermodynamic Analysis of Repression with and without Looping in lac Constructs : Estimates of Free and Local lac Repressor Concentrations and of Physical Properties of a Region of Supercoiled Plasmid DNA in Vivo. *J. Mol. Biol.* 230:161 – 173.

27. Mossing, M., and J. MT Record, 1986. Upstream operators enhance repression of the *lac* promoter. *Science* 233:889–92.

28. Vanzi, F., C. Broggio, L. Sacconi, and F. Pavone, 2006. Lac repressor hinge flexibility and DNA looping: single molecule kinetics by tethered particle motion. *Nucleic Acids Res.earch* 34:3409–3420.

29. Wong, O., M. Guthold, D. Erie, and J. Gelles, 2008. Interconvertible lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biol* 6:e232.

30. Towles, K., J. Beausang, H. Garcia, R. Phillips, and P. Nelson, 2009. First-principles calculation of DNA looping in tethered particle experiments. *Phys Biol* 6:025001.

31. Swigon, D., B. Coleman, and W. Olson, 2006. Modeling the Lac repressor-operator assembly: The influence of DNA looping on Lac repressor conformation. *Proc. Natl. Acad. Sci. USA* 103:9879–84.

32. Zhang, Y., A. McEwen, D. Crothers, and S. Levene, 2006. Analysis of *In-Vivo* LacR-Mediated Gene Repression Based on the Mechanics of DNA Looping. *PLoS ONE* 1:e136.

33. Saiz, L., and J. Vilar, 2007. Multilevel deconstruction of the in vivo behavior of looped DNA-protein complexes. *PLoS ONE* 2:e355.

34. Czapla, L., D. Swigon, and W. Olson, 2008. Effects of the Nucleoid protein HU on the structure, flexibility, and ring-closure properties of DNA deduced from Monte Carlo simulations. *J. Mol. Biol.* 382:353–70.

35. Keseler, I., C. Bonavides-Martinez, J. Collado-Vides, S. Gama-Castro, R. Gunsalus, D. Johnson, M. Krummenacker, L. Nolan, S. Paley, I. Paulsen, M. Peralta-Gil, A. Santos-Zavaleta, A. Shearer, and P. Karp, 2009. EcoCyc: A comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.* 37:D464–70.

36. Czarniecki, D., R. Noel, Jr, and W. Reznikoff, 1997. The -45 region of the *Escherichia coli lac* promoter: CAP-dependent and CAP-independent transcription. *J Bacteriol* 179:423–9.

37. Bremer, H., and P. Dennis, 1996. Modulation of chemical composition and other parameters of the cell by growth rate. *In* F. N. et. al, editor, In Escherichia coli and Salmonella thyphymurium: Cellular and Molecular Biology. American Society for Microbiology, Washington DC, volume 2, 1553–1569.

38. Inada, T., K. Kimata, and H. Aiba, 1996. Mechanism responsible for glucose-lactose diauxie in Escherichia coli: challenge to the cAMP model. *Genes Cells* 1:293–301.

39. Oehler, S., S. Alberti, and B. Müller-Hill, 2006. Induction of the *lac* promoter in the absence of DNA loops and the stoichiometry of induction. *Nucleic Acids Res.* 34:606–12.

40. nad PA Gottlieb, X. Z., 1993. Thermodynamic and Alkylation Interference Analysis of the *lac* Repressor-Operator Substituted with the Analogue 7-Deazaguanine. *Biochemistry* 32:11374–84.

41. Cossart, P., and B. Gicquel-Sanzey, 1985. Regulation of expression of the crp gene of *Escherichia coli* K-12: in vivo study. *J Bacteriol* 161:454–7.

42. Baker, C., S. Tomlinson, A. García, and J. Harman, 2001. Amino acid substitution at position 99 affects the rate of CRP subunit exchange. *Biochemistry* 40:12329–38.

43. Lindemose, S., P. Nielsen, and N. Mollegaard, 2008. Dissecting direct and indirect readout of cAMP receptor protein DNA binding using an inosine and 2,6-diaminopurine in vitro selection system. *Nucleic Acids Res.* 36:4797–807.

44. Liu, M., G. Gupte, S. Roy, R. Bandwar, S. Patel, and S. Garges, 2003. Kinetics of transcription initiation at lacP1. Multiple roles of cyclic AMP receptor protein. *J. Biol. Chem.* 278:39755–61.

45. Strickland, S., G. Palmer, and V. Massey, 1975. Determination of dissociation constants and specific rate constants of enzyme-substrate (or protein-ligand) interactions from rapid reaction kinetic data. *J. Biol. Chem.* 250:4048–52.

46. Saiz, L., J. Rubi, and J. Vilar, 2005. Inferring the *in vivo* looping properties of DNA. *Proc. Natl. Acad. Sci. USA* 102:17642–5.

K_{P1}	$=$	5.6×10^{-7}	M	K_{P2}	$=$	$2.5 \times 10-7$	M
[CAP]	$=$	0.78×10^{-6}	M	[RNAP]	$=$	1.25×10^{-6}	M
K_{P1C1}	$=$	0.2616	M	RT	$=$	0.617	kcal/mol
k_f	$=$	0.12	min^{-1}	k_{fC1}	$=$	1.55	min^{-1}

Table 1: Fixed parameter values

A. Free energy of binding an operator				($^-$ represents a deleted operator)			
ΔG_{O1}	$\simeq$	-14.27	kcal/mol	ΔG_{O1^-a}	$\simeq$	-8.69	kcal/mol
ΔG_{O2}	$\simeq$	-12.91	kcal/mol	ΔG_{O1^-b}	$\simeq$	-7.82	kcal/mol
ΔG_{O3}	$\simeq$	-10.78	kcal/mol	ΔG_{O2^-}	$\simeq$	-4.87	kcal/mol
ΔG_{O3^*}	$\simeq$	-6.33	kcal/mol	ΔG_{O3^-}	$\simeq$	-6.32	kcal/mol
$\Delta G_{O3^*(O1)}$	$\simeq$	-8.73	kcal/mol	$\Delta G_{O3^{*-}}$	$\simeq$	-3.72	kcal/mol
ΔG_{ns}	$\simeq$	-7.0	kcal/mol				
B. Looping energies							
ΔG_{O12}	$\simeq$	8.19	kcal/mol	ΔG_{O23}	$\simeq$	12.99	kcal/mol
ΔG_{O23*}	$\simeq$	8.11	kcal/mol	ΔG_{O13}	$\simeq$	16.14	kcal/mol
ΔG_{O13*}	$\simeq$	4.47	kcal/mol				
C. Cooperativity							
ΔG_{C1loop}	$\simeq$	-1.42	kcal/mol	ΔG_{C1O3}	$\simeq$	9.09	kcal/mol
ΔG^I	$\simeq$	9.81	kcal/mol				

Table 2: Best parameter values computed by the optimization procedure.

description		short loop energy range
Compiled data	Han et al. (Fig 12) (20)	$11 - 17$
Tethered particle	Han et al. (Fig 9) (20)	$10 - 12$
Inferred from data	Saiz et al. (Fig 3) (46)	$7.5 - 10$
HU cyclization model	Czapla et al. (34)	$6 - 13$
DNA looping model	Towles et al. (Fig 9) (30)	> 11.7
DNA looping model	Zhang et al. (32)	$9 - 12$
DNA looping model	Saiz et al. (33)	$8 - 9.5$
DNA looping model	Swigon et al. (5)	$14.7 - 16.1$

Table 3: Short loop free energies comparisons in kcal/mol

	Optimized model predictions		
	cost ξ	ΔG_{O13}	ΔG_{O13*}
	0.5585	7.68	16.86
	0.5227	7.71	12.36
	0.5226	10.87	4.09
	0.5212	11.59	5.95
	0.5278	7.59	9.01
	0.5279	7.57	12.11
	0.5584	7.68	10.97
	0.5281	7.64	16.17
	0.5226	7.71	14.30
	0.5115	16.14	4.47
	0.5225	7.71	12.29
μ		9.08	10.78
σ		2.75	4.43

Table 4: Short loop energy values: The exact looping energies predicted by the optimization, and the associated cost, ξ, are listed below. The mean and standard deviation are listed as the last two rows of the table.

Figure Legends

Figure 1

Cartoon image describing the *lac* repressor coding region and the *lac* operon. In the E. coli genome the DNA coding for a *lac* repressor subunit is preceded by a promoter region, P_i and immediately followed by the *lac* operon. The *lac* operon consists of a regulatory region and the lacZ, lacY, and lacA genes. As shown at the bottom of the figure, the regulatory region is composed of multiple binding sites: the P1 and P2 promoter regions bind RNAP (there are also at least two other sites P3 and P4, but these bind RNAP very weakly); C1 binds CAP; and O1, O2 and O3 bind the *lac* repressor R. O1 and O3 are separated by 92 base pairs, and O1 and O2 are separated by 401 base pairs. Each RNAP produces a mRNA copy of all three *lac* genes lacZ, lacY and lacA.

Figure 2

(a) Repression level curves as a function of the repressor concentration obtained from the model based on biochemical processes 1-10 using the parameters indicated in Table 2. (b) Repression level curves as a function of the repressor concentration for the standard model based on biochemical processes 1-4. The parameter values used for this restricted model are indicated in Table S.9 in the Supplemental Material. The upper figures show the repression curves for all mutants with the O3 operator deleted, while the lower figures show the repression curves for all mutants with the O2 operator deleted, as well as the wild type (WT) (solid black curve).

Figure 3

We minimize the complete model for 20 values of $\vec{x}$ and choose the eleven most minimized vectors as determined by the cost function. As described in the text, we take these values of $\vec{x}$ and their standard deviation for the ith entry of these vectors (Table S.8). We then generate a dataset, $\mathcal{V}$, of 847 vectors, where each vector is one of the original eleven modified by zero, one or two standard deviations of the ith entry of $\vec{x}$. We have sorted $\mathcal{V}$ by the value of cost function and plotted the vector identity (1-828) versus the cost function. The vectors from 829 to 847 have a cost between 8 and 26 and are excluded to better present the lower cost vectors. The inset shows the repression curves associated with the solid dot, after the first break in the grouping of the cost functions.

Figure 4

Short loop energy where each point represents one minimization. We plot ΔG_{O13} versus ΔG_{O13*} for the eleven optimized points which resulted in a cost of $\xi \leq 0.56$. The exact values are listed on the right. The shaded square denotes the region in which looping energies have been measured (20), Table 3.

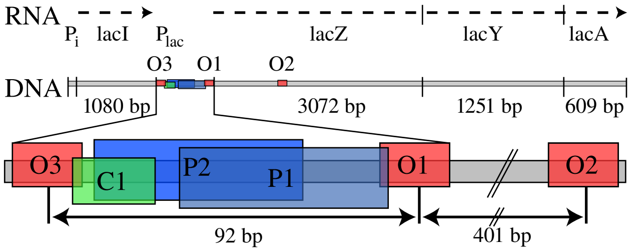
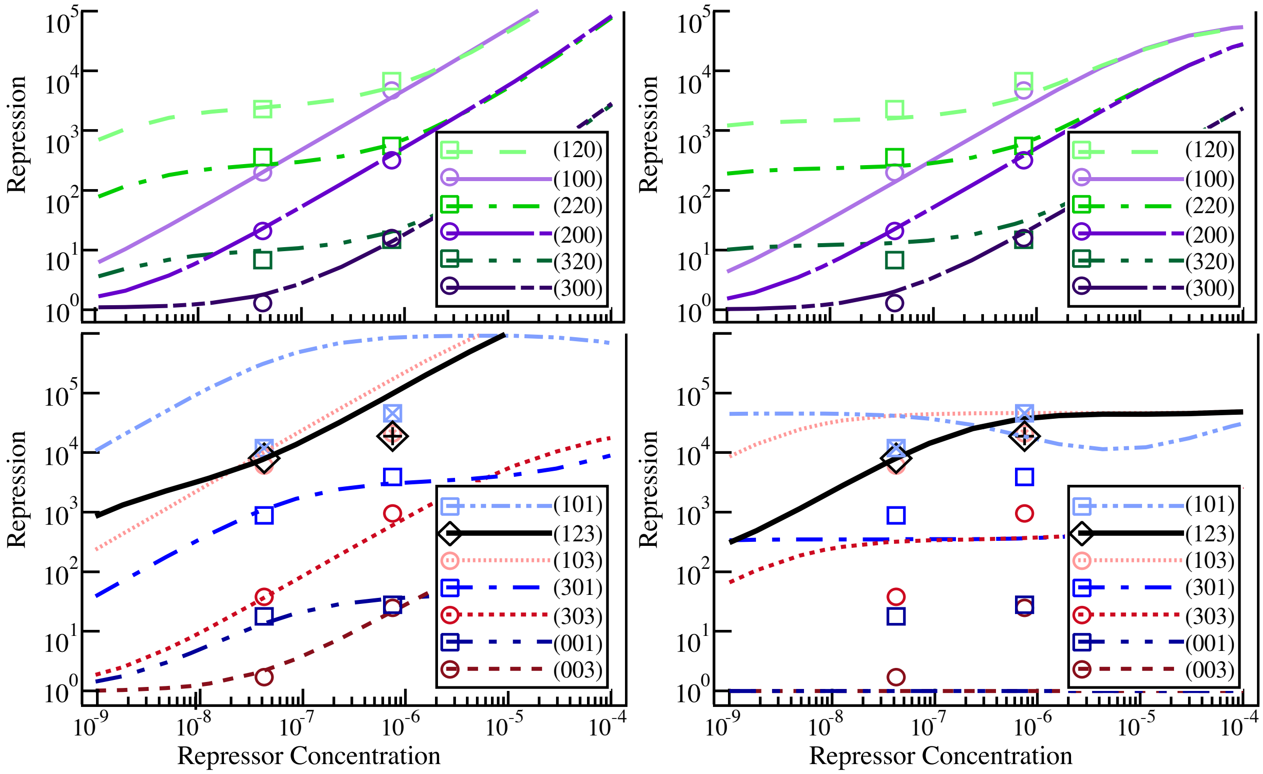

RNA
DNA
P_i
lacI
P_lac
lacZ
lacY
lacA
O3
O1
O2
1080 bp
3072 bp
1251 bp
609 bp
O3
C1
P2
P1
O1
O2
92 bp
401 bp

Figure 1



(a) Full model, $\xi = 0.5115$

(b) Standard model, $\xi = 3.0155$

Figure 2

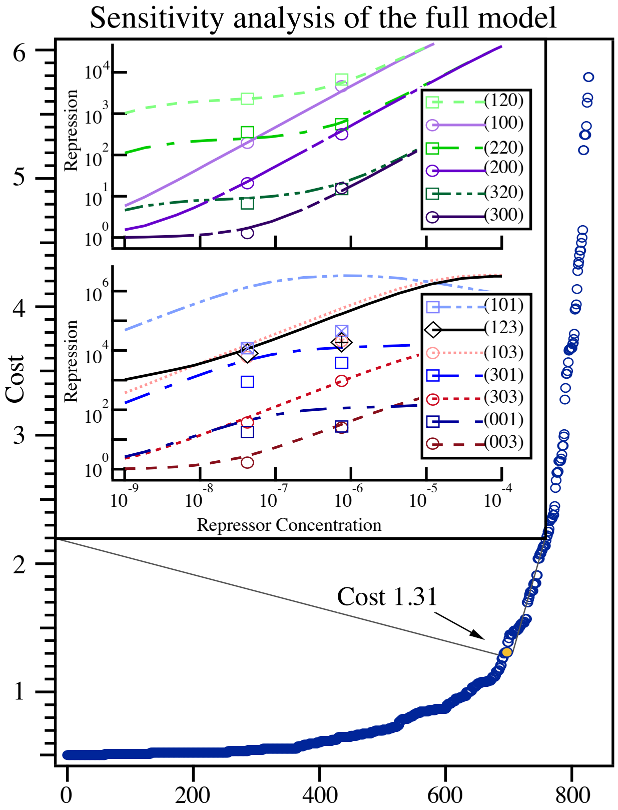

Sensitivity analysis of the full model
Cost
6
5
4
3
2
1
0
200
400
600
800
Cost 1.31
Repression
10^0
10^1
10^2
10^3
10^4
(120)
(100)
(220)
(200)
(320)
(300)
Repression
10^0
10^2
10^4
10^6
(101)
(123)
(103)
(301)
(303)
(001)
(003)
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
Repressor Concentration

Figure 3

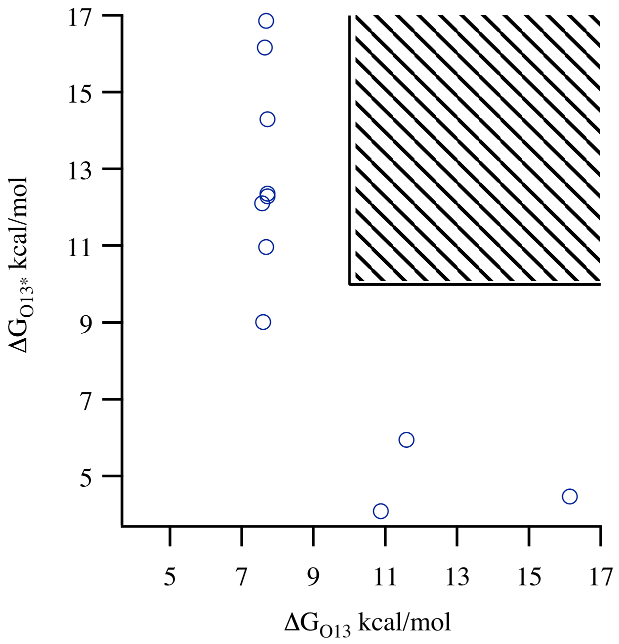

17
15
13
11
9
7
5
ΔG_{O13*} kcal/mol
5
7
9
11
13
15
17
ΔG_{O13} kcal/mol

Figure 4