P H Y S I C A L   R E V I E W   L E T T E R S

How *Xenopus Laevis* Replicates DNA Reliably even though Its Origins of Replication are Located and Initiated Stochastically

John Bechhoefer[*] and Brandon Marshall

Department of Physics, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada

DNA replication in *Xenopus laevis* is extremely reliable, failing to complete before cell division no more than once in 10 000 times; yet replication origin sites are located and initiated stochastically. Using a model based on 1D theories of nucleation and growth and using concepts from extreme-value statistics, we derive the distribution of replication times given a particular initiation function. We show that the experimentally observed initiation strategy for *Xenopus laevis* meets the reliability constraint and is close to the one that requires the fewest resources of a cell.

DNA replication is one of the defining processes of living systems, and evolution has accordingly selected for highly reliable replication mechanisms. The South African clawed frog *Xenopus laevis* is an organism often used to study replication in eukaryotes [1]. The replication of its embryonic cells is particularly interesting, as it corresponds to a "stochastic limit," where the placement and initiation of the sites where DNA replication begins ("replication origins") show significant stochasticity [2]. As with humans, the *Xenopus* genome contains approximately 3×10^9 bases [3]. Just after fertilization, cells divide for 12 generations with an abbreviated cell cycle that is as short as 25 minutes (at 20 °C). The cell cycle is divided into an "*S* phase" of about 20 minutes, when DNA is replicated, and a mitosis phase of about 5 minutes, when chromosomes separate and the cell divides [2]. In order to replicate so many bases in so little time, the cell initiates DNA replication at many [$\sim \mathcal{O}(10^5)$] origins. For these embryonic cells, in contrast to the situation for fully developed somatic cells, there is no sequence dependence to the location of replication origins [2]. In addition, each origin initiates stochastically, with no predetermined time of initiation. The stochasticity in the location and initiation of replication origins leads to a potential difficulty: the typical time for replication is about 20 minutes, but the maximum allowable time is only 25 minutes. In particular, embryonic cells lack the efficient checkpoint mechanisms [4] that somatic cells have to pause the cell cycle to allow for unusually slow replication. The cell must replicate by the time it divides, or die. But empirically, such a "mitotic catastrophe" [5] is rare, $\lesssim 1$ in 10^4 replications [6]. How can one reconcile the variations in *S*-phase duration due to the stochastic placement and initiation of origins with the high reliability of replication?

In the biological literature, the above is known as the "random-completion problem" [3] and has been an unsettled question for over 20 years [4,7,8]. In its simplest form, randomly placed origins imply an exponential distribution of origin separations and, hence, a small number of very large gaps that take a long time to replicate. Two approaches to a solution have been advanced. The first notes evidence that the spacing of origins is not completely random and that any regularity in the spacing of origins will tend to suppress large gaps [3,9]. However, in isolation, such a scenario is fragile: if a single origin fails to initiate, it will create a much larger gap than exists usually. The second approach draws on a recent experimental result that origins initiate throughout *S* phase and, indeed, that the rate of initiation of origins, $I(t)$ (initiations per time per length of unreplicated genome), increases significantly as *S* phase proceeds [10–12]. Intuitively, initiating origins throughout *S* phase allows the cell to "fill in gaps" and avoid unusually long delays.

In this Letter, we first calculate, following theories of nucleation and growth in one dimension [13,14], the distribution of replication times $\rho_{\rm rep}(t)$ given an initiation function $I(t)$ and a constant "fork velocity" v describing the symmetric growth of replication domains. We find that an increasing $I(t)$ can insure replication at the required level of reliability, even in the worst case of completely random origin spacing. We then show that the specific $I(t)$ observed in *in vitro* experiments is close to an optimal $I(t)$ that minimizes the amount of cellular replication machinery (polymerases, helicases, etc.) that a cell is required to supply.

Our derivation of $\rho_{\rm rep}$ uses a model inspired by the Kolmogorov-Johnson-Mehl-Avrami theory of crystallization kinetics [15], which is a stochastic model with three elements: nucleation (initiation) of ordered (replicated) domains; symmetric growth of these domains; and coalescence of domains that grow into each other. (See Fig. 1.) Using such a model, we showed that the fraction f of DNA replicated on an infinite domain at a time t after the start of *S* phase is given by

$$f(t) = 1 - e^{-2vh(t)} \tag{1}$$

where $h(t) = \int_0^t g(t')dt'$ and $g(t) = \int_0^t I(t')dt'$ and $I(t)$ is the initiation function (≥ 0) [16]. Here, v is the fork
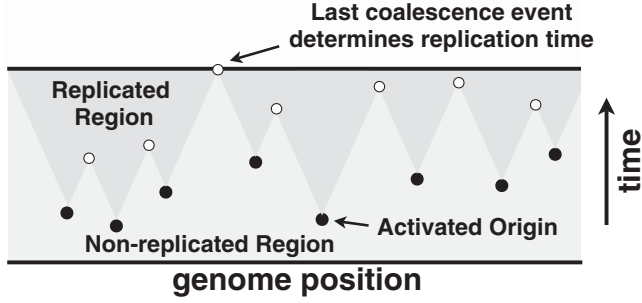
FIG. 1. Schematic of DNA replication model. Space-time diagram showing multiple origins (filled circles), each expanding symmetrically at constant velocity. Domains coalesce when they meet (open circles).

velocity, and $f(t)$ typically has a sigmoidal shape. Equation (1) predicts that it will take infinite time to replicate all the DNA ($f = 1$), but obviously, the replication time should be finite on a finite-length genome. Because the location and time of initiation of origins is stochastic, the time to finish replication will also be a stochastic process.

In order to calculate the distribution of replication times $\rho_{rep}(t)$, we first note that, except for edge effects, there is a one-to-one mapping from replication origins to coalescences of replication domains. (See Fig. 1.) Because the evolution of domains is deterministic once the origin has initiated, one can derive the distribution of coalescence times, $\rho_c(t)$ from the initiation function $I(t)$. In [16], we derived the density of nonreplicated domains ("holes") of size x at time t to be $n_h(x, t) = g^2(t) \exp[-g(t)x - 2vh(t)]$. Since a coalescence event is equivalent to a hole of zero size ($x = 0$), we can write the normalized distribution $\rho_c(t)$ as

$$\rho_c(t) = \frac{2vL}{N_o} g^2(t) e^{-2vh(t)}, \qquad (2)$$

where N_o is the total number of origins along a genome of length L initiated throughout S phase.

As Fig. 1 shows, the time to complete replication corresponds to the last coalescence event. Since there are N_o coalescences, the problem of determining the typical time of the last coalescence is equivalent to asking, "Drawing N_o coalescences from a distribution $\rho_c(t)$, what is the largest time one expects to occur?" Such questions are the subject of the field of extreme-value statistics [17,18], where an analog to the central-limit theorem holds: given a parent distribution whose maximum value is unbounded and whose tail decays asymptotically at least as fast as an exponential (conditions satisfied here), the maximum value drawn in N_o trials will, for N_o large, tend to a Gumbel distribution, $\rho_G(\tau) = (1/\beta) \exp[-\tau - \exp(-\tau)]$, where the scaled time $\tau = (t - t^*)/\beta$, with t^* the mode of the distribution and β its width [17]. An elementary calculation [17] shows that for Eq. (2), the width β is given by $2vg(t^*)$ and the mode t^* by

$$F_c(t^*) = 1 - 1/N_o, \qquad (3)$$

where $F_c(t) = \int_0^t \rho_c(t')dt'$ is the cumulative probability distribution function (CDF) of the probability distribution function (PDF) $\rho_c(t)$. From Eq. (2), the CDF is, asymptotically for large t, given by

$$F_c(t) = 1 - \frac{Lg(t)e^{-2vh(t)}}{N_o}. \qquad (4)$$

Equation (4) is derived by integrating $\rho_c(t)$ by parts and dropping subdominant terms and, with Eq. (3), leads to a transcendental equation for the magnitude of $I(t)$.

In Fig. 2, we show the results of Monte Carlo simulations of the replication-time distribution for various $I(t)$ functions. In all cases, we adjusted the amplitude of $I(t)$ so that the mode of $\rho_{rep}(t)$ is at $t^* = 38$ minutes, which corresponds to the mode deduced from the $I(t)$ measured in the *in vitro* experiments. (For the *in vivo* experiments, $t^* \sim 20$ minutes[4].) The solid lines are fits to a Gumbel distribution. The parameters deduced (the βs) are consistent with the values predicted in the paragraph above.

The striking implication of Fig. 2 is that one can vary the width of the replication-time distribution ρ_c by choosing an initiation function $I(t)$ that increases throughout S phase. Initiating all the origins at the beginning of S phase $[I(t) = I_\delta \delta(t)]$ leads to the broadest possible distribution. Exploring power-law initiation functions $I(t) = I_n t^n$ (with I_n fixed by the t^* constraint), we see that as one progresses from constant ($n = 0$) to linear ($n = 1$) to quadratic ($n = 2$) initiation functions, the width of ρ_c is progressively reduced. The replication-time distribution can also be calculated using the experimental $I(t)$ [12] (not shown). The experimental $I(t)$ is close to a quadratic curve, and its distribution is indistinguishable from the $n = 2$ case.

It would thus appear that the cell can have arbitrarily reliable replication (an arbitrarily narrow distribution ρ_{rep}) simply by arranging for its initiation curve to increase fast
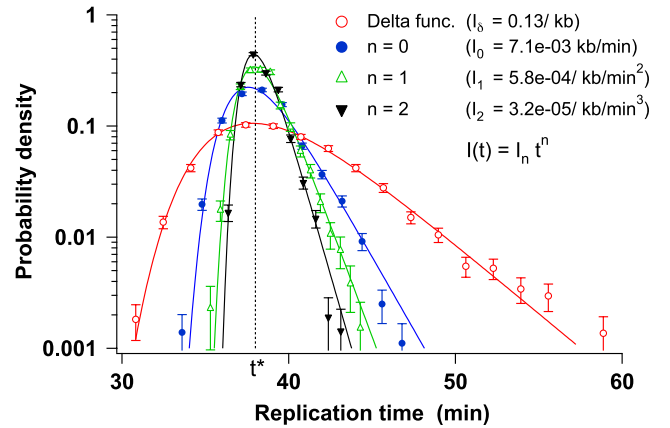

FIG. 2 (color online). Replication-time distribution function, fixing the mode to be $t^* = 38$ minutes. Markers are results from Monte Carlo simulations (3000 trials per simulation); solid lines are fits to the Gumbel distribution.

enough. In fact, the situation is more subtle. Even when all origins are initiated at the beginning of S phase, it is possible to replicate with arbitrary reliability simply by having enough origins. While it is true that there will be a few unusually long gaps that will set the replication time, these gaps may be reduced arbitrarily if one starts with enough replication origins. We thus propose an alternate way of viewing the random-completion problem: Instead of fixing the number of origins and looking at the replication times for different strategies, we fix a time t^{**} at which either a cell has finished replication or it dies. Since evolution selects on the basis of mortality, the replication parameters [$I(t)$, v, the number of potential origins, etc.] should be a consequence of this selection, and not vice versa. Choosing t^{**} to be the cell-cycle time (25) minutes and allowing a failure rate of 10^{-4}, we calculate, for various forms of $I(t)$, the replication parameters required to meet the reliability constraint. (Our results depend only logarithmically on the failure rate.)

In order to compare with experiment, we must confront a further problem. While the *in vivo* replication time is estimated to be 20 minutes, the *in vitro* experiments require nearly twice this time to replicate. We must thus make additional assumptions to translate the *in vitro* experimental results to the *in vivo* situation. In fact, we can do this with one simple assumption. In earlier studies, it was assumed that the replication fork velocity v is constant throughout S phase. The original analysis of the *in vitro* *Xenopus* data thus estimated an average fork velocity of 0.6 kb/ minute. More recent work [19] has shown that the fork velocity starts at 1.1 kb/ minute at the beginning of S phase and then decreases monotonically to 0.3 kb/ minute at the end of S phase. We speculate that the longer time for the *in vitro* S phase is caused by this reduction in fork velocity—perhaps because some protein concentrations are not kept constant. With this single modification—$v = 1.1$ rather than 0.6 kb/ min—we shall find results consistent with the *in vivo* observations.

In Fig. 3, we show results of simulations that constrain the replications to finish by $t^{**} = 25$ minutes, allowing a failure rate of 10^{-4}. We see that it is indeed possible to find amplitudes for $I(t)$ that satisfy the reliability constraint.

While it is always possible to choose an amplitude (e.g., I_δ or I_n) to satisfy the reliability constraint, each choice will have definite implications for the amount of cell resources that are required for its implementation. One may then ask whether there is a "best" strategy for initiating origins (while satisfying the reliability constraint). If so, how close is the experimental $I(t)$ to the optimum?

To answer such questions, one must first define a measure for cell resources. We have considered two possibilities among many that can be imagined: the number of origins initiated throughout S phase and the maximum number of replication forks required. The first choice would be relevant if the origin-initiation proteins were limited. The second would be relevant if the number of polymerases (or other parts of the replication machinery)
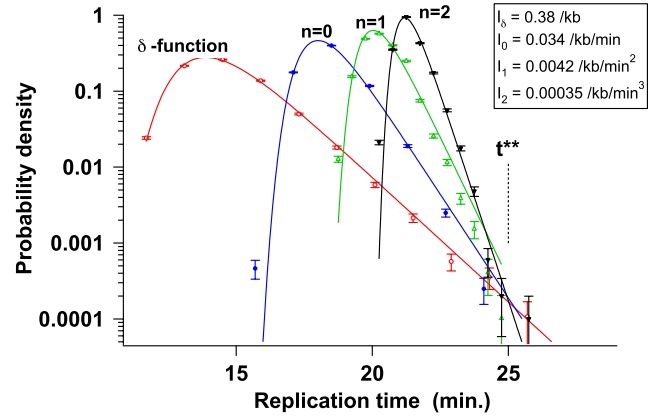

FIG. 3 (color online). Replication-times distribution function, fixing the mortality rate at $t^{**} = 25$ minutes to be 10^{-4}. (The area to the right of the dashed line of each probability distribution function is 10^{-4}.) Markers are results from Monte Carlo simulations (20 000 trials per simulation); solid lines are fits to the Gumbel distribution.

that needed to be active at one time limited the rate of replication [20]. We find qualitatively the same results in both cases [21].

Intuitively, there should be an optimum for the consumption of resources. Within the fork-density scenario, initiating all origins at the beginning leads to a high initial fork density. Holding off initiating until later in S phase helps by allowing the machinery of replication forks to be repeatedly reused. If the cell waits too long to begin replication, then it is essentially shortening S phase, which requires many origins (and forks). Thus, one expects an optimum. We have explored this by calculating the maximum number of forks, n_{max}, required in several cases. First, we calculated it for delta-function initiation ($n_{max} = I_\delta$). Next, we numerically calculate n_{max} for the power-law case. Finally, we use the calculus of variations to calculate the optimal $I(t)$, denoted $I_{opt}(t)$ that minimizes the maximum number of required forks, subject to the reliability constraint. To calculate I_{opt}, we note that the number of replication forks is given by $n(t) = \dot{f}/v = 2g(t) \times \exp[-2vh(t)]$ [16]. One can extract the maximum fork density by writing

$$n_{max}[I(t)] = \lim_{p \to \infty} \left[\int_0^\infty [2g(t)e^{-2vh(t)}]^p dt \right]^{1/p}. \quad (5)$$

The associated Euler-Lagrange equation turns out to be independent of the exponent p. We find

$$\ddot{h}(t) = 2v\dot{h}^2(t), \quad (6)$$

where we recall that $\ddot{h}(t) = I(t)$ and $\dot{h}(t) = g(t)$. Solving Eq. (6) subject to the boundary condition $h(0) = 0$ gives

$$I_{opt}(t) = \frac{1}{2vt^{**}} \left[\delta(t) + \frac{1}{t^{**}} \frac{1}{(1 - t/t^{**})^2} \right]. \quad (7)$$

Equation (7) [with Eq. (1)] implies that the fork density $n = \dot{f}/v = 1/vt^{**}$ is constant throughout S phase.
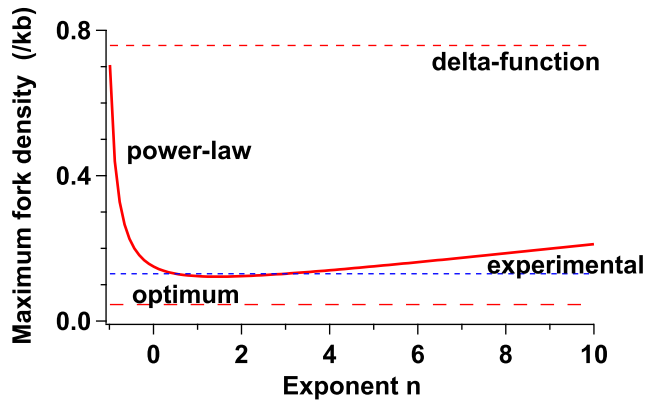
FIG. 4 (color online). Maximum required fork density, for different replication schemes.

In Fig. 4, we summarize the results of these investigations. The dashed line at the top gives the fork density required to make the delta-function $I(t)$ meet the reliability constraint. The solid curve represents the fork density required for power-law initiations. As we anticipated, the curve has a minimum (between $n = 1$ and 2). The fine-dashed line, which lies close to the minimum value of the power-law case, is the experimental maximum fork density [12]. Finally, the broad-dashed line gives the optimal fork density $(1/vt^{**})$.

Although the optimal fork density is lower than that observed, it clearly does not represent a physiologically possible case. It is unrealistic to expect the perfect coordination implied by the delta function at the beginning of S phase. More serious, at the end of S phase, Eq. (7) implies that the rate of initiation diverges, along with the total number of activated origins. Still, we note that the qualitative shape of the curve shares the quadratically increasing form of the experimental result. More generally, it would be surprising if the initiation program were identical to the optimum (even if one were to limit the space of functions to those that are physiologically achievable). We note that the minimum is clearly broad: there is little difference in required fork density between a linear and a quadratic $I(t)$. The main point is that there are some strategies—most notably the initiation of all origins at the beginning of S phase—that are clearly bad, and these differ from the observed $I(t)$.

In conclusion, we have calculated the distribution of replication times $\rho_{\rm rep}$ for the stochastic limit of replication, where origins are placed randomly and initiate stochastically at a rate $I(t)$. Choosing an $I(t)$ that increases with time narrows $\rho_{\rm rep}$ and increases the reliability of replication. Using the known mortality rates and length of the cell cycle, we gave a quantitative interpretation to the random-completion problem and showed that one can meet the reliability constraint using an arbitrary $I(t)$. Different $I(t)$ functions demand different resources from the cell. Measuring this resource use by the maximum required fork density, we show that the experimentally observed form of $I(t)$ is close to optimum. In the future, it would be interesting to consider the effects of any regularity in origin spacing. While we have shown that reliable replication may be achieved even in the worst case of random spacing of origins, there is evidence for some regularity. It would also be interesting to measure the replication-time distribution directly. While determining the time at which the last base (of 3×10^9) replicates is unrealistic, one might be able to determine when a given fraction (e.g., 90 or 95%) of origins have replicated. It is straightforward to generalize the methods presented here to determine the distribution of times required to reach a given replication fraction.

*Electronic mail: johnb@sfu.ca
[1] J. J. Blow, EMBO J. **20**, 3293 (2001).
[2] O. Hyrien and M. Méchali, EMBO J. **12**, 4511 (1993).
[3] J. J. Blow, P. J. Gillespie, D. Francis, and D. A. Jackson, J. Cell Biol. **152**, 15 (2001).
[4] O. Hyrien, K. Marheineke, and A. Goldar, Bioessays (1984-) / BioEssays **25**, 116 (2003).
[5] T. A. Prokhorova, K. Mowrer, C. H. Gilbert, and J. C. Walter, Proc. Natl. Acad. Sci. U.S.A. **100**, 13241 (2003).
[6] C. Hensey and J. Gautier, Dev. Biol. **203**, 36 (1998).
[7] R. A. Laskey, J. Embryol. Exp. Morphol. Supplement **89**, 285 (1985).
[8] N. Rhind, Nat. Cell Biol. **8**, 1313 (2006).
[9] S. Jun, J. Herrick, A. Bensimon, and J. Bechhoefer, Cell Cycle **3**, 223 (2004).
[10] I. Lucas, M. Chevrier-Miller, J. M. Sogo, and O. Hyrien, J. Mol. Biol. **296**, 769 (2000).
[11] J. Herrick, P. Stanislawski, O. Hyrien, and A. Bensimon, J. Mol. Biol. **300**, 1133 (2000).
[12] J. Herrick, S. Jun, J. Bechhoefer, and A. Bensimon, J. Mol. Biol. **320**, 741 (2002).
[13] K. Sekimoto, Physica A (Amsterdam) **125**, 261 (1984); **135**, 328 (1986); Int. J. Mod. Phys. B **5**, 1843 (1991).
[14] E. Ben-Naim and P. L. Krapivsky, Phys. Rev. E **54**, 3562 (1996).
[15] J. W. Christian *The Theory of Phase Transformations in Metals and Alloys, Part I: Equilibrium and General Kinetic Theory* (Pergamon Press, Oxford, 2002), 3rd ed..
[16] S. Jun, H. Zhang, and J. Bechhoefer, Phys. Rev. E **71**, 011908 (2005).
[17] E. J. Gumbel, *Statistics of Extremes* (Columbia Univ. Press, NY, 1958).
[18] S. Kotz and S. Nadarajah, *Extreme Value Distributions* (Imperial College Press, London, 2000).
[19] K. Marheineke and O. Hyrien, J. Biol. Chem. **279**, 28071 (2004).
[20] N. Rhind (private communication).
[21] B. Marshall and J. Bechhoefer (unpublished).