Is the future biology Shakespearean or Newtonian?

Ovidiu Lipan*[a] and Wing H. Wong†[b]

DOI: 10.1039/b607243g

"Cells do not care about mathematics" thus concluded a biologist friend after a discussion on the future of biology. And indeed, why should they care? But if we exchange the word "cell" with "rock", "Moon" or "electrons", do we have to change the sentence also? Starting from this line of thought, we review some recent developments in understanding the stochastic behavior of biological systems. We emphasize the importance of a molecular Signal Generator in the study of genetic networks.

1. Introduction

''Cells do not care about mathematics'' thus concluded a biologist friend after a discussion on the future of biology. And indeed, why should they care? But if we exchange the word ''cell'' with ''rock'', ''Moon'' or ''electrons'', do we have to change the sentence also? Why should the Moon care about mathematics? We know however, that the Moon navigates around the Earth on a mathematical orbit. If we think of how electrons are described in Quantum Electrodynamics, then we can say that electrons have a special taste for sophisticated mathematics. Today, life scientists and computational scientists are engaged in many discussions of this type. In short, the main theme of discussion is actually a question: Is the future biology Shakespearean or Newtonian?

Although far from perfect, a Shakespearean play may serve as an analogy of how biological phenomena were described, studied and understood until very recently. Each character from Romeo and Juliet, for example, can be envisioned to be a molecule which plays a role in a complex cellular process. By studying the process experimentally, we hope to identify the characters in the play and to observe how they interact as the process proceeds. In the end we can describe the complex process just as the text of the play prescribes, for each scene, what each character is supposed to say and how to interact with the other characters on the stage.

However, formatting the biological knowledge in a Shakespeare-like drama is no longer enough. Like in physics or chemistry, the scientific drama must be sustained and accompanied by mathematical equations. The quantitative relations between players are as important as the players themselves. For example, a resistor is very important, but Ohm's law is important as well. The palpable resistor and the abstract Ohm's law cannot be separated if we want to understand an electric circuit. Therefore, the aim of biological research should be twofold: (1) to find the

[a]*Center for Biotechnology and Genomic Medicine, Medical College of Georgia, 1120 15th St. CA-4139, Augusta, GA, 30912, USA. E-mail: olipan@richmond.edu*
[b]*Departments of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA, 94305-4065, USA. E-mail: whwong@stanford.edu*
† Current address: Gottwald Science Center, University of Richmond, VA 23173, USA

Ovidiu Lipan

Ovidiu Lipan is Assistant Professor of physical biology at the University of Richmond, Richmond VA, USA. He was previously (2003–2006) with the Center for Biotechnology and Genomic Medicine from the Medical College of Georgia, GA, USA. Prior to this he was a postdoctoral fellow in computational biology at Harvard University (2000–2003), and a Sherman Fairchild Postdoctoral Fellow in Theoretical Physics at California Institute of Technology (1998–2000). He received his PhD degree in Theoretical Physics from The University of Chicago (1998).

Wing Hung Wong

Wing Hung Wong is a Professor in the Department of Statistics, Department of Health Research and Policy, and (by courtesy) Department of Biological Sciences at Stanford University. Prior to joining Stanford, he taught at the University of Chicago (1980–1993), the Chinese University of Hong Kong (1994–1996), UCLA (1997–2000) and Harvard University (2001–2004). He was trained and worked for many years in mathematical statistics, but recently his research has centered on the application of statistics and mathematics in biological research, especially in the analysis of gene expression microarrays and cis-regulatory sequences.

functional molecules that play a role in biochemical pathways and (2) to describe mathematically the interactions among these molecules and to understand the consequences of these interactions.

While much recent progress has been made towards (1) under the rubric of genome research, the study on (2) is still in its infancy. The emerging field of *Systems Biology* was borne with the (Newtonian) belief that biology can be described and understood in a mathematical language.[1,2] In our analogy, a Newtonian science is not necessarily a deterministic science; in a broader sense, it is a science that recognizes that the book of Nature is written in a mathematical language, as Plato and Galileo stated long ago. What is the nature of this language for life sciences? And what might be a plausible path towards the Newtonian future for biology?

2. How to describe a system: the importance of stochasticity

Computers process information using electrons that flow through billions of semiconductor devices that are highly packed in small areas. Each device can be understood as a basic processing function. For example, a device will take as an input a voltage, and output another voltage that is one hundred times greater. The function of the device is thus to amplify. There are many other possible functions, and the interconnections of so many devices with so many different functions make necessary the use of a design scheme. The theory of electric circuits is fundamental for creating coherent design schemes and it covers topics on positive and negative feedback, oscillations and stability. The behavior of receptors on a cell membrane, or calcium pumps lead us to think of similarities between the flow of information in genetic pathways and electric circuits. Thus it is no surprise that the theory of feedback control circuits in biochemical pathways started not long after the operon concept was introduced by Jacob *et al.* in 1960.[3] Goodwin,[4] gave the first mathematical analysis of operon dynamics followed by Griffith,[5,6] with a more complete analysis. As with the theory of electric circuits, the early theories of biochemical pathways were based on deterministic ordinary differential equations and address questions regarding feedback, stability, hysteresis effects, *etc*. These deterministic theories were ahead of their time, and it is only now, as a result of a speed up in biotechnology discoveries, that we witness the dawn of their importance. But how precise is a response of a cell to a stimulus so we can use deterministic mathematical laws? Can we still believe in a deterministic point of view, from which a biochemical pathway is precise machinery and thus our measurements are inaccurate only because of experimental error? The deterministic point of view fails when we recall that the players in a biochemical pathway are molecules. These molecules bounce one on each other and form complexes, which are more or less stable. We cannot say that with a probability of 1 a complex will be formed. Thus, living systems are not deterministic but inherently noisy and are optimized to function in the presence of stochastic fluctuations. For example, the number of a specific protein varies from cell to cell in a population of cells with an identical genetic background kept under the same environmental conditions. Although noise is usually perceived as being undesirable, some organisms can use it to introduce diversity into a population, like is the case for lysis–lysogeny bifurcation in phage λ. This system was studied by Arkin, Ross and MacAdams,[7] where they used a Monte Carlo simulation to study it. The simulation algorithm used by Arkin, Ross and MacAdams was developed by Gillespie in 1976,[8] and was designed for sequences of coupled chemical reactions. The mathematical foundation of this algorithm is the theory of Markov processes, the basic ingredient being the probability of a chemical reaction to proceed in a specified direction. Such processes are well studied in the context of chemical reactions (Van Kampen[9]). However, the importance of an analytical mathematical approach to study the biological noise present in a genetic network is a recent development.[10] For example, the simplest system to be considered consists of one gene only. This system is specified at any time t by the total number of mRNA molecules r and protein molecules p. The state of the system $q = (r,p)$ changes due to four random transitions: increase or decrease of r by one molecule and likewise for p. Each of these changes can be described by a transition probability rate, which depends on the state q. The mathematical dependence of the transition probability rates on the state q are suggested by the biological system. Once the transition probability rates are known, the probability of the system to be in the state q at time t, $P(q,t)$, is the solution of an equation which is known as the Master Equation.[9] Thattai and van Oudenaarden[10] studied the steady state of this one gene system, a limiting situation when $P(q,t)$ is time independent. They find the mean values for the mRNA, protein and their standard deviation from the mean as a function of the parameters of the system. To experimentally test the theoretical results, the same group[11] used the green fluorescent reporter gene (*gfp*) in the chromosome of *Bacillus subtilis*. The transcriptional efficiency was regulated using an isopropyl-β-D-thiogalactopyranoside (IPTG)-inducible promoter upstream of *gfp* and by varying the concentration of IPTG in the growth medium. Translational efficiency was regulated by constructing a series of *Bacillus subtilis* strains that contained point mutations in the ribosome binding site and initiation codon of *gfp*. From flow cytometry measurements, the mean value and the standard deviation for the protein molecule were computed. As IPTG was varied, the protein noise strength remained constant, confirming the theoretical prediction that the protein noise strength was independent of transcriptional efficiency. From one gene, the next step is to study a gene regulatory network which is composed of many genes in interaction. Consider a particular gene of interest. The noise strength for the protein product of this gene should have two sources: one coming from its transcription and translation, as we discussed above (referred to as ''intrinsic'' noise), and a second one coming from the surrounding molecules that interact with our gene of interest (referred to as ''extrinsic'' noise). These intrinsic and extrinsic contributions to stochasticity in gene expression were studied in 2002 by Swain, Elowitz and Siggia.[12] They demonstrate theoretically that simultaneous measurement of two

identical genes per cell enables discrimination of these two types of noise. Measurements of these two types of noise were performed by Elowitz and collaborators in *Escherichia coli*,[13] and by Raser and O'Shea in *Saccharomyces cerevisiae*.[14] These measurements and theoretical results show that genes must be studied as part of a network. The emerging discipline of *Systems Biology* aims to understand the gene interactions from a global perspective. Levine and Davidson[15] discuss gene regulatory networks for development. To explain spatio-temporal localization of different components of the network they use diagrams that resemble an electric circuit. The conceptual similarity between genetic circuits and electrical circuits is more apparent in studies that aim to construct biomolecular devices. Gardner, *et al.*[16] describe a toggle switch circuit that can be switched between two stable states by transient external signals. An oscillatory circuit, called a repressilator, was designed and constructed by Elowitz and Leibler.[17] We conclude by noting that, at present, a large amount of work on genetic networks deals with biological noise and ideas from electric circuits guide the thoughts in the effort to understand the molecular interactions.

3. How to study the system: the need for a signal generator

From these studies we also realize that the steady state regime is not appropriate for modeling gene circuits. The main reason is that in a living system, signals that vary in time propagate through the system, keeping the system away from a steady state. Also, the need for departing from the steady state is imposed by the desire to find the genetic network connectivity. The analogy with an electrical system helps to illustrate the problem. There, the properties of an unknown system are revealed by applying an input signal generator at one port and then output signals are measured in different points of the electrical system. If sufficient input–output pairs of data are collected, then we can find a network configuration that best explains the data and predicts the outputs for a new set of input signals. The idea of operating with a signal generator upon a stochastic genetic network was proposed by Lipan and Wong.[18] The idea is useful if we can experimentally construct a molecular signal generator. Fortunately, a signal generator can be implemented using a molecular switch based on a two-hybrid assay proposed in 2002 by Quail and collaborators.[19] The main component of this switch is a molecule (phytochrome[19]) which is synthesized in darkness in a form which we will denote by Q1. When the Q1 form absorbs a red light photon (wavelength 664 nm) it is transformed into the form Q2. When Q2 absorbs a far red light photon (wavelength 748 nm) the molecule Q goes back to its original form, Q1. These transitions take milliseconds. The targeted promoter is opened by the Q2 form and the gene is transcribed. After the desired elapsed time, the gene can be turned off by a photon from a far red light source. Using a sequence of red and far red light pulses, the molecular switch can be periodically opened and closed and thus a perturbation can be inserted into the biological system. Another promising method that enables the generation of time-variable data is through the use of microfluidics devices.[20] A Signal Generator is thus an indispensable tool for *Systems Biology*. We believe that in the near future different types of Signal Generators will appear on the market.

4. The study of linear networks

Suppose now that we have implemented a signal generator somewhere in a biochemical pathway and we collect data over a period of time. We measure, using flow cytometry, the responses of 10 000 cells (a number taken just for convenience). The data will reveal the time changes of different enzymes and products, for example. Because of the stochastic nature of the biochemical interactions, the measured data will look like Fig. 1, and will reflect the noisy biological nature of the systems plus experimental errors, and not only a mean value plus experimental errors. Each of the 10 000 cells will have a different response and thus we have to explain an entire histogram, not only its mean value. A histogram is characterized by its moments: the first moment equals the mean value and the standard deviation can be computed from the first two moments. Moreover, the asymmetries of the histogram are captured by the third moment and so on. Closely related to moments are a set of parameters known as cumulants. We can describe a histogram using moments or cumulants, just as we can describe a space using Cartesian or polar coordinates. The simplicity of the mathematical description dictates which type of description is preferred; we found that factorial cumulants are suited for describing stochastic genetic networks.[18,21]

From the perspective that the signal generator creates an input signal, these measured factorial cumulants obtained from the flow cytometry data are to be considered the output signals. Because the biochemical processing function can be inferred from the input–output data pair, the key problem is to find a mathematical description of these input–output relations. For a genetic network with transition probability rates having a linear dependency on the state q (linear genetic network) the Laplace transform of the input–output relations

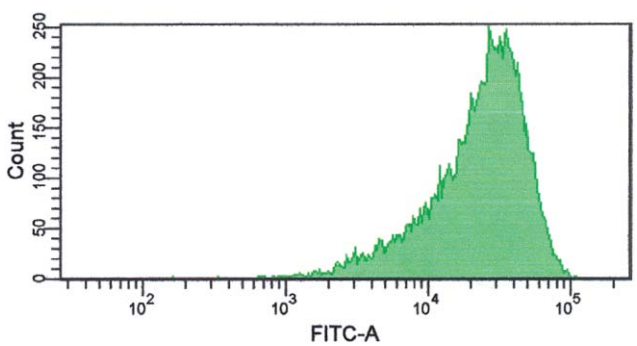

Fig. 1 A histogram from a flow cytometry measurement. FITC-A represents the intensity of a reporter gene, like the green fluorescent protein.

were given in ref. 18 (, *vec* and ⊗ denote respectively the Laplace transform operator on functions, the vectorize and Kroneker product operators on matrices[22]):

$$\mathcal{L}\mu = \frac{1}{(s-H)}\mathcal{L}G. \tag{4.1}$$

$$\mathcal{L}vec(X) = \frac{1}{s-1\otimes H - H\otimes 1} [(1\otimes H + H\otimes 1)L + 2L\Gamma]\frac{1}{s-H}\mathcal{L}G. \tag{4.2}$$

Here G represents the signal generators and H is a matrix that encapsulates the genetic network's parameters. The above relations tell us that inversion and product of matrices are all that we need to find the mean value and the factorial cumulants X. A similar problem was also studied in ref. 23. The advantage of having an analytical expression for a biochemical processing function is that measured data can be fitted to a mathematical model much more easily than using Monte Carlo simulations. Moreover, even if at present the measured data are not informative enough to fit a complex biochemical pathway, at least we can draw semi-quantitative conclusions from the biochemical processing function. In ref. 18 we studied also the advantage of exciting a biochemical pathway with an oscillatory signal. It is well recognized that a periodic pattern programmed into the input signal can be recognized in the output measurements even in the presence of a strong noisy background. One last advantage to be noticed is that the theory for linear biochemical pathways in ref. 18 and ref. 23 is not a mathematical approximation of the biochemical stochastic process. Usually, the Master Equation for the biochemical stochastic process being hard to solve, is transformed into a partial differential equation for which many methods are available. Diverse schemes of approximations are known: Fokker–Plank, Langevine, and Ω-expansion.[9] However, many molecules are present in a cell in low numbers.[13,24] To cover such cases the Master Equation should not be approximated which is possible for linear genetic networks. In conclusion, an advantage of the linear theory is that it is exactly solvable and easy to work with. However, it can be applied only to special situations or as a linear approximation to nonlinear phenomena.

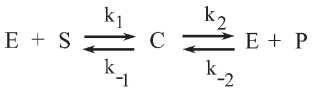

Fig. 2 A catalytic process.

5. The study of nonlinear biochemical pathways

The rate of formation of a biochemical complex depends on the mathematical product of the concentrations of the components that participate in that complex formation. The presence of the mathematical product of concentrations requires the use of nonlinear polynomial functions. Moreover, many biological processes become active only when the concentration of some molecules increases above a threshold. A detailed study of such nonlinear biochemical pathways, under the influence of input generators, was presented in ref. 21. A simple, but fundamental example (Fig. 2) is the enzymatic (E) catalytic process of transforming a substrate (S) into a chemical product (P) through the formation of a complex (C). This type of process is ubiquitous in biology and we can imagine that the enzyme is modulated by some signaling pathway. If the signal that modulates the enzyme is oscillatory in time and it starts at time 0, then the chemical product will be periodically modulated (Fig. 3).

This process is nonlinear because the probability for complex production is proportional to the mathematical product of substrate and enzyme concentration. To illustrate the nature of the equation for the factorial cumulants, we present here the equations for the first order factorial cumulant X_p (which equals the mean value of P) and the second order factorial cumulant X_{pp}.

$$\dot{X}_P = k_2 X_C - \gamma_P X_P - k_{-2}(X_{EP} + X_E X_P) \tag{5.1}$$

$$\dot{X}_{PP} = 2k_2 X_{CP} - 2\gamma_P X_{PP} - 2k_{-2}(X_{EP} X_P + X_E X_{PP}) \tag{5.2}$$

The equations depend on factorial cumulants related to the other components E, C and S. The k's are the coefficients of the transition rates for the corresponding chemical process, presented in Fig. 2. In Fig. 3 we superimpose a series of Monte Carlo simulations of the catalytic process and the numerical solution of the equations for the nonlinear catalytic process.[21] The mean value and standard deviation obtained directly from the simulated data, match very well the ones obtained from numerical solution of the model. Like in the linear case, we can thus use directly the mathematical model to fit experimental data, without using Monte Carlo simulations. The same approach can be applied to many other examples of nonlinear biochemical pathways, driven by signal generators.[21]

6. Challenges in the estimation of complex genetic networks

Newton developed the calculus as the mathematical language for the study of physical systems in continuous space and time. However, the language by itself does not represent scientific knowledge until the exact laws (Newton's laws, Maxwell equations *etc.*) are specified.

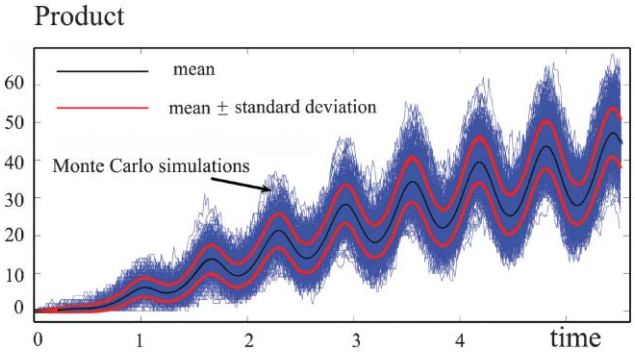

Fig. 3 The transient built up of the chemical product (P). The mean and mean +/− standard deviation are obtained as solutions of a system of equations.[21]

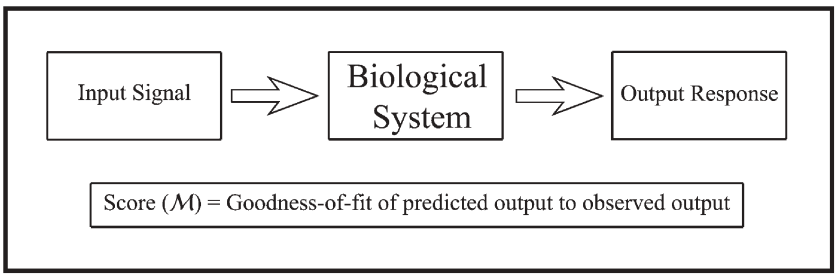

Fig. 4 General approach to estimating a biological system.

The formulation of these laws is guided by the criterion that they should provide explanation and understanding of observed phenomenon and data, and should generate testable predictions. Likewise, in the case of biological systems, Markov processes only provide the mathematical language for their description, and we must rely on experimental data to specify the equations for any concrete system. A general approach that had worked well in electronics and other engineering systems, called the systems approach, can be adapted to study biological systems.

Fig. 4 illustrates the main idea. The general approach is to select some input signal to the system and then measure the response of the system by collecting suitable output data. We assume that the behavior of the biological system can be well represented by a model M that is a member of a given class C of mathematical models. For each model M in the class, we can compute a ''goodness of fit'' score to assess how well the model fits the data—by using M to predict the responses of the system given the particular input signals, and then compare the predictions to the observed output response data. We then use this as the score for M to search for high-scoring models within the model class C. The works discussed in the above sections are useful in several ways. First, they help us to specify the model class C, for example, C may be the class of linear stochastic gene regulatory networks. Second, they provide the mathematical relation between certain key features of the output response and the input signal. For example, it is possible to find the mapping from the factorial moments of the input to the factorial moments of the joint distribution of a subset of variables measured at the output. This then allows us to compute the goodness-of-fit for the model as a function of the parameters in the model, without having to resort to time consuming simulations. Third, the network topology is implicitly captured in the parameter values. For example, in the case of a linear stochastic gene regulatory network, if the evolution of each species of RNA or protein depends on the current values of only a small number of other proteins and RNAs, then the network will assume a sparse network topology.

Although the general approach is clear, the successful implementation of this approach will not be easy. Currently, our ability to monitor the simultaneous responses of many variables at the single-cell level is very limited. The success of this approach, however, depends critically on the availability of output response assays at single-cell level that can provide measurements on a large number of genes or proteins simultaneously (high parallelism). Ideally the assays should monitor real time signals, or at least it should be feasible to perform these assays at high time-sampling rates (high time resolution). Table 1 lists some possible output assays and their advantages and disadvantages.

It is clear that current response assays do not satisfy the requirements of the systems approach. We believe that the development of better assays should be given high priority.

The final ingredient in the systems approach is the computational identification of the set of networks that are consistent with the observed data. For a system with N genes, the number of possible network topologies is in the order of 2 raised to the power N^2. Thus even for a small network with only 30 genes, it is already beyond the capability of current supercomputers to search the space exhaustively. Progress in this direction may require close interaction between computational biologists who develop the search and sampling algorithms for this task, and high-performance computing researchers who develop new hardware architecture and programming models tailored for this problem.

Table 1 Advantages and disadvantages of output assays

Assays	Advantages	Disadvantages
Microarrays	High parallelism	Poor time resolution, non single-cell
Flow cytometry	Single-cell	Poor time resolution, low parallelism
Reporter gene	Real time, single-cell	Expensive, low parallelism

Acknowledgements

The work of WHW was supported by NSF grant DMS0505732. OL is thankful to Jean-Marc Navenot for productive discussions.

References

1 L. Hood, A personal view of molecular technology and how it has changed biology, *J. Proteome Res.*, 2002 Sep-Oct, **1**, 5, 399–409.
2 H. V. Westerhoff and B. O. Palsson, The evolution of molecular biology into systems biology, *Nat. Biotechnol.*, 2004, **22**, 1249.
3 F. Jacob, D. Perrin, C. Sanchez and J. Monod, L'operon: groupe de géne á expression par un operatour, *C. R. Acad. Sci.*, 1960, **250**, 1727.
4 B. Goodwin, Oscillatory behaviour in enzymatic control process, *Adv. Enz. Regul.*, 1965, **3**, 425.
5 J. S. Griffith, Mathematics of cellular control processes. I. Negative feedback to one gene, *J. Theor. Biol.*, 1968, **20**, 202.
6 J. S. Griffith, Mathematics of cellular control processes. II. Positive feedback to one gene, *J. Theor. Biol.*, 1968, **20**, 209.
7 A. Arkin, J. Ross and H. H. McAdams, Stochastic kinetic analysis of developmental pathway bifurcation in phage l-infected *Escherichia coli* cells, *Genetics*, 1998, **149**, 1633, 7.
8 D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, *J. Comput. Phys.*, 1976, **22**, 403.
9 N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, North-Holland, Amsterdam, 1992.
10 M. Thattai and A. van Oudenaarden, Intrinsic noise in gene regulatory

networks, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 8614.
11 E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman and A. van Oudenaarden, Regulation of noise in the expression of a single gene, *Nat. Genet.*, 2002, **31**, 69.
12 P. S. Swain, M. B. Elowitz and E. D. Siggia, Intrinsic and extrinsic contributions to stochasticity in gene expression, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12795.
13 M. B. Elowitz, A. J. Levine, E. D. Siggia and P. S. Swain, Stochastic gene expression in a single cell, *Science*, 2002, **297**, 5584, 1183.
14 J. M. Raser and E. K. O'Shea, Control of stochasticity in eukaryotic gene expression, *Science*, 2004, **304**, 5678, 1811.
15 M. Levine and E. H. Davidson, Gene regulatory networks for development, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 14, 4936.
16 T. S. Gardner, C. R. Cantor and J. J. Collins, Construction of a toggle switch in *Escherichia coli*, *Nature*, 2000, **403**, 339.
17 M. B. Elowitz and S. Leibler, A synthetic oscillatory network of transcriptional regulators, *Nature*, 2000, **403**, 335.
18 O. Lipan and W. H. Wong, The use of oscillatory signals in the study of genetic networks, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 7063.
19 S. Shimizu-Sato, E. Huq, J. M. Tepperman and P. H. Quail, A light-switchable gene promoter system, *Nat. Biotechnol.*, 2002, **20**, 1041.
20 S. Cookson, N. Ostroff, W. L. Wyming, Lee Pang, D. Volfson and J. Hasty, Monitoring dynamics of single-cell gene expression over multiple cell cycles, *Mol. Syst. Biol.*, 2005, **1**, DOI: 10.1038/msb4100032.
21 S. Achimescu and O. Lipan, Signal propagation in nonlinear stochastic gene regulatory networks, *IEE Proceedings–Systems Biology*, 2006, **153**, 3, 120–134.
22 R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1999.
23 C. Gadgil, C. H. Lee and H. G. Othmer, A stochastic analysis of first-order reaction networks, *Bull. Math. Biol.*, 2005, **67**, 5, 901–946.
24 O. G. Berg, A model for the statistical fluctuations of protein numbers in a microbial population, *J. Theor. Biol.*, 1978, **71**, 587.