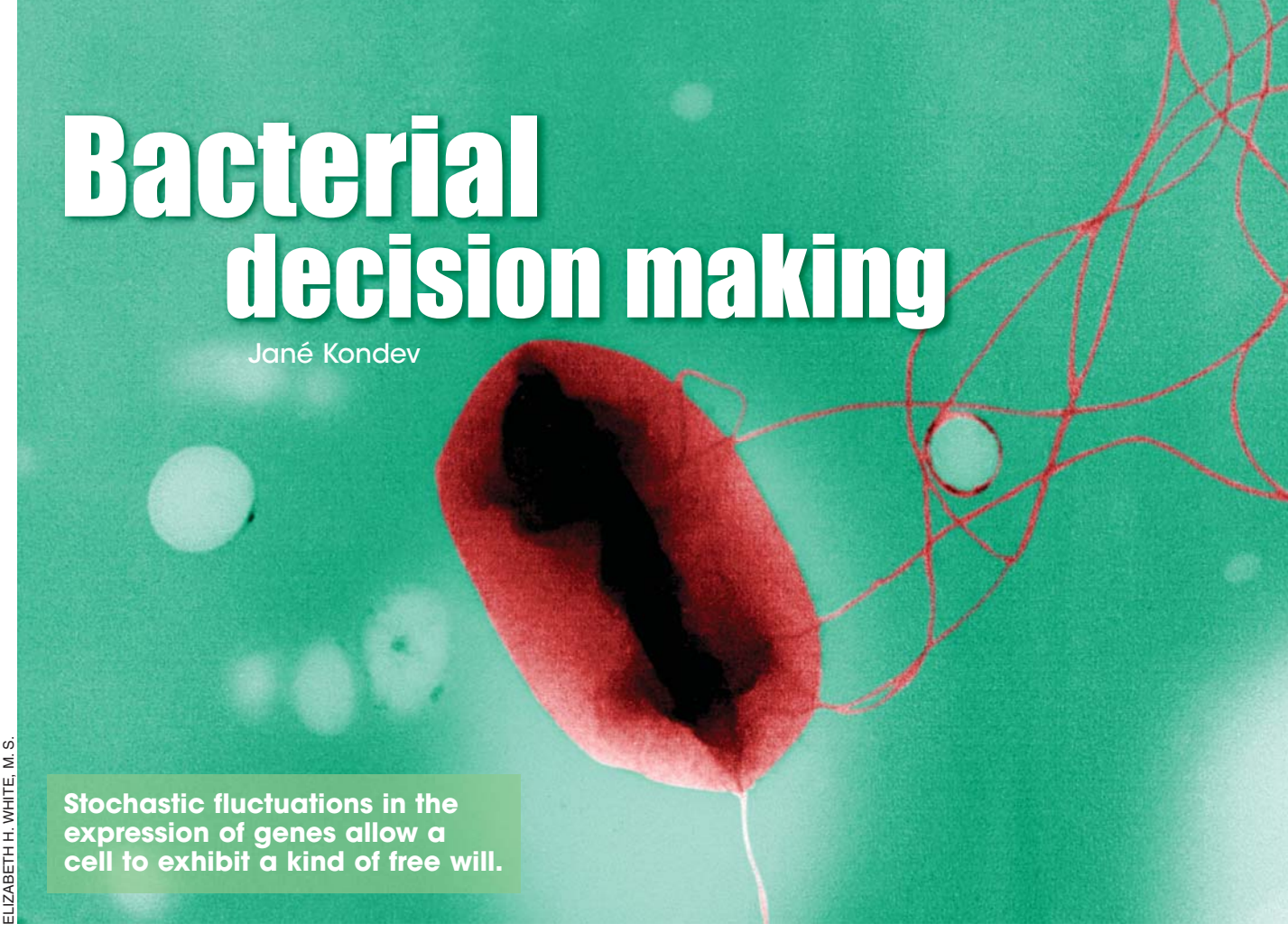
Bacterial decision making

Jané Kondev



ELIZABETH H. WHITE, M. S.

Stochastic fluctuations in the expression of genes allow a cell to exhibit a kind of free will.

Bacteria are single-celled organisms that make decisions all the time about where to swim, what to eat, and when to divide. They are also micron-sized containers filled with a million proteins and a few million base pairs of DNA, as well as RNA molecules, lipids, sugars, inorganic salts, and water. The question of how that bag of chemicals makes seemingly complex decisions has been the focus of biology for more than 50 years. Recently, experiments have begun to yield quantitative data that have led to models of the molecular-scale processes involved.

Physicists have taken the new developments as an invitation to join the party and are helping figure out how cells make decisions in many different contexts. They are bringing to molecular biology the interplay of quantitative models and experiments that has been so critical to the success of physics over the centuries. That interplay is pushing the field in unexpected new directions, and hopes are high that the general principles behind cellular decision making will soon reveal themselves.

To eat or not to eat sugar

In the 1940s *Escherichia coli* became established as the "hydrogen atom" of bacterial decision making during experiments performed at the Pasteur Institute in Paris.[1] At the time, Jacques Monod was working with the bacterium and measuring the growth of a cell culture in the presence of sugar. Typically, the population of bacteria would eat the sugar and double in size every hour or so. But when he measured the growth of *E. coli* in the presence of two different types of sugar, a peculiar thing happened. After a few hours, the exponential growth would pause for about an hour before resuming at a similar pace, a pattern that Monod published in his doctoral thesis and that is shown here in figure 1a. His key observation was to notice that the timing of the pauses was controlled by the ratio of the amounts of the two sugars available to the bacteria. For example, when the primary or preferred sugar, glucose, was present in smaller proportion than a secondary sugar, lactose, the pause would occur earlier.

Monod realized that the bacteria were initially unable to digest lactose, which explained the pause in the colony's growth when all the glucose was eaten. The pause was eventually identified as the time the bacteria required to manufacture the enzyme proteins needed to digest the lactose. The genes for those enzymes are encoded in the *E. coli* DNA, but they are initially inactive. The finding led to Monod's realization that some genes might be turned off in cells and await some chemical cue to switch them on.

In modern times such switching can be observed directly at the level of a single cell.[2] The series of snapshots shown in figure 1b, for instance, captures an *E. coli* cell in the act of producing one of the

Jané Kondev is a professor of physics at Brandeis University in Waltham, Massachusetts.

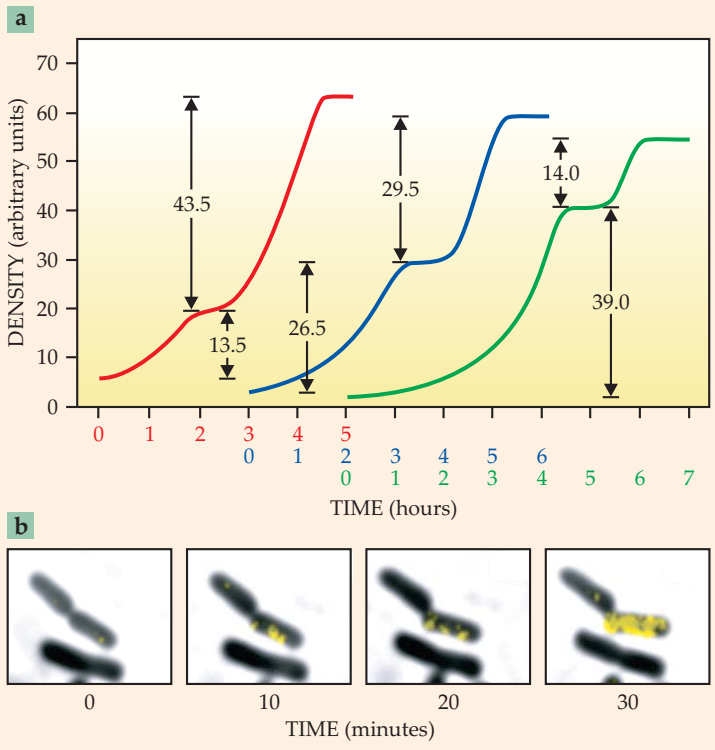


Figure 1. Bacterial decision making in action. (a) The growth of an *Escherichia coli* bacterial culture as a function of time occurs in two phases, separated by a pause. The amount of growth in each phase is proportional to the amount of glucose and lactose in the cells' environment. The two sugars were mixed in 1:3, 1:1, and 3:1 proportions in the three plots shown left to right; the densities of the cell cultures increased in the same proportions in the two phases. (Adapted from ref. 1.) **(b)** In the presence of a lactose surrogate, individual cells can switch from a state in which they are unable to digest lactose to a state in which they are able to consume the secondary sugar. Yellow indicates the amount of a fluorescently labeled protein, lactose permease, which is one of the enzymes needed by the cell to digest lactose. (Adapted from ref. 2.)

proteins associated with lactose digestion. The key question I'll address in this article is, What is the molecular basis by which a cell decides to switch a gene on? Although all the cells in figure 1b are genetically identical and experience the same environment, only one appears to be making the protein. As we'll see, that cellular individuality is a direct consequence of molecular noise that accompanies cellular decision making. The sources of the noise and its biological consequences are currently a hot topic of research. And statistical physics is proving to be an indispensable tool for producing mathematical models capable of explaining data from experiments that look at decisions made by individual cells.

The revolutionary idea that emerged from Monod's experiments is that although cells possess the genetic blueprints for making any number of proteins, they produce only a subset that depends on the conditions in the environment. In other words, gene expression, the process by which genes are read and their contents used to make proteins, is regulated. It is as if every gene has attached to it a spigot that the cell can adjust to control the amount of proteins it produces. Beyond bacterial decision making, the insight that cells turn genes on and off forms the molecular basis for the development of multicellular organisms. In that process, a single fertilized egg cell divides many times to give a collection of cells that are genetically identical yet have widely different properties; compare, for instance, blood cells and nerve cells. At the heart of the processes that generate such diversity is the on–off switching of genes.

Ultimately, the genes and the proteins that turn them on and off are simply molecules diffusing within the cell's interior. How the random thermal motion of those molecules and the interactions between them lead a bacterial cell to make a decision is an intriguing problem that has been addressed by a combination of careful quantitative experimentation and theory. But although the contours of a theory of gene expression are beginning to emerge for specific bacterial genes, the general principles behind the regulation of gene expression remain elusive.

Gene expression as computer

An operon is a stretch of DNA that contains one or more genes under the control of the same regulatory region of "promoter" DNA. A typical protein contains about 300–400 amino acids; since each amino acid is encoded by three bases, a typical gene is about 1000 bases long. The operon model proposed by Monod and his colleague François Jacob in 1961 posited that when lactose is absent the *E. coli* cell turns off the so-called *lac* genes that encode the enzymes necessary for lactose digestion. Only in the presence of lactose and the absence of glucose would that negative control be relieved. The model was initially silent about the identity of the molecular players involved in that negative control, but with the rise of molecular biology in the intervening decades, all the actors were eventually identified.

The key molecular players in the regulation of the *lac* genes are RNA polymerase, the Lac repressor, and CRP (cyclic-AMP receptor protein, where AMP is adenosine monophosphate), which together implement the computation described in figure 2a: Gene expression is turned on in the presence of lactose and the absence of glucose, and it's turned off for other combinations of the two inputs. RNA polymerase is the protein machine that binds to the promoter DNA adjacent to a particular gene and then transcribes the gene by moving along the DNA and assembling an RNA molecule whose sequence is complementary to that of the gene. The RNA molecule is then read by the ribosome, another molecular machine whose job is to assemble proteins from amino-acid building blocks following the genetic code.

The binding of the RNA polymerase to the promoter DNA associated with the *lac* genes is normally very weak, so the genes are not transcribed; gene expression is thus in the off state. The role of the CRP molecule is to increase the probability that the polymerase will bind to the DNA and thus turn the *lac* genes on. It does so by a favorable interaction—one that lowers the total free energy of the system—between it and the RNA polymerase. The

Lac repressor has the opposite effect. When it is bound to the DNA, it prevents the RNA polymerase from binding to the promoter DNA and therefore shuts the *lac* genes off.

The Lac repressor and CRP are "transcription factor" proteins whose role is to implement the logic function described in figure 2a. In the presence of glucose, CRP, like RNA polymerase, binds weakly to promoter DNA and the *lac* genes are off. In the absence of glucose, the cell makes cyclic-AMP molecules, which bind to CRP and alter its shape in such a way that it strongly binds to the promoter DNA and increases the likelihood that RNA polymerase becomes bound at the same time. In that sense, CRP is a glucose detector, although its function is much broader (see PHYSICS TODAY, October 2013, page 10).[3]

When lactose is not present, the *lac* genes are off regardless of whether glucose is present, because then the Lac repressor is bound to promoter DNA and takes up the space that the RNA polymerase would normally occupy before it begins transcription of the *lac* genes. In the presence of lactose, the Lac repressor binds to allolactose, one of the products of lactose digestion, and as a result the Lac repressor's shape is changed such that it no longer binds strongly to promoter DNA. Therefore, the absence of glucose and the presence of lactose is the only situation in which CRP is bound to promoter DNA and the Lac repressor is not, which leads to a high likelihood for polymerase to bind and turn on the *lac* genes.

The description of the *lac* operon as a digital computer that converts its four possible input states into a digital output—with *lac* genes on or off—is simple and powerful in its ability to make qualitative predictions about the system. But it is also incomplete. In particular, the computer model is incapable of predicting how the amount of gene expression, which is quantified by the number of lactose-digesting enzymes produced by *E. coli* per unit time, depends on the amount of glucose and lactose in the environment. Experiments that precisely measure the input–output relation have led to models of the *lac* operon based on statistical mechanics as opposed to Boolean logic.

Figure 2b shows the results of such a quantitative experiment by Terence Hwa and colleagues.[4] In 2007 they measured the amount of *lac*-gene expression as a function of cyclic AMP concentration, which, as described above, is an indicator of the amount of glucose in the environment, and of IPTG concentration (IPTG is a sugar that, like allolactose, binds to Lac repressors and weakens their binding to promoter DNA). While the data are qualitatively consistent with the Boolean description of the *lac* operon—the amount of gene expression is significant only when the concentrations of both cyclic AMP and IPTG are high—they also clearly show the analog nature of the computation done by the *E. coli* cell. The challenge to models of gene regulation is to compute the input–output function in figure 2b.

Statistical mechanics of repression

The *lac* operon is simple compared with the gene regulation that occurs in our cells, but it is still quite complicated in terms of the number of different molecular players involved. An experimental approach developed by molecular biologist Benno Müller-Hill and colleagues in the mid 1990s reduced that complexity by genetically manipulating the *lac* operon to take CRP out of the picture.[5] In such experiments, researchers measured the amount of expression of the *lac* genes solely as a function of the number of Lac repressors in the cell.

The experiments used a synthetic *lac* operon in which the *lac* genes are on in the absence of Lac repressor binding to the promoter DNA sequence and are off when the Lac repressor sits on the promoter,

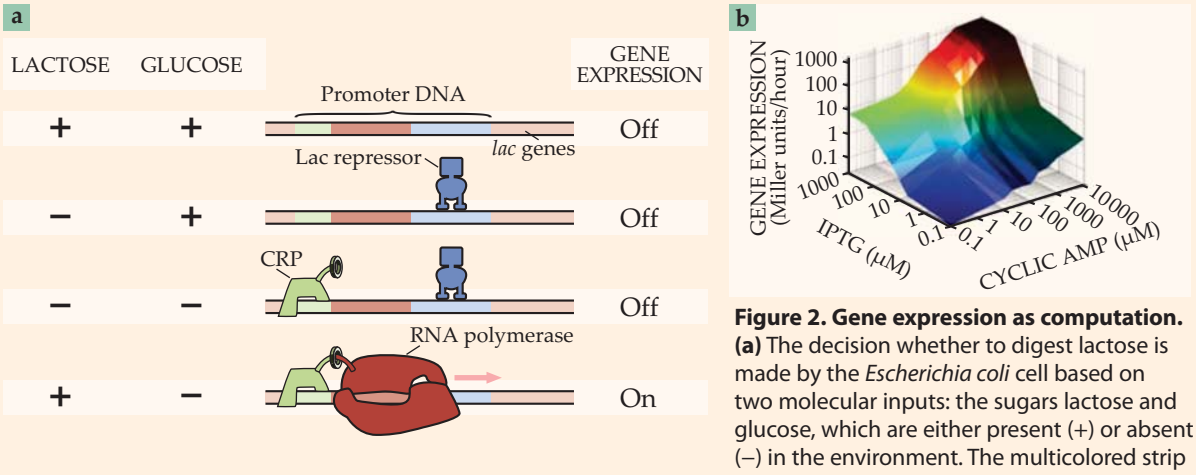


Figure 2. Gene expression as computation. (a) The decision whether to digest lactose is made by the *Escherichia coli* cell based on two molecular inputs: the sugars lactose and glucose, which are either present (+) or absent (–) in the environment. The multicolored strip represents a stretch of "promoter" DNA; the different colors correspond to DNA sequences that bind different proteins. The sugars control the binding of key proteins—Lac repressor, cyclic-AMP receptor protein (CRP), and RNA polymerase—which determine whether the *lac* genes are either expressed, in which case the digestion enzymes are produced, or not. (Adapted from ref. 12.) **(b)** Gene expression is not actually binary but a continuous function of the inputs. This plot shows the measured amount of gene expression as a function of the concentration of cyclic AMP (adenosine monophosphate), which is produced by *E. coli* when glucose is low, and the concentration of IPTG, a common experimental surrogate for lactose. (Adapted from ref. 4.)

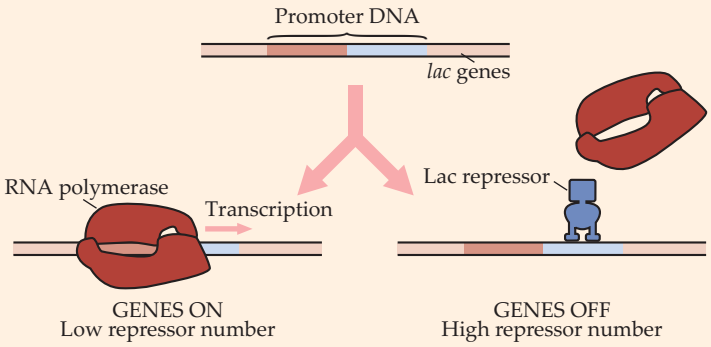


Figure 3. The repression of gene expression, modeled as a two-level system. Transcription, the first step in gene expression, begins with the enzyme RNA polymerase binding to the promoter DNA associated with a particular gene. Once bound, the RNA polymerase can begin reading the gene. The binding of the RNA polymerase to the promoter DNA, however, can be obstructed when a repressor protein is bound to part of the promoter. In that case, the promoter is unavailable to the RNA polymerase, transcription never commences, and gene expression is off. Only when the repressor unbinds from the DNA can the polymerase bind and turn gene expression back on. (Adapted from ref. 12.)

as illustrated by the two-state system in figure 3. Increasing the number of Lac repressors in the cell reduced the expression of the *lac* genes. So did changing the promoter DNA sequence so that it binds the Lac repressor molecules more strongly. The effect of the repressor on gene expression was quantified by computing the ratio of the amount of gene expression in cells with no repressor and the amount measured in cells with repressors. That ratio ranged continuously between 1.3 and 4700, from weak to strong repression, as the number of Lac repressors and the promoter DNA sequence were changed. Clearly, a Boolean description of the gene regulatory process does not explain the outcome of those experiments. What is needed instead is a mathematical function that relates the number of Lac repressors in the cell and their binding affinity with promoter DNA to the amount of repression.

Because transcription, the first step in gene expression, only occurs when the promoter is in the on state—that is, available for RNA polymerase to bind—the amount of gene expression is proportional to the fraction of the time the gene is on. Fortunately, one can turn to statistical mechanics to compute the probability of that state because the binding of the Lac repressor with promoter DNA can effectively be treated as a reaction in equilibrium.

Calculating the probability $p_{On}(R)$ of the *lac* genes being in the on state, where R is the number of repressors in the cell, is a simple exercise in the statistical mechanics of a two-level system. The two states correspond to gene expression being on or off, and the ratio of their probabilities is given by the Boltzmann formula,

$$\frac{p_{On}}{p_{Off}} = e^{-(G_{On} - G_{Off})/k_B T}, \qquad (1)$$

where the exponent contains the free-energy difference between the two states of the promoter DNA. To simplify the calculation of the free energies, let's assume that the repressor is always bound somewhere on the DNA; although that's not crucial to the argument, it happens to be true in the case of the Lac repressor in *E. coli*. When the repressor is attached to the promoter DNA, it has a binding energy ϵ_S, while the repressor's binding energy somewhere else (at a nonspecific site) on the DNA is ϵ_{NS}; the energies satisfy $\epsilon_S < \epsilon_{NS} < 0$ and the specific binding is stronger by virtue of a larger Boltzmann factor. The two promoter states also have different multiplicities—the number of different arrangements of the R repressors along the DNA—and therefore different entropies. When the promoter is in the on state, all R repressors are bound nonspecifically to the DNA and can be on any one of $N_{NS} \approx 5 \times 10^6$ sites corresponding to the number of base pairs that make up the *E. coli* genome; the number of such arrangements is given by the binomial coefficient $N_{NS}!/R!(N_{NS} - R)!$. Therefore, the free energy of the on state is

$$G_{On} = R\epsilon_{NS} - k_B T \ln\left(\frac{N_{NS}!}{R!(N_{NS} - R)!}\right). \qquad (2)$$

Similar reasoning leads to

$$G_{Off} = \epsilon_S + (R-1)\epsilon_{NS} - k_B T \ln\left(\frac{N_{NS}!}{(R-1)!(N_{NS} - R + 1)!}\right) \qquad (3)$$

for the free energy of the off state. The only difference is that one repressor is bound to promoter DNA, thereby switching the promoter off, while $R - 1$ repressors are bound nonspecifically. The difference lowers the energy by $\epsilon_S - \epsilon_{NS}$ and lowers the entropy as well.

With the free energies for the on and off states in hand, one can compute the ratio of their probabilities:

$$\frac{p_{On}}{p_{Off}} = \frac{N_{NS}}{R} e^{\Delta\epsilon/k_B T}, \qquad (4)$$

where $\Delta\epsilon \equiv \epsilon_S - \epsilon_{NS}$ and $N_{NS} \gg R$. The typical number of Lac repressors in an *E. coli* cell is $R = 10$, and the difference in energy between specific and nonspecific binding is about $-15k_B T$, which gives a ratio $p_{On}/p_{Off} \approx 0.15$. To obtain a formula that can be directly compared with experimental data, one can use equation 4 to compute the amount of repression $Rep(R)$, which is defined as the ratio of the amount of gene expression in the absence of Lac repressors in the cell to the amount of gene expression in the presence of R repressors:

$$Rep(R) = \frac{p_{On}(R=0)}{p_{On}(R)} = 1 + \frac{R}{N_{NS}} e^{-\Delta\epsilon/k_B T}. \qquad (5)$$

As first described by Jose Vilar and Stanislas Leibler,[6] that simple theoretical prediction accounts for the experiments by Müller-Hill's group in quantitative detail. More recent experiments by Hernan Garcia and Rob Phillips[7] have led to a precision test of the formula. Using the tools of genetic manipulation, they constructed a number of different *E. coli* strains, each differing in the number of Lac repressors produced by the cell. For each mutant strain, they measured the amount of *lac* gene expression and compared it with a strain with no Lac repressors. Equation 5 was then used to extract the binding energy $\Delta\epsilon$, which, if the statistical mechanics model

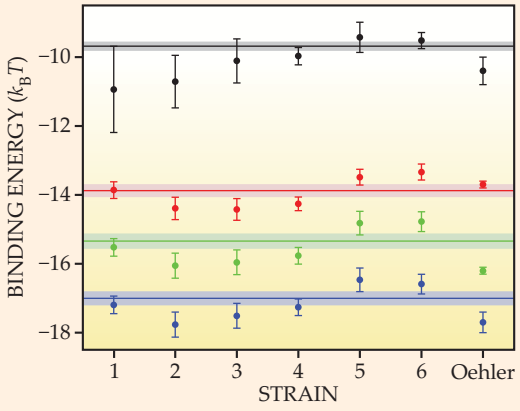


Figure 4. The repression model, tested. Six mutant strains of *Escherichia coli* were constructed so that the cells of each strain contain a different number R of Lac repressor proteins: $R \simeq 10$, 30, 60, 130, 610, and 870 for strains 1 through 6, respectively. For each strain, the amount of gene expression was measured and compared with the amount of gene expression by mutant cells that contained no Lac repressor. The ratio of the two is a measure of the repression of gene expression, from which the binding energy of Lac repressor to promoter DNA can be computed. A specific prediction of the calculation, borne out by the measurements, is that the binding energy should not depend on the strain of *E. coli* used. The different colors in the figure correspond to four different promoter sequences that bind the Lac repressor with different strengths; blue data points correspond to the strongest binding sequence, and black points to the weakest one. (Adapted from ref. 7, with data on the Oehler strain from ref. 5.)

is correct, should not depend on the strain used. To the extent that all the data points plotted in figure 4 lie on horizontal lines, the model is consistent with the data. Some discrepancies between theory and experiment appear in the plots, but none so egregious as to cast doubt on the model.

Does *E. coli* have free will?

In a bacterial colony whose gene expression is described probabilistically, one might expect to observe the stochastic switching between on and off states in real time. In 2005 Ido Golding and colleagues did just that[8] using fluorescently labeled RNA molecules that were produced from a synthetic gene they inserted into the *E. coli* DNA. By monitoring the rise and fall in fluorescence intensity from a single cell, the researchers were able to infer changes in the number of RNA molecules as a function of time over many hours, as shown in figure 5. They noticed that periods of RNA production were interspersed with long periods of no RNA production—an observation consistent with the simple picture of repression outlined in this article.

But simple repression is almost certainly not the cause of the observed switching, as the environment in which the cells were placed had more than enough sugar to ensure that the repressor proteins would be unable to bind to promoter DNA. Indeed, the molecular origins of the observed switching remain unknown. Still, even in the absence of a molecular-scale mechanism, it is interesting to take the switching as a fact of *E. coli* life and try to understand its consequences on the dynamics of gene expression.

We can start with a model of the *E. coli* cell as a simple chemical reactor. A gene in that reactor is transcribed, and corresponding RNA produced, at rate r. That RNA then degrades at a rate γ. (In *E. coli*, typical values are $r = 0.04$ min^{-1} and $\gamma = 0.2$ min^{-1}.)[9] To describe the dynamics of RNA production and degradation, consider the probability distribution $p(n,t)$ that the cell at time t has n RNA molecules of the gene in question. In a small time interval Δt, cells can either produce an additional RNA with probability $r\Delta t$ or have one of its n RNAs degraded with probability $n\gamma\Delta t$; the factor of n accounts for the fact that each molecule decays independently of the others with the same rate. Both processes reduce the probability that the cell has n RNAs, while RNA production in a cell with one less RNA or degradation in a cell with one extra RNA will increase $p(n,t)$. That time evolution of the probability reaches a steady state $p^*(n)$ when there is balance between RNA production and degradation: $rp^*(n-1) = n\gamma\, p^*(n)$. Solving the recursive relation yields the steady-state distribution of RNA molecules,

$$p^*(n) = \frac{(r/\gamma)^n}{n!} e^{-r/\gamma}, \tag{6}$$

which is a Poisson distribution with a mean number $\langle n \rangle = r/\gamma$, around 0.2 for a typical *E. coli* gene, or about one RNA molecule every five cells.[9]

The model of transcription as a simple chemical reactor that produces messenger RNA corresponds to the situation when the gene is always on. If the promoter, as in the case of the *lac* operon, can stochastically switch between a transcriptionally active on state and an inactive off state, the steady-state distribution of mRNA would no longer be Poissonian. It would instead be characterized by a larger variance for the same mean number of RNA molecules. In other words, stochastic switching between different promoter states increases the noise in the transcriptional output of the cell.

Recent research by Harvard University's Sunney Xie and colleagues measured the cell-to-cell variability in the number of RNA molecules for a variety of different genes in *E. coli* and found that for all the genes examined, the noise levels were elevated and quantitatively consistent with their promoters switching between on and off states.[9] Although that's an intriguing result, we're still far from understanding the underlying molecular processes that turn all those genes on and off.

What is needed are the same kinds of theoretically driven experiments that were done to measure the mean amount of gene expression over a cell population. The *lac* operon is poised for an important role in that effort. For example, the model of repression described in figure 3 makes specific predictions about how the noise will change if the number of repressors or the promoter DNA sequence is tuned.[10] The hope is that careful measurements of the noise and comparisons to theory will further elucidate the molecular mechanisms responsible for the regulation of gene expression in cells.

The stochastic nature of gene expression in cells

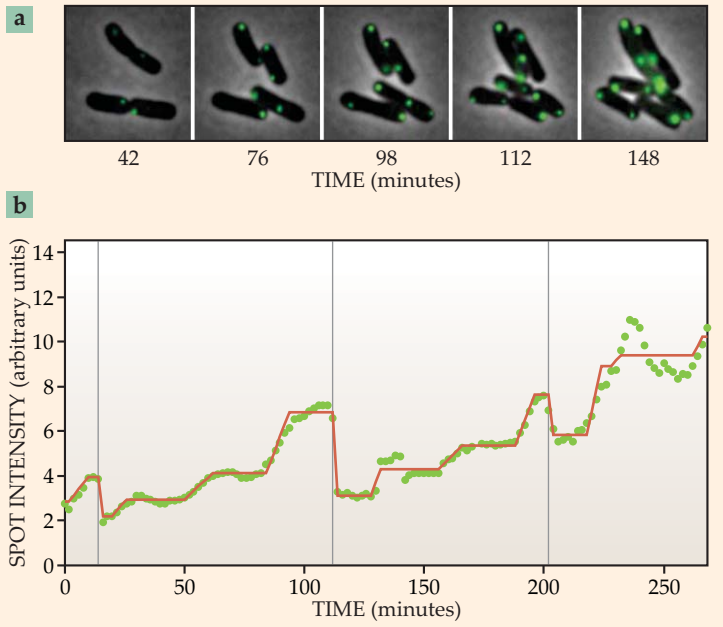


Figure 5. (a) Movie stills of *Escherichia coli* cells with fluorescently labeled RNA molecules capture the dynamics of transcription. Each green spot represents messenger RNA molecules whose number can be inferred from the fluorescence intensity. **(b)** That intensity is measured in a single cell as a function of time. The plot shows periods of RNA production interspersed among periods when the cell is transcriptionally inactive. Plateaus in the time-averaged (red) trace indicate those inactive periods, and the gray vertical lines mark times at which the cell divides, leading to a random partitioning of RNA molecules between the two daughter cells; the trace follows the number of RNAs in just one of the two daughters. (Adapted from ref. 8.)

is interesting not only as a diagnostic for discriminating between different models of regulation of gene expression but also for how it affects the physiology of those cells. The example shown in figure 1b, in which some cells become lactose eaters and others don't in the same environment, is a case in point. The result is a population of cells that, although genetically identical, may behave quite differently. It is as though each *E. coli* cell is free to choose between two diets, glucose or lactose, unencumbered by its genes or the environment.

Experiments in 2008 by Xie and colleagues reveal the molecular nature of that "free will" and trace it to the stochastic expression of *E. coli* genes.[2] One of the expressed *lac* genes codes for a protein, lactose permease, that transports lactose molecules into the cell. That protein therefore provides positive feedback in the expression of *lac* genes: The more lactose permease produced, the more lactose makes it into the cell, and the more likely it is that the *lac* promoter will be in the on state, which, in turn, leads to more expression of *lac* genes and more lactose permease molecules.

Thanks to that positive feedback, *E. coli* cells exist in two different steady states—one in which there are many permeases in the cell (the yellow cell in figure 1b), the other in which the number of permeases is low (the dark cells in 1b). Stochastic fluctuations in the expression of the *lac* genes—fluctuations, for instance, between an on and an off state of the promoter—can flip the switch and turn a lactose noneater to a lactose eater. That explains the movie stills in figure 1b, in which cells end up in two different states. Detailed measurements of the cell-to-cell fluctuations in the amount of lactose permease confirm the model's validity.

As a very different application of that kind of bacterial free will, Leibler and his colleagues demonstrated nearly a decade ago that *E. coli* cells can spontaneously switch to phenotypes that are resistant to antibiotics.[11] What distinguishes that type of antibiotic resistance from the more familiar kind often reported in the news is that the bacteria in this case are not genetic mutants. When the bacterial colony is treated with antibiotics, only the few cells that have switched to the antibiotic-resistant state survive. After the antibiotics have flushed through the system, some of the survivors switch back to the antibiotic-sensitive state, and a new colony eventually arises that is identical to the one present before antibiotics were introduced, with the majority of cells sensitive to antibiotics.

Using that type of bet-hedging strategy, bacteria are able to balance the penalty in reproductive speed that the antibiotic-resistant cells pay with the benefit they provide to the colony by being able to survive antibiotic treatment. How pervasive such strategies are in the living world and what role stochastic gene expression plays in allowing cells to gamble in that way are interesting questions. No doubt a mix of theory and experimentation will continue to yield surprises.

Gene expression by the numbers

Physics-based models are leading to more stringent tests of the molecular mechanisms responsible for gene expression than those provided by the qualitative models presented in biology textbooks. They also pave the way for the design of so-called synthetic genetic circuits, in which the proteins produced by the expression of one gene affect the expression of another. Such circuits hold the promise of bacterial cells capable of producing useful chemicals or combating diseased human cells, including cancerous cells. Whether this foray of physics into biology will lead to fundamentally new biological insights about gene expression remains to be seen.

References

1. B. Müller-Hill, *The* Lac *Operon: A Short History of a Genetic Paradigm*, Walter de Gruyter, New York (1996).
2. P. J. Choi et al., *Science* **322**, 442 (2008).
3. C. You et al., *Nature* **500**, 301 (2013).
4. T. Kuhlman et al., *Proc. Natl. Acad. Sci. USA* **104**, 6043 (2007).
5. S. Oehler et al., *EMBO J.* **13**, 3348 (1994).
6. J. M. G. Vilar, S. Leibler, *J. Mol. Biol.* **331**, 981 (2003).
7. H. G. Garcia, R. Phillips, *Proc. Natl. Acad. Sci. USA* **108**, 12173 (2011).
8. I. Golding et al., *Cell* **123**, 1025 (2005).
9. Y. Taniguchi et al., *Science* **329**, 533 (2010).
10. A. Sanchez et al., *PLOS Comput. Biol.* **7**(3), e1001100 (2011).
11. N. Q. Balaban et al., *Science* **305**, 1622 (2004).
12. R. Phillips et al., *Physical Biology of the Cell*, 2nd ed., Garland Science, New York (2013). ■