Reconstructing DNA replication kinetics from small DNA fragments

Haiyang Zhang and John Bechhoefer*
Department of Physics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada
(Received 20 January 2006; published 5 May 2006)

In higher organisms, DNA replicates simultaneously from many origins. Recent *in vitro* experiments have yielded large amounts of data on the state of replication of DNA fragments. From measurements of the time dependence of the average size of replicated and nonreplicated domains, one can estimate the rate of initiation of DNA replication origins, as well as the average rate at which DNA bases are copied. One problem in making such estimates is that, in the experiments, the DNA is broken up into small fragments, whose finite size can bias downward the measured averages. Here, we present a systematic way of accounting for this bias by deriving theoretical relationships between the original domain-length distributions and fragment-domain length distributions. We also derive unbiased average-domain-length estimators that yield accurate results, even in cases where the replicated (or nonreplicated) domains are larger than the average DNA fragment. Then we apply these estimators to previously obtained experimental data to extract improved estimates of replication kinetics parameters.

DOI: 10.1103/PhysRevE.73.051903 PACS number(s): 87.16.Ac, 02.50.Ey, 05.40.−a, 82.60.Nh

I. INTRODUCTION

In simple prokaryotic organisms such as the bacterium *E. coli*, DNA replication begins at a single well-defined site, the *origin* of replication. The total amount of DNA can be replicated in about 20 min. In higher organisms, although the copying rate of DNA replication is 100 times slower and the genome 1000 times larger, replication can occur as quickly as in bacteria. The apparent paradox is resolved by the observation that the replication of DNA proceeds in parallel at many different sites along the genome. There are thus many different (of order 100 000) replication origins in higher organisms. Some questions immediately arise: where are these replication origins along the genome? Do they "fire" stochastically, and, if so, with what rate?

Recent experiments in cell-free embryos of the often-used frog *Xenopus laevis* have yielded enough data to begin to explore such questions [1,2]. The data from these experiments have been analyzed using a stochastic model originally developed to study crystallization kinetics by Kolmogorov [3], Johnson and Mehl [4], and Avrami [5] (KJMA). The KJMA model is a standard tool in materials science for inferring details about nucleation processes from observations of the fraction of a system that is frozen as a function of time [6]. In the 1980s, Sekimoto observed that the KJMA model could be solved essentially exactly in one spatial dimension [7]. This work was extended by Ben-Naim and Krapivsky [8,9]. Later, Herrick *et al.* applied the KJMA model in one dimension to describe the progress of DNA replication [10–12]. The application to DNA is possible because the KJMA model is a general description of a stochastic process with three elements:

(1) initiation (the random firing of replication origins);
(2) growth (replication *forks* spread out symmetrically);
(3) coalescence (DNA replicates only once per cell cycle; two replicated domains that meet thus coalesce).

In the *Xenopus* experiments, replicated DNA was fluorescently labeled. Altering the label (particularly, its color) at a selected time point during the replication process leads to optical micrographs of fragments of DNA chromosomes that show alternating domains of replicated and nonreplicated regions of DNA at the time when the second type of labeled nucleotide was added. In other words, one has a kind of "snapshot" of the replication state of the DNA at a given time. From many such snapshots acquired at many times, one can infer quantities such as the average size of replicated domains and of nonreplicated domains. The time dependence of these averages then leads to inferred rates of origin initiation and of the fork velocity.

One limitation of the above experiments is that while each chromosome of DNA is a single molecule hundreds of millions of basepairs long, the process of preparing the optical-microscope samples led to DNA fragments that were, on average, only 100 kb long. As long as the typical sizes of replicated and nonreplicated domains are much smaller than this size, the finite size of the fragments will have little effect on any estimates of average domain sizes. But if the average domain size (of either a replicated or nonreplicated domain) is comparable to the size of the fragment of DNA, there will be an obvious bias downward of the inferred average, since we can never observe a domain larger than the fragment itself.

In this paper, we estimate the importance of such effects and propose ways for removing the bias. We then reanalyze the data of Herrick *et al.* [10] using improved estimates of average domain sizes. Although our problem is motivated by the application to DNA replication and we adopt language appropriate to that case, the model itself is quite general and applies to any situation where one-dimensional domains are sampled by finite-size fragments.

II. EFFECT OF FINITE-SIZE FRAGMENTS ON MEASURED LENGTH DISTRIBUTIONS

We first consider the effects that the finite size of DNA fragments will have on domain-length distributions. For a

*Electronic address: johnb@sfu.ca

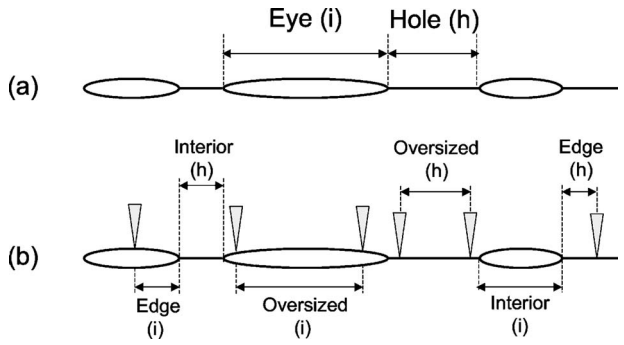

FIG. 1. Definitions of domain types. (a) An infinitely long segment of DNA consists of eyes (replicated regions) and holes (nonreplicated regions). (b) On a finite fragment, there are interior, exterior, and oversized domains for both the eyes (i) and holes (h). Shaded vertical wedges denote cuts in the DNA molecule.

more leisurely discussion, see [13]. The general situation is illustrated in Fig. 1. In part (a), we show a section of a very long DNA molecule, with two different types of domains. For historical reasons having to do with the appearance of partially replicated DNA in electron micrographs, replicated domains are referred to as *eyes* and nonreplicated domains as *holes*. In Fig. 1(b), we see that if one examines a finite fragment of DNA, eye and hole domains are subdivided into one of three categories: A domain that is wholly contained within the fragment is an *interior domain*; one that is cut off by the edge of the fragment is an *edge domain*; and one that covers the entire fragment and more is an *oversized domain*.

We can now formulate our problem more precisely: On the long molecule of DNA, there is an *original distribution* of eye lengths $\varrho(X)$, which is the probability that an observed domain will have length between X and $X+dX$. The DNA molecule is then broken up into fragments. We will consider two situations. In the first, the molecule is broken into segments that all have the same length L (*uniform-cut model*). In the second, the fragments are distributed according to a distribution $\varrho_f(L)$, with average $\langle L\rangle=\int_0^\infty L\varrho(L)dL$ (*general-cut model*). In both cases, we assume that cuts occur randomly with respect to specific locations on the original molecule of DNA.

To simplify notation, we scale all lengths by $\langle L\rangle$, setting $x\equiv X/\langle L\rangle$ and $\ell\equiv L/\langle L\rangle$. This leads to scaled probability distributions, defined by $\rho(x)dx=\varrho(X)dX$ and $\rho_f(\ell)d\ell=\varrho_f(L)dL$.

Given $\rho(x)$, we define the frequency with which one observes an interior domain $n_i(x)\equiv N_i(x)/N_t$, where $N_i(x)dx$ is the number of domains between x and $x+dx$ observed out of a total of $N_t\to\infty$ observed domains. The frequencies of observing edge domains, $n_e(x)$, and oversized domains, $n_o(x)$ are defined similarly. We will consider the relative numbers of interior, edge, and oversized domains in Sec. III later. We also can normalize individually each of these frequency distributions to define probability distributions. For example, $\rho_i(x)=n_i(x)/\int_0^\infty n_i(x)dx$, with $\int_0^\infty\rho_i(x)dx=1$. There are similar definitions for $\rho_e(x)$ and $\rho_o(x)$. Although the probability distributions lack the information on the relative frequencies of the three types of domains that is contained in the frequency distributions, they are useful in evaluating statistics on experimental data, such as the average interior and edge domain sizes $\langle x_i\rangle=\int_0^\infty x\rho_i(x)dx$ and $\langle x_e\rangle=\int_0^\infty x\rho_e(x)dx$.

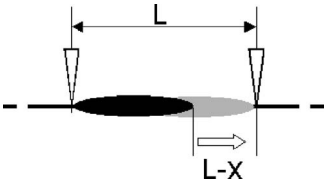

FIG. 2. The *free length* over which an interior domain of size X can be moved within a fragment of length L without being chopped is $L-X$, implying a *free fraction* of $1-x$.

Our question is then as follows: "Given a particular $\rho(x)$, what will be the three observed derived frequency and probability distributions?" Of course, the question may be posed for either hole or eye distributions. We will tackle this problem in two stages, considering first the uniform-cut model and then the general-cut model.

A. Distribution of interior domains

We begin by considering how an arbitrary distribution $\rho(x)$ on the infinite-length DNA generates an interior-domain frequency distribution $n_i(x)$, the expected fraction of domains that are interior domains with size between x and $x+dx$. Again, a domain can be either a hole or an eye. Let us consider a domain of length X and the uniform-cut model with fragments of length L. If $x=X/L>1$, the cut fragment cannot contain the domain, implying that the fraction of interior domains with $x>1$ is zero: $n_i(x>1)=0$. If $x\leqslant 1$, the probability of having an interior domain is proportional to the *free length* within this fragment, i.e., to $1-x$ (Fig. 2). A large domain then has a relatively small likelihood of being contained within a fragment, while a small domain has a relatively large likelihood. A domain that is longer than the fragment cannot be contained at all. Combining these two cases, we have

$$n_i(x)=\begin{cases}(1-x)\rho(x), & x\leqslant 1,\\ 0, & x>1.\end{cases} \tag{1}$$

Normalizing Eq. (1) gives $\rho_i(x)$.

The generalization to the case of a distribution of cut sizes $\rho_f(\ell)$ is straightforward: The probability that a domain of size x is contained within a fragment of size ℓ is $n_i(x,\ell)=\rho_f(\ell)(\ell-x)\rho(x)$, where $\rho_f(\ell)$ is the fraction of fragments of size ℓ. Integrating over all fragment lengths ℓ that are larger than the interior domain, we find

$$n_i'(x)=\rho(x)\int_x^\infty(\ell-x)\rho_f(\ell)d\ell, \tag{2}$$

where the limits of the integral reflect the fact that a segment must be larger than the domain size x in order to contain it. [Equation (2) reduces to (1) by taking $\rho_f(\ell)=\delta(\ell-1)$, with $\delta(\ )$ the Dirac δ function.] Again, one can deduce the corresponding probability distribution $\rho_i'(x)$ by requiring unit nor-

malization. Here, we use primes to denote distributions for the general-cut case.

B. Distribution of edge domains

To understand the distribution of edge domains, we first consider how a domain of size x' is cut. Then we consider the probability density for a resulting edge domain to have a size $x \leqslant x'$. In considering how a domain of size x' is cut, there are two cases: if $x' \leqslant 1$, then the probability density that it was cut was x'. Since the cut position is assumed to be uniformly distributed along the DNA, the probability density that the cut creates a domain of size x is then $2/x'$. The factor of 2 arises because each cut creates *two* edges. Thus, the probability density that the cut produces a domain of size x is $(2/x') \cdot x' = 2$.

In the second case, the original domain x' is larger than 1 and will always be cut. The probability of creating a domain of size x is now uniform over the fragment and is thus 2, just the same result as we found in the first case. Finally, we observe that an edge domain of size x can be created by any domain of size $x' > x$. This leads to the relative frequency of observing an edge domain of size x:

$$n_e(x) = 2\int_x^\infty \rho(x')dx'. \tag{3}$$

Normalizing Eq. (3) leads to the probability distribution $\rho_e(x)$.

To generalize to the case where cuts are distributed as $\rho_f(\ell)$, we follow the logic in Sec. II A, multiplying Eq. (3) by $\rho_f(\ell)$ and integrating over ℓ. Again, only cuts greater than x can lead to an edge of size x. Thus, we have

$$n_e'(x) = 2\int_x^\infty \rho_f(\ell)d\ell \int_x^\infty \rho(x')dx', \tag{4}$$

where $\rho_e'(x)$ may again be determined by normalizing $n_e'(x)$.

C. Distribution of oversized domains

The simplest way to derive the distribution of oversized is to recognize that there is a duality between domains and cut fragments. That is, if one interchanges fragments with domains and cut locations with domain boundaries, then the oversized domains of the original case are just the interior domains of the dual case (Fig. 3).

Applying this analogy to the general-cut case, we have

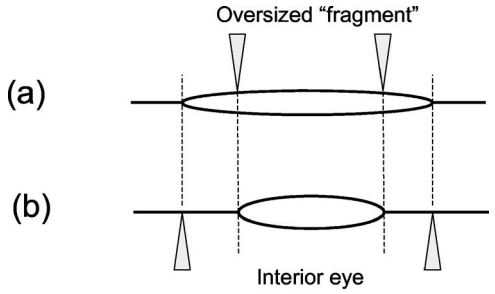

FIG. 3. An illustration of the duality between domains and fragments. (a) Oversized domain; (b) interior domain. (a) and (b) may be derived from each other by interchanging fragments with domains. Shaded vertical wedges denote places where the molecule is cut.

$$n_o'(x) = \rho_f(x)\int_x^\infty (x' - x)\rho(x')dx'. \tag{5}$$

Equation (5) can be derived from Eq. (2) by interchanging $\rho_f(\,)$ with $\rho(\,)$ and changing the variable of integration from ℓ' to x'. To specialize to the uniform-cut case, we simply let ρ_f be a δ function, as there is only a single cut size L. Then

$$n_o(x) = \delta(x - 1)\int_1^\infty (x' - 1)\rho(x')dx'. \tag{6}$$

A distribution of domains $\rho(x)$ on an infinitely long molecule of DNA thus gives rise to three different domain distributions on finite segments of DNA. Table I summarizes the formulas describing the three different domains in the uniform- and general-cut cases.

D. Example

Finally, we illustrate these results by an example (Fig. 4). Let the original distribution be uniform between 0 and 100 units, and let the DNA be cut into fragments uniformly distributed between 50 and 150 units. Then $x = X/100$, and the interior-domain frequencies are given by

$$n_i(x) = \begin{cases} 1 - x, & 0 \leqslant x \leqslant 0.5, \\ (9/8) - (3/2)x + (1/2)x^2, & 0.5 < x \leqslant 1; \end{cases} \tag{7}$$

the edge-domain frequencies are

TABLE I. Summary of the relationships between the fragment domain length distributions and the original domain length distributions, for both uniform- and general-cut models. The θ function is 1 for $x < 1$ and 0 for $x > 1$.

	Fragment distributions	
	Uniform-cut model	General-cut model
Interior	$n_i(x) = \rho(x)(1-x)\theta(1-x)$	$n_i'(x) = \rho(x)\int_x^\infty (\ell - x)\rho_f(\ell)d\ell$
Edge	$n_e(x) = 2\theta(1-x)\int_x^\infty \rho(x')dx'$	$n_e'(x) = 2\int_x^\infty \rho_f(\ell)d\ell \int_x^\infty \rho(x')dx'$
Oversized	$n_o(x) = \delta(x-1)\int_1^\infty (x'-1)\rho(x')dx'$	$n_o'(x) = \rho_f(x)\int_x^\infty (x'-x)\rho(x')dx'$

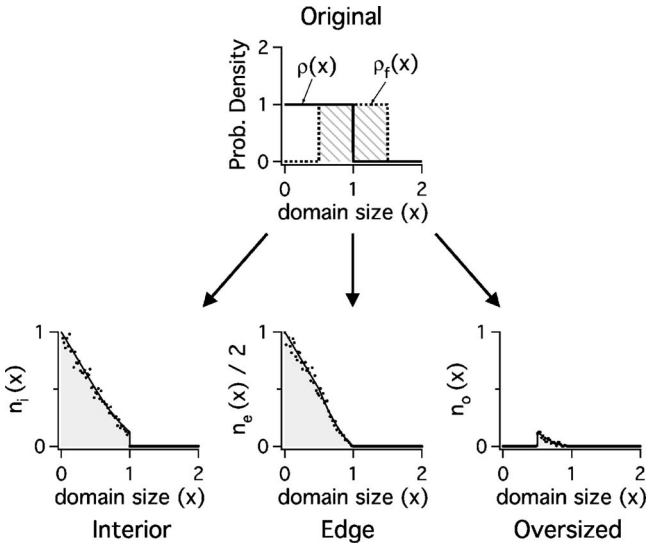

FIG. 4. An example of the transformation of an original uniform distribution of domains when cut into uniformly distributed fragments. Interior, edge, and oversized frequency distributions are shown. Monte Carlo simulations, with a total number of domains $N_i=10^4$, are overlaid. There are no adjustable parameters.

$$n_e(x)=\begin{cases}2(1-x), & 0\leqslant x\leqslant 0.5,\\ 2[(3/2)-x](1-x), & 0.5<x\leqslant 1;\end{cases} \tag{8}$$

and the oversized-domain frequencies are

$$n_o(x)=\begin{cases}0, & 0\leqslant x\leqslant 0.5,\\ (1/2)-x+(1/2)x^2, & 0.5<x\leqslant 1.\end{cases} \tag{9}$$

The frequencies n_i, n_e, and n_o are all zero for $x>1$. More examples are given in [13].

III. UNBIASED ESTIMATORS OF AVERAGE DOMAIN SIZES

In Sec. II we saw how cutting up a long piece of DNA into fragments led to three different types of subdomains: interior, edge, and oversized. We then gave explicit formulas for calculating the frequency distributions of each of these domains, given an original distribution $\rho(x)$. In experiments, one is faced with the reverse problem: given experimentally measured frequency distributions $n_i(x)$, $n_e(x)$, and $n_o(x)$, what can one infer about the original distribution $\rho(x)$?

In principle, from Eq. (2), we can already invert measurements of the fragment distribution $\rho_f(\ell)$ and interior distribution $n_i(x)$ to recover at least part of the original distribution $\rho(x)$; however, unless there is a great deal of data, $\rho(x)$ will be poorly determined [13]. It turns out, though, that the algorithms for inferring replication initiation for DNA—the motivation for our study—require knowledge only of the average replicated and unreplicated lengths. These are just $\langle x\rangle \equiv \int_0^\infty x'\rho(x')dx'$ in our notation. In other words, we need only to estimate the average domain size. In earlier work, we estimated this average using what we will call here the *interior* estimator: $\bar{x}_i \equiv \Sigma_{j=1}^n (x_i)_j/n$, where $(x_i)_j$ is the jth *interior* domain measured on the fragments. (We use overbars for statistical estimators derived from data and angle brackets for averages over the distributions.) For a large amount of data, $\bar{x}_i \to \langle x_i\rangle$. Obviously, since $\rho_i(x)\neq\rho(x)$, the estimator will in general not converge to the true average domain size. Indeed, we expect that $\langle x_i\rangle \leqslant \langle x\rangle$, since domains larger than the fragment size are excluded from the interior distribution. (In the example above, $\langle x_i\rangle = 63/200 = 0.315$, which is less than $\langle x\rangle = 0.5$.) In other words, $\bar{x}_i$ is a biased estimator of $\langle x\rangle$. Can one do better?

In this paper, we present two different unbiased estimators. The first can be derived by integrating the frequency distributions. For simplicity, we focus on the uniform-cut case. Integrating the densities, we define $n_i=\int_0^\infty n_i(x)dx$, with similar expressions for n_e and n_o. Then

$$n_i = \int_0^1 \rho(x)dx - \int_0^1 x\rho(x)dx$$

$$n_e/2 = \int_1^\infty \rho(x)dx + \int_0^1 x\rho(x)dx$$

$$n_o = \int_1^\infty x\rho(x)dx - \int_1^\infty \rho(x)dx. \tag{10}$$

We see that $n_o+n_e/2=\langle x\rangle$ and $n_i+n_e/2=1$. Going back to dimensional units, we have

$$\bar{x}_1 = \frac{\bar{X}_1}{L} = \frac{N_o+N_e/2}{N_i+N_e/2}, \tag{11}$$

where N_i, N_e, and N_o are the total number of observed interior, edge, and oversized domains. The result for $n_e/2$ comes from integration by parts. From Eq. (10), we see that $\bar{X}_1$ is an unbiased estimator of $\langle X\rangle$. Equation (11) is a remarkable result: to estimate $\langle X\rangle$, we merely need to know the fragment size and to count the numbers of interior, edge, and oversized domains. For example, if the domain size is a constant ten times the cut size, each original domain will be cut into nine oversized domains, two edge domains, and no interior domains. Equation (11) then gives $\langle x\rangle_1=[9+(2/2)]/[0+(2/2)]=10$.

The generalization to a distribution of cuts follows the same logic as our previous derivation. We integrate n_i', $n_e'/2$, and n_o' and add them, as above. After repeated integration by parts, we find again that $n_o'+n_e'/2=\langle x\rangle$ and $n_i'+n_e'/2=1$. We then have the same result as Eq. (11), with L replaced by $\langle L\rangle$ (see the Appendix). In the example at the end of Sec. II C, we have $n_i'=25/48$, $n_e'=23/24$, and $n_o'=1/48$. Equation (11) then gives $\langle x\rangle=[(1/48)+(23/48)]/[(25/48)+(23/48)]=1/2$, which is the expected result.

By looking at the first moments of the frequency distributions, we can derive a second unbiased estimator. For the single-cut case, we define $x_i^{tot}=\int_0^\infty x n_i(x)dx$ and have

$$x_i^{tot} = \int_0^1 x\rho(x)dx - \int_0^1 x^2\rho(x)dx,$$

$$x_e^{tot} = \int_1^\infty \rho(x)dx + \int_0^1 x^2\rho(x)dx. \quad (12)$$

Because there is only one size of the oversized domain, $x_o^{tot} = n_o$, as given by Eq. (10). We then see that $x_i^{tot} + x_e^{tot} + x_o^{tot} = \langle x \rangle$. In dimensional units, we have

$$\bar{x}_2 = \frac{\bar{X}_2}{L} = \frac{X_i^{tot} + X_e^{tot} + X_o^{tot}}{N_i + N_e/2}, \quad (13)$$

where X_i^{tot} is the total length of all observed domains, with analogous definitions for X_e^{tot} and X_o^{tot}. Intuitively, the sum of these quantities is just the total length of all the domains, whether they be interior, edge, or oversized. Dividing this by the total number of domains $(N_i + N_e/2)$ then gives our estimator.

As before, the generalization to a distribution of fragment sizes is straightforward and leads simply to replacing L by $\langle L \rangle$ in Eq. (13). In the example at the end of Sec. II C, one finds $x_i^{tot} = 71/384$, $x_e^{tot} = 29/96$, and $x_o^{tot} = 5/384$. Summing these gives 1/2, as before.

Both estimators $\bar{x}_1$ and $\bar{x}_2$ are unbiased. Since $\bar{x}_1$ merely involves counting numbers of domains, it is clearly simpler to measure from the data. However, at small x, the second estimator has a lower variance [13]. This is reasonable, as $\bar{x}_2$ incorporates additional information concerning the lengths of domains. Since the main interest of both estimators is their lack of bias for moderate and large x, they are, in practice, interchangeable.

We discuss one additional biased estimator,

$$\bar{x}_{ie} = \frac{\bar{X}_{ie}}{\langle L \rangle} = \frac{X_i^{tot} + X_e^{tot}}{N_i + N_e/2}, \quad (14)$$

which throws out the information from oversized domains. (If the minimum fragment size $L_{min} > X_{max}$, the maximum domain size, then there are no oversized domains and $\bar{x}_{ie}$ will be unbiased, but this condition will usually not be true.) As we discuss in Sec. IV below, the experiments we analyze give information about the interior and edge domains, but not the oversized domains. As Fig. 5 illustrates, $\bar{x}_{ie}$, while biased, is more accurate than $\bar{x}_i$. Again, we distinguish between $\langle x_{ie} \rangle$, which is derived from moments of $\rho_i(x)$ and $\rho_e(x)$, and $\bar{x}_{ie}$, which is a statistic defined on experimental data.

Finally, we note that it is possible to consider another biased estimator that simply inverts Eq. (1) [13]. This *adjusted-interior* estimator is based only on the interior distribution but corrects for the reduced "phase space" available to larger interior domains. Of course, such an estimator cannot give much information about domains larger than the average fragment size. As one might expect, this estimator performs better than the naive interior estimator but worse than the interior-edge estimator discussed in the previous paragraph.

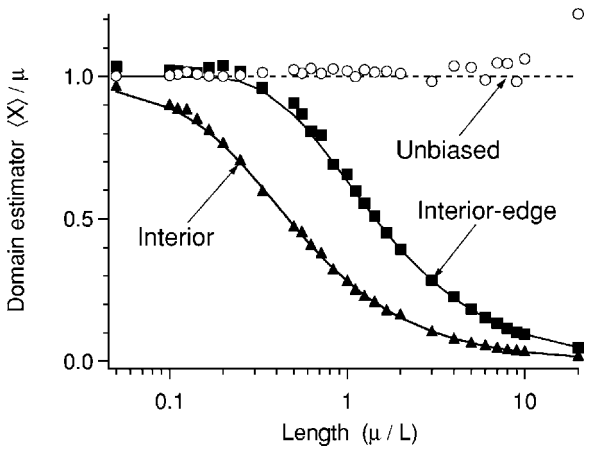

FIG. 5. Performance of different estimators, based on an exponential distribution of domain sizes $\rho(x) = (1/\mu)\exp(-x/\mu)$, with average μ. All lengths are scaled by L, the (unique) fragment size. Unbiased, interior, and interior-edge estimators are shown, with Monte-Carlo simulation data overlaid. Points from the two unbiased estimators are indistinguishable. To guarantee a 10% bias, the interior estimator requires $x < 0.08$, while the interior-edge estimator requires only that $x < 0.4$.

IV. APPLICATION TO A *XENOPUS* REPLICATION EXPERIMENT

In this section, we apply our methods for treating finite-size effects to the experimental data of Herrick *et al.* [1,10] on *in vitro* replication in *Xenopus* cell embryos. As discussed in the Introduction, in this experiment, newly replicated DNA is fluorescently labeled. At a given time point during replication, one changes the color of the label, so that DNA replicated after that time has a different label (color) than the DNA previously replicated. At the end of replication, the DNA is combed out onto a glass slide [14] (a process that breaks up the DNA into short fragments typically 100 kb long) and observed via two-color fluorescence microscopy. Each fragment of DNA may be viewed as a "snapshot" of the state of the DNA at the time the second dye was added. In optical microscopy, the segments appear as alternating domains of red and green, with the former color representing regions of DNA that had replicated before the second dye was added (*eyes*) and the latter the regions that had not yet replicated (*holes*). In the analysis proposed by Jun *et al.* [10,12], the key quantity is the average eye and hole size as a function either of time or of replication fraction f. In that work, all estimates of the average lengths of eyes and holes were made using the interior estimator $\bar{x}_i$. Bias was avoided by simply discarding data where the average eye or hole size exceeded 10% of the average fragment size.

In addition to the limitations imposed by the finite size of fragments, our main focus here, the experiment had two other important limitations. First, the optical resolution limited the observable domain size to about 2 kb. Second, the DNA analyzed came indiscriminately from approximately 20 000 cells, whose replication starting times were only approximately synchronized. This lack of synchrony meant that the time point of the observation of the DNA fragment was

not a reliable indication of the DNA's stage of overall replication. A simple way to get around this problem is to sort data by replication fraction rather than by time. In effect, we use the local replication fraction of each segment as the "clock." A more sophisticated way to deal with the data involves estimating the starting-time distribution and deconvoluting its effects [10].

Since our theory relies heavily on the assumption that cuts on the DNA occur with equal probability anywhere along the molecule, we first show how to test this assumption on the data. Next, we recompute the relevant domain averages using the interior-edge estimator $\bar{x}_{ie}$ introduced previously. Finally, we use the new data on domain averages to recompute quantities of interest in DNA replication.

A. Is the DNA cut randomly?

As mentioned previously, our theory for reconstructing domain averages from fragmented data assumes that the original DNA molecule is broken up randomly. (We do not assume a particular distribution of fragments, merely that the fragmentation process itself is random.) We have tested for two seemingly plausible scenarios. The first is that there is a tendency to fragment at domain boundaries. The second is that there is a preference to fragment in either hole or eye domains.

Let us consider the case where cuts occur randomly. To simplify the discussion, we again assume a uniform-cut size. We consider the ratio $r \equiv \langle x_e \rangle / \langle x_i \rangle$ of average edge-domain size to average interior-domain size. From Table I, we compute $\rho_e(x)$ and $\rho_i(x)$ by normalizing $n_e(x)$ and $n_i(x)$. Then we compute $\langle x_e \rangle$ and $\langle x_i \rangle$ by taking the first moments of ρ_e and ρ_i. Explicitly, we have

$$r = \frac{\int_0^1 dx(x) \int_x^\infty \rho(x')dx'}{\int_0^1 dx \int_x^\infty \rho(x')dx'} \cdot \frac{\int_0^1 dx(1-x)\rho(x)}{\int_0^1 dx(x)(1-x)\rho(x)}. \quad (15)$$

In the limit $\langle x_e \rangle$ and $\langle x_i \rangle \gg 1$ (domains bigger than the fragment size), then $r \to 3/2$ for *all* reasonable distributions $\rho(x)$. This may be seen by noting in this limit that $\rho(x)$ is safely approximated by $\rho(0)$. The integrals may then be evaluated explicitly.

In the limit $\langle x_e \rangle$ and $\langle x_i \rangle \ll 1$ (domains smaller than the fragment size), the ratio's value depends on the shape of $\rho(x)$. If we consider the exponential distribution of Fig. 5, then the ratio tends to 1. Both limits have identical behavior in the general-cut model as long as the tails of the ρ_f distribution decay fast enough (e.g., exponentially).

We focus on the exponential distribution because the KJMA model predicts that if replication origins are distributed randomly along the genome, then holes will have an exponential size distribution, with a mean whose size shrinks during replication [11]. At the start of replication, there are only small, isolated replicated domains, implying holes are large. At the end, there are only small isolated holes left, with a correspondingly small average. At any time, though, the size distribution should be exponential. In order to test whether holes are in fact exponentially distributed, we reanalyze the data of Herrick *et al.* Figure 6 shows the size distribution of fragments, along with the distribution of replication fraction f as evaluated on each fragment.

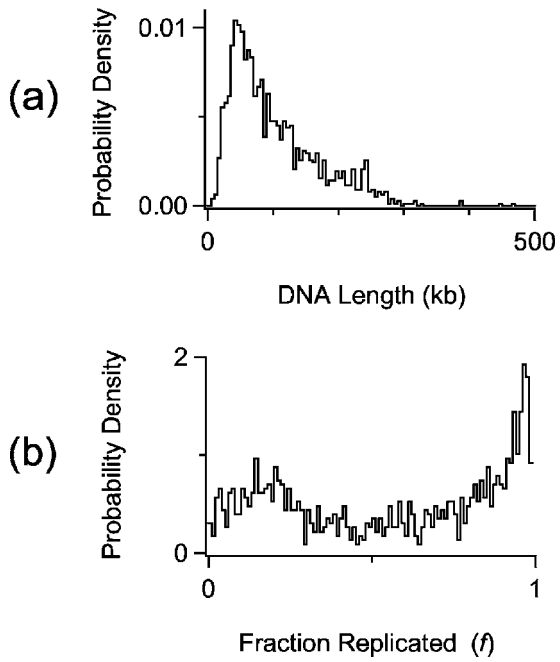

FIG. 6. (a) Distribution of lengths of combed DNA fragments. The average length is 102 kb, with a standard deviation of 75 kb. There are 1142 fragments in the dataset. (b) Distribution of the replication fraction of the DNA fragments.

To study the distribution of hole lengths, we sort data as follows: First, we avoid extremely long or short segments and consider data only between 80 and 240 kb. This gives us 612 fragments (a majority of the total number). Second, we bin segments according to replication fraction f, as discussed previously. To have adequate statistics, we use 20% bins for f. Figure 7 shows edge and interior domains in the 0% to 20% bin. Except for the first point, the histograms are well fit by exponentials. (We can account qualitatively for the lack of domains in the 0–5 kb range by the limited optical resolution of the microscope observations.) The decay constants of the exponential distributions give $\langle x_e \rangle$ and $\langle x_i \rangle$. Figure 8 shows the ratio $r = \langle x_e \rangle / \langle x_i \rangle$ as a function of the replication fraction. As we see, $r \approx 3/2$ for small f (where domains are large) and ≈ 1 at large f (where domains are small). Note that the test is only qualitative since we plot r as a function of f, and not a normalized domain size. We cannot simply do the latter, as we have to use our theory (which assumes the result we test) to evaluate the domain sizes and evaluate a null hypothesis. Still, Fig. 8 suggests that the assumption of random cuts is reasonable.

The second test is to see whether there is a preference for cuts to occur on either holes or eyes. As discussed in [13], this test also leads to the qualitative conclusion that the cuts are consistent with being randomly spaced. We thus conclude that it is reasonable to apply the theory developed in this paper to the dataset of Herrick *et al.* [1,10]. Physically, this

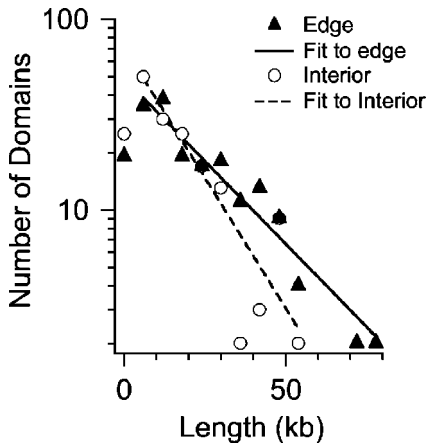

FIG. 7. Histograms of interior and edge hole domains, for replication fraction f between 0 and 0.2. Solid lines are fits to exponential distributions, excluding the first point (see the discussion in the text).

conclusion makes sense: in the experimental protocol, there is essentially no physical difference between replicated and nonreplicated domains. Since visualization occurs after full completion of the replication cycle, both domains have in fact been fully replicated, and what distinguishes them is simply the type of fluorescent label that is attached. Still, it is interesting that one can do independent tests about the nature of the cuts.

B. Reanalysis of the data

We may now proceed to redo the analysis of the data presented in Herrick *et al.* [10]. Because the replication starting times of individual DNA molecules is unknown, we cannot use the data from oversized domains, as we have no way of knowing what replication fraction to assign them to. As a result, we cannot use the unbiased estimators discussed earlier; however, we can use the interior-edge estimator.

In Fig. 9, we show the average eye and hole lengths, as a function of the replication fraction, as computed using the interior estimator (as originally done by Herrick *et al.* in [10]) and with our interior-edge estimator. As we can see, the main differences occur at $f<0.2$ for the holes and $f>0.8$ for

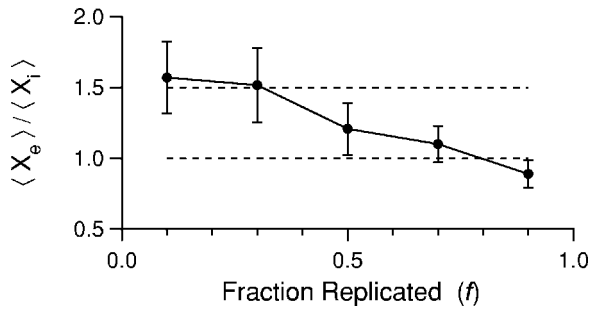

FIG. 8. Ratio of average edge domain length to average interior domain length as a function of the replication fraction. The dashed lines show the expected limits, 1.5 for small f, or large domains, and 1 for large f, or small domains.

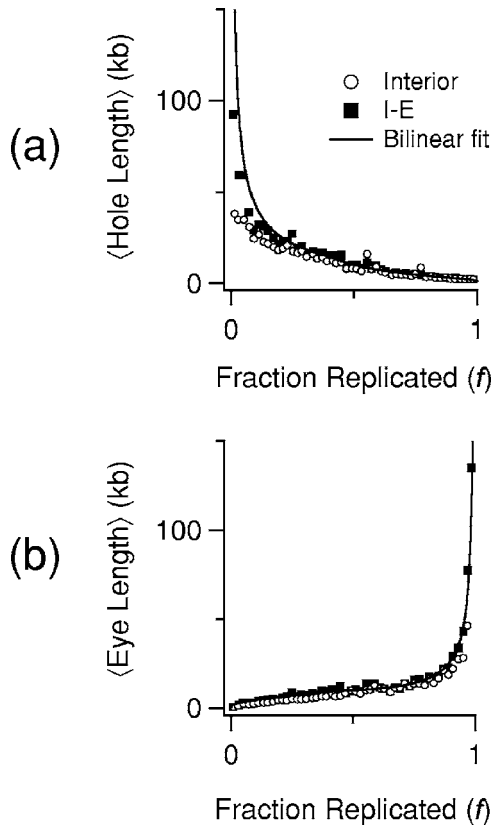

FIG. 9. Average hole (a) and eye (b) lengths as a function of replication fraction f. Open circles are computed with the interior estimator, filled squares with the interior-edge estimator. The solid curve is a fit based on the theory of Herrick *et al.* [10].

the eyes. In both cases, the limit is where the average domain size is large.

In the earlier analysis of Herrick *et al.* [10], the problematic parts of the data were simply truncated and curve fits on the data were done on partial data ranges. In Fig. 9, we have overlaid a curve based on a fit to the theory of Herrick *et al.* [10]. In that theory, a parameter-free inversion is first done to estimate the initiation rate $I(t)$ of replication origins. This function is given (for both the interior and interior-edge estimators) in Fig. 10. Noting that the form of $I(t)$ is well approximated by straight-line segments, we fit two line segments to the $I(t)$ data, as shown. (We neglect the decreasing data in the third segment, since one can show that finite-size effects lead to a systematic undercount of initiation rates at long times, where there is little data [12].) This bilinear form for $I(t)$ is then used to generate the curve that is superposed on the domain-length data in Fig. 9. As we can see, there is little qualitative difference in the two estimates of $I(t)$. The quantitative parameters, which differ (particularly the slope I_2 of the second segment), are given in Table II. We can trace the overall similarity of the two results back to the fact that the two domain-length estimators agree over a large range in Fig. 9 that the extrapolation to the missing data done in the original analysis of the interior-domain data is accurate.

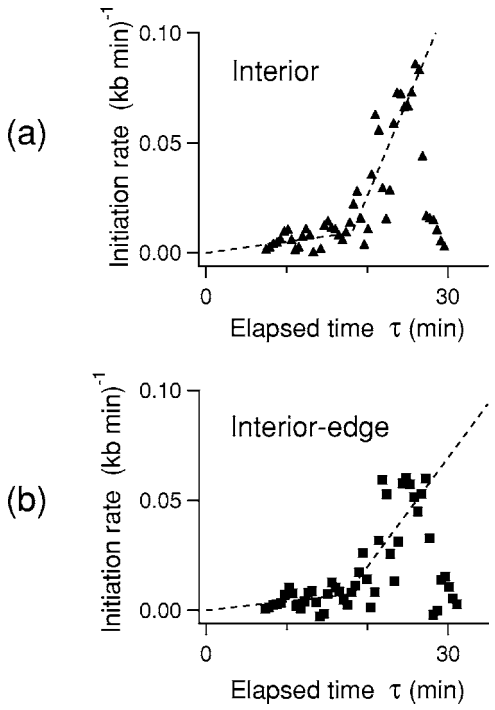

FIG. 10. Extracted initiation rate $I(t)$ as a function of time t using data estimated from the interior estimator (a) and interior-edge estimator (b).

V. CONCLUSIONS

In this paper, we have studied the effects of cutting long molecules of DNA into short fragments on the distributions of domain lengths in DNA experiments. We have seen how the fragmentation of the DNA leads to three different types of domains—interior, edge, and oversized. From an original distribution of domain lengths $\rho(x)$, we can calculate the frequency distributions of these three fragment domains. We can also use statistics such as the average of interior, edge, and oversized domains to make unbiased estimations of the average domain size. The lack of starting-time synchrony in the DNA experiment that motivated this work meant that we had information on only two types of domains; interior and edge. We used those quantities to create an *interior-edge* estimator that, while biased, was more accurate than the previously used interior estimator. In the particular example of an exponential distribution (valid for holes in the replication experiments), the interior-edge estimator increased the maximum usable domain size by a factor of 5. In the end, the results of the reanalysis of the data of Herrick *et al.* [10] led to results that were very close to those originally obtained. We conclude that robust results may be obtained, even with a "naive" analysis of DNA replication.

TABLE II. Extracted nucleation rate $I(t)$ and fork velocity v from data analyzed using two estimators of average domain sizes. The slopes I_1 and I_2 refer to the first and second of the straight-line segments. The time T_c refers to the time marking the end of segment 1 and beginning of segment 2.

Estimator	v (kb/min)	I_1	I_2	T_c (min)
Interior	0.58	5.1×10^{-4}	8.8×10^{-3}	18.1
Interior edge	0.62	4.1×10^{-4}	5.0×10^{-3}	17.4

Future experiments, however, may have to rely on the more sophisticated analysis given in this paper. The type of replication studied in the *Xenopus* experiments was quite special, in that the experiments were done on DNA in cell embryos just after conception and before the process of cell differentiation had begun. At that stage of life, little if any proteins need to be synthesized (they are stored in the egg), and DNA replication occurs rapidly, with an extremely large number of origins. The average interorigin separation inferred by Herrick *et al.* was 6.3 kb, compared to an average domain size of $\approx$100 kb [10]. Thus, the experiments were conducted in a limit where finite-size effects could be expected to be small. (The interorigin separation sets a minimum length scale for average domain sizes and is relevant in the middle portion of replication. Individual domains, particularly at the beginning and end of the replication process, can be much larger than this size.) By contrast, in human developed cells, the average interorigin separation can be as much as 100 kb [15]. While there is some potential for experimentally increasing the average fragment size in the combing technique, it is clear that finite-size effects will be much more important in such work.

Finally, as noted earlier, while this work was motivated by the desire to understand more fully an experiment on DNA replication, the specific theory developed here to treat finite-size effects is a general one that will apply to any situation where one-dimensional domains on a long substrate are cut into fragments.

ACKNOWLEDGMENTS

We thank Suckjoon Jun for many useful discussions and for help in making contact with his previous results. This work was supported by NSERC (Canada).

APPENDIX: DERIVATION OF THE UNBIASED ESTIMATORS

Here, we sketch the derivation of the unbiased estimator of Eq. (11) for the general-cut model. The derivation mostly consists of repeated integration by parts. In all cases, the boundary terms are zero.

From Eq. (2), we have

$$n_i' = \int_0^\infty dx\, \rho(x) \int_x^\infty d\ell\ \ell \rho_f(\ell) - \int_0^\infty dx\, x\rho(x) \int_x^\infty d\ell\ \rho_f(\ell). \tag{A1}$$

From Eq. (4), we have

$$n_e'/2 = \int_0^\infty dx \int_x^\infty d\ell\ \rho_f(\ell) \int_x^\infty dx' \rho(x') \tag{A2}$$

$$= \int_0^\infty dx\, x\rho(x) \int_x^\infty d\ell\ \rho_f(\ell) + \int_0^\infty d\ell\ \ell\rho_f(\ell) \int_\ell^\infty dx\ \rho(x) \tag{A3}$$

$$= \int_0^\infty dx\, x\rho(x) \int_x^\infty d\ell\ \rho_f(\ell) + \int_0^\infty dx\, \rho(x) \int_0^x d\ell\ \ell\rho_f(\ell), \tag{A4}$$

where we have integrated by parts and used $\int_0^\infty dx\, \rho(x)=1$. After again integrating Eq. (A4) by parts, we add Eqs. (A1) and (A4) and find

$$n_i' + n_e'/2 = \int_0^\infty dx\, \rho(x) \left[\int_0^x d\ell\ \ell\rho_f(\ell) + \int_x^\infty d\ell\ \ell\rho_f(\ell) \right]$$
$$= \int_0^\infty dx\, \rho(x) = 1, \tag{A5}$$

since $\langle \ell \rangle = 1$ in our scaled units.

Similarly, we start from Eq. (5) and integrate each term by parts to find

$$n_o' = \int_0^\infty d\ell\ \rho_f(\ell) \int_\ell^\infty dx\, x\rho(x) - \int_0^\infty d\ell\ \ell\rho_f(\ell) \int_\ell^\infty dx\, \rho(x) \tag{A6}$$

$$= \int_0^\infty dx\, x\rho(x) \int_0^x d\ell\ \rho_f(\ell) - \int_0^\infty dx\, \rho(x) \int_0^x d\ell\ell\rho_f(\ell). \tag{A7}$$

Then

$$n_o' + n_e'/2 = \int_0^\infty dx\, x\rho(x) \left(\int_0^x d\ell\ \rho_f(\ell) + \int_x^\infty d\ell\ \rho_f(\ell) \right)$$
$$= \int_0^\infty dx\, x\rho(x)(1) = \langle x \rangle. \tag{A8}$$

The result for the second unbiased estimator, Eq. (13), $x_i^{tot} + x_e^{tot} + x_o^{tot} = \langle x \rangle$, is established in a similar manner.

[1] J. Herrick, P. Stanislawski, O. Hyrien, and A. Bensimon, J. Mol. Biol. **300**, 1133 (2000).

[2] J. J. Blow, P. J. Gillespie, D. Francis, and D. A. Jackson, J. Cell Biol. **152**, 15 (2001).

[3] A. N. Kolmogorov, Izv. Akad. Nauk SSSR, Ser. Fiz. **1**, 335 (1937) [Bull. Acad. Sci. USSR, Phys. Ser. (Engl. Transl.) **1**, 335 (1937)].

[4] W. A. Johnson and P. A. Mehl, Trans. AIME **135**, 416 (1939).

[5] M. Avrami, J. Chem. Phys. **7**, 1103 (1939); **8**, 212 (1940); **9**, 177 (1941).

[6] J. W. Christian, *The Theory of Phase Transformations in Metals and Alloys, Part I: Equilibrium and General Kinetic Theory*, 3rd ed. (Pergamon, Oxford, 2002).

[7] K. Sekimoto, Physica A **125A**, 261 (1984); **135A**, 328 (1986); Int. J. Mod. Phys. B **5**, 1843 (1991).

[8] E. Ben-Naim and P. L. Krapivsky, Phys. Rev. E **54**, 3562 (1996).

[9] P. L. Krapivsky, Phys. Rev. B **45**, 12699 (1992).

[10] J. Herrick, S. Jun, J. Bechhoefer, and A. Bensimon, J. Mol. Biol. **320**, 741 (2002).

[11] S. Jun, H. Zhang, and J. Bechhoefer, Phys. Rev. E **71**, 011908 (2005).

[12] S. Jun and J. Bechhoefer, Phys. Rev. E **71**, 011909 (2005).

[13] H. Zhang, M.Sc. thesis, Simon Fraser University, 2005.

[14] A. Bensimon, A. Simon, A. Chiffaudel, V. Croquette, F. Heslot, and D. Bensimon, Science **265**, 2096 (1994).

[15] Y. Jeon, S. Bekiranov, N. Karnani, P. Kapranov, S. Ghosh, D. MacAlpine, C. Lee, D. Hwang, T. Gingeras, and A. Dutta, Proc. Natl. Acad. Sci. U.S.A. **102**, 6419 (2005).