Review

Special issue – CellBio-X

Transcription by the numbers redux: experiments and calculations that surprise

Hernan G. Garcia[1], Alvaro Sanchez[2], Thomas Kuhlman[3], Jane Kondev[4] and Rob Phillips[5]

[1] Department of Physics, California Institute of Technology, Pasadena, CA 91125, USA
[2] Graduate Program in Biophysics and Structural Biology, Brandeis University, Waltham, MA 02454, USA
[3] Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA
[4] Department of Physics, Brandeis University, Waltham, MA 02454, USA
[5] Departments of Applied Physics and Bioengineering, California Institute of Technology, Pasadena, CA 91125, USA

The study of transcription has witnessed an explosion of quantitative effort both experimentally and theoretically. In this article we highlight some of the exciting recent experimental efforts in the study of transcription with an eye to the demands that such experiments put on theoretical models of transcription. From a modeling perspective, we focus on two broad classes of models: the so-called thermodynamic models that use statistical mechanics to reckon the level of gene expression as probabilities of promoter occupancy, and rate-equation treatments that focus on the temporal evolution of the activity of a given promoter and that make it possible to compute the distributions of messenger RNA and proteins. We consider several appealing case studies to illustrate how quantitative models have been used to dissect transcriptional regulation.

Introduction

The very existence of this special themed issue on CellBio-X hints at a growing belief in what one might call a Bio-X effect – the idea that somehow by attacking biological problems from a physical or quantitative perspective we will either refine our understanding of established biological processes or discover completely new effects of mechanisms. One way to view the possible significance of this emphasis on biological numeracy is by analogy to the different kinds of catch fisherman can expect when using nets or hooks of different types. Certain nets are sure to catch some fish and not others. By introducing new ways of fishing, or by casting these nets or hooks in new places, a different ocean is revealed. We argue here that the types of approaches reflected in this special issue provide a complementary biological net that can reveal things that are impossible to see using traditional verbal and pictorial descriptions.

The type of quantitative approaches in biology argued for above have been ballyhooed far and wide, whether in the pages of learned reports [1,2], new online resources [3], in a variety of books and articles [4–9] or via the establishment of new programs or courses at universities around the world [10–12]. But what is the basis for this growing enthusiasm for biological numeracy and what rewards has it delivered (or might it deliver in the future) in our understanding of cellular decision-making in particular?

Even as relative newcomers to the study of transcription, it is clear to all of us that, with the passing of each year, the rapid pace of technological advances is resulting in a new generation of impressive and beautiful experiments that are painting a much more nuanced picture of the regulatory steps exploited by cells as they make decisions. One thing is clear: many of these experiments challenge the conventional verbal and pictorial representations of gene expression. With this increasing reliance on systematic and precise measurements of gene expression [13–15] comes the possibility of asking entirely new classes of questions about how regulation works. Further, these approaches are beginning to suggest how regulatory networks can be engineered to create entirely new biological functions, one of the signature achievements of the synthetic biology approach. With the new-found emphasis on reporting the results of these experiments quantitatively a growing trend has emerged to use models that are described in the same quantitative language as the data.

To make this claim concrete, consider the example of transcriptional repressors that bind at two sites on the DNA simultaneously, thereby looping the intervening fragment of the genome. Elegant experiments have measured how the level of gene expression depends upon the length of the DNA loops in the *lac* operon, resulting in the authors noting serious differences between the *in vitro* and *in vivo* signatures of the underlying DNA mechanics [16]. Indeed, these and other similar experiments have served as the basis of more than a decade of effort aimed at a deeper understanding of biological action at a distance and, specifically, trying to reconcile the *in vitro* and *in vivo* views of DNA mechanics [17–24]. One of the ambitions of the present paper is to provide a series of examples of precisely this type where biological numeracy serves as the basis for

Corresponding author: Phillips, R. (phillips@pboc.caltech.edu).

0962-8924/$ – see front matter doi:10.1016/j.tcb.2010.07.002 Trends in Cell Biology, December 2010, Vol. 20, No. 12

asking new kinds of biological questions. The history of modern biology is replete with examples of this kind: Mendel counting peas with different traits, Morgan and Sturtevant tracking the frequencies of mutations in flies, Delbrück and Luria measuring the fluctuations in the number of bacteria resistant to viral infection, Hodgkin and Huxley measuring the electrical currents across cell membranes, to name but a few. In all these cases analysis of quantitative data from a quantitative perspective led to new biological insights.

In an earlier set of papers [19,20] we explored biological numeracy in the context of transcription using thermodynamic models [25,26]. Here we extend the arguments made there from the vantage point of the impressive experimental advances which have characterized the field since those articles were written. Some of these experimental advances include the direct observation of transcription at the single-molecule level [27,28], single-cell measurements on transcription which yield protein and mRNA distributions in a population of cells [29–31], high-throughput methods which permit the analysis of many architectural motifs, and an explosion of synthetic biology transcriptional architectures [32–35].

As a result of these powerful experimental advances there has also been a new round of model building aimed at responding to this next generation of measurements. It is now becoming routine to see extremely complicated diagrams of 'genetic networks' with vague and hopeful analogies to electronic circuits. What marks our understanding of such circuits and the electronic components that constitute them, however, is a reliable understanding of their input–output properties (or transfer function) [36]. Part of our mission is to explore the interplay between experimental and theoretical strategies for dissecting transcriptional regulation in a way that comments on the fruitfulness of such analogies.

One of the many ways in which new experimental methods are sharpening the questions we can ask about transcription centers on the fact that it is now possible to measure the distribution of gene products in a population of cells by watching cellular decision making at the single-cell level [14,37,38]. We argue that distributions provide yet another way to probe the mechanistic underpinnings of observed patterns of gene expression. Although the details are themselves fascinating, our primary emphasis here will be on the style of quantitative thinking used in attacking these problems. Further, with apologies to the many scientists whose work has propelled the field forward, we will focus on a small number of instructive case studies which we find are most sympathetic for illustrating our main arguments, and with no attempt at being comprehensive in our coverage of the literature.

In the next section we provide an overview of the use of thermodynamic models to study cellular decision making. The main purpose of this section is to show how the thermodynamic models have honed the questions we can ask about regulatory networks and have clarified our understanding, while at the same time bringing into relief certain surprises and paradoxes. The second main section focuses on both measurements and models in which time figures explicitly. Experiments have now reached the point where it is possible to watch the synthesis of individual mRNAs, for example, on a cell-by-cell basis. Both the individual trajectories and the distributions obtained by tallying up the behavior of many cells together pose challenges which fall outside the scope of the thermodynamic models but can be explored using rate equations that reckon how the transcriptional state of the system will change during a small instant of time, Δt.

Equilibrium models of gene expression

The ability to perform systematic experimental manipulation of the various parameters (such as transcription factor binding-site positions, strengths and concentrations) highlighted in Figure 1 has resulted in a variety of different measurements of the level of gene expression for a spectrum of promoters [33,39–44], although our discussion will often focus on the classic *lac* operon which has become a central quantitative testbed [18,45–50]. Within the framework of the thermodynamic models which compute the probability that RNA polymerase will occupy the promoter of interest, the simplest way to make a direct comparison between the measurements and models is through the vehicle of the fold-change which gives the ratio of the level of gene expression in the presence and absence of regulatory elements whose abundance serves as an experimental control knob. For the special case of simple repression considered in Figure 2A, the fold-change can be written simply as

$$\text{fold}-\text{change} = \frac{\text{gene expression}(R \neq 0)}{\text{gene expression}(R = 0)} \approx \left(1 + \frac{[R]}{K}\right)^{-1} \quad [1]$$

where $[R]$ is the concentration of repressors and K is an effective dissociation constant which is a measure of the affinity of repressors for their target binding sites. The origins of this formula are illustrated schematically in Figure 2B which shows how to take the cartoon representation of the various states of the promoter and to find their associated statistical weights as prescribed by the Boltzmann factor from equilibrium statistical mechanics [8,19,20]. Note that the concentration of polymerase does not enter Equation 1 because we are considering the 'weak promoter' approximation in which the affinity of RNA polymerase for the promoter is very weak [19,20]. With the Boltzmann factors in hand we can then compute the level of gene expression on the assumption that promoter occupancy and gene expression are linearly related [8,19,20].

How can we explore the potency of a formula such as that given in Equation 1 and the many other similar formulas highlighted in Figure 3? Several important case studies have been carried out using well-characterized bacterial promoters which permit a direct and meaningful comparison between the measurements and the result (and similar calculations and measurements have been made for more complex regulatory architectures as shown with a few examples in Figure 2). Note that we are vehemently opposed to the idea that the goal of a model is to 'fit the data'. Instead, in addition to the central aim of having a

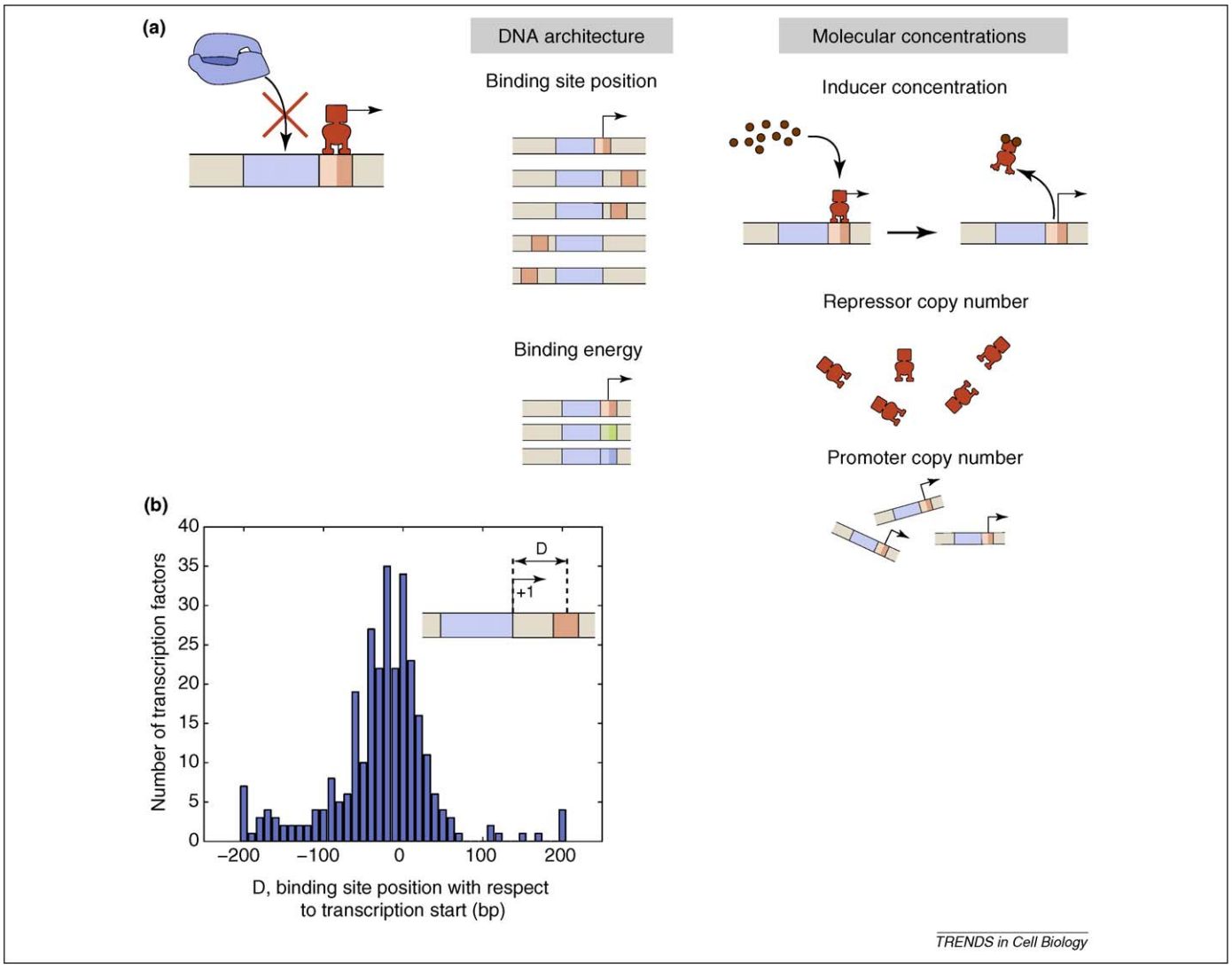

Figure 1. Transcriptional control knobs for experimentalists and theorists alike. **(a)** Schematic diagram illustrating the simple-repression architecture used as a case study in this paper. In this architecture a repressor bound to a binding site near the promoter excludes RNA polymerase from binding. The figure shows the ways in which key parameters such as the position and strength of binding sites, the copy numbers of genes and their associated transcription factors and inducer concentrations can be tuned to elicit different biological responses. **(b)** Experimental census of repression architectures in *E. coli*. The figure shows the distribution of binding site positions with respect to their target promoters for simple repressors in *E. coli*. This plot was generated based on data available on RegulonDB [87].

coherent 'story' about entire suites of data and the mechanisms that underlie them, a much more useful outcome of model building of the kind we describe here is that it leads to some surprise or paradox, which in turn might imply that the original cartoon representation of the regulatory process of interest is incomplete or flawed.

Some of the most complete quantitative examples of this overall strategy have taken place in the *lac* operon where, by eliminating the auxiliary operators, it is possible to construct a genetic circuit with the kind of simple repression highlighted in Figure 1A. Indeed, all of the 'control knobs' highlighted in that figure have been systematically altered experimentally and the resulting level of gene expression has been characterized as shown in Figures 4A and 4B.

In one of the most thorough studies to date, the *lac* operon was probed in quantitative detail by using the thermodynamic framework to dissect the way in which the molecular factors responsible for activation and repression interact. There is an unparalleled depth of knowledge and quantitative data available for all the molecular players and interactions responsible for the output of the *lac* system. This provides a unique opportunity to challenge the quantitative modeling perspective with real experimental data and thereby demonstrate that this classic and well-characterized biological system can have a new life as a proving ground for the techniques of physical biology. This case study is highlighted in Figure 4B. Here, through the judicious construction of a variety of mutants, the response of the *lac* system to each of its molecular components was carefully isolated and measured [42,44]. By comparing the results of these experiments to a thermodynamic model formulated based on the known properties and interactions of the system, it was shown how the complete output of the operon can be explained in quantitative detail as the result of the accumulation of multiple known interactions between the individual components. In the language of electronic circuits introduced above, this can be likened to predicting the properties of the circuit based upon the known quantitative characteristics of its constituent capacitances, resistances and so on.

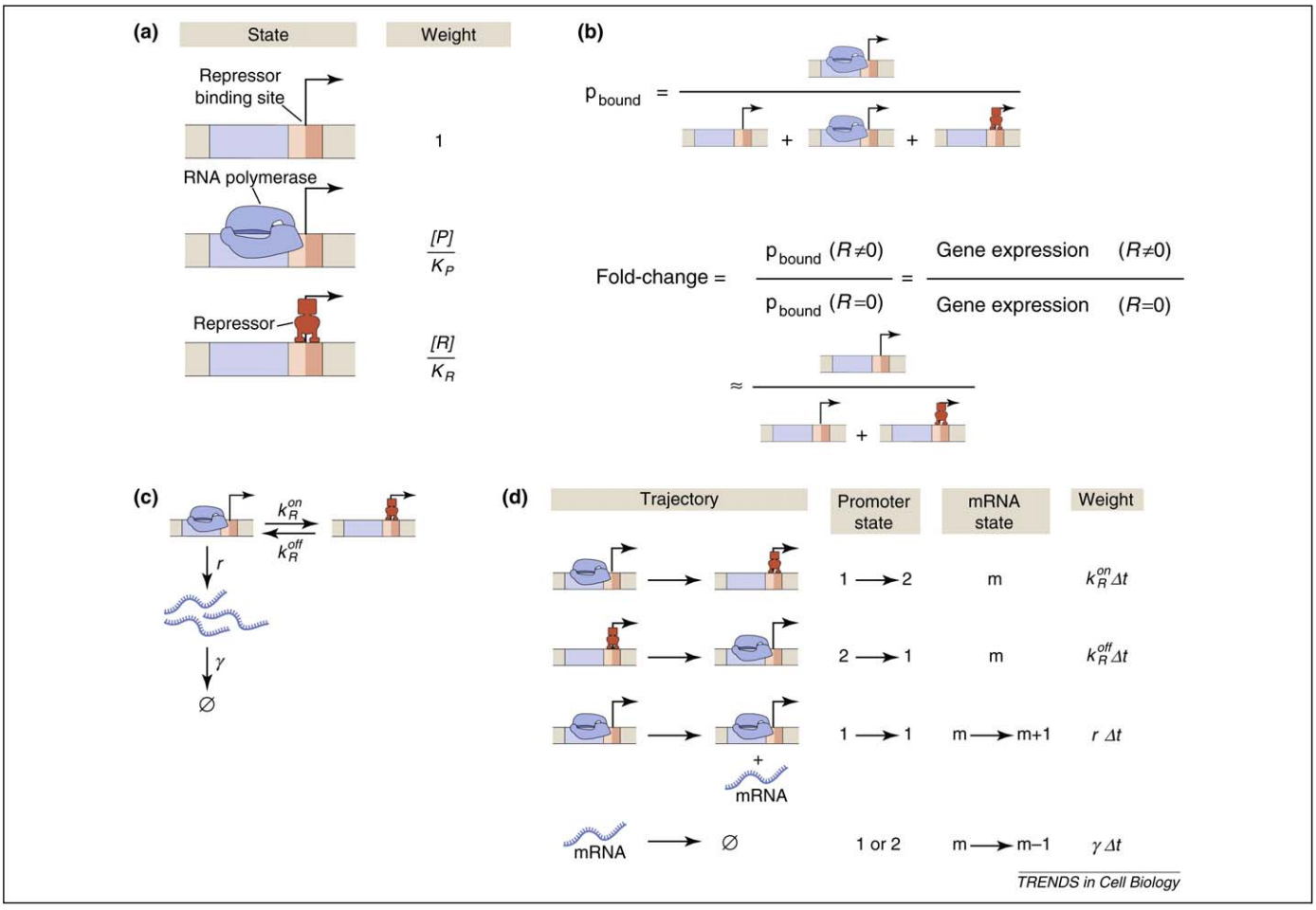

Figure 2. Modeling framework for simple repression. **(a)** States and weights for simple repression. In the thermodynamic models for promoter activity in simple repression there are three competing states. The parameters are the concentration of RNA polymerase, [*P*], its dissociation constant to the promoter K_P, the concentration of repressor, [*R*], and its dissociation constant to its operator, K_R. **(b)** Probability of finding RNA polymerase bound to the promoter (top) and resulting fold-change in gene expression (bottom) using the weak promoter approximation in which the probability of polymerase bound is negligible compared to the other two states. **(c)** Kinetic model of simple repression. When RNA polymerase is bound to the promoter it produces transcripts at a rate *r*, which decay at a rate γ. The promoter is switched off and on as a result of repressor binding with rates k_R^{on} and k_R^{off}. Note that in this version of the model we do not consider the 'empty' state in which the promoter has neither polymerase nor repressor. **(d)** Trajectories and weights for simple repression. The kinetic model shown in **(c)** is characterized by a number of different transitions that can occur in a time step Δt, each of which has a probability determined by the relevant rate constants. During each of these processes the state of both the promoter and the mRNA can change.

Conversely, one can imagine the characterization of a system in which much less is known about the constituents and their interactions. By comparing the results of experimental characterization to the predictions of a simple model capturing the known properties of the system, inconsistencies that arise can be a signal that our understanding of the system is incomplete. For example, the wild-type response of the *lac* system to changing concentrations of its repressor is extremely sensitive: the output serves essentially as a switch – it is completely off at high levels of repressor and abruptly switches on as concentrations are lowered. There is nothing inherently surprising about this observation, and such behavior might be expected from a cartoon model of the action of a repressor. However, when the sensitivity of the response is compared quantitatively to the prediction of simple modeling, it is seen that such a high level of sensitivity cannot result from the action of the repressor alone. It is only through the combined action and interaction of the repressor, positive feedback, and DNA looping that the high sensitivity can be explained.

Once a gene is transcribed it can be subject to further regulation before it is finally present in the cell as an active protein. One way in which genes can be post-transcriptionally regulated is through interaction with small untranslated RNAs, or sRNAs [51–53]. sRNAs can bind to the transcribed mRNAs and block their availability to the translational machinery or mark them for degradation. To understand these mechanisms the same thermodynamic ideas introduced above have recently been played out in the context of RNA regulation. Quantitative dissection of this kind of regulation [54,55] shows that the stoichiometric co-degradation of sRNA with their targets results in different quantitative regulatory characteristics than does regulation by protein transcription factors (which are not consumed during regulation and act catalytically), and thermodynamic modeling conveys a deeper understanding of this mode of regulation and its advantages and disadvantages relative to regulation by proteins.

One of the frustrating features of the experimental strategy used in the case studies described above, where the idea is to measure the gene regulation function (or the

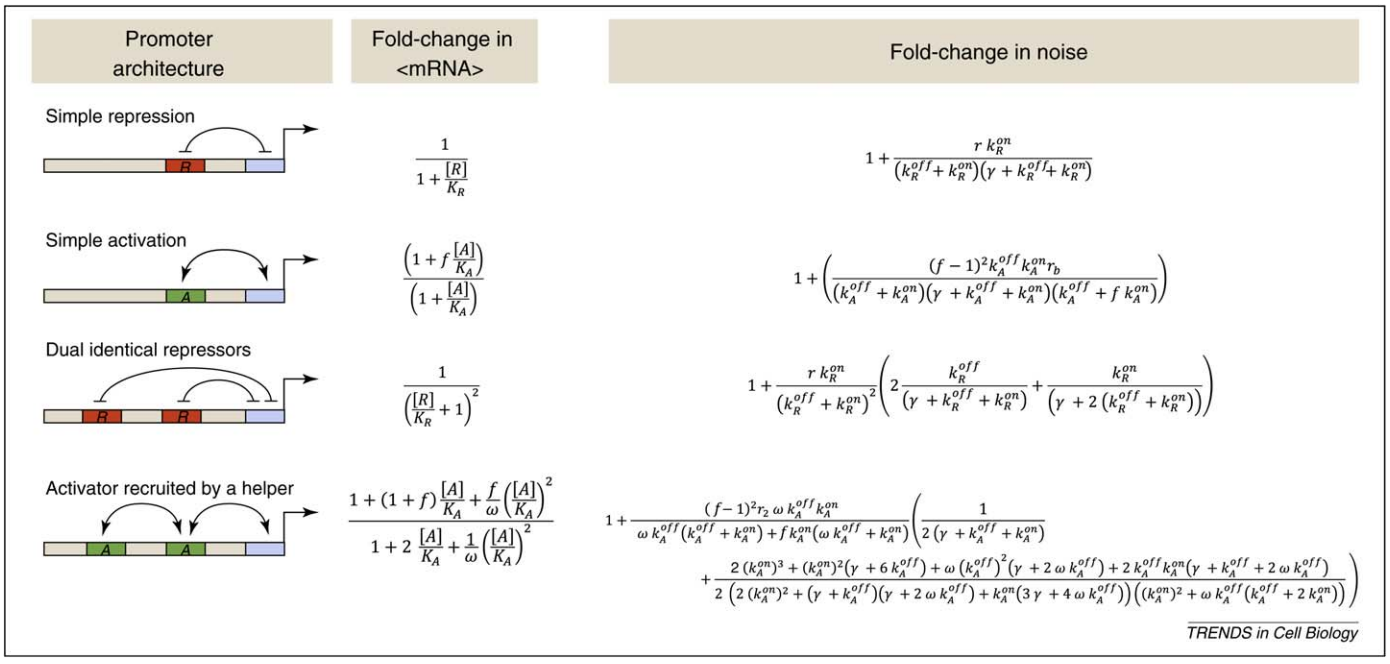

Figure 3. Formulae for transcriptional response. For each regulatory motif the thermodynamic models result in a simple expression for the fold-change in gene expression as a function of key parameters such as the concentration of the repressors and activators ([*R*] and [*A*], respectively) and the dissociation constants (K_R and K_A). The parameter ω accounts for cooperativity between transcription factors whereas f is the increase in transcription rate due to the presence of an activator [19,20]. Similarly, for each regulatory motif the kinetic models described in the text permit a calculation of the fold-change in the noise (defined as the ratio between the normalized variance for a regulated promoter, and the normalized variance [74] for an unregulated, Poisson promoter) as shown in the third column. The parameter ω captures the effect of cooperativity in reducing the rate of dissociation of one activator due to the presence of the second activator. The fold-change in noise strength is computed using a stochastic kinetic model of gene regulation [78], and is a function of the kinetic rates of transcription and degradation of mRNA (r and γ, respectively), and of the rates of binding (k_R^{on} and k_A^{off}) and dissociation (k_R^{off} and k_R^{on}), which are assumed to be identical for all the operators in the table. The main objective of this part of the table is to illustrate that within this class of models it is possible to explicitly compute different measures of variability.

fold-change), is that it requires a new strain every time we want to change the number of repressors, for example. That is, each of the black data points in Figure 4A corresponds to a different strain. Is there a more systematic way to tune the repressor concentration without resorting to the construction of new strains? A recent set of clever experiments (just one of many illustrations of the amazing experimental advances in recent years) found a way to circumvent this limitation by allowing the dilution of the repressor molecules as the cells divide. When a mother cell containing N repressors divides, each of the daughters should get roughly N/2 repressors and in subsequent generations this results in roughly $N/2^n$ repressors in the daughter cells when the original mother cell has undergone n rounds of division [56]. The significance of this fact is that the level of repression is thereby systematically titrated with each new generation. In turn, the regulated gene increases its level of protein production with each subsequent generation. One beauty of this method is that it permits a direct determination of the number of repressors that mediate the fold-change, a fundamental prerequisite for any direct comparison between the thermodynamic models and their experimental realization as shown in Figure 4A. Interestingly, this example feeds directly into the next section of the article because it illustrates some of the nuance that comes on the heels of knowing something about the fluctuations in a system as opposed to only mean values.

Experimentally, by far the most common way of exerting control of the binding of transcription factors to DNA is by using inducer molecules [42,44]. Although this approach allows for tuning the strength of DNA binding, in this case an extra layer of knowledge and modeling is required to explicitly link theory and experiment. Unless the intracellular concentration of inducer (which can be taken up by the cell in either an active or passive manner) as well as the parameters of inducer–transcription factor interactions are all known, it is very hard to relate the extracellular inducer concentration to an effective concentration of transcription factors that are able to bind DNA.

Another way in which the transcription factor copy number is tuned in multicellular organisms is to exploit the naturally occurring spatial variation in their concentration that arises in different parts of a developing embryo. At different stages of the developmental process different spatial patterns of transcription factor concentrations are established. Recent quantitative experimental efforts in the developing fruit fly embryo are in the process of paving the way to the same sorts of systematic theory–experiment interplay already enjoyed in the study of transcription in bacteria [57–63]. For example, by measuring the spatially-dependent expression of a reporter gene that is under the control of transcription factors that have a concentration gradient along the anterior–posterior axis of the embryo, a first cut has been made at the input–output relation between the *hunchback* and *bicoid* genes as shown in Figure 4C [59]. Building on earlier work in flies that explored the so-called minimal stripe element [64], recent experiments have adopted the synthetic biology approach by placing different repressor binding sites at different locations on the genome and then measuring the resulting fold-change in a way that makes it possible to compare to

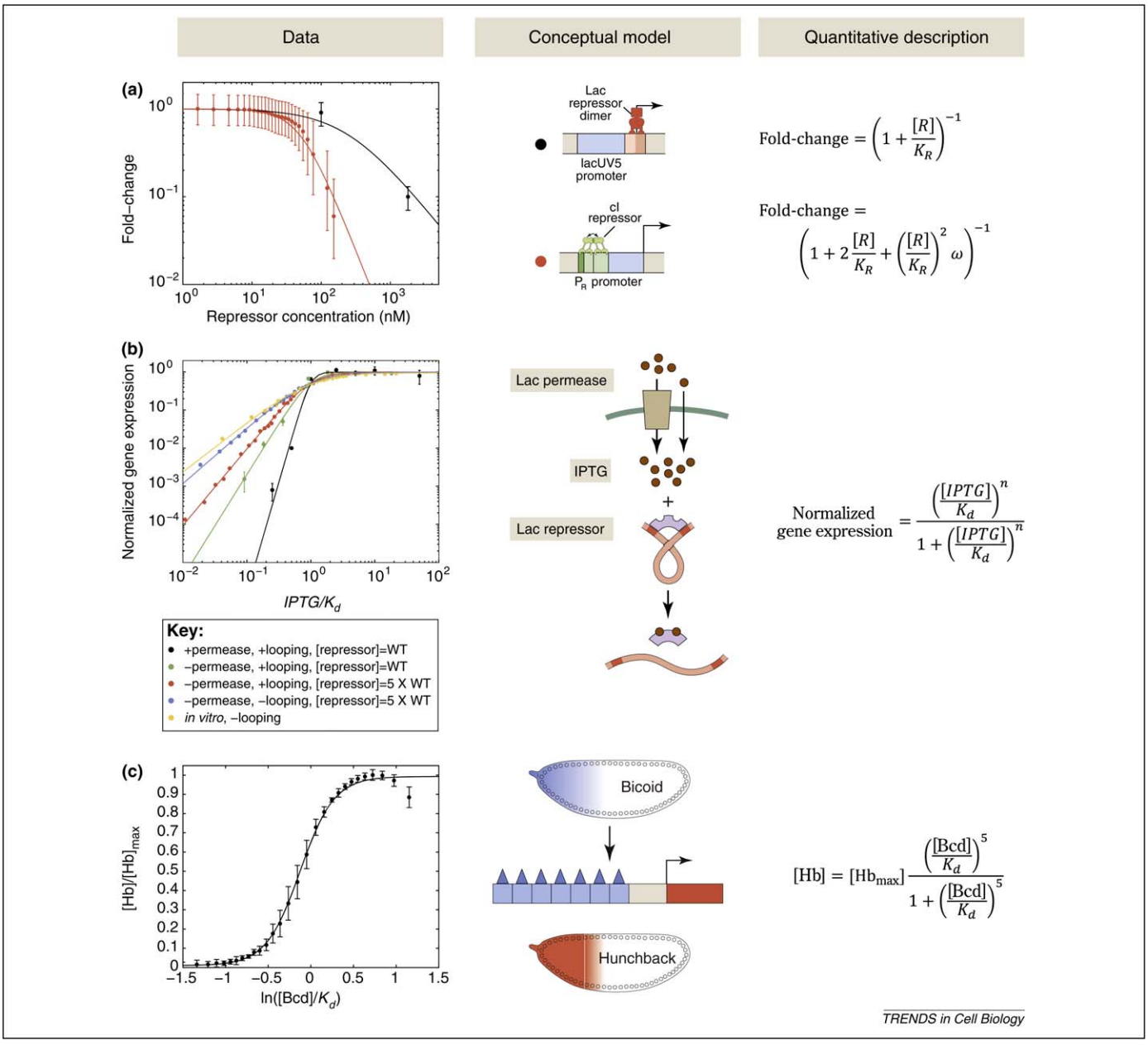

Figure 4. Confrontation of thermodynamic models and experiments. In each case the data are juxtaposed with an equation that serves as a prism through which to view the data. **(a)** Fold-change as a function of the number of repressors for several different repression examples [39,56]. Center panel: above, the *lac* (lactose utilization) promoter; below, rightward promoter P_R from bacteriophage lambda. The equations describe the fold-change in terms of the repressor concentration, [*R*], and dissociation constant, K_d [19]. **(b)** Level of gene expression as a function of inducer concentration for a series of different mutants of the *lac* operon [42,44]. As different elements of the system were deleted the sensitivity of the induction neared that of the purified *in vitro* system. This sensitivity can be quantified by fitting to a thermodynamically-inspired functional form such as the Hill function shown here. Each curve has been normalized to its corresponding maximum in gene expression. **(c)** *hunchback* gene expression as a function of Bicoid concentration which varies from high to low along the anterior–posterior axis of the developing fly embryo. The data are plotted in such a way that anterior is to the right of the curve. This data is fitted to a Hill function with a sensitivity of five although in fact there are seven binding sites for Bicoid in the *hunchback* enhancer [59]. Abbreviations: cI, the bacteriophage lambda repressor; Bcd, Bicoid; Hb, Hunchback; IPTG, *lac* operon inducer isopropyl β-D-thiogalactoside; *lacUV5*, cAMP-independent variant of the *lac* promoter.

first-generation thermodynamic models for these complex systems [62].

A crucial assumption of the thermodynamic model approach is the use of an equilibrium framework for describing the competition between RNA polymerase and the factors that regulate it for the same piece of genomic real estate. One of the ways to judge the merits of this approach is by appealing to the relative timescales of the processes that mediate regulation in comparison with the rate of transcription initiation itself. For promoters where there is a clear separation of timescales for these two classes of processes – regulation on one side and initiation of RNA production on the other – the mean number of messenger RNAs produced by the cell is proportional to the equilibrium probability of the promoter being in a transcriptionally active state. In that limit, when the processes accompanying regulation are fast compared to those associated with initiation of RNA production, transcription factors and RNA polymerase will have enough time to reach binding equilibrium with promoter DNA, and RNA pro-

duction initiates from this equilibrium state. In the opposite limit, in the case of fast transcription initiation, the slow switching between different promoter states is not affected by RNA production and the mean RNA number reflects the mean time the promoter spends in the active state. As an example, *in vitro* and *in vivo* studies of the *lac* promoter have found that the typical time taken for the Lac repressor to bind to and dissociate from the promoter DNA is on the order of minutes [65,66], whereas the events that lead to transcription when the repressor is not present take place at second or sub-second timescales [67,68], thus justifying the equilibrium assumption.

The same concrete interplay between systematic measurements and thermodynamic models described in this section has been played out again and again for a range of different prokaryotic and eukaryotic promoters. Although there are reasons to be skeptical as to whether insights as dramatic as those garnered in the early days of gene regulation will emerge from these kinds of quantitative approaches, the fact that so many researchers are now using these ideas signals a growing consensus that we can only claim to really understand what is going on when we can construct a quantitative framework that mirrors what is observed experimentally. Perhaps even more significantly, this kind of detailed quantitative understanding might serve as the most useful jumping-off point for those trying to engineer new architectures using more than enlightened empiricism.

Despite their broad reach, the thermodynamic models are relatively silent when it comes to the growing mass of temporal measurements which examine the regulatory responses of individual cells over time, or for those measurements in which cell-to-cell variability or mRNA and protein distributions are reported. For these phenomena we must turn to a different class of models.

Putting the dynamics back in transcription

No matter how appealing the simplicity of the descriptions introduced in the previous section, there are now an increasing number of single-cell experiments that are delivering not only the entire distributions (as opposed to the means that are the central focus of the thermodynamic models), but also that yield the stochastic trajectories of mRNA (and protein) concentrations as a function of time, as shown in Figure 5. These kinds of data call for theoretical models that go beyond the thermodynamic framework.

One general class of models used to respond to such data are built using rate equations or master equations (these approaches have important differences, but we focus on their common features). These models tell us how in a small time-increment the population of the chemical species of interest (e.g. mRNA or protein) or the probability distributions themselves will vary [69–72]. The key assumption of these models is that one can define distinct states of the promoter, as in the thermodynamic models, and then describe the temporal evolution of promoter activity as a biased random-walk between the different states, as shown in Figures 2C and 2D. The transitions from one state to the next are characterized by rate constants – namely the probabilities per unit time that the specific transition of interest will occur [29,70,72–81]

If we interest ourselves in the temporal evolution of mRNA levels, the idea in these time-dependent approaches is that the amount of mRNA found at time $t+\Delta t$ can be obtained by considering the amount at time t and then summing up all the ways that mRNAs can be gained and lost during that small increment of time Δt. For example, there will be loss of mRNA due to both degradation and cell division, whereas there will also be terms tending to increase the amount of mRNA as a result of transcription itself (and the average rate of transcription will depend in turn upon the concentrations of regulatory proteins such as activators and repressors). The simplest model for the transcription process posits a mean production rate per unit time r, and a mean degradation rate per mRNA γ, resulting in a steady-state average mRNA number of $<mRNA> = r/\gamma$.

However, even for this simple model, if we consider mRNA number as a function of time the instantaneous number will not always be equal to this predicted mean value. Because the arrival and binding of individual RNA polymerase molecules at the promoter is an inherently random event, at any given time there can be fluctuations resulting in slightly more or less mRNA than the predicted mean. The size of these fluctuations can be quantified by the ratio between the variance of the distribution $\{Var(mRNA)\}$ and the square of its mean $\{<mRNA>\}$. For the simple model outlined above, the fluctuations are characterized by

$$\eta^2 = \frac{Var(mRNA)}{\langle mRNA\rangle^2} = \frac{1}{r/\gamma} = \frac{1}{\langle mRNA\rangle} \qquad [2]$$

This simple model of stochastic mRNA production and decay implies that mRNA is made stochastically in uncorrelated transcription events that are independent. The prediction of the model is that the mRNA number is described by a Poisson distribution, for which the variance is equal to the mean.

One of the powerful insights that emerges from experimental data such as those shown in Figure 5A is that they reveal that the most naïve model of mRNA dynamics described above is not borne out experimentally. Whereas the simplest model is predicated on the idea of a uniform rate of mRNA production, we see even for a simple regulatory architecture that the mRNA production is 'bursty', with brief periods of time in which the promoter is active and multiple mRNAs are produced, followed by long periods of time in which transcription is turned off. In a case such as this the governing equations are more involved because one has to track how the probability of being in either the active or inactive state changes in a time Δt [69–72,74,78]. However, even with this more complex two-state model it is possible to compute the expected mean and the variance; the resulting expressions are shown in Figure 5A and more generally in Figure 3. Consistent with the observations, the variance and the mean are not equal, as the initial naïve model predicts.

Rather than focusing solely on the lowest orders moments of the mRNA distribution, recent measurements and models have even permitted a determination of the entire distribution [70,76]. One particularly interesting case study in yeast is highlighted in Figure 5B. The num-

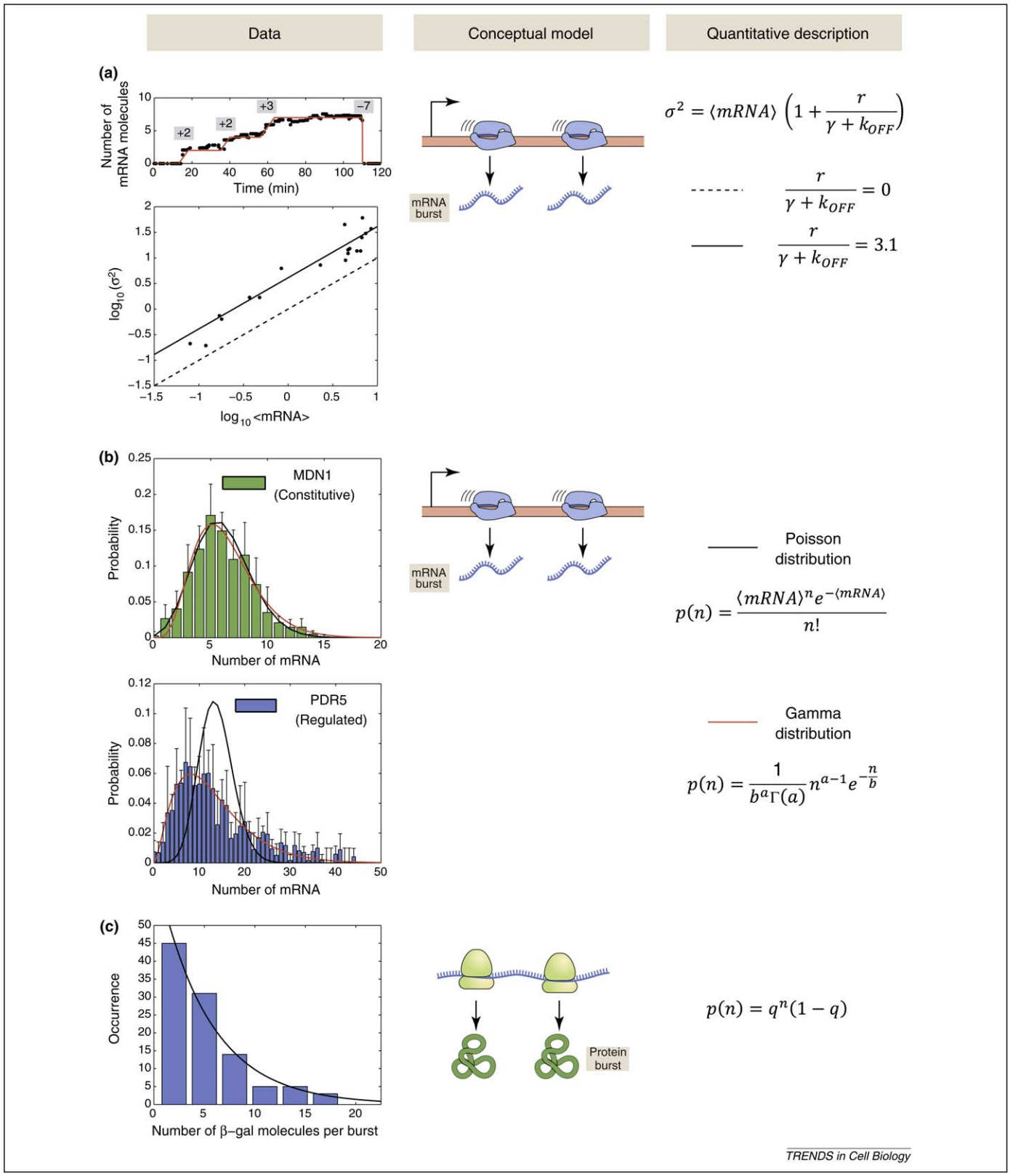

Figure 5. Transcription and translation dynamics and distributions. **(a)** mRNA dynamics and steady state distribution in *E. coli* [28]. A single mRNA production time trace is shown (top) together with the variance σ^2 of the mRNA distribution as a function of the mean, <mRNA>, (bottom). The dashed line corresponds to a model where the initiation of mRNA transcription is a stochastic Poisson process. The solid line corresponds to a model where mRNA is produced in bursts. The promoter switches stochastically from an active to an inactive state with rate k_{OFF} and from inactive to active with a rate k_{ON}. We make the assumption that $k_{OFF} >> k_{ON}$. The transcription rate is *r*, and the mRNA decay rate is γ [79]. **(b)** mRNA distribution in yeast for a constitutive (MDN1) and regulated gene (PDR5). The continuous curves are fits to the distributions using either a Poisson model or a gamma distribution function which accounts for the bursting nature of the transcription process. *n* is the number of mRNA molecules, and *a* and *b* are interepreted as the mRNA burst frequency and burst size, respectively [30]. **(c)** Bursts in expression of the enzyme β-galactosidase (β-gal) corresponding to the translation of single mRNA molecules. The data are consistent with a geometrical distribution where the probability of translation of an mRNA molecule resulting in *n* proteins is given by *q* and the probability of the molecule decaying is given by (1 - *q*) [88].

ber of mRNA molecules being actively transcribed in individual cells was determined using state-of-the-art single molecule techniques. By measuring the entire mRNA distribution, quantitative information about the processes that must be responsible for generating the observed distribution, and even the rates at which they take place, can be determined.

As shown in this section, recent experiments are now routinely generating data that call for theoretical analysis beyond the thermodynamic models. As a result, ideas based on rate equations have stepped into the breach and are themselves producing a range of falsifiable predictions that not only guide experiments but have also altered our picture of the transcription process itself.

Conclusions

The amazing progress in biology in the last half century seems in many ways analogous to progress in astronomy after the invention of the telescope. The expansion of our factual understanding of living matter is staggering. Further, it seems that the analogy to astronomy goes deeper. Just as quantitative observations of the motions of celestial bodies called for theoretical underpinnings, allied with the development of this new generation of biological facts has come a concomitant need for theoretical frameworks that allow us to tell stories about these facts in a way that brings them under the same theoretical roof and in a way that suggests fruitful directions for further experimentation.

The attempt to cast our understanding of biological processes such as transcription in purely quantitative terms as reviewed in this article is only in its infancy. Indeed, many challenges stand in the way of making this approach more generally applicable – including ignorance of the complete set of molecular players and linkages in many networks of interest and an unruly proliferation of parameters even in those cases where the relevant molecular actors and linkages are known. It is no accident that much of our discussion focused on the seemingly overworked example of the *lac* operon. This reflects the fact that, to make quantitative progress as advocated here, it is necessary to have a well-characterized system – and few if any systems have been subjected to the same level of experimental scrutiny as the *lac* operon. Our Figure 3 is an attempt to make more generic predictions about other common regulatory architectures to break away from a *lac* operon-dominated mindset. It is in a similar spirit that several other key case-studies in yeast and flies have been brought to bear on the much more challenging case studies to be found in eukaryotes where other factors such as nucleosomes add another level of complexity to the problem. Our sense is that an important way to make continued progress is the selection of certain key case studies which will be characterized by depth rather than breadth. In these cases, the acid test should remain the ability to make testable predictions about how certain key 'control knobs' alter the level of expression, and the fundamental mantra of the quantitative approach is that failure of the predictions of such models is an opportunity to learn something new.

Although the discussion in this paper has centered on transcription, we could have written a similar story using the same two frameworks (i.e. thermodynamic models and rate equations) for discussing signal transduction in bacterial chemotaxis, for example, and much work in this vein is already underway [82–85]. The same could be said for a variety of other interesting problems in biology. In that sense, this paper should be seen more broadly as reflecting several useful strategies with much broader biological reach than merely the fascinating topic of transcription. In each of these cases the underlying argument is the same. As noted by Abraham Pais in his discussion of Einstein's role in the emergence of the modern quantum theory of solids, 'In order to recognize an anomaly, one needs a theory or a rule or at least a prejudice' [86]. In that sense, the approach advocated here is to use quantitative models to build prejudices which can then serve as a scalpel to dissect experiments in a way that the traditional verbal and pictorial descriptions cannot, and which reveal anomalies that can help us better understand and ultimately control living matter.

Acknowledgments

We are grateful to Rob Brewster, Robert Sidney Cox III, Ido Golding, Thomas Gregor, Daniel Jones, Justin Kinney, Dan Larson, Ron Milo, Nigel Orme, Linda Song and several anonymous reviewers for stimulating discussions, providing data, help with figures and/or critical evaluation of the manuscript. H.G. and R.P. are also extremely grateful to the National Institutes of Health (NIH) for support through the NIH Director's Pioneer Award (DP1 OD000217), RO1 GM085286 and RO1 GM085286-01S. T.K. acknowledges grant support from the NIH (GM078591, GM071508) and the Howard Hughes Medical Institute (52005884), A.S. and J.K. acknowledge funding support from the National Science Foundation through grant DMR-0706458. A.S. was also supported by grants GM81648 and GM43369 from the NIH.

References

1 National Research Council (US) Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century. (2003) *Bio 2010: Transforming Undergraduate Education for Future Research Biologists*, National Academies Press
2 Committee of the Association of American Medical Colleges (AAMC) and the Howard Hughes Medical Institute (HHMI) (2009) *Scientific Foundations for Future Physicians*, HHMI–AAMC
3 Milo, R. *et al.* (2010) BioNumbers – the database of key numbers in molecular and cell biology. *Nucleic Acids Res* 38, D750–753
4 Berg, H.C. (1993) *Random Walks in Biology*, Princeton University Press
5 Cohen, J.E. (2004) Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol* 2, e439
6 Nelson, P.C. *et al.* (2004) *Biological Physics: Energy, Information, Life*, W.H. Freeman
7 Alon, U. (2007) *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC
8 Phillips, R. *et al.* (2009) *Physical Biology of the Cell*, Garland Science
9 Phillips, R. and Milo, R. (2009) A feeling for the numbers in biology. *Proc. Natl. Acad. Sci. U. S. A.* 106, 21465–21471
10 Bialek, W. and Botstein, D. (2004) Introductory science and mathematics education for 21st-Century biologists. *Science* 303, 788–790
11 Wingreen, N. and Botstein, D. (2006) Back to the future: education for systems-level biologists. *Nat. Rev. Mol. Cell Biol.* 7, 829–832
12 Crouch, C.H. *et al.* (2010) Physics for Future Physicians and Life Scientists: a moment of opportunity. *APS News* 19, 8
13 Raj, A. and van Oudenaarden, A. (2009) Single-molecule approaches to stochastic gene expression. *Annu. Rev. Biophys.* 38, 255–270
14 Larson, D.R. *et al.* (2009) A single molecule view of gene expression. *Trends Cell Biol.* 19, 630–637
15 Locke, J.C. and Elowitz, M.B. (2009) Using movies to analyse gene circuit dynamics in single cells. *Nat. Rev. Microbiol.* 7, 383–392

16 Muller, J. *et al.* (1996) Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. *J. Mol. Biol.* 257, 21–29
17 Rippe, K. (2001) Making contacts on a nucleic acid polymer. *Trends Biochem. Sci.* 26, 733–740
18 Vilar, J.M. and Leibler, S. (2003) DNA looping and physical constraints on transcription regulation. *J. Mol. Biol.* 331, 981–989
19 Bintu, L. *et al.* (2005) Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev.* 15, 125–135
20 Bintu, L. *et al.* (2005) Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* 15, 116–124
21 Saiz, L. *et al.* (2005) Inferring the *in vivo* looping properties of DNA. *Proc. Natl. Acad. Sci. U. S. A.* 102, 17642–17645
22 Zhang, Y. *et al.* (2006) Analysis of *in-vivo* LacR-mediated gene repression based on the mechanics of DNA looping. *PLoS One* 1, e136
23 Swigon, D. *et al.* (2006) Modeling the Lac repressor–operator assembly: the influence of DNA looping on Lac repressor conformation. *Proc. Natl. Acad. Sci. U. S. A.* 103, 9879–9884
24 Garcia, H.G. *et al.* (2007) Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity. *Biopolymers* 85, 115–130
25 Ackers, G.K. *et al.* (1982) Quantitative model for gene regulation by lambda phage repressor. *Proc. Natl. Acad. Sci. U. S. A.* 79, 1129–1133
26 Buchler, N.E. *et al.* (2003) On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5136–5141
27 Bertrand, E. *et al.* (1998) Localization of ASH1 mRNA particles in living yeast. *Mol Cell.* 2, 437–445
28 Golding, I. *et al.* (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123, 1025–1036
29 Blake, W.J. *et al.* (2006) Phenotypic consequences of promoter-mediated transcriptional noise. *Mol. Cell* 24, 853–865
30 Zenklusen, D. *et al.* (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.* 15, 1263–1271
31 Raj, A. *et al.* (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* 5, 877–879
32 Voigt, C.A. and Keasling, J.D. (2005) Programming cellular function. *Nat. Chem. Biol.* 1, 304–307
33 Guido, N.J. *et al.* (2006) A bottom-up approach to gene regulation. *Nature* 439, 856–860
34 Cox, R.S., 3rd *et al.* (2007) Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.* 3, 145
35 Gertz, J. *et al.* (2009) Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature* 457, 215–218
36 Horowitz, P. and Hill, W. (1989) *The Art of Electronics*, Cambridge University Press
37 Kaern, M. *et al.* (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6, 451–464
38 Xie, X.S. *et al.* (2008) Single-molecule approach to molecular biology in living bacterial cells. *Annu. Rev. Biophys.* 37, 417–444
39 Oehler, S. *et al.* (1994) Quality and position of the three *lac* operators of *E. coli* define efficiency of repression. *EMBO J.* 13, 3348–3355
40 Dodd, I.B. *et al.* (2005) Revisited gene regulation in bacteriophage lambda. *Curr. Opin. Genet. Dev.* 15, 145–152
41 Becker, N.A. *et al.* (2005) Bacterial repression loops require enhanced DNA flexibility. *J. Mol. Biol.* 349, 716–730
42 Oehler, S. *et al.* (2006) Induction of the *lac* promoter in the absence of DNA loops and the stoichiometry of induction. *Nucleic Acids Res.* 34, 606–612
43 Zaslaver, A. *et al.* (2006) A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nat. Methods* 3, 623–628
44 Kuhlman, T. *et al.* (2007) Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 6043–6048
45 Wong, P. *et al.* (1997) Mathematical model of the *lac* operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnol. Prog.* 13, 132–143
46 Mahaffy, J.M. and Savev, E.S. (1999) Stability analysis for a mathematical model of the *lac* operon. *Q. Appl. Math.* 57, 37–53
47 Yildirim, N. and Mackey, M.C. (2003) Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. *Biophys J.* 84, 2841–2851
48 Vilar, J.M. *et al.* (2003) Modeling network dynamics: the *lac* operon, a case study. *J. Cell Biol.* 161, 471–476
49 Santillan, M. and Mackey, M.C. (2004) Influence of catabolite repression and inducer exclusion on the bistable behavior of the *lac* operon. *Biophys. J.* 86, 1282–1292
50 Dreisigmeyer, D.W. *et al.* (2008) Determinants of bistability in induction of the *Escherichia coli lac* operon. *IET Syst. Biol.* 2, 293–303
51 Gottesman, S. (2004) The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.* 58, 303–328
52 Gottesman, S. (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.* 21, 399–404
53 Storz, G. *et al.* (2005) An abundance of RNA regulators. *Annu. Rev. Biochem.* 74, 199–217
54 Levine, E. *et al.* (2007) Quantitative characteristics of gene regulation by small RNA. *PLoS Biol* 5, e229
55 Mehta, P. *et al.* (2008) A quantitative comparison of sRNA-based and protein-based gene regulation. *Mol. Syst. Biol.* 4, 221
56 Rosenfeld, N. *et al.* (2005) Gene regulation at the single-cell level. *Science* 307, 1962–1965
57 Gregor, T. *et al.* (2005) Diffusion and scaling during early embryonic pattern formation. *Proc. Natl. Acad. Sci. U. S. A.* 102, 18403–18407
58 Houchmandzadeh, B. *et al.* (2005) Precise domain specification in the developing *Drosophila* embryo. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 72, 061920
59 Gregor, T. *et al.* (2007) Probing the limits to positional information. *Cell* 130, 153–164
60 Gregor, T. *et al.* (2007) Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell* 130, 141–152
61 Barkai, N. and Shilo, B.Z. (2009) Robust generation and decoding of morphogen gradients. *Cold Spring Harb. Perspect. Biol.* 1, a001990
62 Fakhouri, W.D. *et al.* (2010) Deciphering a transcriptional regulatory code: modeling short-range repression in the *Drosophila* embryo. *Mol. Syst. Biol.* 6, 341
63 Ben-Zvi, D. and Barkai, N. (2010) Scaling of morphogen gradients by an expansion–repression integral feedback control. *Proc. Natl. Acad. Sci. U. S. A.* 107, 6924–6929
64 Small, S. *et al.* (1992) Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. *EMBO J.* 11, 4047–4057
65 Elf, J. *et al.* (2007) Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* 316, 1191–1194
66 Wong, O.K. *et al.* (2008) Interconvertible *lac* repressor–{DNA} loops revealed by single-molecule experiments. *PLoS Biol.* 6, e232
67 Kennell, D. and Riezman, H. (1977) Transcription and translation initiation frequencies of the *Escherichia coli lac* operon. *J. Mol. Biol.* 114, 1–21
68 Record, M.T., Jr *et al.* (1996) *Escherichia coli* RNA polymerase (sigma70) promoters and the kinetics of the steps of transcription initiation. In *Escherichia coli and Salmonella Cellular and Molecular Biology* (Vols 1–2) (Neidhardt, F.C. et al., eds), pp. 792–821, American Society for Microbiology Press
69 Berg, O.G. (1978) A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.* 71, 587–603
70 Peccoud, J. and Ycart, B. (1995) Markovian modeling of gene product synthesis. *Theor. Popul. Biol.* 48, 222
71 McAdams, H.H. and Arkin, A. (1998) Simulation of prokaryotic genetic circuits. *Annu. Rev. Biophys. Biomol. Struct.* 27, 199–224
72 Kepler, T.B. and Elston, T.C. (2001) Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys. J.* 81, 3116–3136
73 Blake, W.J. *et al.* (2003) Noise in eukaryotic gene expression. *Nature* 422, 633–637
74 Paulsson, J. (2004) Summing up the noise in gene networks. *Nature* 427, 415–418
75 Raser, J.M. and O'Shea, E.K. (2004) Control of stochasticity in eukaryotic gene expression. *Science* 304, 1811–1814
76 Raj, A. *et al.* (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 10, e309
77 Kim, H.D. and O'Shea, E.K. (2008) A quantitative model of transcription factor-activated gene expression. *Nat. Struct. Mol. Biol.* 15, 1192–1198
78 Sanchez, A. and Kondev, J. (2008) Transcriptional control of noise in gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 105, 5081–5086

79 Shahrezaei V, S.P. (2008) Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 105, 17256–17261
80 Boeger, H.G., J., Kornberg, R.D. (2008) Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. *Cell* 133, 716–726
81 Choi, P.J. *et al.* (2008) A stochastic single molecule event triggers phenotype switching of a bacterial cell. *Science* 322, 442–445
82 Barkai, N. and Leibler, S. (1997) Robustness in simple biochemical networks. *Nature* 387, 913–917
83 Mello, B.A. and Tu, Y. (2005) An allosteric model for heterogeneous receptor complexes: understanding bacterial chemotaxis responses to multiple stimuli. *Proc. Natl. Acad. Sci. U. S. A.* 102, 17354–17359
84 Keymer, J.E. *et al.* (2006) Chemosensing in *Escherichia coli*: two regimes of two-state receptors. *Proc. Natl. Acad. Sci. U. S. A.* 103, 1786–1791
85 Tu, Y. *et al.* (2008) Modeling the chemotactic response of *Escherichia coli* to time-varying stimuli. *Proc. Natl. Acad. Sci. U. S. A.* 105, 14855–14860
86 Pais, A. (2005) *'Subtle is the Lord. . .': The Science and the Life of Albert Einstein*, Oxford University Press
87 Gama-Castro, S. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* 36, D120–124
88 Cai, L. *et al.* (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature* 440, 358–362