

Comparison of the theoretical and real-world evolutionary potential of a genetic circuit

M Razo-Mejia^{1,2}, J Q Boedicker², D Jones², A DeLuna³, J B Kinney⁴
and R Phillips²

¹ Ingenieria Biotecnologica, Instituto Politecnico Nacional, Av Mineral de Valenciana No 200 Col Fracc Industrial Puerto Interior, Silao de la Victoria, Guanajuato, 36275, Mexico

² Department of Applied Physics, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA

³ Laboratorio Nacional de Genomica para la Biodiversidad (Langebio), Centro de Investigacion y de Estudios Avanzados (Cinvestav), 36821 Irapuato, Guanajuato, Mexico

⁴ Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

E-mail: phillips@pboc.caltech.edu

Received 7 November 2013, revised 19 February 2014


Accepted for publication 5 March 2014

Published 1 April 2014

Abstract

With the development of next-generation sequencing technologies, many large scale experimental efforts aim to map genotypic variability among individuals. This natural variability in populations fuels many fundamental biological processes, ranging from evolutionary adaptation and speciation to the spread of genetic diseases and drug resistance. An interesting and important component of this variability is present within the regulatory regions of genes. As these regions evolve, accumulated mutations lead to modulation of gene expression, which may have consequences for the phenotype. A simple model system where the link between genetic variability, gene regulation and function can be studied in detail is missing. In this article we develop a model to explore how the sequence of the wild-type *lac* promoter dictates the fold-change in gene expression. The model combines single-base pair resolution maps of transcription factor and RNA polymerase binding energies with a comprehensive thermodynamic model of gene regulation. The model was validated by predicting and then measuring the variability of *lac* operon regulation in a collection of natural isolates. We then implement the model to analyze the sensitivity of the promoter sequence to the regulatory output, and predict the potential for regulation to evolve due to point mutations in the promoter region.

Keywords: thermodynamic models, *lac* operon, evolutionary potential, transcriptional regulation, natural variability

 Online supplementary data available from stacks.iop.org/PhysBio/11/026005/mmedia

1. Introduction

Despite efforts to understand genotypic variability within natural populations [1] and recent interest in fine-tuning genetic circuits for synthetic biology [2], it still remains unclear how, with base pair resolution, the sequence of a gene regulatory region can be translated into output levels of

gene expression [3]. Generally, classical population genetics has treated regulatory architectures as changeless parameters, rather than potential evolutionary variables, focusing on changes in protein structure rather than gene regulation. However, genetic regulatory architecture can also determine the variation of traits, and thus the evolutionary potential of

these genes [4]. After all, the structure of bacterial promoters dictates interactions among the transcriptional apparatus, and through the modification of this structure, regulatory circuits can be modified to potentially allow cells to occupy different niches [5, 6].

Thermodynamic models of gene regulation have been widely used as a theoretical framework to dissect and understand genetic architectures [7–11]. Such dissections have led to a quantitative understanding of how parameters such as binding energies, transcription factor copy numbers, and the mechanical properties of the DNA dictate expression levels. Recently the development of experimental techniques combining these types of models with cell sorting and high-throughput sequencing have made it possible to understand gene regulation at single-base pair resolution [12–14], as well as to deliberately design promoter architectures with desired input–output functions [15]. These models connect the sequence of a promoter to the output phenotype, making it possible to predict variability and evolutionary potential of gene regulatory circuits.

The *lac* operon has served as a paradigm of a genetic regulatory system for more than 60 years [16, 17]. This operon contains the molecular machinery that some bacterial species, including the model organism *E. coli*, use to import and consume lactose. Extensive quantitative characterization of the regulation of this genetic circuit [18, 19], as well as of the link between fitness and expression of the operon [20–24] make it an ideal system for exploring the evolutionary potential of a regulatory circuit. With previous exhaustive description and quantification of the parameters controlling the expression level of this genetic circuit [19, 25–27] we now have what we think is a nearly complete picture of the regulatory *knobs* that can modify the expression level, shown schematically in figure 1(a). In this article we build upon this understanding by directly linking the sequence of the promoter region with these control parameters, thereby creating a map from genotype to transcriptional output.

Within a collection of *E. coli* isolated from different host organisms we observe significant variability for the regulation of the *lac* operon, as shown in figure 1(b). By characterizing the variability of the regulatory control parameters shown in figure 1(a) within these strains, we identified evolutionary trends in which certain parameters or subsets of parameters are seen to vary more often than others within this collection of natural isolates. Using the map of promoter sequence to transcriptional output, we demonstrated that the regulatory input–output function for the *lac* promoter could account for most of the natural variability in regulation we observed. We then implement the map to explore the theoretical potential for this regulatory region to evolve. This level of analysis gives us clues as to how selection could fine tune gene expression levels according to the environmental conditions to which cells are exposed.

2. Results

2.1. Quantitative model of the natural parameters that regulate gene expression

Thermodynamic models of gene regulation have become a widely used theoretical tool to understand and dissect different regulatory architectures [3, 12, 19, 26, 27, 31]. The *lac* promoter is one such regulatory architecture that has been studied in detail [32]. Models have been constructed and experimentally validated for both the wild-type *lac* promoter and synthetic promoter regions built up from the *lac* operon’s regulatory components [12, 15, 19, 26, 27, 32–37].

In a simple dynamical model of transcription the number of messenger RNA (mRNA) is proportional to the transcription rate and the degradation rate of the mRNA,

$$\frac{dm}{dt} = -\gamma \cdot m + \sum_i r_i \cdot p_i, \quad (1)$$

where γ is the mRNA degradation rate and m is the number of transcripts of the gene per cell; r_i and p_i are the transcription rate and the probability of state i respectively. We can think of p_i as a measure of the time spent in the different transcriptionally active states. Thermodynamic models assume that the gene expression level is dictated by the probability of finding the RNA polymerase (RNAP) bound to the promoter region of interest [7–9]. With a further quasi-equilibrium assumption for the relevant processes leading to transcription initiation, we derive a statistical mechanics description of how parameters such as transcription factor copy number and their relevant binding energies, encoded in the DNA binding site sequence, affect this probability [10]. Quantitative experimental tests of predictions derived from equilibrium models have suggested the reasonableness of the assumption [15, 19, 26, 27], although caution should be used as the equilibrium assumption is not necessarily valid in all cases. The validity of this equilibrium assumption relies on the different time-scales of the processes involved in the transcription of a gene. Specifically the rate of binding and unbinding of the transcription factors and the RNAP from the promoter region should be faster than the open complex formation rate; if so, the probability of finding the RNAP bound to the promoter is given by its equilibrium value [9, 38]. For the case of the Lac repressor, the rate of unbinding from the operator is 0.022 s^{-1} [39], and the binding of an unoccupied operator with ten repressors per cell occurs at a similar rate [40]. Open complex formation, a rate limiting step in promoter escape, has been measured at a rate of $2 \times 10^{-3} \text{ s}^{-1}$ [41]. Promoter escape is about an order to magnitude slower than the binding and unbinding of the Lac repressor, and this separation of time-scales supports the equilibrium assumption for this particular case. We enumerate the possible states of the system and assign statistical weights according to the Boltzmann distribution as shown in figure 2.

From these states and weights we derive an equation describing the probability of finding the system in a transcriptionally active state, and therefore the production term from equation (1),

$$\sum_i r_i p_i = \sum_i r_i \frac{W_i}{Z_{\text{tot}}}, \quad (2)$$

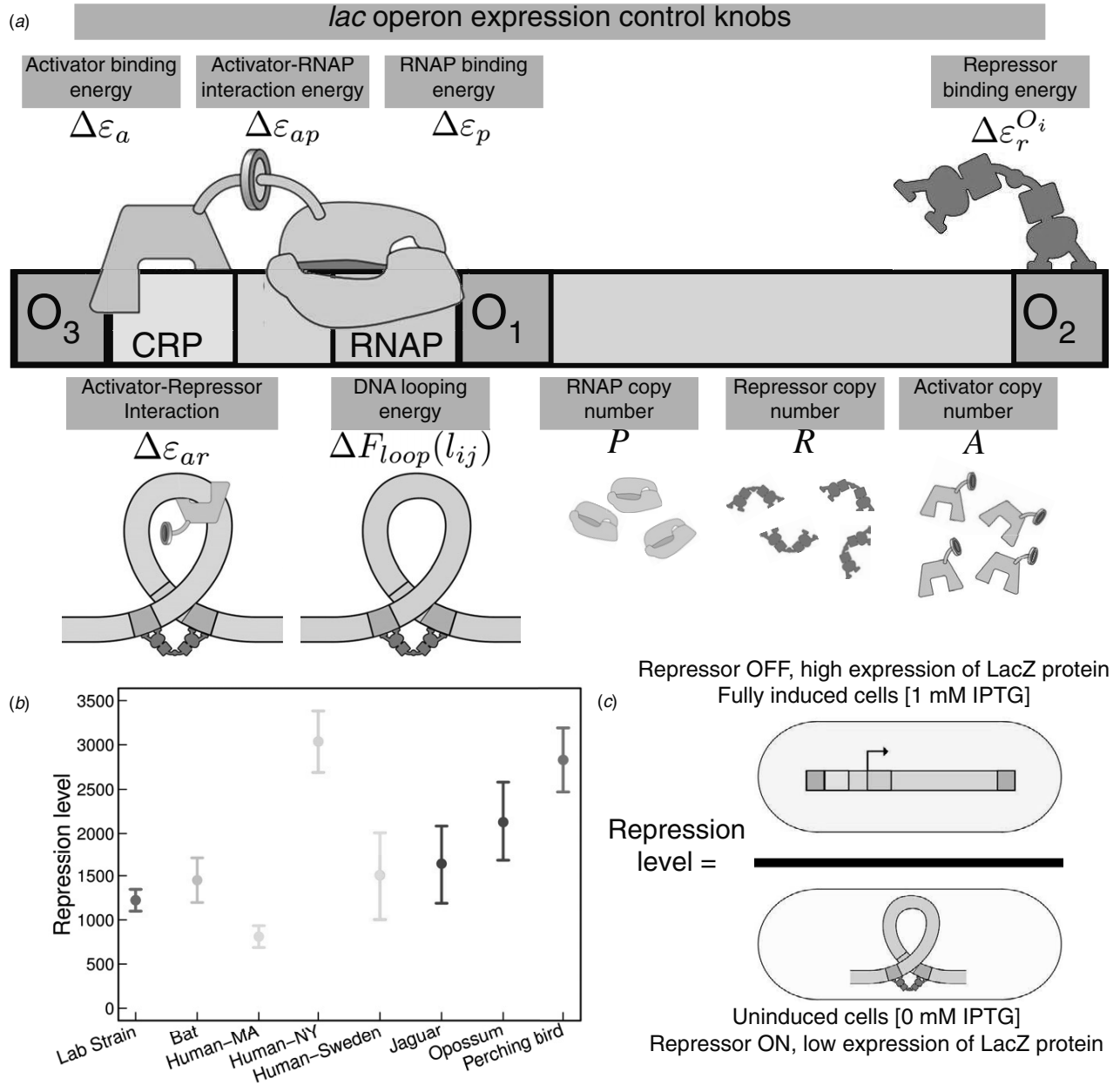


Figure 1. (a) Regulatory knobs that control the expression of the *lac* operon and the symbols used to characterize these knobs in the thermodynamic model. The activator CRP increases expression, the Lac repressor binds to the three operators to decrease expression, and looping can lock the repressor onto O_1 leading to increased repression. The interaction energy between RNAP and CRP reflects the stabilization of the open complex formation due to the presence of the activator [28], and the interaction between the Lac repressor and CRP stabilizes the formation of the upstream loop [29]. (b) Variability in the repression level of *E. coli* natural isolates and the lab control strain MG1655. Strains are named after the host organism from which they were originally isolated [30]. Error bars represent the standard deviation from at least three independent measurements. (c) Schematic representation of the repression level, in which the role of the repressor in gene regulation is experimentally measured by comparing the ratio of LacZ proteins in cells grown in the presence of 1 mM IPTG to cells grown in the absence of IPTG. LacZ protein concentrations were measured using a colorimetric assay.

where W_i is the statistical weight of states in which the polymerase is bound, which are assumed to lead to the transcription of the operon (shaded blue in figure 2), and $Z_{\text{tot}} = \sum_{\text{All states}} W_{\text{state}}$ is the partition function, or the sum of the statistical weights of all states. We connect this model to experimental measurements of repression, that is the ratio of

gene expression in the absence of the active repressor to gene expression in the presence of active repressor, using:

$$\text{repression} = \frac{\text{gene expression}(R = 0)}{\text{gene expression}(R \neq 0)}, \quad (3)$$


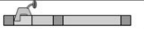



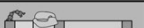





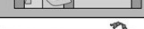






















State	Weight	State	Weight
	1		$\frac{A}{N_{NS}} e^{-\beta \Delta \epsilon_a}$
	$\frac{P}{N_{NS}} e^{-\beta \Delta \epsilon_p}$		$\frac{(A)(P)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_p + \Delta \epsilon_{ap})}$
	$\frac{2R(P)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O2} + \Delta \epsilon_p)}$		$\frac{2R(P)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O3} + \Delta \epsilon_p)}$
	$\frac{4R(R-1)(P)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3} + \Delta \epsilon_p)}$		$\frac{2R(A)(P)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_r^{O2} + \Delta \epsilon_p + \Delta \epsilon_{ap})}$
	$\frac{2R(A)(P)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_p + \Delta \epsilon_r^{O3})}$		$\frac{4R(R-1)(A)(P)}{N_{NS}^4} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_p + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3})}$
	$\frac{2R}{N_{NS}} e^{-\beta \Delta \epsilon_r^{O1}}$		$\frac{2R}{N_{NS}} e^{-\beta \Delta \epsilon_r^{O2}}$
	$\frac{2R}{N_{NS}} e^{-\beta \Delta \epsilon_r^{O3}}$		$\frac{4R(R-1)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2})}$
	$\frac{4R(R-1)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O3})}$		$\frac{4R(R-1)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3})}$
	$\frac{8R(R-1)(R-2)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3})}$		$\frac{2R(A)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_r^{O1})}$
	$\frac{2R(A)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_r^{O2})}$		$\frac{4R(R-1)(A)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2})}$
	$\frac{2R(A)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_r^{O3})}$		$\frac{4R(R-1)(A)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O3})}$
	$\frac{4R(R-1)(A)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3})}$		$\frac{8R(R-1)(R-2)(A)}{N_{NS}^4} e^{-\beta(\Delta \epsilon_a + \Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3})}$
	$\frac{2R}{N_{NS}} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta F_{loop}(l_{12}))}$		$\frac{2R}{N_{NS}} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O3} + \Delta F_{loop}(l_{13}))}$
	$\frac{2R}{N_{NS}} e^{-\beta(\Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3} + \Delta F_{loop}(l_{23}))}$		$\frac{4R(R-1)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3} + \Delta F_{loop}(l_{12}))}$
	$\frac{4R(R-1)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3} + \Delta F_{loop}(l_{13}))}$		$\frac{4R(R-1)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3} + \Delta F_{loop}(l_{23}))}$
	$\frac{2R(A)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta \epsilon_a + \Delta F_{loop}(l_{12}))}$		$\frac{4R(R-1)(A)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3} + \Delta \epsilon_a + \Delta F_{loop}(l_{12}))}$
	$\frac{2R(A)}{N_{NS}^2} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O3} + \Delta \epsilon_a + \Delta \epsilon_{ar} + \Delta F_{loop}(l_{13}))}$		$\frac{4R(R-1)(A)}{N_{NS}^3} e^{-\beta(\Delta \epsilon_r^{O1} + \Delta \epsilon_r^{O2} + \Delta \epsilon_r^{O3} + \Delta \epsilon_a + \Delta \epsilon_{ar} + \Delta F_{loop}(l_{13}))}$

Figure 2. Thermodynamic model of gene regulation. The table shows all states permitted within the model and their respective statistical weights as obtained using statistical mechanics. In these weights P = number of RNAP per cell, R = number of repressor molecules per cell, A = number of activator molecules per cell, $\Delta \epsilon_r^{O_i}$ = binding energy of Lac repressor to the i th operator, $\Delta \epsilon_p$ = binding energy of RNAP to the promoter, $\Delta \epsilon_a$ = activator binding energy, $\Delta F_{loop}(l_{ij})$ = looping free energy between operator O_i and O_j , N_{NS} = number of nonspecific binding sites on the genome, $\Delta \epsilon_{ap}$ = interaction energy between the activator and the RNAP, $\Delta \epsilon_{ar}$ = interaction energy between the activator and the repressor, and β = inverse of the Boltzmann constant times the temperature (see supplemental material, available from stacks.iop.org/PhysBio/11/026005/mmedia). States with blue background are assumed to lead to transcription of the operon.

where R is the number of repressor molecules per cell. The experimental equivalent of repression is depicted in figure 1(c). In experiments, isopropyl β -D-1-thiogalactopyranoside (IPTG) is used to inactivate the Lac repressor, preventing it from binding to the genome with high affinity [19]. Repression, as defined in equation (3), has been a standard metric for the role of transcription factors, including the Lac repressor, on gene expression [7, 42]. By measuring the ratio of steady-state levels of a gene reporter protein, here LacZ, we are able to isolate the role of the repressor in gene regulation, as described further in section S8 of the supplemental material (available from stacks.iop.org/PhysBio/11/026005/mmedia).

Various models of the wild-type *lac* promoter have been reported in the past using this simple structure. Our work builds upon the work by Kinney *et al* [12]. Kinney and collaborators combined a thermodynamic model of regulation with high-throughput sequencing to predict gene expression from statistical sequence information of the cAMP-receptor protein (CRP) and the RNAP binding sites. To predict how the sequence of the entire regulatory region influences expression, we adapted this model to account for how the binding site sequence and copy number of the Lac repressor modulate gene expression. Our model also takes into account growth rate effects, captured in the RNAP copy number [43, 44].

Based on previous work done on the *lac* operon [12, 19], we assumed that the presence of the activator does not affect the rate of transcription (r_i from equation (1)), but instead influences the probability of recruiting the polymerase to the promoter (p_i from equation (1)). Previous experimental characterization of the repressor binding energy to the different operators [26], the looping free energy for the upstream loop between $O_1 - O_3$ [27], activator concentration and its interaction energy with RNAP [19], RNAP binding energy [15] and RNAP copy number as a function of the growth rate [44], left us only with three unknown parameters for the model. One of these missing parameters, a decrease in the looping free energy when CRP and Lac repressor are bound at the same time, is a consequence of the experimental observation that the presence of CRP stabilizes the formation of the loop between $O_1 - O_3$ [29, 45]. The remaining two parameters, the looping energies for the $O_1 - O_2$ and $O_3 - O_2$ loops are not well characterized. These looping energies may differ from upstream loops due to the absence of the RNAP binding site which modifies the mechanical properties of the loop [46]. We fit these parameters for our model using Oehler *et al* repression measurements on *lac* operon constructs with partially mutagenized or swapped binding sites [42, 47] (see section S5 of the supplemental material

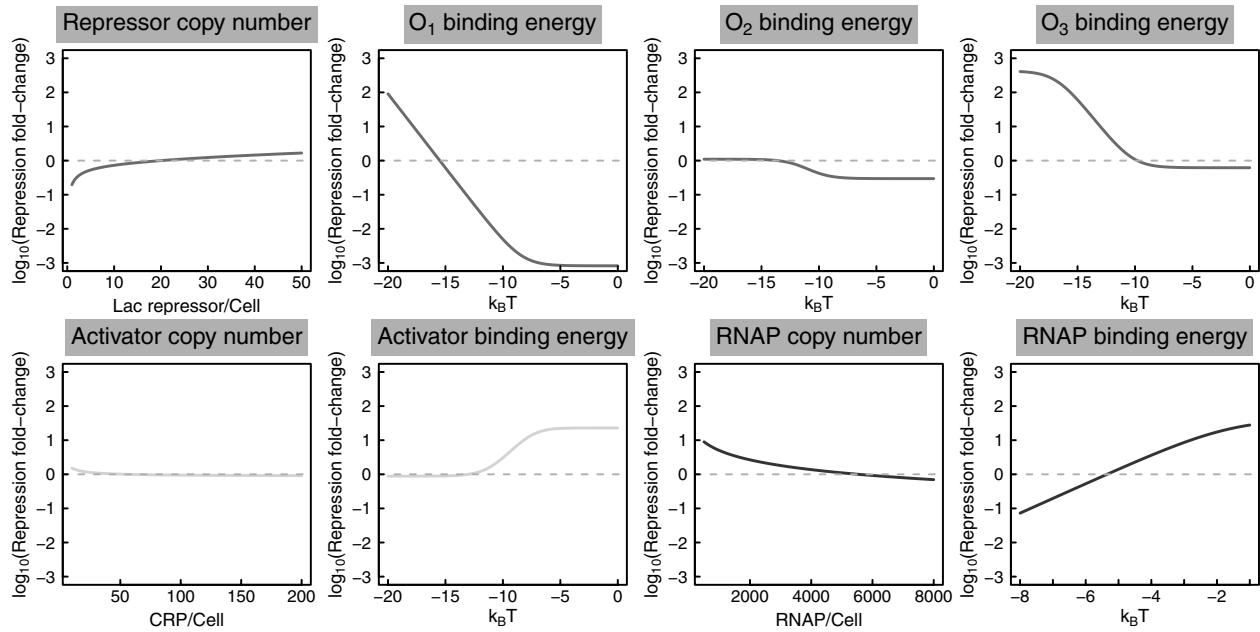


Figure 3. Sensitivity of phenotype to the parameters controlling the gene expression level. Each graph shows how a specific model parameter changes the level of gene expression. The \log_{10} ratio of repression is calculated with respect to the predicted repression for the lab strain MG1655. The vertical axis spans between 1000 fold decrease to 1000 fold increase in repression with respect to this strain. The gray dotted line indicates the reference value for the lab strain MG1655. Values above this line indicate the operon is more tightly repressed and values below this line have a leakier expression profile (see table S1, available from stacks.iop.org/PhysBio/11/026005/mmedia for further detail on the reference parameters).

(available from stacks.iop.org/PhysBio/11/026005/mmedia) for further details). Using these parameters the model is consistent with previous measurements (figure S4, available from stacks.iop.org/PhysBio/11/026005/mmedia). We emphasize that having the 14 parameters of the model characterized (see table S1, available from stacks.iop.org/PhysBio/11/026005/mmedia) provides testable predictions without free parameters that we compare with our experimental results.

2.2. Sensitivity of expression to model parameters

As an exploratory tool, the model can predict the change in regulation due to modifications in the promoter architecture. Figure 3 shows the fold-change in the repression level as a function of each of the parameters, using the lab strain MG1655 as a reference state (see supplemental material, available from stacks.iop.org/PhysBio/11/026005/mmedia for further detail on these reference parameters). We have reported parameters using strain MG1655 as a reference strain because this strain served as the basis for which most parameter values were determined and the gene expression model was derived.

From this figure we see that within the confines of this model, modifications in the O_1 binding energy have the most drastic effect on the repression of the operon. For the case of O_2 we see that increasing its affinity for the repressor does not translate into an increased ability to turn off the operon; but by decreasing this operator affinity the model predicts a reduction in the repression with respect to the reference strain.

Surprisingly the repression level is predicted to be insensitive to activator copy number. The same cannot be said about the affinity of the activator, since decreasing the activator binding energy greatly influences the repression level.

2.3. Mapping from sequence space to level of regulation

Recent developments of an experimental technique called *sort-seq*, involving cell sorting and high-throughput sequencing, have proved to be very successful in revealing how regulatory information is encoded in the genome with base pair resolution [12]. This technique generates energy matrices that make it possible to map from a given binding site sequence to its corresponding binding energy for a collection of different proteins and binding sites. Combining these energy matrices with thermodynamic models enables us to convert promoter sequence to the output level of gene expression. Recently these energy matrices have been used to deliberately design promoters with a desired expression level, demonstrating the validity of these matrices as a design tool for synthetic constructs [15]. We use the matrices for CRP and RNAP published previously [12]. We experimentally determined the matrix for the LacI operator using previously published methods [12], as discussed in section 4. Figure 4(a) shows a schematic representation of the relevant protein binding sites involved in the regulation of the *lac* operon and their respective energy matrices. Implementing these matrices into the thermodynamic model gives us a map from genotype to phenotype. We use this map to calculate the fold-change in repression relative to MG1655 for all possible point mutations

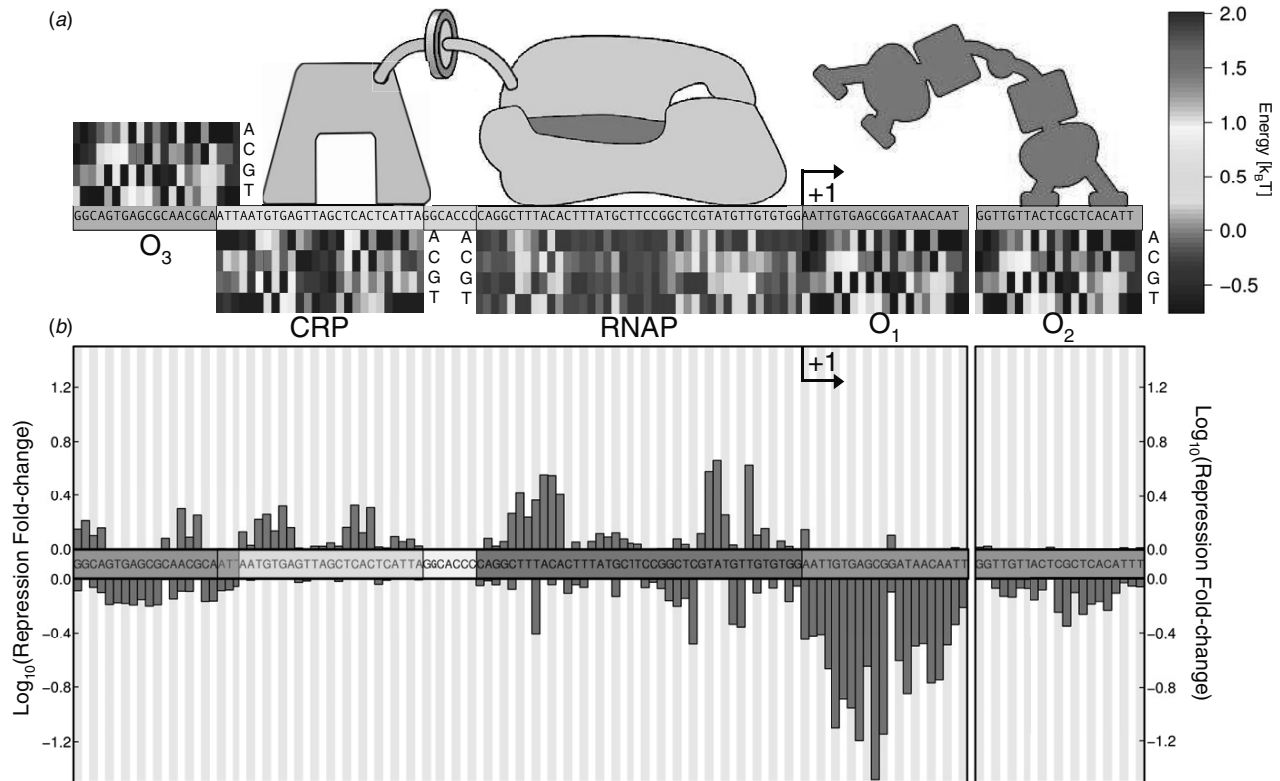


Figure 4. Mapping from promoter sequence to regulatory level. (a) Energy matrices for the relevant transcription factors (blue—RNAP, green—CRP, red—Lac repressor). These matrices allow us to map from sequence space to the corresponding binding energy. The contribution of each base pair to the total binding energy is color coded. The total binding energy for a given sequence is obtained by adding together the contribution of each individual base pair. (b) Using the energy matrices from (a) and the model whose states are depicted in figure 2, the \log_{10} repression change was calculated for all possible single point mutations of the promoter region. The height of the bars represents the biggest possible changes in the repression level (gray bars for biggest predicted decrease in repression, orange bar for biggest predicted increase in repression) given that the corresponding base pair is mutated with respect to the reference sequence (*lac* promoter region of the lab strain MG1655). The black arrows indicate the transcription start site.

in this region. Figure 4(b) shows the fold-changes in repression levels for the two base pair substitutions at each position that result in the largest predicted increase or decrease in repression.

Again we see that mutations in the O_1 binding site have the largest effect on regulation since a single-base pair change can lower the ability of the cell to repress the operon by a factor of ≈ 20 . With only two relevant mutation that could significantly increase the repression level, this map reveals how this operator and its corresponding transcription factor diverged in a coordinated fashion; the wild-type sequence has nearly maximum affinity for the repressor [48]. It is known that the non-natural operator O_{id} binds more strongly than O_1 [42]. O_{id} is one base pair shorter than O_1 and current maps made with *sort-seq* cannot predict changes in binding affinity for binding sites of differing length, although accounting for length differences in binding sites is not a fundamental limitation of this method.

For the auxiliary binding sites, the effect discussed in section 2.2 is reflected in this map: increasing the Lac repressor affinity for the O_2 binding site does not increase repression. Mutations in almost all positions can decrease repression, and no base pair substitutions significantly increase the repression

level. Mutations in the O_3 binding site have the potential to either increase or decrease the repression level. With respect to the RNAP binding site, we can see that, as expected, the most influential base pairs surround the well characterized -35 and -10 boxes. The CRP binding site overlaps three base pairs with the upstream Lac repressor auxiliary operator. As the heat-map reveals, the binding energy is relatively insensitive to changes in those base pairs, so we assume independence when calculating the binding energy and capture the synergy between the Lac repressor bound to O_3 and CRP with an interaction energy term.

The construction of the sequence to phenotype map enables us to predict the evolvability of the *lac* promoter region. We calculated the effect that all possible double mutations would have in the regulation of the operon, again with respect to the predicted repression level of the reference strain MG1655. Figure 5 shows what we call the ‘phenotype change distribution’ obtained by mutating one or two base pairs from the reference sequence, under the assumption of same growth rate and transcription factor copy numbers as the reference strain. The distribution peaks at zero for both cases, meaning that the majority of mutations are predicted not to change the repression level with respect to the reference strain,

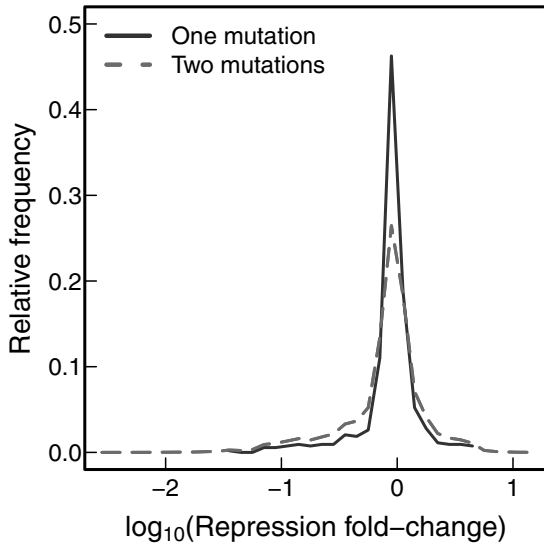


Figure 5. Phenotype change distribution. Relative frequency of the predicted changes in repression level by mutating one (solid blue line) or two (dashed red line) base pairs from the reference sequence (MG1655 promoter region).

and would result in genetic drift. However it is interesting to note that the range of repression values predicted by the model with only one mutation varied between 30 times lower and 4.6 times higher than the reference value, and with two mutations the repression varied between 345 times lower and 15 times higher than the reference value. This suggests that regulation of this operon could rapidly adapt and fine tune regulation given appropriate selection.

2.4. Promoter sequence variability of natural isolates and available sequenced genomes

In order to explore the natural variability of this regulatory circuit, we analyzed the *lac* promoter region of 22 wild-type *E. coli* strains which were isolated from different organisms [30], along with 69 fully sequenced *E. coli* strains (including MG1655) available online (http://ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). Figure 6 summarizes the sequencing results; for comparison, we plot the ‘genotype to phenotype map’ from figure 4(b) to gain insight into how the sequence variability influences regulation in these strains. Figure 6(b) shows the relative frequency of single nucleotide polymorphisms with respect to the consensus sequence. Qualitatively we can appreciate that the mutations found in these strains fell mostly within base pairs which according to the model weakly regulated expression. To quantify this observation we mapped the sequences to their corresponding binding energies. As shown in figure 6(c) the distribution of parameters is such that the observed mutations result in relatively small changes to the binding energies, less than $1 k_B T$ relative to the reference sequence, except for the O_3 binding energy that is predicted to increase $> 1 k_B T$ in 16 strains.

Table 1. Lac repressor copy number as measured with the immunodot blots and doubling time of the eight strains with measured repression level shown in figure 1(b). The errors represent the standard error of three independent experiments.

Strain	Repressor/cell	Doubling time (min)
Lab strain	21 ± 4	29.1 ± 0.2
Bat	12 ± 1	27.5 ± 0.2
Human-MA	20 ± 4	35.6 ± 0.6
Human-NY	23 ± 4	41.5 ± 0.4
Human-Sweden	28 ± 1	34.2 ± 0.3
Jaguar	21 ± 3	32.0 ± 0.2
Opossum	26 ± 2	33.5 ± 0.2
Perching bird	24 ± 4	30.2 ± 0.3

2.5. Does the model account for variability in the natural isolates?

Next we further characterized the eight strains from figure 1(b) in order to determine if the observed variability in regulation could be accounted for in the model (see section S2, available from stacks.iop.org/PhysBio/11/026005/mmedia for details on the 16S rRNA of this subset of strains). In particular, we measured the *in vivo* repressor copy number with quantitative immunoblots (see section 4) and the growth rate. Table 1 shows the measured repressor copy number and the doubling time for these strains.

Using the thermodynamic model by taking into account the repressor copy number, the promoter sequence and the growth rate, we predict the repression level for each of the isolates measured in figure 1(b). In figure 7 we plot these predicted values versus the experimental measurements. We find that the model accounts for the overall trends observed in the isolates, with the predictions for six of eight strains falling within two standard deviations of the measurements. A few of the measured repression values fall outside of the prediction, suggesting that the model may not capture the full set of control parameters operating in all of the strains.

2.6. Exploring the variability among different species

We extended our analysis to different microbial species with similar *lac* promoter architectures. After identifying bacterial species containing the *lac* repressor, we used the *sort-seq* derived energy matrices shown in figure 4(a) to identify the positions of the transcription factor binding sites in each of these candidate strains. We identified a set of eight species whose *lac* promoter architecture was similar to *E. coli*. Figure 8 shows the 16S rRNA phylogenetic tree for these strains. The predicted change in regulation was calculated for these strains using the model whose states are shown in figure 2, the energy matrices in figure 4(a), and assuming all strains have the same growth rate and transcription factor copy numbers as the lab strain MG1655. The repression level relative to *E. coli* among these species is predicted to increase as much as a factor of ≈ 20 and decrease as much as a factor of ≈ 4 . Regulation of the operon seems to follow phylogenetic patterns in the 16S rRNA tree, with *E. coli* relatives having a similar predicted repression level, *Citrobacter* evolved to increase repression, and *Salmonella* evolved to decrease repression.

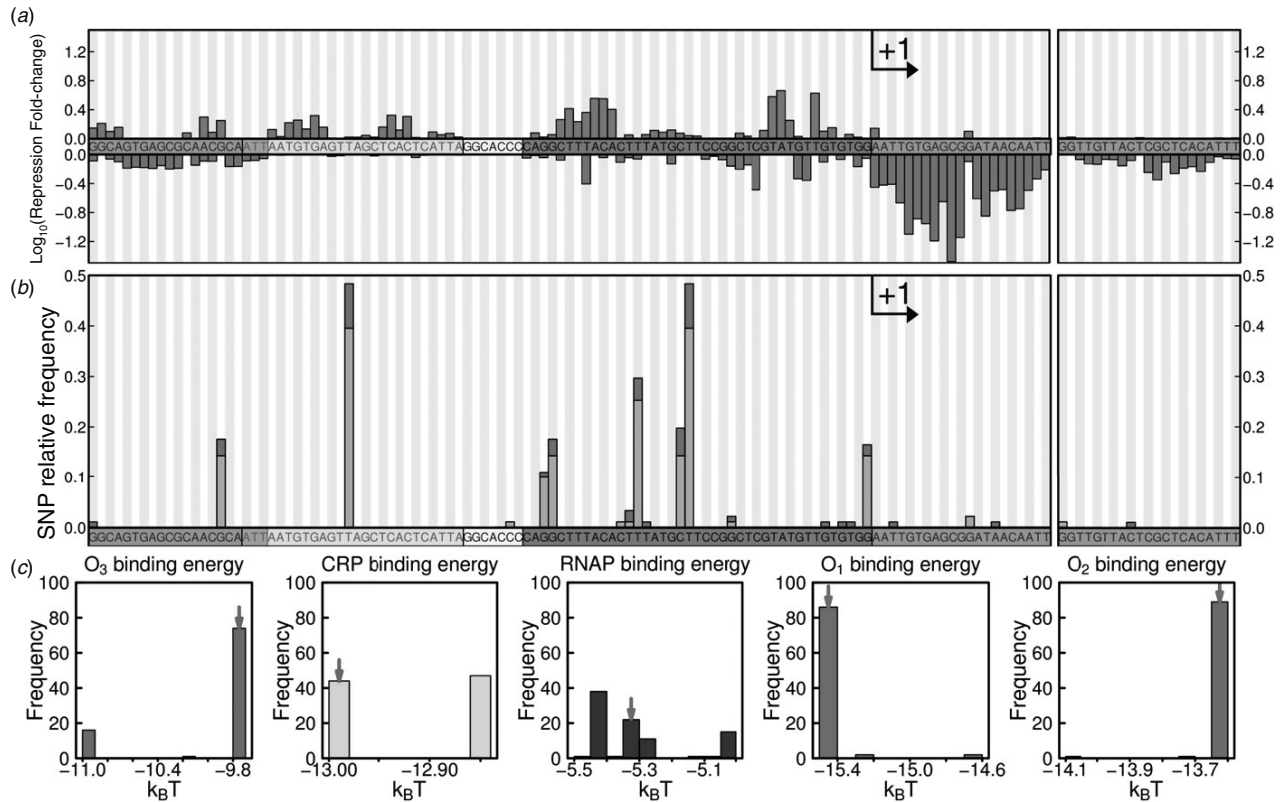


Figure 6. Mutational landscape of the regulatory region of the *lac* operon. (a) The genotype to phenotype map is reproduced from figure 4(b) in order to show how each base pair in the region influences gene regulation. (b) Comparing the sequence of the *lac* promoter from 91 *E. coli* strains identifies which base pairs were mutated in this region. The height of the bars represent the relative frequency of a mutation with respect to the consensus sequence. The red part of each bar represents the 22 natural isolates from different hosts [30] and the light blue part of these bars represents the 69 fully sequenced genomes (http://ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). Color coding of the binding sites and the transcription start site is as in figure 4. (c) Using the energy matrices of figure 4(a), we calculate the variability of protein binding energies for all sequences. The red arrow indicates reference binding energies for control strain MG1655.

3. Discussion

The approach presented here combines thermodynamic models of gene regulation with energy matrices generated with *sort-seq* to produce a single-base pair resolution picture of the role that each position of the promoter region has in regulation. These types of models based on equilibrium statistical mechanics have been used previously for the *lac* operon [19, 25], here we expanded the model to account for important cellular parameters such as growth rate, the binding site strengths of all transcription factors, and the binding site strength of RNAP. Thermodynamic models are functions of the natural variables of the system as opposed to the widely used phenomenological Hill functions [49], where it is less straightforward how changes to a promoter region translate to changes in regulatory parameters such as K_M , the half saturation constant, and n , the Hill coefficient. Currently our model assumes that protein–protein interactions and DNA looping energies are kept constant, but these variables could also be a function of the promoter sequence, affecting the positioning of the transcription factors and therefore their interactions with the other molecules involved.

The underlying framework developed here can be applied to any type of architecture. Here we use the *lac* operon because

it is well characterized. There is no reason to believe that this approach could not be extended to other regulatory regions, however such an effort would require extensive quantitative characterization of the control parameters of each genetic circuit, such as protein copy numbers, interaction energies, and binding affinities. Although this level of characterization requires additional experimental effort, we believe that developing such predictive, single-base pair models of gene regulation can lead to significant insights into how genetic circuits function, interact with each other, and evolve.

The majority of the natural variability found among the sequenced promoters tended to fall in bases predicted to have low impact on overall regulation, as shown in figure 6. As an example the highly conserved mutation in the CRP binding energy or the mutations along the RNAP binding site are predicted to change the binding energy by less than $1 k_B T$, having a very low impact on the repression level. With respect to the repressor binding sites, among the sequenced natural isolates only one mutation was found in the O_2 binding site. Unlike the O_1 and O_3 operators, the evolution of O_2 may be constrained given that its sequence encodes both gene regulatory information and is part of the coding region of the β -galactosidase gene.

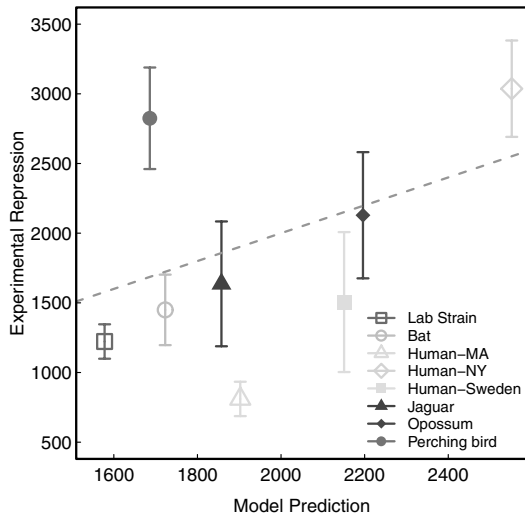
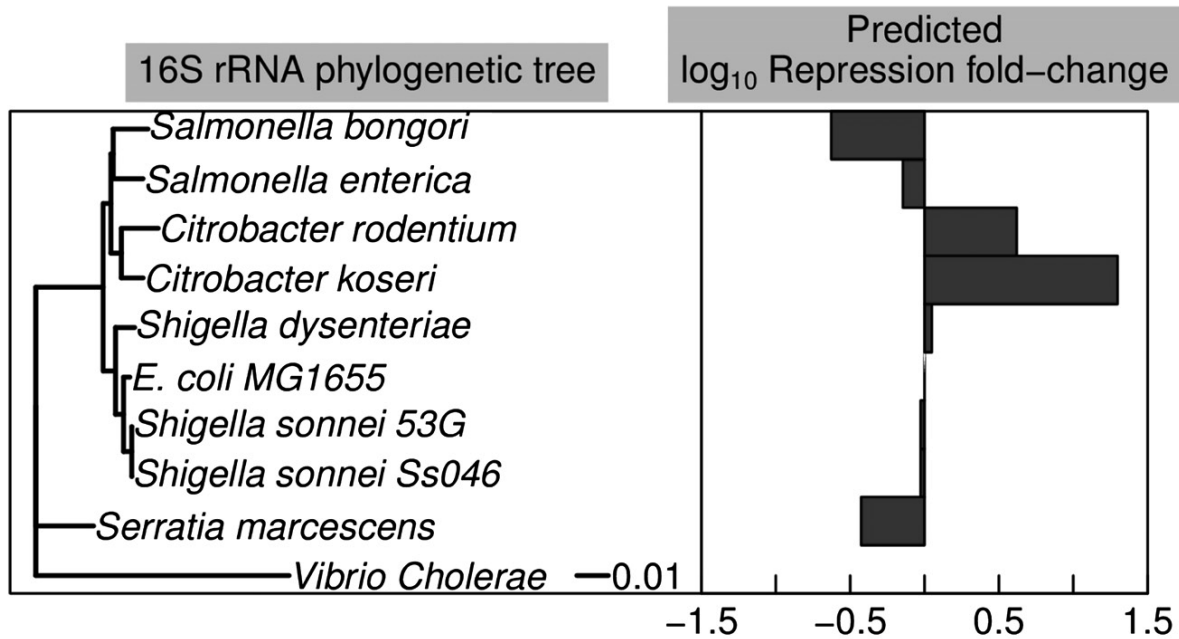


Figure 7. Comparison of model predictions with experimental measurements. Error bars represent the standard deviation of at least three independent measurements each with three replicates. The dotted line plots $x = y$.

As shown in figure 7, after taking into account the variability in the promoter sequence, changes in the repressor copy number, and changes in the growth rate the model accounts for most of the variability in regulation for the majority of the isolates. Linear regression of the entire experimental dataset weighted by the inverse of their standard deviation gives a slope of 1.26 with an R^2 of 0.24. It can be seen that many of the points fall close to or on the $x = y$ line, indicating that the poor fit is a result of a few outliers within the dataset. Removing the outliers (Perching bird, Human-MA, and Human-NY) results in a best fit line of slope 1.05 with R^2 0.74, reiterating that the model is consistent with the phenotype of five of eight isolates. It is interesting that the three isolates whose regulatory outputs were predicted poorly by the model (Perching bird, Human-MA, and Human-NY in figure 7) all have identical promoter sequences, which is the consensus promoter sequence as shown in figure S1 (available from stacks.iop.org/PhysBio/11/026005/mmedia). Although these three strains have identical sequences, two strains repressed more than predicted and the other strain repressed less. This indicates there are likely other cellular

(a)



(b)

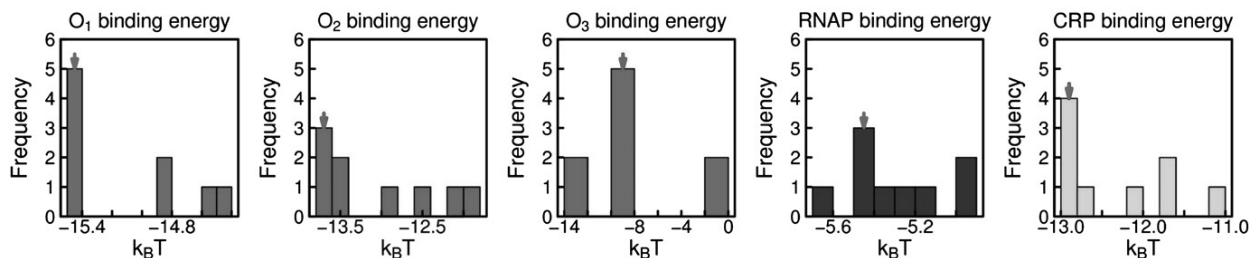


Figure 8. Predicted variability among different microbial species based on genome sequences and our model for regulation derived for *E. coli*. (a) On the left a 16S rRNA phylogenetic tree of diverse species with a similar *lac* promoter architecture done with the neighbor-joining algorithm. *Vibrio cholerae* was used as an outgroup species. The scale bar represents the relative number of substitutions per sequence. On the right the predicted \log_{10} fold-change in repression with respect to *E. coli* MG1655 assuming the same growth rate and transcription factor copy numbers. The outgroup species fold-change was not calculated. (b) Parameter distribution calculated using the promoter region sequence and the energy matrices. The red arrow indicates the MG1655 reference value. Strains lacking a binding site were binned as zero.

parameters that influence gene expression levels that are not included in the model. Currently the model cannot take into account variation in the protein structure of the transcription factors or the RNAP and its sigma factors. Changes in these proteins could account for some of the discrepancies between the model and the observed levels of regulation. It is likely that some global parameters that modulate transcriptional outputs which are not accounted for in the model also contribute to the disagreement with model predictions. We note that repression is a measurement of expression relative to expression in the absence of the repressor. This definition enables us to isolate the role of a particular transcription factor in regulation. Therefore, as discussed in section S8 (available from stacks.iop.org/PhysBio/11/026005/mmedia), some global regulatory parameters such as ribosomal binding sites of the relevant genes and variables such as the ribosome copy number should not impact repression levels.

From an evolutionary perspective, it is interesting that the regulation seems to be more sensitive to changes in the activator binding energy than to the activator protein copy number, as shown in figure 3. This result might be attributed to the nature of this transcription factor. CRP is known to be a 'global' transcription factor that regulates >50% of the *E. coli* transcription units [50]. Given its important global role in the structure of the transcriptome, changing the copy number of CRP would have a global impact on expression whereas tuning its binding affinity at a particular regulatory region has a local impact on one promoter. The regulatory knob of CRP copy number not influencing expression at the *lac* operon indicates this regulatory region may have evolved to be robust against changes in this global regulatory parameter.

The fact that the O_3 operator has the possibility to change in both directions (greater or lower affinity) as reflected in figure 4(b) suggests plasticity of the operon, allowing it to evolve according to environmental conditions. In fact this parameter changed the most among the related microbial species as shown in figure 8(b), having species such as *Citrobacter koseri* with an operator predicted to be $5 k_B T$ stronger than the reference value, and other species such as *Salmonella bongori* that completely lost this binding site. Although we do not yet know whether these regulatory predictions will be borne out in experimental measurements, this analysis demonstrates the utility of our sequence-to-phenotype map in interpreting the consequences of variability within the regulatory regions of sequenced genomes.

To the best of our knowledge figure 5 shows the first quantification of how easily regulation can change given one or two point mutations along the entire promoter region. Previous studies were limited to a subset of base pairs in the Lac repressor operators and two amino acid substitutions in the Lac repressor [51]. The distribution of predicted phenotypes is very sharp close to the reference value, as a consequence the majority of the possible mutations would not be selected on. But given that regulation can change by an order of magnitude or more in both directions (increased or decreased repression) with only two mutations, changing the regulatory region of the gene could function as a fast response strategy of adaptation.

It is known from previous work that *lac* operon expression can have an impact on cell fitness [20–22, 24]. Under

laboratory conditions, high expression of the *lac* operon resulted in loss of fitness due to expression of *lacY*, a transporter which imports lactose into the cell. This would suggest regulation is essential to avoid the negative consequences of *lacY* overexpression, and tight regulation would be selected. However it is possible that natural selection would act also to modulate the magnitude of the response. Strains exposed to environments with periodical bursts of lactose could trigger instantly a high gene dosage, resulting in a steeper slope on an induction curve, while strains rarely exposed to lactose would have a moderate response, i.e. a less steep induction curve. Our exploration and prediction of regulatory phenotypes in sequenced genomes shows that the biggest changes in regulation were found to increase repression (see figure 6(c)), suggesting that lactose might not be present regularly in the natural environment of some strains.

The combination of thermodynamic models with *sort-seq* generated energy matrices presented here promises to be an useful tool to study the evolution of gene regulation. This theoretical framework allows us to explore the effect that the modification of control parameters can have on the expression levels, and to predict how point mutations in gene promoter regions enable cells to evolve their gene regulatory circuits.

4. Materials and methods

4.1. Growth conditions

Unless otherwise indicated, all experiments started by inoculating the strains from frozen stocks kept at -80°C . Cultures were grown overnight in Luria Broth (EMD, Gibbstown, NJ) at 37°C with shaking at 250 rpm. In all of the experiments these cultures were used to inoculate three replicates for each of the relevant conditions, diluting them 1:3000 into 3 mL of M9 buffer (2 mM MgSO_4 , 0.10 mM CaCl_2 , 48 mM Na_2HPO_4 , 22 mM KH_2PO_4 , 8.6 mM NaCl, 19 mM NH_4Cl) with 0.5% glucose and 0.2% casamino acids (here referred to as 'supplemented M9'). Cells were cultured at 37°C with shaking at 250 rpm and harvested at the indicated OD_{600} .

4.2. Gene expression measurements

To perform the LacZ assay we followed the protocol used by Garcia and Phillips [26]. Strains were grown in supplemented M9 for approximately ten generations and harvested at an OD_{600} around 0.4. A volume of the cells was added to Z-buffer (60 mM Na_2HPO_4 , 40 mM NaH_2PO_4 , 10 mM KCl, 1 mM MgSO_4 , 50 mM β -mercaptoethanol, pH 7.0) for a total volume of 1 mL. For fully induced cells we used 50 μL and for uninduced cultures we concentrated the cells by spinning down 1 mL of culture and resuspending in Z-buffer. The cells were lysed by adding 25 μL of 0.1% SDS and 50 μL of chloroform and vortexing for 15 s. To obtain the readout, we added 200 μL of 4 mg mL^{-1} 2-nitrophenyl β -D-galactopyranoside (ONPG). Once the solution became noticeably yellow, we stopped the reaction by adding 200 μL of 2.5 M Na_2CO_3 .

To remove cell debris we spun down the tubes at $13000 \times g$ for 3 min. 200 μL of the supernatant were read at OD_{420} and

OD_{550} on a microplate reader (Tecan Safire2). The absolute activity of LacZ was measured in Miller units as

$$\text{MU} = 1000 \times \frac{OD_{420} - 1.75 \times OD_{550}}{t \times v \times OD_{600}} \times 0.826, \quad (4)$$

where t is the time we let the reaction run and v is the volume of cells used in mL. The factor of 0.826 adjusts for the concentration of ONP relative to the standard LacZ assay.

4.3. Measuring *in vivo* lac repressor copy number

To measure the repressor copy number of the natural isolates we followed the same procedure reported by Garcia and Phillips [26]. Strains were grown in 3 mL of supplemented M9 until they reached an $OD_{600} \approx 0.4 - 0.6$. Then they were transferred into 47 mL of warm media and grown at 37 °C to an OD_{600} of 0.4–0.6. 45 mL of culture were spun down at $6000 \times g$ and resuspended into 900 μL of breaking buffer (0.2 M Tris-HCl, 0.2 M KCl, 0.01 M Magnesium acetate, 5% glucose, 0.3 mM DTT, 50 mg/100 mL lysozyme, 50 $\mu\text{g L}^{-1}$ phenylmethanesulfonyl fluoride, pH 7.6).

Cells were lysed by performing four freeze-thaw cycles, adding 4 μL of a 2000 Kunitz/mL DNase solution and 40 μL of a 1 M MgCl_2 solution and incubating at 4 °C with mixing for 4 h after the first cycle. After the final cycle, cells were spun down at $13\,000 \times g$ for 45 min at 4 °C. We then obtained the supernatant and measured its volume. The pellet was resuspended in 900 μL of breaking buffer and again spun down at $15\,000 \times g$ for 45 min at 4 °C. In order to review the quality of the lysing process, 2 μL of this resuspended pellet was used as a control to ensure the luminescent signal of the resuspension was <30% of the sample.

To perform the immunoblot we pre-wet a nitrocellulose membrane (0.2 μM , Bio-Rad) in TBS buffer (20 mM Tris – HCl, 500 mM NaCl) and left it to air dry. For the standard curve a purified stock of Lac repressor tetramer [46] was serially diluted into HG105 (ΔlacI strain) lysate. 2 μL were spotted for each of the references and each of the samples. After the samples were visibly dried the membrane was blocked using TBST (20 mM Tris Base, 140 mM NaCl, 0.1% Tween 20, pH 7.6) +2% BSA +5% dry milk for 1 h at room temperature with mixing. We then incubated the membrane in a 1:1000 dilution of anti-LacI monoclonal antibody (from mouse; Millipore) in blocking solution for 1.5 h at room temperature with mixing. The membrane was gently washed with TBS \approx five times. To obtain the luminescent signal the membrane was incubated in a 1:2000 dilution of HRP-linked anti-mouse secondary antibody (GE Healthcare) for 1.5 h at room temperature with mixing and washed again \approx 5 times with TBS. The membrane was dried and developed with Thermo Scientific Super-Signal West Femto Substrate and imaged in a Bio-Rad VersaDoc 3000 system.

4.4. Constructing the *in vivo* lac repressor energy matrix

The energy matrix was inferred from *sort-seq* data in a manner analogous to methods described in Kinney PNAS 2010 [12]. Briefly, a library of mutant lac promoters was constructed in which the region [–100 : 25] (where coordinates are

with respect to the transcription start site) was mutagenized with a 3% mutation rate. The transcriptional activity of each mutant promoter was measured by flow cytometry using a GFP reporter. To fit the LacI energy matrix, we used a Markov chain Monte Carlo algorithm to fit an energy matrix to the LacI O_1 binding site by maximizing the mutual information between energies predicted by the matrix and flow cytometry measurements. The justification for maximizing mutual information is described in detail in [12, 52].

Acknowledgments

We would like to acknowledge Ron Milo, Niv Antonovsky, Adrian Jinich, Sushant Sundaresh, Joanna Robaszewski and Hernan Garcia for useful discussions. We are grateful to Valeria Souza (UNAM) for her kind donation of the *E. coli* strains. This work was supported by the National Institutes of Health, grant numbers DP1 OD000217A (Directors Pioneer Award), R01 GM085286 and R01 GM085286B (www.nih.gov). This work was also supported by the Donna and Benjamin M Rosen Center for Bioengineering at Caltech. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- [1] Thompson J R, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt D E, Benoit J, Sarma-Rupavtarm R, Distel D L and Polz M F 2005 Genotypic diversity within a natural coastal bacterioplankton population *Science* **307** 1311–3
- [2] Zelcbuch L *et al* 2013 Spanning high-dimensional expression space using ribosome-binding site combinatorics *Nucl. Acids Res.* **41** e98
- [3] Segal E and Widom J 2009 From DNA sequence to transcriptional behaviour: a quantitative approach *Nature Rev. Genet.* **10** 443–56
- [4] Hansen T E 2006 The evolution of genetic architecture *Annu. Rev. Ecol. Syst.* **37** 123–57
- [5] McAdams H H, Srinivasan B and Arkin A P 2004 The evolution of genetic regulatory systems in bacteria *Nature Rev. Genet.* **5** 169–78
- [6] Perez J C and Groisman E A 2009 Evolution of transcriptional regulatory circuits in bacteria *Cell* **138** 233–44
- [7] Ackers G K, Johnson A D and Shea A M 1982 Quantitative model for gene regulation by lambda phage repressor *Proc. Natl Acad. Sci. USA* **79** 1129–33
- [8] Buchler N E, Gerland U and Hwa T 2003 On schemes of combinatorial transcription logic *Proc. Natl Acad. Sci. USA* **100** 5136–41
- [9] Bintu L, Buchler N E, Garcia H G, Gerland U, Hwa T, Kondev J and Phillips R 2005 Transcriptional regulation by the numbers: models *Curr. Opin. Genet. Dev.* **15** 116–24
- [10] Bintu L, Buchler N E, Garcia H G, Gerland U, Hwa T, Kondev J, Kuhlman T and Phillips R 2005 Transcriptional regulation by the numbers: applications *Curr. Opin. Genet. Dev.* **15** 125–35
- [11] Sherman M S and Cohen B A 2012 Thermodynamic state ensemble models of cis-regulation *PLoS Comput. Biol.* **8** e1002407
- [12] Kinney J B, Murugan A, Callan C G and Cox E C 2010 Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence *Proc. Natl Acad. Sci. USA* **107** 9158–63

- [13] Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A and Segal E 2012 Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters *Nature Biotechnol.* **30** 521–30
- [14] Melnikov A *et al* 2012 Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay *Nature Biotechnol.* **30** 271–7
- [15] Brewster R C, Jones D L and Phillips R 2012 Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli* *PLoS Comput. Biol.* **8** e1002811
- [16] Wilson C J, Zhan H, Swint-Kruse L and Matthews K S 2007 The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding *Cell. Mol. Life Sci.* **64** 3–16
- [17] Reznikoff W S 1992 The lactose operon-controlling elements: a complex paradigm *Mol. Microbiol.* **6** 2419–22
- [18] Setty Y, Mayo A E, Surette M G and Alon U 2003 Detailed map of a cis-regulatory input function *Proc. Natl Acad. Sci. USA* **100** 7702–7
- [19] Kuhlman T, Zhang Z, Saier M H and Hwa T 2007 Combinatorial transcriptional control of the lactose operon of *Escherichia coli* *Proc. Natl Acad. Sci. USA* **104** 6043–8
- [20] Dean A M 1989 Selection and neutrality in lactose operons of *Escherichia coli* *Genetics* **123** 441–54 (PMID: 2513251)
- [21] Dekel E and Alon U 2005 Optimality and evolutionary tuning of the expression level of a protein *Nature* **436** 588–92
- [22] Perfeito L, Ghozzi S, Berg J, Schnetz K and Lässig M 2011 Nonlinear fitness landscape of a molecular pathway *PLoS Genet.* **7** 1–10
- [23] Poelwijk F J, Heyning P D, de Vos M G J, Kiviet D J and Tans S J 2011 Optimality and evolution of transcriptionally regulated gene expression *BMC Syst. Biol.* **5** 128
- [24] Eames M and Kortemme T 2012 Cost-benefit tradeoffs in engineered *lac* operons *Science* **336** 911–5
- [25] Vilar J M G 2010 Accurate prediction of gene expression by integration of DNA sequence statistics with detailed modeling of transcription regulation *Biophys. J.* **99** 2408–13
- [26] Garcia H G and Phillips R 2011 Quantitative dissection of the simple repression input–output function *Proc. Natl Acad. Sci. USA* **108** 12173–8
- [27] Boedicker J Q, Garcia H G and Phillips R 2013 Theoretical and experimental dissection of DNA loop-mediated repression *Phys. Rev. Lett.* **110** 018101
- [28] Tagami H and Aiba H 1995 Role of CRP in transcription activation at *Escherichia coli lac* promoter: CRP is dispensable after the formation of open complex *Nucl. Acids Res.* **23** 599–605
- [29] Hudson J M and Fried M G 1990 Co-operative interactions between the catabolite gene activator protein and the *lac* repressor at the lactose promoter *J. Mol. Biol.* **214** 381–96
- [30] Souza V, Rocha M, Valera A and Eguarte L E 1999 Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents *Appl. Environ. Microbiol.* **65** 3373–85 (PMID: 10427022)
- [31] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U and Gaul U 2008 Predicting expression patterns from regulatory sequence in *Drosophila* segmentation *Nature* **451** 535–40
- [32] Saiz L and Vilar J M G 2008 *Ab initio* thermodynamic modeling of distal multisite transcription regulation *Nucl. Acids Res.* **36** 726–31
- [33] Vilar J M G and Leibler S 2003 DNA looping and physical constraints on transcription regulation *J. Mol. Biol.* **331** 981–9
- [34] Vilar J M G and Saiz L 2005 DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise *Curr. Opin. Genet. Dev.* **15** 136–44
- [35] Saiz L and Vilar J M G 2007 Multilevel deconstruction of the *in vivo* behavior of looped DNA-protein complexes *PLoS One* **2** e355
- [36] Saiz L and Vilar J M G 2006 DNA looping: the consequences and its control *Curr. Opin. Struct. Biol.* **16** 344–50
- [37] Saiz L, Rubi J M and Vilar J M G 2005 Inferring the *in vivo* looping properties of DNA *Proc. Natl Acad. Sci. USA* **102** 17642–5
- [38] Rydenfelt M, Cox R, Garcia H and Phillips R 2014 Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration *Phys. Rev. E* **89** 012702
- [39] Nelson H C and Sauer R T 1985 Lambda repressor mutations that increase the affinity and specificity of operator binding *Cell* **42** 549–58
- [40] Elf J, Gene-Wei Li and Xie X S 2007 Probing transcription factor dynamics at the single-molecule level in a living cell *Science* **316** 1191–4
- [41] Friedman L J and Gelles J 2012 Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation *Cell* **148** 679–89
- [42] Oehler S, Eismann E R, Krämer H and Müller-Hill B 1990 The three operators of the *lac* operon cooperate in repression *EMBO J.* **9** 973–9 (PMID: 2182324)
- [43] Bremer H and Dennis P P 1996 Modulation of chemical composition and other parameters of the cell by growth rate *Escherichia coli and Salmonella: Cellular and Molecular Biology* ed F C Neidhardt *et al* (Washington, DC: ASM) pp 1553–69
- [44] Klumpp S and Hwa T 2008 Growth-rate-dependent partitioning of RNA polymerases in bacteria *Proc. Natl Acad. Sci. USA* **105** 20245–50
- [45] Vossen K M, Stickle D F and Fried M G 1996 The mechanism of CAP-*lac*repressor binding cooperativity at the *E. coli* lactose promoter *J. Mol. Biol.* **255** 44–54
- [46] Johnson S, Lindén M and Phillips R 2012 Sequence dependence of transcription factor-mediated DNA looping *Nucl. Acids Res.* **40** 7728–38
- [47] Oehler S, Amouyal M, Kolkhof P, Von Wilcken-Bergmann B and Müller-Hill B 1994 Quality and position of the three *lac* operators of *E. coli* define efficiency of repression *EMBO J.* **13** 3348–55 (PMID: 8045263)
- [48] Poelwijk F J, Kiviet D J and Tans S J 2006 Evolutionary potential of a duplicated repressor-operator pair: simulating pathways using mutation data *PLoS Comput. Biol.* **2** e58
- [49] Santillán M 2008 On the use of the hill functions in mathematical models of gene regulatory networks *Math. Modelling Nat. Phenom.* **3** 85–97
- [50] Martínez-Antonio A and Collado-Vides J 2003 Identifying global regulators in transcriptional regulatory networks in bacteria *Curr. Opin. Microbiol.* **6** 482–9
- [51] Dawid A, Kiviet D J, Kogenaru M, de Vos M and Tans S J 2010 Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape *Chaos* **20** 026105
- [52] Kinney J B, Tkacik G and Callan C G 2007 Precise physical models of protein-DNA interaction from high-throughput data *Proc. Natl Acad. Sci. USA* **104** 501–6