Phys. Biol. **7** (2010) 016003 (9pp) doi:10.1088/1478-3975/7/1/016003

# The simplicity of completion time distributions for common complex biochemical processes

**Golan Bel**[1,3], **Brian Munsky**[1,3] **and Ilya Nemenman**[2]

[1] Center for Nonlinear Studies and the Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[2] Departments of Physics and Biology and Computational and Life Sciences Strategic Initiative, Emory University, Atlanta, GA 30322, USA

E-mail: golanbel@gmail.com, brian.munsky@gmail.com and ilya.nemenman@emory.edu

Received 28 August 2009
Accepted for publication 12 November 2009
Published 21 December 2009
Online at stacks.iop.org/PhysBio/7/016003

**Abstract**
Biochemical processes typically involve huge numbers of individual reversible steps, each with its own dynamical rate constants. For example, kinetic proofreading processes rely upon numerous sequential reactions in order to guarantee the precise construction of specific macromolecules. In this work, we study the transient properties of such systems and fully characterize their first passage (completion) time distributions. In particular, we provide explicit expressions for the mean and the variance of the completion time for a kinetic proofreading process and computational analyses for more complicated biochemical systems. We find that, for a wide range of parameters, as the system size grows, the completion time behavior simplifies: it becomes either deterministic or exponentially distributed, with a very narrow transition between the two regimes. In both regimes, the dynamical complexity of the full system is trivial compared to its apparent structural complexity. Similar simplicity is likely to arise in the dynamics of many complex multistep biochemical processes. In particular, these findings suggest not only that one may not be able to understand individual elementary reactions from macroscopic observations, but also that such an understanding may be unnecessary.

## 1. Introduction

Considering the ever increasing quantity of known biochemical reactions, one cannot help but be amazed and daunted by the incredible complexity of implied cellular networks. For example, just a handful of different proteins can form a combinatorially large number of interacting molecular species, such as in the case of immune signaling [1], where multiple receptor modification sites result in a model with 354 distinct chemical species. One must then ask: When do all details of this seemingly incomprehensible complexity actually matter, and when is there a smaller set of coarse-grained dynamical variables, parameters, and reactions that approximate the salient features of the system's dynamics? What determines which features are relevant and which are not? And if the networks have a simple equivalent dynamics, did nature choose to make them so complex in order to fulfill a specific biological function? Or is the unnecessary complexity a 'fossil record' of the evolutionary heritage?

In this paper, we begin investigation of these questions in the context of certain biochemical kinetics networks, namely a reversible linear pathway, a kinetic proofreading (KPR) scheme [2], their combination and an extension to a much more general multistep completion process. These motifs are common in a variety of cellular processes—including DNA synthesis and repair [3, 4], protein translation [2, 5], molecular transport [6], receptor-initiated signaling [7–12] and other processes—where assembly of large biochemical structures requires multiple reversible steps. However, in this paper, we leave aside the functional behavior of these networks and focus instead on a different question: Do these complex

[3] Contributed equally

 

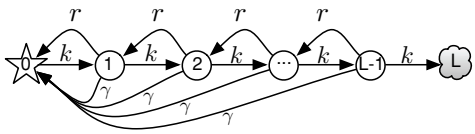


**Figure 1.** Schematic description of the model. The process begins at the site $i = 0$, represented with a star. At each site, the process may transition one step to the right with the forward rate $k$, one step to the left with the backward rate $r$, or all the way back to the origin with the return rate $\gamma$. The right-most site, $i = L$ is an absorbing site (cloud) at which the process is completed.

kinetic schemes have a simplified, yet accurate description? Since multistep structural complexity (see figure 1) is crucial for kinetic proofreading, the KPR process is an ideally suited example for this analysis, but our conclusions will extend to numerous other complex biochemical processes.

We show analytically and numerically that, over broad ranges of parameters, different kinetic schemes exhibit the behavior of either a deterministic process, or a single-step exponential-waiting-time process. We also propose intuitive arguments for the result, which leads us to believe that similar simplifications of complex behavior may be widespread, and even universal. We support this conjecture by numerically studying more complex systems, but leave a general mathematical proof of this conjecture to future work.

### *1.1. The model*

For this study, we begin with a general KPR (gKPR) model [2], for which many properties can be computed analytically. The model is represented by the Markov chain in figure 1. At time $t = 0$, the dynamics begins at the point represented by the star ($i = 0$). The process can leave this state at some exponentially distributed waiting time, defined by a *forward rate* $k$, and the process can continue in the forward direction with rate $k$ until it reaches the final absorbing point (cloud) at $i = L$. At each interior point, $i \in \{1, 2, \ldots, L-1\}$, the process can also move one step to the left with a *backward rate* $r$ or all the way back to the origin with a *return* or *proofreading rate* $\gamma$. The forward and the backward rates emphasize the reversibility of all reactions, and the return rate corresponds to a catastrophic failure, after which the whole process must start anew. For example, in immune signaling, $\gamma$ would represent the rate of receptor-ligand dissociation, which destroys receptor cross-linking and prevents future forward events for a relatively long period of time [1].

This model is substantially simplified compared to detailed models of real biological processes [1] in that, in nature, all three rates may depend on $i$, and the nodes may not form a single linear chain. Even so, a detailed understanding of this simplified model provides an excellent starting point in the process of understanding these more complicated systems. Indeed, we will also show here that all qualitative conclusions made for the gKPR scheme also hold in numerical studies of more complicated systems in which rates are site dependent and where the connections of the nodes are much more varied than a simple linear chain.

### *1.2. The relevant features*

To determine if a kinetic model can be well approximated by a simpler one, we must first decide which of its features must be retained. To illustrate this question, consider the activation of a signaling cascade by an extracellular ligand (as represented in figure 1). The ligand binding initiates the process, bringing it from state $i = 0$ to state $i = 1$. With the exception of this transition, the extracellular environment does not affect the process. Similarly, the downstream signaling pathways are only affected when the signaling construct attains its fully activated state at $i = L$. Thus, as far as the rest of the cell is concerned, only the times of process initiation and completion are controllable, observable or otherwise important. That is, the system can be characterized by the distribution of the *first passage* or the *escape time* between the release at $i = 0$ at $t = 0$ and the completion at $i = L$. Analysis of this distribution and showing its very simple limiting behavior is the main contribution of our work.

We note that, even though a lot is known about the first passage times in different scenarios [13–19] and about temporal dynamics of KPR schemes [10, 11, 20], to our knowledge, the distribution of the first passage time for KPR-type process has not yet been analyzed rigorously and little is known regarding how this first passage time depends upon biochemical parameters such as system size and reaction rates.

## 2. Results

In the following subsections, we provide analytical solution for the Laplace transform of the completion time distribution for the full model followed by precise analyses of three different cases of the gKPR scheme depicted in figure 1, each corresponding to a different continuous time/discrete space Markov chain with exponential transition times (our results can be generalized to the case of non-exponentially distributed transition times using the methods of [21]). First is a normal random walk process (that is $\gamma = 0$) with an absorbing boundary at $i = L$ and a reflecting boundary at $i = 0$. This model is denoted as the transmission mode (TM) process [13]. The second model is the directed KPR (dKPR) scheme where ($k > 0, r = 0, \gamma > 0$). The third model is the full gKPR process, where all rates are non-zero. For each model, we provide exact solutions for the escape time distributions in the Laplace domain and explicit expressions for the mean ($\mu$) and variance ($\sigma^2$) of the escape times. By considering the squared coefficient of variation, $\mathrm{CV}^2 \equiv \sigma^2/\mu^2$, for these processes (see figures 3 and 5), we explore how these distributions change as the system parameters are adjusted and expose the fact that all three processes exhibit similar, yet not identical, behavior. In particular, we find that all three processes exhibit sharp transitions from near-deterministic ($\mathrm{CV}^2 \ll 1$) to exponential ($\mathrm{CV}^2 = 1$) completion times distribution as the critical parameters change, but that the actual location of this transition differs between the TM and dKPR processes. Furthermore, we observe that all these processes have the same limiting behaviors on either side of the transition, and that the transition from one behavior to the other becomes

sharper as the system size increases. Finally, in subsections 2.6 and 2.7, we also numerically explore the first passage time properties for more complicated cases where the reaction rates are site dependent and where more complicated reaction events are possible. For these processes, we again observe the same simplifying behavior in the process dynamics and sharp transitions that depend on the size of the system (see figures 8 and 9).

### 2.1. Analytical solution of the general model

Let the vector $\mathbf{p} = [p_0(t), p_1(t), \ldots, p_L(t)]^T$ denote the probabilities of each state in the kinetic diagram shown in figure 1. This distribution evolves according to the master equation (ME), which can be written: $\dot{\mathbf{p}}(t) = \mathbf{A}\mathbf{p}(t)$, where the infinitesimal generator matrix $\mathbf{A}$ is

$$A_{ij} = \begin{cases} -k & \text{for} \quad i = j = 0, \\ -k-\gamma-r & \text{for} \quad 0 < i = j \leqslant L-1, \\ \gamma + r & \text{for} \quad (i,j) = (0,1), \\ \gamma & \text{for} \quad i = 0 \ \text{ and } \ 2 \leqslant j \leqslant L-1, \\ r & \text{for} \quad i = j-1 \ \text{ and } \ 2 \leqslant j \leqslant L-1, \\ k & \text{for} \quad i = j+1 \ \text{ and } \ 2 \leqslant j \leqslant L-1, \\ 0 & \text{everywhere else.} \end{cases} \tag{1}$$

By applying the Laplace transform,

$$P_i(s) = \int_0^\infty p_i(t)\, \mathrm{e}^{-st}\, \mathrm{d}t, \tag{2}$$

one can convert the ME to a set of linear algebraic equations:

$$(s - \mathbf{A})\mathbf{P}(s) = \mathbf{p}(t=0) = [1\ 0 \ldots 0]^T. \tag{3}$$

Note that this equation includes the specification of the initial condition, $p_i\ (t=0) = \delta_{i,0}$, where $\delta$ is the Kronecker delta.

We now construct a general solution for this equation in the form

$$P_i(s) = C_1\lambda_1^i + C_2\lambda_2^i. \tag{4}$$

Inserting this into the expression for $0 < i < L-1$, one finds that the space-independent parameters $\lambda_{1,2}$ satisfy

$$\frac{k}{s+k+\gamma+r} + \frac{r}{s+k+\gamma+r}\lambda_\mu^2 - \lambda_\mu = 0. \tag{5}$$

Similarly, the coefficients $C_1$ and $C_2$ must obey the equations for $P_0(s)$ and $P_{L-1}(s)$ in (3), which can be written as

$$(s+k)(C_1+C_2) = 1 + r\,(C_1\lambda_1 + C_2\lambda_2) + \gamma\left(C_1\left[\frac{1-\lambda_1^L}{1-\lambda_1} - 1\right] + C_2\left[\frac{1-\lambda_2^L}{1-\lambda_2} - 1\right]\right) \tag{6}$$

$$C_1\lambda_1^{L-1} + C_2\lambda_2^{L-1} = \frac{k}{s+k+r+\gamma}\left(C_1\lambda_1^{L-2} + C_2\lambda_2^{L-2}\right), \tag{7}$$

where we have applied the geometric series identity, $\sum_{i=1}^{L-1}\lambda^i = \frac{1-\lambda^L}{1-\lambda} - 1$.

Since $p_L(t)$ is the cumulative probability that the system has reached the absorbing state, the first passage time probability density, $f(t) = \mathrm{d}p_L(t)/\mathrm{d}t$, can be written in the Laplace domain as

$$F(s) = k\,P_{L-1}(s). \tag{8}$$

Once this quantity is known, all uncentered moments of the escape time are easily derived as

$$T^{(m)} = \int_0^\infty t^m f(t)\,\mathrm{d}t = (-1)^m \frac{\mathrm{d}^m F(s)}{\mathrm{d}s^m}\bigg|_{s=0}. \tag{9}$$

With this in mind, we now consider the three special cases in the following subsections.

### 2.2. Transmission mode (TM)

The first case to be considered is the transmission mode: the continuous time, discrete space random walk, where the process can only move forward (with rate $k$) or backward (with rate $r$) to its nearest neighbor. Applying the boundary conditions as expressed in equation (7) yields the expressions for $C_1$ and $C_2$:

$$C_1 = \frac{1}{(s+k-r\lambda_2)\left[\frac{\lambda_2-1}{\lambda_1-1} - \left(\frac{\lambda_1}{\lambda_2}\right)^L\right]} \quad \text{and} \quad C_2 = -C_1\frac{\lambda_1^L}{\lambda_2^L}, \tag{10}$$

where $\lambda_1$ and $\lambda_2$ are obtained from equation (5):

$$\lambda_{1,2} = \frac{s+k+r \pm \sqrt{(s+k+r)^2 - 4kr}}{2r}. \tag{11}$$

Following simple algebra, the Laplace transform of the first passage time probability density function (PDF) then becomes

$$F(s) = C_1 k \lambda_1^{L-1}\left(1 - \frac{\lambda_1}{\lambda_2}\right), \tag{12}$$

from which all moments of the first passage time can be extracted. In particular, the mean escape time and the coefficient of variation can be written as

$$\mu_{\mathrm{TM}} = \frac{1}{k}\frac{L - (L+1)\theta + \theta^{L+1}}{(1-\theta)^2}, \tag{13}$$

$$\mathrm{CV}^2_{\mathrm{TM}} = \frac{L - 4\theta - (L+1)\theta^2 + 4(L - L\theta + 1)\theta^{L+1} + \theta^{2L+2}}{(L - L\theta + \theta[\theta^L - 1])^2}, \tag{14}$$

where we have used the definition: $\theta = r/k$. For a deterministic process, CV $= 0$, and for an exponentially distributed one, CV $= 1$. This makes the coefficient of variation a useful property characterizing a distribution.

Figure 2(*A*)–(*C*) shows the effects that changes in the parameters $\theta$ and $L$ have on the distribution of the escape time. In order to show the distribution for diverse parameters simultaneously, time has been rescaled by the mean $\mu$ for each curve, $\tau = t/\mu$. This leads to the probability density $f(\tau) = \mu f(t)$. Figure 2(*A*) shows that, for a fixed $L$, as $\theta$ increases, the distribution becomes broader and approaches an exponential distribution, while as $\theta$ decreases, the distribution approaches a $\Gamma$-distribution, $\Gamma(L, 1/k)$. In order to quantify these behaviors, we provide the trends of the mean and the coefficient of variation for the corresponding regimes.

$$\mu_{\mathrm{TM}}(L,\theta) \approx \begin{cases} \theta^{L-1}/k & \text{for} \quad \theta \gg 2, \\ L/k & \text{for} \quad \theta \ll \frac{L}{L-1}, \end{cases} \tag{15}$$

$$\mathrm{CV}^2_{\mathrm{TM}}(L,\theta) \approx \begin{cases} 1 - 2(L-1)/\theta^L & \text{for} \quad \theta \gg \frac{L+2}{L-1}, \\ 1/L & \text{for} \quad \theta \ll \frac{L}{2(L-1)}. \end{cases} \tag{16}$$

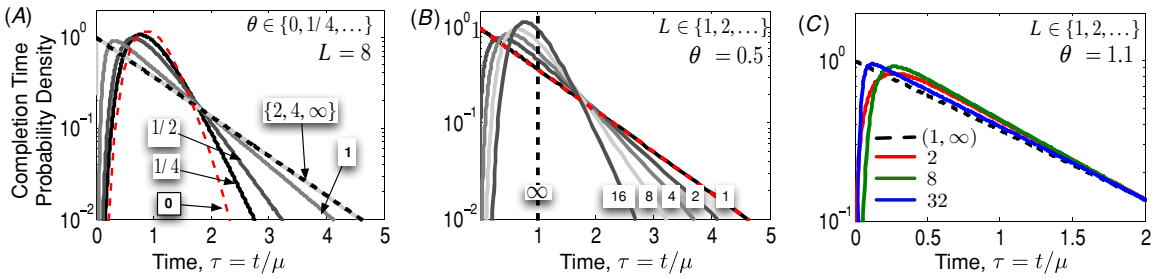


**Figure 2.** Effect of changing $\theta = r/k$ and $L$ on the first passage time distribution for the TM process. The time has been rescaled for each curve as $\tau = t/\mu$. (*A*) First passage time distribution for different values of the backward rate, $r$, and a fixed length $L = 8$. Here $r$ ranges from $k/4$ to $4k$, as denoted in the boxes for the solid lines. The two dashed lines correspond to the limiting cases, $\theta = 0, \infty$ ($\Gamma$-distribution and an exponential, respectively). (*B*, *C*) Effect of changing the length $L$ on the escape time distribution (*B*) for $\theta = 0.5$ and (*C*) for $\theta = 1.1$. For $\theta < 1$, the limiting behavior as $L \to \infty$ is a delta function; for $\theta > 1$, the limiting distribution is the exponential.

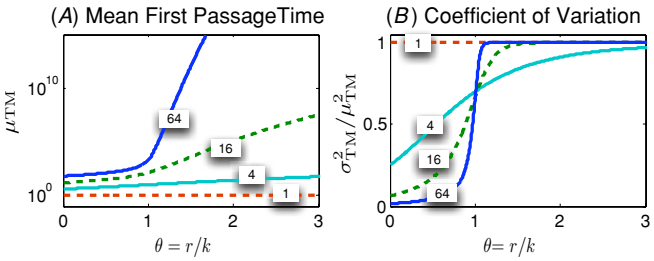


**Figure 3.** Effect of changing the length and the backward rate, $r$, on the mean (*A*) and the squared coefficient of variation (*B*) of the TM process first passage times. The curves have been computed using equations (13) and (14) and are plotted for increasing values of $L = \{1, 2, 4, 8, 16, 32\}$.

It is worth mentioning that $\theta = 1$ means an unbiased random walk, while $\theta < 1 (>1)$ means a walk biased toward the exit (entry) point.

Figures 2(*B*) and (*C*) show that changes in $L$ have different effects on the escape time distribution depending upon the value of $\theta$. When $\theta < 1$, the limiting distribution as $L$ becomes large is a $\delta$-function at $t = L/[k(1-\theta)]$, whereas for $\theta > 1$, the limiting distribution is an exponential with $\mu_{TM} = \theta^{L+1}/[k(1-\theta)^2]$.

Figure 3 illustrates the effect that changes in $L$ and $\theta$ have on $\mu_{TM}$ and $CV^2_{TM}$, as given by equations (13), (14). It is of particular interest to examine these as the chain becomes long. From equation (14), we see that, as $L$ increases, $CV^2_{TM}$ converges point-wise to the step function

$$\lim_{L\to\infty} CV^2_{TM}(L, \theta) = u(\theta - 1) = \begin{cases} 0 & \text{for} \quad \theta < 1, \\ 1 & \text{for} \quad \theta > 1. \end{cases} \tag{17}$$

Numerical analysis of equation (14) around $\theta = 1$ shows that the maximum slope of $CV^2_{TM}$ (to leading order in $L$) occurs at a point that approaches $\theta = 1$ at a rate:

$$1 - \arg\max_{\theta} \frac{dCV^2_{TM}}{d\theta} = \frac{21}{2L^2} + \mathcal{O}(L^{-3}). \tag{18}$$

The slope at $\theta = 1 - 21/(2L^2)$ is

$$\max_{\theta} \frac{dCV^2_{TM}}{d\theta} = \frac{4}{45}L + \mathcal{O}(1). \tag{19}$$

Thus for a given large $L$, the range of $\theta$ over which the first passage time changes from a narrow $\Gamma$-distribution to a broad exponential distribution is centered just left of $\theta = 1$, and it becomes increasingly narrow as $L$ increases.

### *2.3. Directed kinetic proofreading (dKPR)*

The second case we consider is that of directed kinetic proofreading, in which the backward transition rate is neglected, $r = 0$, but the return rate is non-zero, $\gamma > 0$. In this case, the solution is much simpler and can be written as

$$\tilde{p}_i(s) = C_1 \lambda^i, \tag{20}$$

where $\lambda$ is the single root of equation (5) given by

$$\lambda = \frac{k}{s + k + \gamma}, \tag{21}$$

and the coefficient $C_1$ is reduced to

$$C_1 = \frac{1}{s + k - \gamma\left(\frac{1-\lambda^L}{1-\lambda} - 1\right)}. \tag{22}$$

In this case, the Laplace transform of the first passage time is given by

$$f(s) = k p_{L-1}(s) = \frac{k}{s + k - \gamma\left(\frac{1-\lambda^L}{1-\lambda} - 1\right)} \lambda^{L-1}. \tag{23}$$

Defining $\psi = \gamma/k$, the mean and the coefficient of variation of the first passage times can be determined from equation (23):

$$\mu_{dKPR} = \frac{1}{k\psi}[(1+\psi)^L - 1], \tag{24}$$

$$CV^2_{dKPR} = \frac{(1+\psi)^{2L} - 2\psi L(1+\psi)^{L-1} - 1}{(1+\psi)^{2L} - 2(1+\psi)^L + 1}. \tag{25}$$

Figures 4(*A*) and (*B*) show the effects that changes in $\psi$ and $L$ have on the distribution of the waiting times for the dKPR process. As in the previous section, time has been rescaled by $\mu$ for each curve. For a fixed $L$, as $\psi$ changes, the distribution again approaches either an exponential distribution or $\Gamma$-distribution for $\psi \to \infty, 0$, respectively. Unlike for the TM process, the limiting distribution as $L \to \infty$ is exponential for any value of $\psi > 0$.

In figure 5, we illustrate the dependence of $\mu_{dKPR}$ and $CV^2_{dKPR}$ on $L$ and $\psi$. From equations (24) and (25), their limiting behaviors are

$$\mu_{dKPR}(L, \psi) \approx \begin{cases} \psi^{L-1}/k & \text{for} \quad \psi \gg L, \\ L/k & \text{for} \quad \psi \ll L/2, \end{cases} \tag{26}$$

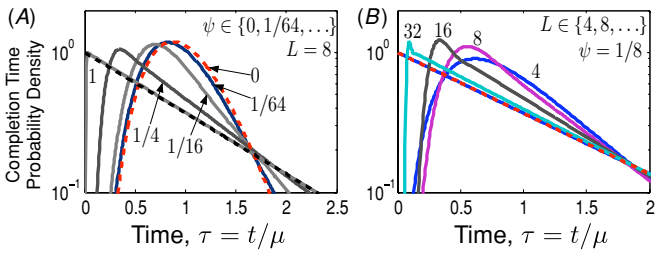


**Figure 4.** Effect of changing $\psi = \gamma/k$ and $L$ on the first passage time distribution (normalized by its mean) for the dKPR process. (*A*) The first passage time distribution for different values of the return rate, $\gamma$ and a fixed length $L = 8$. The parameter $\psi$ ranges from 1/64 to 1 as denoted in the figure. The two dashed lines correspond to the limiting cases, where $\psi = 0, \infty$. The former results in a $\Gamma$-distribution, and the latter in an exponential distribution. (*B*) Effect of changing the length $L$ on the first passage time distribution for $\psi = 1/8$. For any value of $\psi > 0$, the limiting behavior as $L \to \infty$ is an exponential distribution.

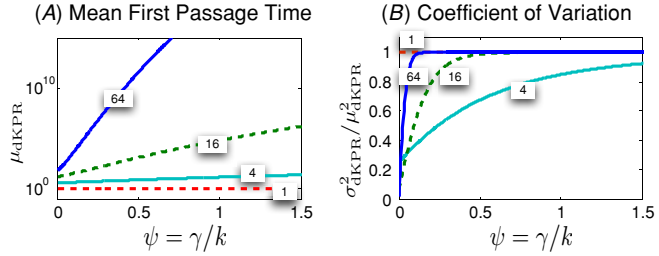


**Figure 5.** Effect of changing the length and the proofreading rate, $\gamma$, on the mean (*A*) and the squared coefficient of variation (*B*) of the escape time for the dKPR system. The curves have been computed analytically using equations (24) and (25) and are plotted for increasing values of $L = \{1, 4, 16, 64\}$.

$$\mathrm{CV}^2_{\mathrm{dKPR}}(L, \psi) \approx \begin{cases} 1 - 2(L-1)/\psi^L & \text{for} \quad \psi \gg 2L, \\ 1/L & \text{for} \quad \psi \ll 3/L^2. \end{cases} \tag{27}$$

Furthermore, as $L$ grows, the coefficient of variation tends to converge point-wise to a step function at $\psi = 0$:

$$\lim_{L\to\infty} \mathrm{CV}^2_{\mathrm{dKPR}} = \begin{cases} 0 & \text{for} \quad \psi = 0, \\ 1 & \text{for} \quad \psi > 0. \end{cases} \tag{28}$$

As in the TM process, this convergence can be studied by examining the maximum slope of the coefficient of variation. Since the second derivative of $\mathrm{CV}^2_{\mathrm{dKPR}}$ is always negative for $\psi \geqslant 0$, this maximum slope occurs at $\psi = 0$. Taking the derivative of equation (25) at the point $\psi = 0$ yields an exact expression for the maximal slope,

$$\max_{\psi} \frac{\mathrm{dCV}^2_{\mathrm{dKPR}}}{\mathrm{d}\psi} = \frac{\mathrm{dCV}^2_{\mathrm{dKPR}}}{\mathrm{d}\psi}\bigg|_{\psi=0} = \frac{L^2 - 1}{3L}. \tag{29}$$

These trends are readily apparent in figure 5(*B*), where as $L$ or $\psi$ increases, $\mathrm{CV}^2$ approaches unity.

### 2.4. Comparison between the TM and the dKPR models

The TM and the dKPR processes exhibit very similar behaviors in their first passage time distributions: for a fixed large $L$, increases in $\theta$ or $\psi$ result in sharp transitions from deterministic to exponential completion times. Moreover, the two processes have *quantitatively* the same limiting behaviors on either side of the transition: the means and the CVs are asymptotically the same functions of $\theta$ and $\psi$ [cf equations (15), (16), (26), (27)].

However, the similarity between the limits of both processes is not exact. For the TM, the deterministic-to-exponential transition (defined by the point of the maximum slope of $\mathrm{CV}^2$) is near $\theta = 1$, approaching it as $L$ grows [cf equation (18)], while the same transition for the dKPR is always at $\psi = 0$. Moreover, although for both models the width of the transition region, as defined by the maximum slope of $\mathrm{CV}^2$, is inversely proportional to the system size (for $L \gg 1$), the width is 15/4 times larger for the TM process. Finally, while the small/large $\theta$ and $\psi$ limits are the same in both models, the terms *small* and *large* themselves have different meanings. In particular, for the TM model the meanings are effectively independent of the system size (equation (16)), while for the dKPR model the meanings strongly depend on $L$ (equation (27)).

### 2.5. General kinetic proofreading (gKPR)

In this case, all the rates $k$, $\gamma$ and $r$ are non-zero, and equation (5) has two solutions

$$\lambda_{1,2} = \frac{s + k + r + \gamma \pm \sqrt{(s + k + r + \gamma)^2 - 4kr}}{2r}. \tag{30}$$

By applying the boundary conditions in equation (7), we obtain the expressions for $C_1$ and $C_2$:

$$C_1 = \frac{1}{r(\lambda_2 - 1) - \gamma\frac{1-\lambda_1^L}{1-\lambda_1} + \left(\frac{\lambda_1}{\lambda_2}\right)^L \left(r(1-\lambda_1) + \gamma\frac{1-\lambda_2^L}{1-\lambda_2}\right)} \tag{31}$$

$$C_2 = -C_1 \left(\frac{\lambda_1}{\lambda_2}\right)^L, \tag{32}$$

with which one can define the Laplace transform of the first passage time PDF:

$$F(s) = C_1 k \lambda_1^{L-1} \left(1 - \frac{\lambda_1}{\lambda_2}\right). \tag{33}$$

Once again, it is possible to derive the mean and variance of the escape time in this scheme

$$\mu_{\mathrm{gKPR}} = \frac{1}{2k\psi}\left[\frac{1 - \theta + \psi}{\sqrt{(1 + \theta + \psi)^2 - 4\theta}}\left(l_+^L - l_-^L\right)\theta^L + \left(l_+^L + l_-^L\right)\theta^L - 2\right], \tag{34}$$

where $l_\pm$ are defined as

$$l_\pm\theta = \frac{1 + \theta + \psi \pm \sqrt{(1 + \theta + \psi)^2 - 4\theta}}{2}. \tag{35}$$

The first passage time variance in this case is given by

$$k^2\psi^2\sigma^2_{\mathrm{gKPR}} = \frac{1}{2}\theta^{2L}\left(l_-^{2L} + l_+^{2L}\right) - 1 + \frac{\theta^{2L-1}(\theta - 1 - \psi)\left(l_-^{2L} - l_+^{2L}\right) + 2L\psi\theta^{L-1}\left(l_-^L - l_+^L\right)}{2(l_+ - l_-)}$$

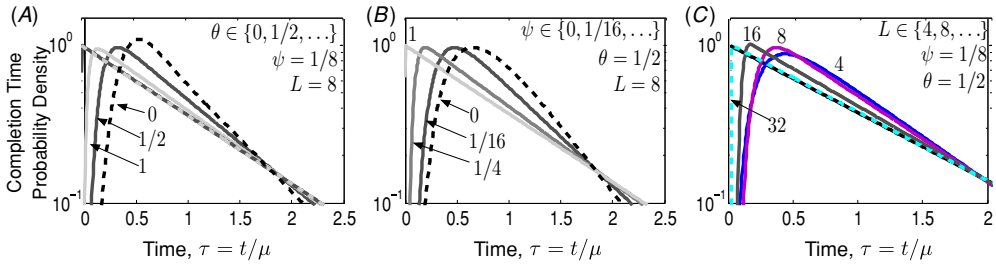


**Figure 6.** The escape time probability density function for the gKPR scheme. (*A*) $\psi = \gamma/k = 1/8$, $L = 8$, and variable of $\theta = r/k$. (*B*) $\theta = 1/2$, $L = 8$ and variable $\psi$. (*C*) $\theta = 1/2$, $\psi = 1/8$, and variable $L$. In all cases, the limiting behavior is an exponential as $L$, $\theta$ or $\psi$ grow.

$$+\psi\frac{2\theta - L\left(l_-^L + l_+^L\right)(-\theta+1+\psi)\theta^{L-2} - \theta^{2L-1}\left(l_-^{2L} + l_+^{2L}\right)}{(l_+ - l_-)^2}$$
$$-\frac{2\psi(1-\theta^{L-1})}{(l_+ - l_-)^2} + \frac{2\theta^{L-2}\psi(\theta - 1 + \psi)\left(l_-^L - l_+^L\right)}{(l_+ - l_-)^3}. \quad (36)$$

Figure 6 illustrates the probability distribution for the exit times of the gKPR process for different $\theta$, $\psi$ and $L$. Based upon the previous results, it is not surprising that the escape time distributions converge to an exponential distribution as $\psi$ or $\theta$ is large (cf figures 6(*A*) and (*B*)), or to a $\Gamma$-distribution when $\psi = \theta = 0$. It is also not surprising that the gKPR first passage time distribution converges to an exponential distribution when $\gamma > 0$ and $L$ is large (cf figure 6(*C*)). What is surprising is how neatly the two constituent processes, TM and dKPR, combine to define the trends of the gKPR process.

Figures 7(*A*)–(*D*) show the mean and the coefficient of variation of the first passage time distributions for this process under various conditions. In panel A, we plot $\mu_{\mathrm{gKPR}}$ as a function of $\theta$ and $\psi$ for a fixed system size of $L = 8$, and panel B shows the corresponding $\mathrm{CV}^2_{\mathrm{gKPR}}$. Panels C and D show the same information, but for $L = 16$. We see that the general trend for the increase in the mean passage time and the convergence of the $\mathrm{CV}^2$ are determined in the same manner as those for the TM and dKPR processes. In particular, we find that the contour lines for both $\mu_{\mathrm{gKPR}}$ and $\mathrm{CV}^2_{\mathrm{gKPR}}$ are almost linear. However, this linearity is not exact—the actual contour lines for $\mu_{\mathrm{gKPR}}(\psi, \theta)$ are slightly concave and the contour lines for $\mathrm{CV}^2_{\mathrm{gKPR}}(\psi, \theta)$ are slightly convex. From figures 3 and 5 above, we see that changes in $L$ have a large effect on the first passage time of the TM and dKPR processes particularly around $\theta = 1$ and $\psi = 0$, respectively. In the gKPR process, these effects correspond to changes in the endpoints, and therefore the slopes of the contour lines in figures 7(*A*)–(*D*).

With explicit expressions for the mean and coefficient of variation, one can again examine their limiting behaviors for growing $\psi$ and $\theta$. In particular, we find that these are equal to those of the TM and the dKPR models when $\theta \to \infty$ or $\psi \to \infty$, respectively. Further, if $L$ is large and $\psi > 0$, the mean first passage is

$$\lim_{L\to\infty} \mu_{\mathrm{gKPR}} \approx \frac{(l_+\theta)^L}{2k\psi}\left(1 + \frac{1-\theta+\psi}{\sqrt{(1+\theta+\psi)^2 - 4\theta}}\right). \quad (37)$$

Finally, the coefficient of variation, $\mathrm{CV}^2_{\mathrm{gKPR}}$, approaches unity for all values except when $\psi = 0$ and $\theta < 1$, and

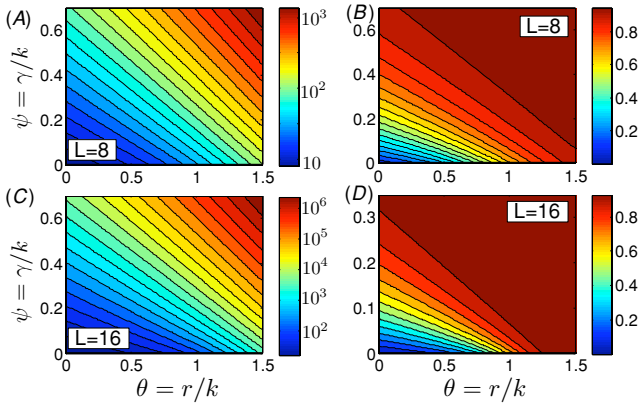


**Figure 7.** Effects of parameter variation on the escape time distribution for the gKPR process. (*A*) Mean completion time versus $\theta$ and $\psi$ for $L = 8$. (*B*) Coefficient of variation, $\mathrm{CV}^2_{\mathrm{gKPR}}$ versus $\theta$ and $\psi$ for $L = 8$. (*C*, *D*) the same for $L = 16$.

$$\mathrm{CV}^2_{\mathrm{gKPR}}(L,\theta) \approx \begin{cases} 1 - \frac{2(L-1)}{(\psi+\theta)^L} & \text{for} \quad \psi \gg 2L \quad \text{and} \quad \theta \gg 4, \\ 1/L & \text{for} \quad \psi \ll \frac{L^2}{3} \quad \text{and} \quad \theta \ll \frac{1}{2}. \end{cases} \quad (38)$$

This shows that, for large proofreading and backward rates, the two effects have equal influences on the distribution of the completion time. However, one should bear in mind that, again, the meaning of small/large $\theta$, $\psi$ is different.

### *2.6. Kinetic proofreading with site-dependent rates*

The previous subsections have shown that the TM, dKPR and gKPR processes all exhibit a similar simplification of behavior when all rates are the same at every intermediate state in the process. In reality, these rates may vary from one site to the next since each transition may correspond to a different physical reaction. In the case of the dKPR, one can still derive expressions for the first passage time distributions (the expression is omitted here), and in the case of more complicated processes, one can explore these distributions numerically. To illustrate the effects of such variation, we have numerically explored a gKPR process where every rate is different, but chosen from some relatively broad lognormal distribution. Figure 8 shows how such site-dependent rates affect the coefficient of variation for the gKPR process. Here all forward and backward rates, $\{r_i, k_i, \gamma_i\}$ have been generated from the same distribution, and then the backward rates $\{\gamma_i, r_i\}$

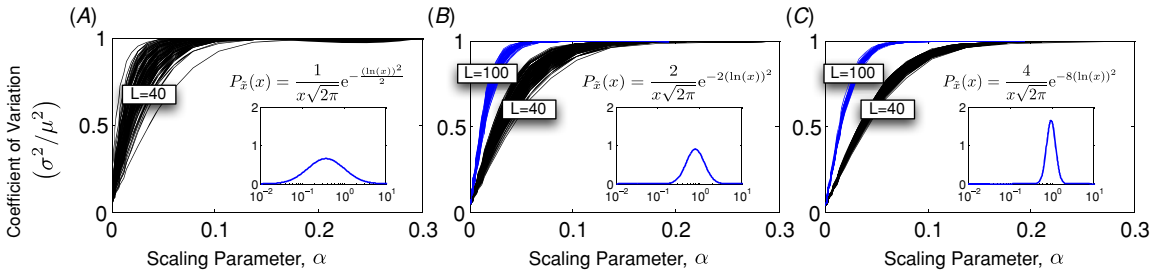


**Figure 8.** Coefficient of variation for a gKPR process with random parameters versus backward to forward bias. The length of the process is either 40 (black lines) or 100 (blue lines) and all rates $k_i$, $\gamma_i/\alpha$ and $r_i/\alpha$ are taken independently from the lognormal distribution shown in the inset. The three panels correspond to three increasingly narrow distributions for the parameters.

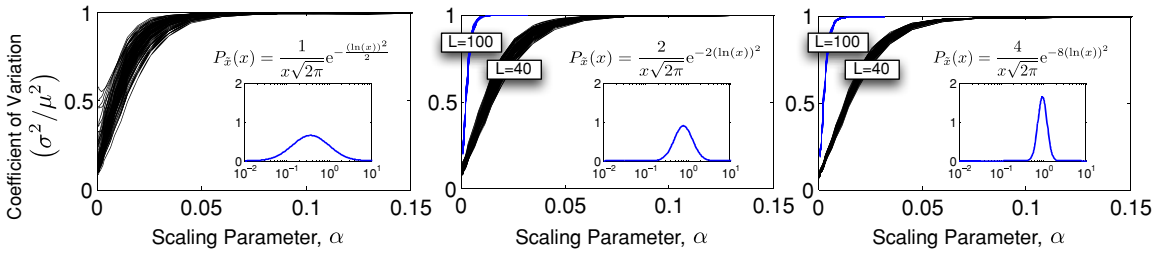


**Figure 9.** Coefficient of variation for an arbitrary kinetic proofreading like process with random parameters versus backward to forward bias. The master equation for this process is $\dot{\mathbf{p}}(t) = (\alpha\mathbf{B} + \mathbf{F})\,\mathbf{p}(t)$, where $\mathbf{F}$ is banded such that the system can move 1, 2 or 3 steps forward in a single jump, and $\mathbf{B}$ is upper triangular such that the process can move any number of steps backwards. The length of the process is either 40 (black lines) or 100 (blue lines), and each non-zero element of $\mathbf{F}$ and $\mathbf{B}$ is randomly chosen from the lognormal distribution plotted in the inset. The three panels correspond to three increasingly narrow distributions for the parameters.

have been scaled uniformly by a parameter, $\alpha$ that has been used to adjust the bias from completely forward $\alpha = 0$ to backward $\alpha \gg 1/L$. From figure 8, we see once again that there is a sharp transition from when the coefficient of variation is small at $\alpha = 0$ to when the coefficient of variation is near one when the bias is backward. As in the previous systems, this transition depends upon on the length of the system—longer lengths correspond to sharper transitions. Furthermore, as the lengths increase, variation in the parameters appears to be less important as can be seen by comparing the variation in the curves corresponding to $L = 100$ (blue curves) to those for a smaller length of $L = 40$ (black curves).

### *2.7. Multiple leap completion processes*

In addition to the gKPR scheme illustrated by figure 1, we also explore a much more general set of multistep completion processes where reactions can take the system not just one, but many steps toward the completion state or toward the initial state. In terms of chemical processes, these multiple step jumps could correspond to additions or removals of different multi-molecular complexes rather than just individual molecules. In this case, there are now many different interconnected pathways by which the process can travel from state $i = 0$ to $i = L$. In such systems, the master equation, $\mathrm{d}\mathbf{p}/\mathrm{d}t = \mathbf{A}\mathbf{p}(t)$, has an infinitesimal generator, $\mathbf{A}$ given by $\mathbf{A} = \alpha\mathbf{B} + \mathbf{F}$, where the 'backward' matrix, $\mathbf{B}$, is upper-triangular and represents reactions that allow the system to return an arbitrary number of states backward with certain site-dependent rates, and the 'forward' matrix $\mathbf{F}$ is a lower-triangular banded matrix, which allows for different forward jumps of lengths $m < L$, again with site-dependent rates. Since $m$ is constrained to be less than $L$, there is always a minimum of about $L/m$ jumps necessary to complete the process.

In the expression of the infinitesimal generator, $\alpha$ controls the bias, and we show once again that there is a sharp threshold between an almost deterministic and an exponential behavior as $\alpha$ grows. For this general process, we have randomly generated hundreds of realizations each with different site-dependent rates taken from a broad lognormal distribution, and we find that for every such parameter set, there is a sharp transition from a narrow 'deterministic' to a broad exponential waiting time distribution as can be seen in figure 9. Furthermore, despite drastic differences in the randomly chosen parameters, we find that the dynamical behaviors of the systems are so close that it is difficult to distinguish one parameter set from the next based solely on the waiting time. Finally, we find the same dependence of this transition on the size of the system as has been observed for dKPR and gKPR processes (compare the process with 40 steps (black lines) to the process with 100 steps (blue lines) in figure 9).

## 3. Discussion

The results for the coefficient of variation of the escape time distribution, as well as the shapes of the distributions themselves, clearly show that the kinetic proofreading process and other multistep completion processes have two simple limiting behaviors as the system size increases. First, when the overall bias is forward, the completion time becomes narrowly distributed. Second, when the overall bias is backward, the

escape time distribution approaches an exponential. Both of these behaviors are substantially simpler than one could have expected from the original complex kinetic diagram, implying that the observable behavior of this complex system can be approximated accurately by a single-parameter equivalent, corresponding either to a deterministic reaction or a simple two-state Markov chain. Interestingly, the approach to the deterministic regime as the system size grows is well understood (see, for example, [22] on the discussion of this effect in the context of reproducibility of responses of rod cells to single photon capture events). However, the exponential regime has not been explored extensively before, even though it is the more robust of the two, emerging for any $\psi > 0$.

Both limiting behaviors of these systems are explainable by simple intuitive arguments. First, a system with a forward bias completes the entire process in a certain characteristic time, and the relative standard deviation of this time scales as $1/\sqrt{\text{number of steps}}$, as is always the case for the addition of independent identically distributed random variables. In the opposite case, the backward bias ensures that the process repeatedly returns to the initial state, from which many *independent* escape attempts are made. Due to the independence, the number of such attempts before a success has a geometric distribution (the discrete analog of an exponential distribution), and its form effectively defines the first passage time distribution. In other words, the system tries to climb out of a free energy well (with the ground state near the entry point), and escape times in such cases are usually exponentially distributed.

Although the KPR models most rigorously analyzed here are relatively simple linear chain processes with site-independent transition rates, our numerical studies strongly suggest that the conclusions we make generalize to more complicated systems. We have shown numerically that our conclusions do not change when the kinetic rates $k, r, \gamma$ are site-specific and/or when the reactions allow for certain states to be skipped and for there to be many different interconnected pathways by which the process may be completed. Similarly, if biochemical processes involve multiple independent pathways, each with exponential/deterministic waiting times, then the first of these pathways to complete will also be exponential/deterministic. Furthermore, first passage times for higher dimensional random walks also frequently exhibit simplified dynamics, as has been shown via reductions to a stochastic model of the genetic toggle switch [23]. Finally, the 'free energy well' argument says that the overall bias of a system's motion will control the choice between the exponential (Markovian) and the deterministic behaviors even for more complex systems. In particular, it is clear that any KPR-like system, where a strong backward bias is required to undo potential mistakes, is likely to fall in the exponential escape time distribution regime.

Given that so much structural complexity is used to achieve a very simple dynamics in these processes, it is natural to ask why the complexity is used at all. One hypothesis is that such agglomeration of multiple independent kinetic parameters into a few coarse-grained variables means that multiple chemotypes can result in the same phenotype. Thus, the system possesses *many* situationally *sensitive* knobs with which it can compensate for environmental changes and maintain *a few* simple behaviors. Such adaptive flexibility has been observed in a variety of contexts [24–26]. An alternative hypothesis may be that these extra elements are vestigial network components to which the cell is *insensitive* in its current evolutionary or developmental situation. The current work provides a starting point to evaluate these possibilities via parametric sensitivity analysis.

Finally, the fact that the KPR process, as well as many others, has such simple limiting behaviors has important consequences for the modeling of biochemical systems. The bad news is that it is unreasonable to hope to characterize individual molecular reactions with observations of the input-to-output responses—many different internal organizations will result in equivalent observable behaviors. The good news is that, when attempting to understand such processes in a wider cellular context, it is often unnecessary to explicitly treat every individual step—a coarse-grained model with only a handful of aggregate parameters may be sufficient. This result suggests an explanation for why simple phenomenological Markovian reaction rate models of complicated processes, such as transcription, translation, enzyme activation and others, have had such a great success in explaining biological data.

## Acknowledgments

We thank N Sinitsyn and N Hengartner for discussions during early stages of this work. We also thank B Goldstein, R Gutenkunst and M Monine. We also thank the Center for Nonlinear Studies for providing an excellent collaborative environment within Los Alamos National Lab. This work was partially funded by LANL LDRD program.

## References

[1] Faeder J, Hlavacek W, Reischl I, Blinov M, Metzger H, Redondo A, Wofsy C and Goldstein B 2003 Investigation of early events in FceRI-mediated signaling using a detailed mathematical model *J. Immunol.* **170** 3769–81

[2] Hopfield J 1974 Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity *Proc. Natl Acad. Sci. USA* **71** 4135–9

[3] Yan J, Magnasco M and Marko J 1999 A kinetic proofreading mechanism for disentanglement of DNA by topoisomerases *Nature* **401** 932–5

[4] Sancar A, Unsal-Kacmaz L L-B amd K and Linn S 2004 Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints *Annu. Rev. Biochem.* **73** 39–85

[5] Blanchard S, Gonzalez J R, Kim H, Chu S and Puglist J 2004 tRNA selection and kinetic proofreading in translation *Nat. Struct. Mol. Biol.* **11** 1008–14

[6] Jovanovic-Talisman T, Tetenbaum-Novatt J, McKenney A, Zilman A, Peters R, Rout M and Chait B 2008 Artificial nanopores that mimic the transport selectivity of the nuclear pore complex *Nature* **457** 1023

[7] McKeithan T 1995 Kinetic proofreading in T-cell receptor signal transduction *Proc. Natl Acad. Sci. USA* **92** 5042–6

[8] Rabinowitz J, Beeson C, Lyons D, Davis M and McConnell H 1996 Kinetic discrimination in T-cell activation *Proc. Natl Acad. Sci. USA* **93** 1401–5

[9] Rosette C *et al* 2001 The impact of duration versus extent of TCR occupancy of T cell activation: a revision of the kinetic proofreading model *Immunity* **15** 59–70
[10] Liu Z, Haleem-Smith H, Chen H and Metzger H 2001 Unexpected signals in a system subject to kinetic proofreading *Proc. Natl Acad. Sci. USA* **98** 7289–94
[11] Goldstein B, Faeder J and Hlavacek W 2004 Mathematical and computational models of immune-receptor signalling *Nat. Rev. Immunol.* **4** 445–56
[12] Hlavacek W, Redondo A, Wofsy C and Goldstein B 2002 Kinetic proofreading in receptor-mediated transduction of cellular signals: receptor aggregation partially activated receptors, and cytosolic messengers *Bull. Math. Biol.* **64** 887–911
[13] Redner S 2001 *A Guide To First-Passage Processes* (Cambridge: Cambridge University Press)
[14] D'Orsogna M and Chou T 2005 First passage and cooperativity of queuing kinetics *Phys. Rev. Lett.* **95** 170603
[15] Shaevitz J, Block S and Schnitzer M 2005 Statistical kinetics of macromolecular dynamics *Biophys. J.* **89** 2277–85
[16] Lee C L, Stell G and Wang J 2003 First passage time distribution and non-makovian dynamics of protein folding *J. Chem. Phys.* **118** 959
[17] Leite V B P, Alonso L C P, Newton M and Wang J 2005 Single molecule electron transfer dynamics in complex environments *Phys. Rev. Lett.* **95** 118301
[18] Lu T, Shen T, Zong C, Hasty J and Wolynes P 2006 Statistic of cellular signal transduction as a race to the nucleus by multiple random walkers in compartment/phosphorylation space *Proc. Natl Acad. Sci. USA* **103** 16752–7
[19] Xu W L, Xue K and Wang E K 2008 Exploring the origin of power law distribution in single-molecule conformation dynamics: energy landscape perspectives *Chem. Phys. Lett.* **463** 405–9
[20] Ninio J 1987 Alternative to the steady-state method: derivation of reaction rates from first-passage times and pathway probabilities *Proc. Natl Acad. Sci. USA* **84** 663–7
[21] Bel G and Barkai E 2006 Random walk to a nonergodic equilibrium concept *Phys. Rev.* E **73** 016125
[22] Doan T, Mendez A, Detwiler P, Chen J and Rieke F 2006 Multiple phosphorylation sites confer reproducibility of the rod's single-photon responses *Science* **313** 530
[23] Munsky B and Khammash M 2008 Transient analysis of stochastic switches and trajectories with applications to gene regulatory networks *IET Syst. Biol.* **2** 323–33
[24] Stern S, Dror T, Stolovicki E, Brenner N and Braun E 2007 Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge *Mol. Syst. Biol.* **3** 106
[25] Ziv E, Nemenman I and Wiggins C 2007 Optimal information processing in small stochastic biochemical networks *PLoS ONE* **2** e1077
[26] Gutenkunst R, Waterfall J, Casey F, Brown K, Myers C and Sethna J 2007 Universally sloppy parameter sensitivities in systems biology *PLoS Comput. Biol.* **3** e189