Exact probabilistic solution of spatial-dependent stochastics and associated spatial potential landscape for the bicoid protein

David Lepzelter[1] and Jin Wang[1,2,3,*]

[1]*Department of Physics and Astronomy, State University of New York at Stony Brook, Stony Brook, New York 11794, USA*

[2]*Department of Chemistry, State University of New York at Stony Brook, Stony Brook, New York 11794, USA*

[3]*State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin 130022, People's Republic of China*

(Received 26 September 2007; revised manuscript received 13 March 2008; published 22 April 2008)

We investigated the spatial-dependent stochastic effects originating from the finite number of *bicoid* proteins in *Drosophila melanogaster*, which are crucial to cell development. We obtained an exact solution to the spatial-dependent stochastic chemical master equation and recovered the usual reaction-diffusion solution for the average of the bicoid concentration, valid in the bulk. We also used the steady state probability to get the spatial potential landscape. The stochastic effects are captured by the Poisson distribution; so, as the average of the bicoid concentration decreases from the anterior (A) to the posterior (P) of the embryo, the statistical fluctuations also decrease. An alternative way of interpreting this is that the shape of the spatial potential landscape shrinks from A to P. While the mathematical result is known, we offer a simple approach to understanding why it is what it is and give associated physical intuitions. The approach can be generalized and applied to any problem with a particle that diffuses, decays, and has a stochastic source.

DOI: 10.1103/PhysRevE.77.041917 PACS number(s): 87.18.−h, 82.39.−k

Cell functions are often realized by complex networks of proteins and genes interacting with one another. Researchers have generally used chemical kinetic equations to represent the interactions in these networks. However, while conventional chemical kinetic equations work perfectly well under the bulk conditions, they do not always give accurate results in the cell due to large statistical fluctuations caused by the relatively small number of molecules involved [1–4]. Furthermore, an organism is not a homogeneous system; spatial dependence is crucial for many biological functions. One concrete example of a system for which these details are important involves the embryo development of *Drosophila melanogaster*, the common fruit fly.

D. melanogaster is a common organism for genetic and developmental biology for several reasons. It is easy to perform experiments on, and a large amount of background knowledge exists on it. Also, until late in its development as an embyro, the organism lacks distinct cells; each embryo has a large number of nuclei, but there are no cell membranes to block the diffusion of proteins from one nucleus to another. This last piece of information makes *D. melanogaster* ideal for refining ideas of diffusion-related spatial pattern formation in an embryo.

Of the proteins and genes in the embryo, a few stand out for having large effects on development. One of these is the *bicoid* protein. It is useful to observe for four main reasons. First, it seems to have a direct regulatory effect on many developmental genes [5,6]. Second, its production is independent of the presence or absence of other developmental proteins. Third, its average concentration level at any given spatial point is essentially constant in time for much of gastrulation. Fourth, its concentration and effects on other proteins are very obviously subject to statistical fluctuations, which make the use of stochastics absolutely necessary for a realistic understanding of the system [7,8]. Specifically, the internal noise of the system, the variability due to finite numbers of proteins, has caused significant debate on how the embryo can so accurately determine the spatial location of the sudden jump in the concentration of a protein, called *hunchback*, which is dependent on *bicoid* concentration. The *hunchback* gradient, in turn, is central in determining the location of the head of the fly, and its spatial precision is greater than one might expect [7,8].

These aspects of the protein have inspired numerical calculations using implementations of the chemical master equation for the system [7,9]. Such calculations have generally been in one spatial dimension because there is an easily recognizable gradient in the anterior-posterior direction which has a definite effect on development. The dorsal-ventral axis, in contrast, has a much smaller gradient, and at least in the case of *bicoid*, it seems to have less of an effect on the initial stages of development.

Even these numerical calculations need some assumptions, however. We will show in this paper that the same simple assumptions which make the problem calculable numerically or using field theory (as in [10]) also make an exact and straightforward analytic solution possible for the bicoid probability distribution in one spatial dimension, and offer arguments as to why the same methods should work in other geometries. This is more than simply a continuation of a trend away from the bulk average concentration calculations done in the past, though it is that as well; even with an exact solution already known from [10], this analysis is important because it significantly clarifies our understanding of the system and similar systems. It offers a simple global characterization of the system, as opposed to local approaches or field theoretic characterizations.

The basic assumptions of our approach involve the three processes which govern the protein's behavior. First, the pro-

[*] Corresponding author. jin.wang.1@stonybrook.edu

duction of *bicoid* (which occurs in a highly localized part of the embryo) is assumed to be stochastic in nature. Second, movement of the protein through the embryo is assumed to behave according to traditional (random-walk-type) stochastic diffusion. Third, *bicoid* degradation is assumed to be a stochastic event, i.e., through a decay reaction $\mathrm{Bcd}\xrightarrow{k}\emptyset$.

These three assumptions lead to a spatial-dependent chemical master equation, which is a complete description of the probabilities involved in the system assuming no other effects. The importance of the chemical master equation to a gene-protein network can be compared to that of the Schrödinger equation for an atom: it forms the fundamental basis for further detailed characterization [11]. While the nonlinear rate equations provide a quantitative description of cellular networks on the average level, often showing complex behavior, the probabilistic description obeys the linear master equation. Linearity can make the probabilistic description more regular even in more chaotic deterministic cases. While the chemical kinetics gives reasonable description in the bulk, the probabilistic description provides the foundation for the mesoscopic intracellular network. The chemical kinetics gives the deterministic trajectories with a probability of one, but the probabilistic description provides a distribution of possible protein concentrations. In other words, when one knows the probability distribution, one knows the weights of individual states in protein concentration space. It is in this sense we can call it a probabilistic landscape in protein concentration space.

Landscape concepts have been introduced to the biology community in the areas of molecular and developmental biology [12,13] and population dynamics [14,15]. The landscape is quantified in the areas of protein dynamics [16] and protein folding [17] while the potential energy landscape is known *a priori* with quasiequilibrium assumptions. For nonequilibrium cellular networks, the potential landscape is not known *a priori*. One can, however, obtain the information by finding the probabilistic distribution through solving the master equation. A generalized potential U corresponding to the probabilistic description P for the nonequilibrium networks can be defined as $U=-\ln P$ in analogy with the Boltzman relationship in equilibrium statistical mechanics [11,18–24]. Once the landscape can be quantified this way, it can give a global characterization of the network, providing the weight distribution in the protein concentration space and quantifying the importance of each state (in terms of weight). The stability, robustness, and function of the network can be now studied in a global and physical way from landscape perspectives [11,18–24].

When the concentration has spatial dependence, as in the developmental process, the probability distribution in protein concentration space becomes a probabilistic functional of protein concentrations which themselves also depend on space. It is in that sense a statistical probabilistic field theory representation (field being the protein concentrations which depend on space). Therefore by solving the probabilistic functional, we can map out the spatial-dependent landscape of the cellular network. This is crucial for unraveling the origin of stability, robustness, and function of spatial-dependent cellular networks.

It should be noted that one complicating factor generally not included in master equation calculations is external noise, which can represent anything from environmental temperature fluctuations to diffusion from outside the embryo, and is not explicitly accounted for in this model. Ignoring such effects, one arrives at a chemical master equation. This equation is most easily expressed in terms of a vector, $\vec{n}$, whose components $\vec{n}=(n_0,n_{\Delta x},n_{2\Delta x},\ldots)=(\{n_x\})$ correspond to the number of *bicoid* proteins at evenly spaced spatial positions $x=0,\Delta x,2\Delta x,\ldots$, with Δx an essentially arbitrary constant. The equation is [25,26]

$$g[P(\vec{n}-\hat{0})-P(\vec{n})],$$

$$\begin{aligned}\frac{dP(\vec{n})}{dt} &= +k\sum\nolimits_x[(n_x+1)P(\vec{n}+\hat{x})-n_xP(\vec{n})]\\ &\quad + D\sum\nolimits_{xy}[(n_x+1)P(\vec{n}+\hat{x}-\hat{y})-n_xP(\vec{n})],\end{aligned} \qquad (1)$$

where $P(\vec{n})$ is the probability that number and position of proteins is described exactly by $\vec{n}$. g is the rate of protein generation, $\hat{0}$ is a unit vector in the 0 space (representing a single protein at the origin, spatial point 0), and the term multiplying g represents the process of a protein being created at the origin. k is the rate of degradation, $\hat{x}$ represents a single protein at point x, and the term multiplying k represents the protein decay at any spatial position. D is the finite-volume diffusion rate, and the term multiplying it gives diffusion from each spatial point to its neighbors. The sums over x are over all space $x=0,\Delta x,2\Delta x,\ldots$, and over y are all spatial neighbors of x ($y=x\pm\Delta x$).

The next step in this process would be to find a time independent steady-state solution, $\frac{dP(\vec{n})}{dt}=0$ for all $\vec{n}$. It should be noted that the deterministic form of this problem can be easily solved; $0=\frac{\partial C}{\partial t}=D\frac{\partial^2 C}{\partial x^2}+g\delta(x)-kC$ yields $C(x)=(g/\sqrt{kD})e^{-x\sqrt{k/D}}$. This corresponds to the reaction diffusion equation and its associated solution, often used in bulk studies. However, the uncertainties in concentration due to the finite number of molecules can only be found by solving the master equation. While the master equation itself does not immediately suggest a solution, the assumptions made do strongly suggest the use of Green's function techniques often encountered in physics and chemistry. Each individual protein has no interactions of any kind with any other protein; its creation, diffusion, and decay are all completely independent of any other effects. Therefore we propose an ansatz in a format slightly different from that of the master equation,

$$P=\sum_{n=0}^{\infty}\frac{e^{-g/k}(g/k)^n}{n!}\prod_{m=1}^{n}G(x_m), \qquad (2)$$

where n is the total number of proteins present in the system, m is a representation of each protein in the system, and $G(x_m)$ is actually a multidimensional generating function describing the chance that protein m is at the spatial point x_m. One can understand the probability expression above as the decomposition of the generation functions in Poisson space.

In order to prove the validity of the ansatz, we must first match its form more closely with the notation used in the master equation. Let us consider the spatial point x. For any given total number of proteins n, there are n proteins each with probability distribution G. Let G_x be the discrete version of $G(x)$. Then for a given n, the probability of n_x proteins existing at point x should be $P(n_x) = \binom{n}{n_x}(G_x)^{n_x}$.

Combining this with the simple Poisson probability of n proteins existing, we find

$$P(\vec{n}) = \prod_{x}^{\text{all space}} \frac{e^{-gG_x/k}(gG_x/k)^{n_x}}{n_x!}. \tag{3}$$

We note that, while we implicitly used a vector $\vec{n}$ which began at the spatial point 0, the solution takes the form of Eq. (3) for other geometries as well. We also note that the form of the solution for the probability of *bicoid* concentration $P(\vec{n})$ is simply that of a Poisson distribution with average value gG_x/k for each point in space, without spatial correlations. This is suggested, but not explored in detail, by a solution to a different problem in [27]. Others, e.g., [10], state a Poisson solution, but we believe that this approach offers a useful contribution to the understanding of the problem because it is relatively simple and straightforward.

Given the form of the ansatz, we will define the Poisson distribution for the point x, $\mathcal{P}_x(n_x) = \frac{e^{-gG_x/k}(gG_x/k)^{n_x}}{n_x!}$, and note that $\mathcal{P}_x(n_x+1) = \frac{e^{-gG_x/k}(gG_x/k)^{(n_x+1)}}{(n_x+1)!} = \frac{gG_x/k}{n_x+1}\mathcal{P}_x(n_x)$.

Then inserting the ansatz into the master equation,

$$\frac{dP(\vec{n})}{dt} = \left[\begin{array}{c} g\left(\dfrac{n_0}{gG_0/k} - 1\right) \\ + k\sum_x [(gG_x/k) - n_x] \\ + D\sum_{xy}\left[\left(n_y \dfrac{G_x}{G_y}\right) - n_x\right] \end{array} \right] \prod_{x'}^{\text{space}} \mathcal{P}_{x'}(n_{x'}).$$

Using $\Sigma_x G_x = 1$, and rearranging a sum,

$$\frac{dP(\vec{n})}{dt} = \left[\begin{array}{c} \dfrac{kn_0}{G_0} - g + k\dfrac{g}{k} - k\sum_x n_x \\ + D\sum_{xy}\left[\left(n_x \dfrac{G_y}{G_x}\right) - n_x\right] \end{array} \right] \prod_{x'}^{\text{space}} \mathcal{P}_{x'}(n_{x'}).$$

Again, all space in this geometry is $x = 0, \Delta x, 2\Delta x, \ldots$, and the neighbors x are $y = x \pm \Delta x$, except at $x = 0$ where y can only be Δx. Therefore

$$\frac{dP(\vec{n})}{dt} = \left[\begin{array}{c} \dfrac{kn_0}{G_0} - kn_0 + Dn_0\left(\dfrac{G_{\Delta x}}{G_0} - 1\right) \\ - k\sum_{x=1}^{\infty} n_x\left[1 + \dfrac{D}{k}\left(\dfrac{G_{x+\Delta x}}{G_x} + \dfrac{G_{x-\Delta x}}{G_x} - 2\right)\right] \end{array} \right] \times \prod_{x'=0}^{\infty} \mathcal{P}_{x'}(n_{x'}). \tag{4}$$

Since we are interested in the steady state solution, we solve for $\frac{dP(\vec{n})}{dt} = 0$. As n_x can in theory be any finite number, to ensure that the right-hand side of Eq. (4) is 0 we must ensure that the coefficients of each n_x are 0.

$$\frac{k}{G_0} - k + D\left(\frac{G_{\Delta x}}{G_0} - 1\right) = 0$$

$$-k + D\left(\frac{G_{x+\Delta x}}{G_x} + \frac{G_{x-\Delta x}}{G_x} - 2\right) = 0, \quad x > 0$$

Defining for convenience $z \equiv (1 + \frac{k}{2D} - \sqrt{\frac{k^2}{4D^2} + \frac{k}{D}})$, the solution,

$$G_0 = 1 - z,$$

$$G_x = zG_{x-\Delta x} = z^{x/\Delta x}(1 - z) = (1 - z)e^{\ln(z)x/\Delta x},$$

is simple. Since the mean of the distribution should be given by gG_x/k, it is reassuring to note that it corresponds to a decaying exponential function ($\ln z < 0$), the same form expected from both experiment and nonstochastic theory. It should be noted that this does not correspond exactly to the expected $e^{-x\sqrt{k/D}}$; this is because the definition of D is not precisely the same for finite-volume spaces, and because within each space the solution is assumed to be well-mixed. However, both of these issues can be avoided by using small enough distances between spatial points.

Substituting G_x into our formulation of $P(\vec{n})$, we obtain the final analytical expression for the probability:

$$P(\vec{n}) = \prod_{x}^{\text{all space}} \frac{e^{-gz^{x/\Delta x}(1-z)/k}[gz^{x/\Delta x}(1-z)/k]^{n_x}}{n_x!}. \tag{5}$$

We note that if this form is used in the original chemical master equation, it does in fact give $\frac{dP}{dt} = 0$, and therefore the ansatz is the correct and exact analytical solution to the steady-state problem.

Both the mean values and the noise given by this model, which decay exponentially from anterior (A) to posterior (P), seem to match current experimental data (see Figs. 1 and 2), with some caveats regarding the effective diffusion constant [8]. In both figures, the first 20% of the embryo is assumed to be part of a diffuse source of unknown local concentration and is therefore not considered part of the overall Green function fit. The main portion of Fig. 1 has a line with predicted values, and two more with predicted uncertainties from both intrinsic and experimental noise.

The inset shows probability distributions with only intrinsic (nonexperimental) noise for nuclei at 47% and 49% embryo length. These two locations mark the end of the head and the beginning of the body, and in spite of significant overlap in the probability distributions of *bicoid* concentrations (and additional issues with subsampling), the embryo is capable of almost perfectly distinguishing on which side of the 48% embryo length boundary they fall. This fact is at the center of current debate on the subject of how precisely such accuracy is achieved; some, e.g., [8], favor a simple averaging scheme over neighboring spatial points by means of *hunchback* self-interaction, while others suggest a more complex system involving more kinds of proteins and some pattern formation [28,29]. In either case, it is useful to know that the minimum reasonable noise, that of a Poisson distri-

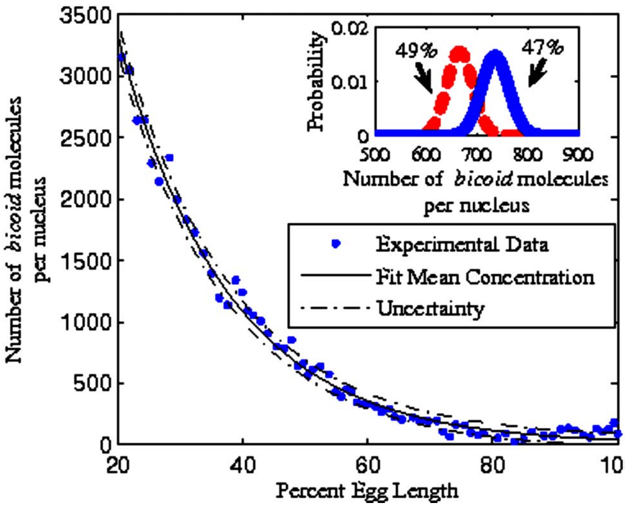

FIG. 1. (Color online) Calculation of expected distribution versus data, courtesy of Dr. Tomas Gregor, from an embryo. Error bars include intrinsic Poisson noise from proteins, photon counting noise, and a small constant Gaussian noise intended to account for focal plane alignment. Errors from nuclear identification are not included. This fit gives $\chi^2/\text{degree of freedom}=1.26$.

bution, can be considered correct and exact given the basic assumptions mentioned previously.

These statistical fluctuations, given the Poisson form of the solution, are easy to calculate: $\sigma=\sqrt{gG_x/k}=\sqrt{ge^{\ln(z)x/\Delta x}(1-z)/k}$. We see that, since $\ln z<0$, the size of the fluctuation decays from A to P (clearly shown in Fig. 2). Adding expected experimental noise from photon counting and focal plane alignment gives a larger (and no longer purely Poisson) noise. The inset shows fractional uncertainty, σ/C, where C is the number of *bicoid* molecules, and the trend of increasing total (experimental plus intrinsic) fractional uncertainty from A to P agrees with experiment.

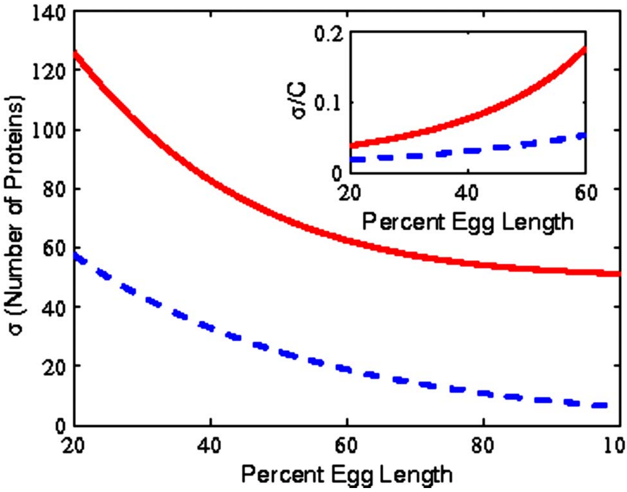

FIG. 2. (Color online) Calculated noise from Fig. 1; dotted blue line shows intrinsic noise only, while the solid red line shows both intrinsic and predicted experimental noise. Inset shows the predicted total experimental standard deviation divided by the mean, with dotted blue and solid red lines having the same meaning. Both solid lines follow roughly the trends as in [8], though without errors from nuclear identification they are somewhat smaller than the real experimental uncertainties.

It should be noted that these experimental results are likely accurate, but still somewhat preliminary; other results, with more data but obtained using different methods involving difficult calibrations, may support a different distribution [7]. However, should other distributions prove more realistic, the examined model should give valuable insight into the actual mechanisms in *D. melanogaster*: non-Poisson generation, nonmonomer decay, or some other important process not previously mentioned would be vital in forming the shape of the distribution.

While the precise mathematics have involved a one-dimensional problem with a source exactly at one end, it would not be difficult to prove the validity of the same kind of solution with a different geometry. Another boundary condition, a moved or spread-out source, and an additional dimension or two should make it less easy to find the solution for G_x by hand, but the problem is not difficult with a computer. In any case, the validity of the general solution, with a Poisson distribution at every point in space, can be applied in any situation for which there are particles which diffuse, decay, and have one or multiple Markovian (Poisson-type) sources.

It is important that, even though diffusion relates the concentration at one point in space with a concentration at another, it does not cause spatial correlations in this system. This is an important result because, while experimenters and theorists have always assumed Poisson-type intrinsic noise was the minimum possible, additional intrinsic noise and correlations have been thought possible [7]. In this system, they do not exist because each protein's existence and location are independent of every other protein's existence and location. Spatial correlations may exist in cases where protein generation is non-Poisson, protein decay is nonmonomer, or spatial transport does not have the traditional $\nabla^2 C$ form. Of these cases, this paper's methods should be most easily generalized to non-Poisson protein generation.

Now we turn to the discussion of spatial landscape, a different way to view the probability distributions involved. We use generalized potential landscape $U=-\ln P$ to relate with the steady-state probabilistic functional obtained by the exact solution of the spatial-dependent master equation above. In Fig. 3, we show the landscape in concentration and space. We can see from the bottom panel that the shape of the landscape at each spatial point is like a funnel with the bottom of the lowest potential corresponding to the peak of the probabilistic distribution at that location. This is also clear from the two-dimensional representations of the potential versus protein number shown above the main graph at 20%, 50%, and 80% egg length. The widths of the funnels are measured by the variances in potential at each spatial location. A funneled landscape implies that the network is stable and robust. In this way, it can perform its biological function effectively and reliably. As we can see the funneled landscape becomes narrower from anterior to posterior. This implies varying stability and robustness distributed along spatial locations.

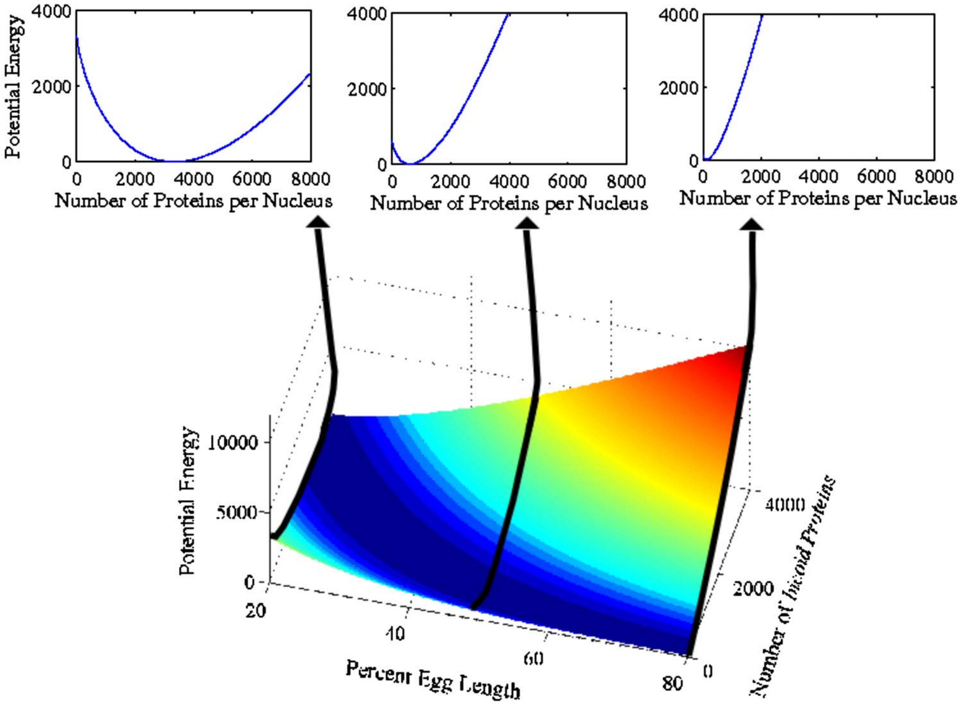

FIG. 3. (Color online) Potential versus number of proteins over space. The main graph shows the complete figure, in which each point in space (percent egg length) has its own potential energy function. Three of these are shown explicitly above the main graph, at 20%, 50%, and 80% egg length.

In summary, we used a relatively common model of protein production, diffusion, and degradation to solve exactly and analytically for the stochastic distribution of the *bicoid* protein in *Drosophila melanogaster*. The probabilistic solution is a Poisson distribution at each point in space, with the mean of the Poisson distribution decaying exponentially away from the source, and matches current experimental data well. The intrinsic fluctuations, noise due to a finite number of molecules in the system and which do not exist in the bulk, decrease away from the source at a slower rate than the mean. We also discussed how to uncover the underlying spatial landscape from the probabilistic distribution. The landscape provides a global and physical foundation of quantitatively addressing the critical issues of stability, robustness, and function of the spatial-dependent cellular networks. The methodology used here can be easily generalized to more dimensions and different boundary conditions and can be applied to any stochastic system with similar creation, diffusion, and decay processes.

D.L. would like to thank John Reinitz and Tomas Gregor for making valiant efforts to convey to him the current understanding of the biology behind the problem. We would like to acknowledge financial support from the National Science Foundation and the Petroleum Research Fund of the American Chemical Society, and partial support from the National Natural Sciences Foundation of China with Grants No. 20735003, No. 90713022, and 973 Project No. 2007CB714500.

[1] H. McAdams and A. Arkin, Proc. Natl. Acad. Sci. U.S.A. **94**, 814 (1997).

[2] M. Elowitz and S. Leibler, Nature (London) **403**, 335 (2000).

[3] M. Thattai and A. van Oudenaarden, Proc. Natl. Acad. Sci. U.S.A. **98**, 8614 (2001).

[4] J. Paulsson, Nature (London) **427**, 415 (2004).

[5] W. Driever and C. Nüsslein-Volhard, Cell **54**, 95 (1988).

[6] G. Struhl, K. Struhl, and P. M. Macdonald, Cell **57**, 1259 (1989).

[7] J. Reinitz (unpublished).

[8] T. Gregor, E. Wieschaus, A. McGregor, W. Bialek, and D. Tank, Cell **130**, 141 (2007).

[9] J. Hattne, D. Fange, and J. Elf, Bioinformatics **21**, 2923 (2005).

[10] F. Tostevin, P. R. ten Wolde, and M. Howard, PLOS Comput. Biol. **3**, e78 (2007).

[11] M. Sasai and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **100**, 2374 (2003).

[12] M. Delbruck, *Unites Biologiques Douees de Continuite Genetique Colloques Internationaux du Centre National de la Recheche Scientifique* (CNRS, Paris, 1949).

[13] C. H. Waddington, *Strategy of the Gene* (Allen and Unwin, London, 1957), p. 290.

[14] R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930), p. 251.

[15] S. Wright, in *Proceedings of the Sixth International Congress on Genetics*, edited by D. F. Jones (Brooklyn Botanical Garden, Brooklyn, 1932), Vol. 1, p. 356.

[16] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).

[17] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).

[18] N. G. V. Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier Science, New York, 1992), p. 465.

[19] P. Ao, J. Phys. A **37**, L25 (2004).

[20] H. Qian and D. A. Bear, Biophys. Chem. **114**, 213 (2005).

[21] J. Wang, B. Huang, X. Xia, and Z. R. Sun, Biophys. J. **91**, L54 (2006).

[22] J. Wang, B. Huang, X. Xia, and Z. R. Sun, PLOS Comput. Biol. **2**, e147, 1371 (2006).

[23] K. Kim and J. Wang, PLOS Comput. Biol. **3**, e60 (2007).

[24] B. Han and J. Wang, Biophys. J. **92**, 3755 (2007).

[25] D. Gillespie, J. Phys. Chem. **81**, 2340 (1977).

[26] C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer-Verlag, Berlin, 1985).

[27] C. Gardiner and S. Chaturvedi, J. Stat. Phys. **17**, 429 (1977).

[28] J. Jaeger, D. H. Sharp, and J. Reinitz, Mech. Dev. **124**, 108 (2007).

[29] J. Reinitz, Nature (London) **448**, 420 (2007).