Origin of lognormal-like distributions with a common width in a growth and division process

Kazufumi Hosoda,[1] Tomoaki Matsuura,[1,2] Hiroaki Suzuki,[1,2] and Tetsuya Yomo[1,2,*]

[1]*Department of Bioinformatic Engineering, Graduate School of Information Science and Technology,*
Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
[2]*ERATO, JST, Suita, Osaka 565-0871, Japan*

Lognormal statistical distributions are observed in a variety of scientific fields. The widths of these distributions in the log scale are often similar, but the underlying mechanism that maintains these widths within a small range has not been well explained. We show that a stochastic process of halving followed by addition can yield a stationary distribution that resembles the universal lognormal distribution with a certain width. The mechanism that we propose here would provide insight into the essence of why lognormal-like distributions in many systems have a common width.

I. INTRODUCTION

The shape of a frequency distribution curve provides clues to the underlying processes that form the distribution. The lognormal distribution is a type of probability distribution in which the logarithm of a variable is normally distributed [1, 2]. Lognormal distributions are observed in various fields of science including the life sciences [3], materials sciences, and social sciences [4]. Interestingly, although the kinetics that govern the systems represented by such distributions are likely to be diverse, the width of observed lognormal distributions on the logarithmic scale is often similar across fields when they are normalized, as seen in some of the examples from literatures plotted in Fig. 1(a). More specifically, the variance of the natural logarithm of the variable (V) ranges from about 0.2 to 0.5 in these examples [4–14], and in this paper we refer to it as the "common width."

Here we are specifically interested in the fact that the sizes of the living cells from bacteria to mammalian cells also fall into lognormal distributions with the common width [Fig. 1(a)]. Cell size distributions should be determined by the growth and division processes [5], and the kinetics of consecutive division (fragmentation) processes are well studied and known to yield lognormal distributions under certain conditions [15,16]. In this paper, we consider a simple growth and division model mimicking bacteria as a basis for the investigation of the underlying mechanism that produces a lognormal distribution with the common width. We show that a simple process involving the stochastic decrease of a variable by halving (binary division) followed by an increase by growth can yield a lognormal-like distribution with the common width [$V \approx (\ln 2)/2 = 0.35$] as a stationary probability distribution (SPD). We also discuss the universality and applicability of this model, by showing that this characteristic distribution appears regardless of the values of some set of parameters that govern the dynamics.

II. MODEL

A bacterium grows in size and divides into two individuals, which initially are each half the size of the original. Mimicking

---

*yomo@ist.osaka-u.ac.jp

these dynamics, we assume an object with a positive variable x (e.g., the size of bacteria) that grows at a rate $h_g(x)$ and divides into b different individuals, each of which has a valuable that is $1/b$ of the original ($b > 1$; $b = 2$ for halving) at a probability $h_d(x)$ per unit time [Fig. 1(b)]. Then, we derive the rate equation

$$\frac{\partial f(x,t)}{\partial t} = b^2 h_d(bx) f(bx,t) - h_d(x) f(x,t) - \frac{\partial h_g(x) f(x,t)}{\partial x},$$

where $f(x,t)$ is the frequency of the objects with variable x at time t. The first term on the right-hand side, $b^2 h_d(bx) f(bx,t)$, represents the increase in the frequency due to cell division [Fig. 1(b), arrow i]. Because the range of x for the parental objects [$b\Delta x$ in Fig. 1(b)] is b-fold greater than that of the daughter objects (Δx) and the division of a single parental object bx produces b daughter objects of x, this term is multiplied by b^2. The second term, $-h_d(x) f(x,t)$, represents the decrease in the objects' frequency due to division [Fig. 1(b), arrow ii]. The third term, $-\frac{\partial h_g(x) f(x,t)}{\partial x}$, represents the flux of the frequency due to continuous growth [Fig. 1(b), difference of arrows iii and iv]. In this model, the total frequency $N(t) = \int_0^\infty f(x,t)\,dx$ increases over time as a result of the increase in the number of objects by the division process; therefore, we normalized $f(x,t)$ to obtain the probability density function $p(x,t) = f(x,t)/N(t)$. By integrating $\partial f(x,t)/\partial t$ with respect to x, we obtain $dN(t)/dt = (b-1)v(t)N(t)$, where $v(t) = \int_0^\infty h_d(x)\,p(x,t)\,dx$. Taking this increase in the total frequency into account, we derive the rate equation of $p(x,t)$:

$$\frac{\partial p(x,t)}{\partial t} = b^2 h_d(bx)p(bx,t) - h_d(x)p(x,t)$$
$$- (b-1)v(t)p(x,t) - \frac{\partial h_g(x)p(x,t)}{\partial x}. \quad (1)$$

This is the master equation for the general model.

As representative functions for $h_g(x)$ and $h_d(x)$ to be analyzed, we examined cases in which the growth rate and the division probability are proportional to arbitrary powers of x (n_g and n_d, respectively), i.e., $h_g(x) = k_g x^{n_g}$ and $h_d(x) = k_d x^{n_d}$, where k_g and k_d are constants. These are monotonic functions with a positive value for $x > 0$. This description can generally cover various possible cases; for
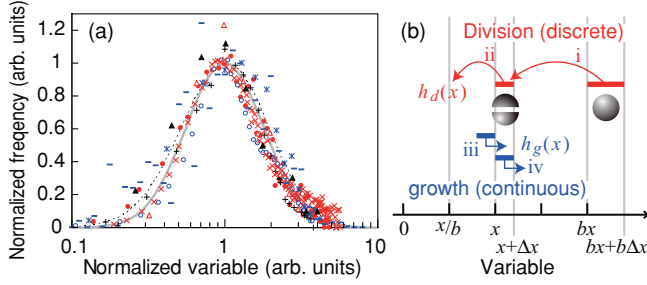
FIG. 1. (Color online) (a) Probability distributions of the variables for various objects reported in the literature. From these data, the variance of the natural logarithm of the variable (V) was derived by fitting to a normalized lognormal distribution $\exp[-X^2/(2V)]$, where X is the natural log of the normalized variable. Systems include the following: *E. coli* size by forward scattering ● [6] ($V = 0.35$), *E. coli* volume from a Coulter counter ○ [7] (0.28), mouse lymphoblast volume + [9] (0.29), human platelet volume × [10] (0.33), mouse islets of Langerhans size (cube root of cell number) ∗ [11] (0.48), liposome surface area (square of the diameter) ▲ [12] (0.41), nanoparticle length △ [13] (0.32), and number of words in a sentence – [14] (0.48). The dashed black line shows the SPD obtained from the numerical simulation using the bacterial model [Eq. (1) with $h_g(x) = 2.3 \times 10^{-7}x, h_d(x) = 3.1 \times 10^{-3}x, b = 2$], which gave $V = 0.43$. The bold gray line shows the lognormal curve with $V = (\ln 2)/2 = 0.35$. Our model gives a value for V that roughly falls within the range described in the literature. The normalized variable on the horizontal axis was calculated by subtracting a value for the invalid range of variables where the frequency was zero from the original value. The resulting value was then normalized so that the variable at the maximum frequency (i.e., the mode) of the moving average of three data points was one. The frequency of the normalized variable on the linear scale from the literature was converted to frequency on the logarithmic scale by multiplying by the variable itself. (b) Schematic diagram of the general model described by Eq. (1).

example, when $n_g < 0$, the object with a larger value of x grows more slowly, and when $n_g > 0$, the object with a larger value of x grows faster. For such cases, Eq. (1) is rewritten as

$$\frac{\partial p(x,t)}{\partial t} = b^{n_d+2}k_d \, x^{n_d} p(bx,t) - k_d x^{n_d} p(x,t)$$
$$- (b-1)k_d\mu_{n_d}(t)p(x,t) - k_g \frac{\partial x^{n_g} p(x,t)}{\partial x}, \quad (2)$$

where $\mu_i(t)$ is the ith raw moment of $p(x,t)$.

III. RESULTS

A. Distribution of the variable derived from the bacterial model

Let us first show a simple case regarding bacterial size. We assume that both of the growth rate and the division probability are proportional to x, i.e., $n_g = n_d = 1$ in Eq. (2) [17]. We also assume that the division is a halving, i.e., $b = 2$. We designate this condition as the bacterial model. We have used the finite difference method to numerically solve Eq. (2). Because x spans multiple orders of magnitude, we used a logarithmic scale (grid width in the simulation was set to create a constant bin for the logarithm of x) in the calculation. As a result, we obtained a lognormal-like SPD [Fig. 1(a), black dashed line]. Again, the shape of this SPD is similar not only to that of

the size distributions of bacteria [5–8] but also to that of the size distributions of other single-celled microorganisms [8], mammalian cells [8–10], and organs [11] as well as nonliving material such as liposomes [12] and nanoparticles [13] and the number of words in a sentence [14] [Fig. 1(a), symbols]. Thus, the bacterial model representation of Eq. (2) might possess an intrinsic feature that is common to a variety of examples.

B. Dependence of the shape of the SPDs on the constants

To show how the model applies to more general cases other than the bacterial model, we analyze the dependence of the shape of the SPD on the constants in Eq. (2) (k_g and k_d, b, and n_g and n_d).

1. Dependence on k_g and k_d

Primarily, we show that the shape of the SPD is independent of the rate constants k_g and k_d when $n_g = n_d$ (including the bacterial model). We will prove this by transforming the equation. Let us define $q(z)$ as the SPD for a normalized variable $z = x/a$, where $a = k_g/k_d$, which satisfies $p(x) \, dx = q(z) \, dz$. In this case, the shape of the probability distribution for $\ln(z)$ is the same as that for $\ln(x)$, with the only difference being in the scale. At the stationary state ($\partial p(x,t)/\partial t = 0$), Eq. (2) is transformed to

$$b^2(bz)^{n_d}q(bz) - z^{n_d}q(z) - (b-1)q(z)\int_0^\infty z^{n_d}q(z)\,dz$$
$$- a^{n_g-n_d}\frac{dz^{n_g}q(z)}{dz} = 0.$$

In this formulation, a is canceled out when $n_g = n_d$. Therefore, in the bacterial model and in its extended model, the shape of the SPD is independent of the rate constants k_g and k_d.

2. Dependence on b

We performed a numerical simulation of Eq. (2) for various values of b when $n_g = n_d = 1$ [Fig. 2(a)]. The resultant SPDs fit a lognormal distribution; however, when $b \neq 2$, the widths deviated from the common width. We calculated the variance of $\ln(x)$ (defined as VV) that represents the width of the SPD in the log scale and plotted it against b [Fig. 2(b), ○]. The plot shows a monotonic increase of V with increasing b. We have thus shown that $b = 2$ is an important parameter that determines the common width.

3. Dependence on n_g and n_d

We again performed a numerical simulation and plotted the SPDs for various values of n_g and n_d with $b = 2$ as a constant [Fig. 2(c)]. The results show that the SPDs are similar when $n_g = n_d = n$ ($n = -1\sim4$), although the SPD becomes slightly narrower as n recedes from 1 [Fig. 2(c)]. When $n_g \neq n_d$ [$n_g = 1.5$ and $n_d = 1$; Fig. 2(c), ■], the SPD is still lognormal-like, but there is a significant change in the width of the distribution. Therefore, $n_g = n_d$ is also an important condition that determines the common width.
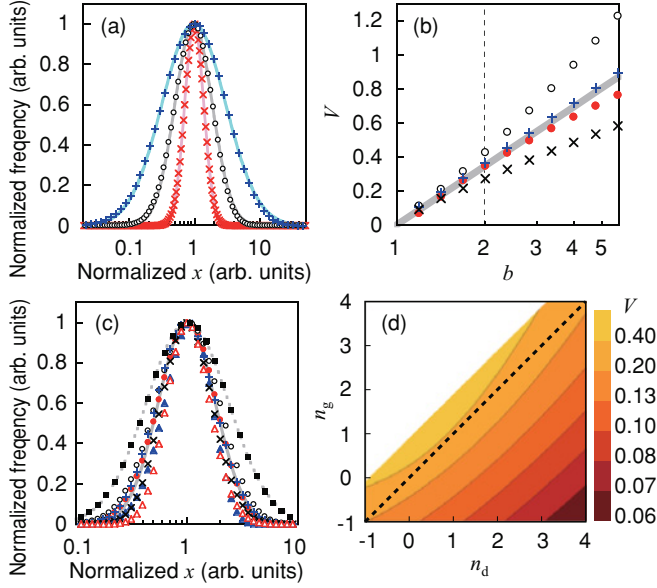
FIG. 2. (Color online) The dependency of SPD on b, n_g, and n_d. (a) SPDs derived from numerical simulations of Eq. (2) with $n_g = n_d = n = 1$ and $b = 1.2$ ($\times$), 2 ($\circ$), and 5.7 (+). The solid lines show the fit of each SPD to the lognormal. (b) The b dependency of the variance of $\ln(x)$ (V). V was calculated from the SPDs derived from the numerical simulation of Eq. (2) with various b values when $n = 0$ ($\bullet$), 1 ($\circ$), 2 (+), and 3($\times$). The solid gray line shows $V = (\ln b)/2$. (c) SPDs with various n_g and n_d. The SPDs were derived from numerical simulations of Eq. (2) with $b = 2$ for $n_g = 1.5$ and $n_d = 1$ ($\blacksquare$) and for $n_g = n_d = n = -1$ ($\blacktriangle$), 0 ($\bullet$), 1 ($\circ$), 2 (+), 3 ($\times$), and 4($\triangle$). The gray dashed line shows the fit of $\blacksquare$ to the lognormal, and the solid gray line shows the lognormal with $V = (\ln 2)/2$. (d) The dependency of V on n_g and n_d. The V values were calculated from Eq. (3) with $b = 2$ and $n_d - n_g > -1$ because $n_d - n_g > -1$ is a sufficient requirement for the existence of the SPD. The dashed line shows $n_g = n_d$.

4. Approximated analytical solution of V

To further investigate the dependence of the width of the lognormal-like distribution on the constants, let us derive an approximate analytical solution of V. By multiplying both sides of Eq. (2) by x^m and integrating both sides with respect to x ($0 \sim \infty$), we obtain an equation for the raw moments of the SPD in the stationary state: $k_d(b^{1-m} - 1)\mu_{n_d+m} - (b-1)k_d\mu_m\mu_{n_d} + k_g m\mu_{n_g+m-1} = 0$. From this equation, we obtain two equations for $m = 1$ and $m = 2$. Then, using the lognormal approximation with the mean of $\ln(x)$ M and the variance of $\ln(x)$ V, the ith raw moment of the lognormal can be expressed as $\mu_i = \exp(iM + i^2V/2)$. From these three equations we obtain the equation for V as

$$b + e^{2n_d V} = 2be^{(n_g-1)V}. \qquad (3)$$

In Eq. (3), the rate constants k_g and k_d are again canceled out. When $n_g = n_d = 1$, V is obtained as $(\ln b)/2$, which agrees with the value of V derived from the numerical simulation [Fig. 2(b), $\circ$] and shows that V is approximately proportional to $\ln b$. As b increases, the deviation of V from $(\ln b)/2$ increases because the approximation for V is based on the raw moments of the linear variable x; the disagreement is

enhanced when the width of the SPD in the log scale increases. We then plotted a contour map of V for n_g and n_d as derived from Eq. (3) with $b = 2$ [Fig. 2(d)]. This plot clearly shows that V is smaller when n_g is smaller (objects with a smaller variable grow relatively faster) and n_d is larger (objects with a larger variable divide relatively more frequently). More importantly, the isocontours are parallel to the line $n_g = n_d$. This result implies that V is nearly constant when $n_g = n_d$, regardless of the values of each. We also used Eq. (3) to estimate an approximate analytical solution for V when $n_g = n_d$. By assuming that V is proportional to $\ln b$ for $\ln b$ near 0 and for $n_g = n_d = n$, differentiating both sides of Eq. (3) with $\ln b$ gives $V = (\ln b)/2$. Indeed, the value of V from the SPD derived from the numerical simulation is insensitive to n and approximately $(\ln b)/2$ even when $n \neq 1$ [Fig. 2(b)]. As mentioned above, V is approximately $(\ln b)/2$ regardless of k_g, k_d, and n as long as $n_g = n_d$; that is, the common width appeared [$V \approx (\ln 2)/2$] when the division is halving and the growth rate and the division probability are proportional.

C. When the division is not exactly even

We have so far assumed that the variables of daughter objects after division are exactly $1/b$ of the parental object. Realistically, resultant variables after division often fluctuate. To consider the effect of fluctuations in the division process, we revise the division term in Eq. (2) with the conditions $n_g = n_d = n$ and $b = 2$. When the probability of the variable after division is set to the uniform distribution function, the first term on the right side in Eq. (2) becomes $2\int_x^\infty k_d s^{n-1} p(s,t)\,ds$, where s is the integration variable. This type of formulation of the division has been described as a fragmentation of a cluster [16,18,19]. In this case of the uniform distribution function, the SPDs derived from the numerical simulation were skewed in the log scale, and their widths were greater than $V = (\ln 2)/2$ [Fig. 3(a)]. This result indicates that there might be some limit to the variability of the division process. We next examined the case when the division fluctuates within a Gaussian probability distribution. In this case, objects with variable x divide into two daughters with variable ξ ($0 < \xi < x$) and $x - \xi$, where the probability of ξ follows a Gaussian distribution function with a mean value $x/2$ and a standard deviation σx [$G(x/2, \sigma x, \xi)$] [Fig. 3(b), insets]. This function was truncated to satisfy the expression $\int_0^x G(x/2, \sigma x, \xi)\,d\xi = 1$. Thus, the first term on the right side of Eq. (2) becomes $2\int_x^\infty k_d s^n G(s/2, \sigma s, x) p(s,t)\,ds$. For the bacterial model ($n = 1$), the results of the numerical simulation for various values of σ show that the skewness of $\ln(x)$ is close to zero for a value of σ up to 0.1 [Fig. 3(b)]. This result implies that if the fluctuation in b is Gaussian and moderate ($\sigma < 0.1$), the shape of the SPD is insensitive to fluctuations in the division process.

IV. DISCUSSION

Our numerical and analytical investigations show that lognormal-like size distributions with the common width [$V \approx (\ln 2)/2$; Fig. 1(a), bold gray line] result from a balance between growth (continuous additive increase) and stochastic division (discrete multiplicative decrease) processes that have
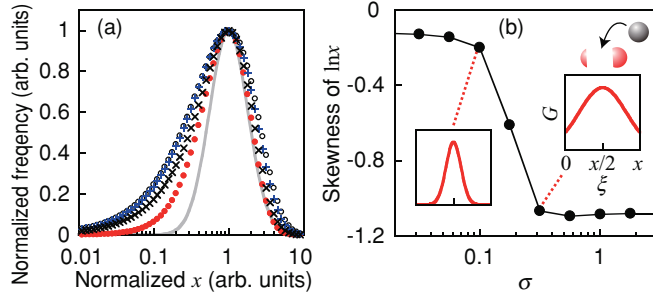
FIG. 3. (Color online) The effect on the model when the variables for daughter objects after halving are not exactly half of the parental object. (a) SPDs derived from numerical simulation when the probability of the variable after division is set to the uniform distribution with $n = 0$ (●), 1 (○), 2 (+), and 3 (×). The solid gray line shows the lognormal with $V = (\ln 2)/2$. By differentiating both sides of the master equation by x in the stationary state, the SPD can be solved only when n is 0 or 1 as $4a^{-2}x\exp(-2a^{-1}x)$ or $a^{-1}\exp(-a^{-1}x)$, respectively. (b) The σ dependency of the skewness of $\ln(x)$ for SPDs calculated from the numerical simulation when the probability of the variable after division is set to the truncated normal distribution and when $n = 1$. The insets show the probability distribution of the daughter variable ξ (0 to x) for each σ.

the following characteristics: (i) division is a halving process ($b = 2$) and (ii) growth rate and division probability are proportional ($n_g = n_d$). We will now discuss the mechanism that produces a stationary distribution such as size homeostasis. Let us consider the average rates of increase and decrease for the variables controlling the growth and division processes. Because division occurs stochastically, the average decrease rate should be the product of the division probability and the amount of decrease in a single division. Here, the amount of decrease in a single division is proportional to the variable (i.e., $x - x/b = (1 - 1/b)x$] because the division is a multiplicative process [Fig. 1(b)]. From requirement (ii) above, the growth rate $h_g(x)$ and the division probability $h_d(x)$ are proportional such that the difference between the average increase rate $h_g(x)$ and the average decrease rate $h_d(x)(1 - 1/b)x$ is $h_d(x)(c - (1 - 1/b)x)$ [where c is a proportionality constant $c = h_g(x)/h_d(x)$]. For smaller variables, the average increase rate then becomes greater than the average decrease rate, while for larger variables, the reverse applies. Because of this feature, the variable remains within a range that forms a stationary distribution. If the amount of decrease in a single division is constant for all ranges of the variable, then the ratio of the increase rate and the decrease rate is constant for all ranges of the variable. In such a case, the distribution diffuses due to the

stochasticity in the division, and there is then no stationary distribution. As mentioned above, when requirement (ii) is true, the major factor for producing stationary distribution is the multiplicative property of the division. For a rigorous proof of the existence of the stationary distribution, further studies are required.

Next, we discuss the factors that produce the lognormality and the common width of the SPD. In our model, the variability of the value of x results only from the discreteness of the stochastic division process. Because the amount of the discrete decrease in a single division in the log scale is constant for all ranges of the variable, i.e., $\ln(x) - \ln(x/b) = \ln b$, the distribution appears as a normal-like shape in the log-scale curve. This aspect is in accordance with existing fragmentation theories [15,16]. An important finding of our model is that the distribution becomes stationary over time due to the presence of the continuous growth term, and the result for the width [$V \approx (\ln b)/2$] is determined by the amount of the discrete decrease in a single division in the log scale.

The two prerequisites of our model for producing a lognormal distribution with the common width are reasonable assumptions in nature. Halving might be applicable to a process where division appears to be $b > 2$ because division can be thought of as a sequence of halvings. Furthermore, the proportionality of the growth rate and the division probability is not a strict assumption because growth and division dynamics may be governed by the same factors. In reality, specific dynamics could be involved in producing each frequency distribution. For instance, specific regulatory mechanisms might play important roles in determining the size of specific bacteria or of mammalian cells. Thus, the applicability of our model to particular examples should be examined in each case. However, the consideration of growth and division processes would provide general insight into the essence of the underlying processes that form lognormal-like distributions with the common width. The combination of the processes of addition and halving might contribute to the ubiquity of these distributions in nature.

ACKNOWLEDGMENTS

[1] E. L. Crow and K. Shimizu, *Lognormal Distributions: Theory and Applications* (M. Dekker, New York, 1988).

[2] N. Balakrishnan and W. S. Chen, *Handbook of Tables for Order Statistics from Lognormal Distributions with Applications* (Kluwer Academic Publishers, Dordrecht; Boston, 1999), p. xiii.

[3] A. L. Koch, J. Theor. Biol. **12**, 276 (1966).

[4] E. Limpert, W. A. Stahel, and M. Abbt, Bioscience **51**, 341 (2001).

[5] A. L. Koch, J. Gen. Microbiol. **45**, 409 (1966).

[6] M. A. McClain *et al.*, Anal. Chem. **73**, 5334 (2001).

[7] R. J. Harvey, A. G. Marr, and P. R. Painter, J. Bacteriol. **93**, 605 (1967).

[8] H. G. Kilian *et al.*, Eur. Phys. J. E: Soft Matter **17**, 307 (2005).

[9] A. Tzur *et al.*, Science **325**, 167 (2009).

[10] J. M. Paulus, Blood **46**, 321 (1975).

[11] J. Jo, M. Y. Choi, and D. S. Koh, Biophys. J. **93**, 2655 (2007).

[12] B. G. Tenchov, and T. K. Yanev, J. Colloid Interface Sci. **111**, 1 (1986).

[13] C. G. Granqvist *et al.*, Phys. Rev. Lett. **37**, 625 (1976).

[14] C. B. Williams, Biometrika **31**, 356 (1940).

[15] A. N. Kolmogorov, Dokl. Akad. Nauk. SSSR **31**, 99 (1941).

[16] Z. Cheng and S. Redner, J. Phys. A.: Math. Gen. **23**, 1233 (1990).

[17] Y. Wakamoto, J. Ramsden, and K. Yasuda, Analyst **130**, 311 (2005).

[18] R. M. Ziff and E. D. McGrady, J. Phys. A: Math. Gen. **18**, 3027 (1985).

[19] Z. Cheng and S. Redner, Phys. Rev. Lett. **60**, 2450 (1988).