

Dynamics of DNA Replication in Yeast

Renata Retkute and Conrad A. Nieduszynski

Centre for Genetics and Genomics, University of Nottingham, Nottingham NG7 2UH, United Kingdom

Alessandro de Moura

Institute of Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, AB24 3UE United Kingdom

(Received 15 December 2010; published 4 August 2011)

We present a mathematical model for the spatial dynamics of DNA replication. Using this model we determine the probability distribution for the time at which each chromosomal position is replicated. From this we show, contrary to previous reports, that mean replication time curves cannot be used to directly determine origin parameters. We demonstrate that the stochastic nature of replication dynamics leaves a clear signature in experimentally measured population average data, and we show that the width of the activation time probability distribution can be inferred from this data. Our results compare favorably with experimental measurements in *Saccharomyces cerevisiae*.

DOI: [10.1103/PhysRevLett.107.068103](https://doi.org/10.1103/PhysRevLett.107.068103)

PACS numbers: 87.14.gk, 87.10.Ca, 87.10.Mn

DNA replication starts at specific locations in the chromosome called replication origins. Most bacterial genomes are replicated from a single origin, but the much greater size of eukaryotic genomes requires multiple origins per chromosome to ensure that the replication process does not take too long. Genome replication has been comprehensively studied in the model organism *S. cerevisiae* (brewers' yeast). Origin locations in *S. cerevisiae* are determined by specific DNA sequences and are thus fixed in every yeast cell [1]. The positions of origins in yeast have been comprehensively catalogued [2]. However, a given origin may not be active during replication, because origins must be licensed before the start of *S* phase (the part of the cell cycle where DNA replication takes place). Licensing consists of a series of specific protein complexes binding at origin locations, culminating in the loading of pairs of Mcm2-7 molecules. If in a given cell licensing of a certain origin is not completed by the time *S* phase starts, the origin is unable to function [3].

High-throughput methods have allowed the measurement of replication times as a function of chromosomal position for the whole genome [4]. These experiments yield average replication times over large cell populations (typically $>10^7$ cells) and therefore can mask the cell-to-cell variability present in the system [5]; to date single cell and single molecule studies are not able to measure the kinetics of whole genome replication [5,6]. The low abundance of the molecules involved in triggering origin activation strongly suggests that origins have stochastic activation times [7]. This is often ignored in the biological

literature, where there is a pervasive notion of a “replication program,” in which origins are considered to be programmed to fire following a precisely controlled order. This idea has frequently led to erroneous interpretations of replication time profiles (reviewed in [8]).

There has been much interest recently in mathematical modeling of DNA replication. Two different modeling approaches have been used: simulations to capture the replication dynamics at a single cell level [9–12], and probabilistic models that characterize the dynamics of replication at a population level [13,14]. Some models of DNA replication [14,15] are closely related to Kolmogorov's classical model of nucleation processes [16]. Our model can be regarded as an inhomogeneous model of nucleation with quenched disorder, where nucleation starts at specific sites. Inhomogeneous models of nucleation have been studied in the context of statistical physics and have relevance to surface science and other areas [17,18].

Reference [14] is particularly relevant for this work because it proposes that origins fire stochastically in time, and goes on to show that this can lead to reproducible replication dynamics population wide, without the need to invoke a “replication program.” Although valuable insights have been gained from previous works, they ignore the possibility that origins can fail to license, and we will show that this has a crucial effect on the system's dynamics. In addition, most of the existing models are numerical. In this work, we introduce an analytical model of eukaryotic DNA replication which fully takes into account the stochastic nature of both origin activation and the licensing process. Using a simple two-origin chromosome, we illustrate how replication time curves from measurements are influenced by the stochasticity of origin activation as well as by the possibility that licensing fails, reinforcing results we obtained previously by direct simulations [8].

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

We establish that the shape of the average replication time profile has a signature of the stochasticity of the replication process, even though it is a quantity defined by a population average. We derive an analytical expression relating the replication time in the regions between two origins to the standard deviation Δt of the activation time distribution of the origins. This is a valuable result, since few single cell experiments have been able to give direct information about the stochastic properties of the replication dynamics. Our results allow Δt to be obtained from widely available population-wide measurements. We apply this result to measured replication time data, and estimate Δt from the data, obtaining a result in agreement with current estimates in the literature.

In our model we consider a chromosome with N origins, where each origin i is defined by the following: its chromosomal position x_i ; the probability q_i that the origin achieves licensing (in a given cell within a population) and is thus capable of activating; and the activation time probability distribution $p_i(t)$, which is the probability density of origin i activating and starting bidirectional replication forks at time t . Since an origin may not be competent in every cell within the population, in general $q_i < 1$, and p_i satisfies $\int_{-\infty}^{+\infty} p_i(t)dt = q_i$. The fundamental quantity from which all statistical properties of this system can be calculated is the probability density $P(x, t)$, defined such that $P(x, t)dt$ is the probability that chromosomal position x is replicated between times t and $t + dt$. If only origin i were present, P would be given by $P(x, t) = p_i(t - |x - x_i|/v)$, where v is the fork velocity, which we assume to be a constant.

In the presence of all N origins, the calculation of $P(x, t)$ is complicated by the fact that position x can be replicated by forks originated from any of the origins. Let us assume that position x is replicated between times t and $t + dt$ by a fork from origin i . This requires that (i) origin i activated at time $t - |x - x_i|/v$, so that the fork arrives at x at time t ; and (ii) all other origins $j \neq i$ have either not activated or they have activated but their forks would arrive at x later than t . The probability density for event (i) is $p_i(x, t) = p_i(t - |x - x_i|/v)$, and the probability for event (ii) is $Q_i(x, t) = \prod_{j \neq i} M_j(x, t)$, where M_i is the probability that a fork from origin i arrives later than t , or fails to activate: $M_i(x, t) = s_i + \int_t^{+\infty} p_i(x, y)dy$, where $s_i = 1 - q_i$ is the probability of origin i not being competent. Therefore, the probability density $P_i(x, t)$ that position x is replicated by origin i at time t is

$$P_i(x, t) = p_i(x, t)Q_i(x, t). \quad (1)$$

Finally, the probability density that position x is replicated at time t , irrespective of which origin the fork started from, is

$$P(x, t) = \sum_{i=1}^N P_i(x, t). \quad (2)$$

One of the most important quantities for comparison with experimental data is the average replication time $T(x)$ at position x , which is given in terms of P as

$$T(x) = \frac{1}{1 - s_1 s_2 \cdots s_N} \int_{-\infty}^{+\infty} t P(x, t) dt, \quad (3)$$

where $1 - s_1 s_2 \cdots s_N = \int_{-\infty}^{+\infty} P(x, t) dt$ is the probability that at least one of the origins will activate. The average replication time across whole chromosomes [$T(x)$ curves] has been measured in a number of organisms. However, caution is required when interpreting $T(x)$ curves. In some of the biological literature, $T(x)$ curves are used to directly infer origin parameters [4]. For example, it is widely accepted that the values of T at x_i are the average activation times of origins. However, Eq. (3) shows that $T(x)$ is determined collectively by all origins [8]. This suggests that simple interpretations of $T(x)$ are not justifiable.

We want to use the general theory presented above to study replication dynamics in a simple setting. From now on we focus on the case of a hypothetical linear chromosome with just two origins. We define the chromosomal coordinates so that one of the origins has position $x_1 = 0$; the other origin has position $x_2 = D$. We assume for simplicity that each origin can activate within a time window Δt with uniform probability; we will argue later that our conclusions are largely independent of the precise shape of the probability distribution. We select origin activation times so that the average activation time of the first origin is 0. The other origin has an average activation time τ , and we assume without loss of generality that $\tau \geq 0$. Thus the activation time distributions are

$$p_i(t) = \frac{q_i}{\Delta t}, \quad \text{if } t \in \left[t_i^{\text{av}} - \frac{\Delta t}{2}, t_i^{\text{av}} + \frac{\Delta t}{2} \right], \quad (4)$$

where $i = 1, 2$, $t_1^{\text{av}} = 0$, and $t_2^{\text{av}} = \tau$. p_1 and p_2 are set to zero outside the stated intervals.

Using Eqs. (3) and (4), we can write analytical expressions for the probability density $P(x, t)$ and the average replication time $T(x)$. From Eq. (4) and (2), $P(x, t)$ vanishes outside the intervals I_1 and I_2 given by

$$I_i = \left[\frac{|x_i - x|}{v} + t_i^{\text{av}} - \frac{\Delta t}{2}; \frac{|x_i - x|}{v} + t_i^{\text{av}} + \frac{\Delta t}{2} \right], \quad (5)$$

where $x_1 = 0$ and $x_2 = D$. In total there are five scenarios (depending on the relative values of τ , Δt , D , and v) that differ in the dynamics of how the chromosome is replicated. From here on we will consider just the case where the condition $\tau + \Delta t < D/v$ is satisfied, since this is the case for many pairs of origins in real chromosomes. This means that the variations in the activation time Δt are small enough that a fork from one origin can only replicate the other origin if that origin is not competent. The expression for $T(x)$ is then

$$T(x) = \begin{cases} \left(\frac{Dq_2s_1 - q_2s_1x}{v} + \frac{q_2s_1\tau + q_1|x|}{v} \right) / (1 - s_1s_2), & \text{if } x < \frac{D+v(\tau-\Delta t)}{2}; \\ \frac{(D-2x+v\tau)\{q_1[D+v(\tau-\Delta t)]+q_2s_1[D+v(\tau+\Delta t)]\}}{2\Delta tv^2(1-s_1s_2)} + 2(D-\Delta tv-2x+v\tau)\left(\frac{q_1q_2[D^2+\Delta t^2v^2+(v\tau-2x)^2]}{12\Delta t^2v^3(1-s_1s_2)}\right) \\ + \frac{\Delta tv\{4q_1q_2x+[-3q_1+(q_1-3)q_2]v\tau\}}{12\Delta t^2v^3(1-s_1s_2)} + \frac{D\{\Delta t[-3q_1+(q_1-3)q_2]v+2q_1q_2(v\tau-2x)\}}{12\Delta t^2v^3(1-s_1s_2)}, & \text{if } \frac{D+v(\tau-\Delta t)}{2} \leq x < \frac{D+v\tau}{2}; \\ \frac{(D-2x+v\tau)\{-q_2[D+v(\tau-\Delta t)]-q_1s_2[D+v(\tau+\Delta t)]\}}{2\Delta tv^2(1-s_1s_2)} - 2[D-2x+v(\tau+\Delta t)]\left(\frac{q_1q_2[D^2+\Delta t^2v^2+(v\tau-2x)^2]}{12\Delta t^2v^3(1-s_1s_2)}\right) \\ + \frac{\Delta tv[-4q_1q_2x+(-3q_1-3q_2+5q_1q_2)v\tau]}{12\Delta t^2v^3(1-s_1s_2)} + \frac{D[\Delta t(-3q_1-3q_2+5q_1q_2)v+2q_1q_2(-2x+v\tau)]}{12\Delta t^2v^3(1-s_1s_2)}, & \text{if } \frac{D+v\tau}{2} \leq x \leq \frac{D+v(\tau+\Delta t)}{2}; \\ \left(\frac{q_1s_2x}{v} + q_2\tau + \frac{q_2|D-x|}{v} \right) / (1 - s_1s_2), & \text{if } x > \frac{D+v(\tau+\Delta t)}{2}. \end{cases} \quad (6)$$

A plot of $T(x)$ for different values of q_1 is shown in Fig. 1. We see that $T(x)$ has discontinuous derivatives at the origin locations, because the forks originate there. At the origins, the mean replication times are

$$\begin{aligned} T(x_1) &= T(0) = q_2s_1(D/v + \tau)/(1 - s_1s_2); \\ T(x_2) &= T(D) = (q_1s_2D/v + q_2\tau)/(1 - s_1s_2). \end{aligned} \quad (7)$$

It is commonly assumed in the replication literature that $T(x)$ has a minimum at an origin, and that the value of this minimum directly gives the average activation time for the origin. However, Eq. (7) shows that this is not the case and, in fact, $T(x_i) \geq t_i^{\text{av}}$: the mean replication time at an origin location is equal to or greater than the origin's average activation time. Only when an origin has $q_i = 1$ can $T(x_i) = t_i^{\text{av}}$, because if an origin fails to activate in a given cell, the DNA at the origin location will not be replicated until a fork from another origin arrives. This means that T_i is higher for origins that are more likely to fail, as seen directly in Fig. 1. Another important conclusion from Eq. (7) is that even when both origins have the same average activation time ($\tau = 0$), generally we have $T(x_1) \neq T(x_2)$. This is again due to the possibility of origins not activating. Therefore, the origin with the lower minimum of $T(x)$ does not necessarily activate earlier than the other origin: minima of $T(x)$ cannot be used to draw conclusions on the relative activation times of the corresponding origins, as previously assumed [4,10]. Equations (1)–(6) show that in general $T(x)$ at any point depends collectively on the parameters of all origins. However, if an origin is highly competent, early activating, and isolated from other origins, $T(x)$ at that origin's position will be close to the origin's average activation time.

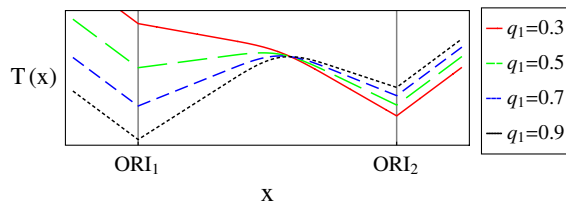


FIG. 1 (color online). Replication time curves for differing values of competence q_1 . Parameters values: $q_2 = 0.7$, $\tau = 1$, $\Delta t = 4$, $D = 10$, $v = 1$.

Equation (6) challenges the assumption that origins are located at minima of $T(x)$. From Eq. (6) the expression for the slope of T near the first origin (for $x > 0$):

$$T'(x) = \frac{q_1 - q_2s_1}{v(1 - s_1s_2)}. \quad (8)$$

This expression shows that the slope is a function of the competencies q_i of both origins as well as the fork velocity v . For the origin at $x = 0$ to be a minimum of $T(x)$, we must have $T' > 0$ for $x > 0$, from which we get the condition $q_1 > \frac{q_2}{1+q_2}$. This shows that if an origin has low competence compared to its neighbor, it may not be a minimum of $T(x)$, which can be seen in Fig. 1. This phenomenon has been observed in experimental data [4]. Note that if $q_1 > 1/2$, this condition is always satisfied and a minimum is guaranteed for this two-origin system. In addition, Eq. (8) shows that the fork velocity is not given by the slope of $T(x)$, an assumption widely used in the literature [4].

Figure 1 shows that $T(x)$ has sharp corners at origin locations. The reason for this is that in every cell the forks always start at the same locations (the origins), which causes a discontinuous change in the proportion of left-propagating compared to right-propagating replication forks, which in turn causes the discontinuity in the derivative of $T(x)$ at the origins. In contrast, Fig. 1 shows that the local maximum of $T(x)$ between two origins is a smooth curve. The reason is that in different cells in a population forks meet each other and terminate at different locations on the DNA, because of the stochastic variations in activation times. This reasoning suggests that the shape of the maxima of T could be used to infer information about the width Δt of the activation time distribution. We expect that sharp maxima should correspond to forks meeting within a narrow time window, and consequently a small value of Δt ; conversely, a broad maximum corresponds to a high Δt . This can be seen in Fig. 2, where $T(x)$ is plotted for various

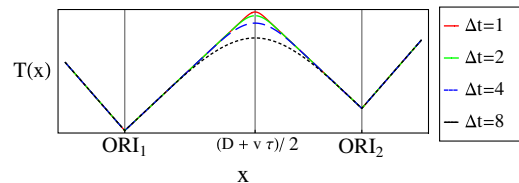


FIG. 2 (color online). Replication time curves for different widths of the activation time window.

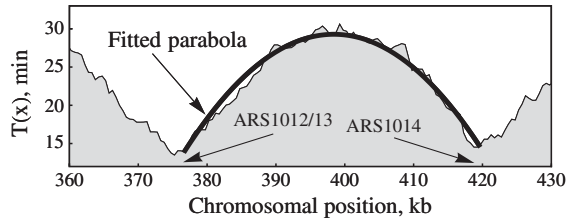


FIG. 3. Replication time curves [4] for *S. cerevisiae* chromosome X with a fitted parabola [375–420 kilobase pairs (kb)].

values of Δt . In order to investigate this more quantitatively, we use the modulus of the second derivative of $T(x)$ at the maxima to measure how broad the maxima is—low values of $|T''|$ correspond to broad peaks. We now use Eq. (6) to find the relationship between $|T''|$ and the origin parameters. At the maximum of $T(x)$, we find

$$|T''| = \frac{4q_1q_2\sqrt{1 - \left|\frac{1}{q_2} - \frac{1}{q_1}\right|}}{v^2\Delta t(1 - s_1s_2)}. \quad (9)$$

Thus $|T''|$ is inversely proportional to Δt . Notice also that $|T''|$ does not depend on τ , which means it is independent of the origins' average activation times. This expression can be used to calculate Δt from an experimental replication time profile $T(x)$, if the origin competencies and the fork velocity are known. This is a very useful result because it allows the determination of a quantity characterizing stochastic properties of the system Δt from $T(x)$, which is defined by a population average. This is valuable because experiments to directly measure Δt are technically difficult, and there are few results available [5,6]. We note that this does not require assuming that all cells in the population are synchronized, since in each individual cell in an asynchronous population, the statistics of the relative activation times of origins remain unaltered [8].

Equation (9) was obtained using a simple uniform distribution for $p_i(t)$. However, we expect it to be a good approximation for any single-peaked distribution function $p_i(t)$, since Eq. (9) only involves the second moment (the variance) of the distribution, and the replication dynamics are mostly determined by the average activation time and the width of the activation distribution—the first and second moments of $p_i(t)$. To test this assumption we used Eqs. (1)–(6) to numerically compute $T(x)$ for pairs of origins with either a Gaussian or a skewed distribution that lead to sigmoidal cumulative distributions [14]. Choosing parameters such that all these distributions have the same mean and variance, we find that in all cases T'' never differs between distributions by more than 10%.

Despite the fact that we have been considering a hypothetical two-origin chromosome, we expect Eq. (6) to be a good approximation for chromosomes with many origins when two neighboring origins are relatively isolated from

other origins. To test this, we looked at experimental data [4] for *S. cerevisiae* chromosome X, specifically the region containing origins ARS1012/13 and ARS1014 (Fig. 3). The smoothness of the curve—ignoring the fluctuations caused by experimental noise—is direct evidence for stochastic origin activation, in agreement with other results [5,6]. We fitted a parabola through the data points and from this determined the value of $|T''|$. Using Eq. (9) we estimate the value of Δt as 10.8 min [19]. This value is in agreement with the limited number of single cell measurements that have been made at other *S. cerevisiae* origins [6].

We thank M. Hawkins for valuable discussions. This work has been supported through the Biotechnology and Biological Sciences Research Council (Grants No. BB/E023754/1, No. BB/G001596/1, and No. BB-G010722).

- [1] C. A. Nieduszynski, Y. Knox, and A. D. Donaldson, *Genes Dev.* **20**, 1874 (2006).
- [2] C. A. Nieduszynski, S. A. P. Hiraga, C. J. Benham, and A. D. Donaldson, *Nucleic Acids Res.* **35**, D40 (2007).
- [3] J. J. Blow and P. J. Gillespie, *Nat. Rev. Cancer* **8**, 799 (2008).
- [4] M. K. Raghuraman, E. A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D. J. Lockhart, R. W. Davis, B. J. Brewer, and W. L. Fangman, *Science* **294**, 115 (2001).
- [5] S. Tuduri, H. Tourriere, and P. Pasero, *Chrom. Res.* **18**, 91 (2010).
- [6] E. Kitamura, J. J. Blow, and T. U. Tanaka, *Cell* **125**, 1297 (2006).
- [7] K. L. Friedman, B. J. Brewer, and W. L. Fangman, *Genes Cells* **2**, 667 (1997).
- [8] A. P. S. de Moura, R. Retkute, M. Hawkins, and C. A. Nieduszynski, *Nucleic Acids Res.* **38**, 5623 (2010).
- [9] J. Lygeros, K. Koutroumpas, S. Dimopoulos, I. Legouras, P. Kouretas, C. Heichinger, P. Nurse, and Z. Lygerou, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 12 295 (2008).
- [10] T. Spiesser, E. Klipp, and M. Barberis, *Mol. Genet. Genomics* **282**, 25 (2009).
- [11] M. D. Sekedat, D. Fenyő, R. S. Rogers, A. J. Tackett, J. D. Aitchison, and B. T. Chait, *Mol. Syst. Biol.* **6**, 353 (2010).
- [12] O. Hyrien and A. Goldar, *Chrom. Res.* **18**, 147 (2010).
- [13] H. Luo, J. Li, M. Eshaghi, J. Liu, and R. K. M. Karuturi, *BMC Bioinf.* **11**, 247 (2010).
- [14] S. C.-H. Yang, N. Rhind, and J. Bechhoefer, *Mol. Syst. Biol.* **6**, 404 (2010).
- [15] S. Jun and J. Bechhoefer, *Phys. Rev. E* **71**, 011909 (2005).
- [16] A. Kolmogorov, *Izv. Akad. Nauk. SSSR* **1**, 335 (1937).
- [17] Y. A. Andrienko, N. V. Brilliantov, and P. L. Krapivsky, *Phys. Rev. A* **45**, 2263 (1992).
- [18] A. Al-Mahboob, Y. Fujikawa, J. T. Sadowski, T. Hashizume, and T. Sakurai, *Phys. Rev. B* **82**, 235421 (2010).
- [19] Parameter values used were $q_1 = 0.74$, $q_2 = 0.78$ and $v = 1.9$ kb/min, estimated in [8].