J. Mol. Biol. (1979) **128**, 219–245

Spatial Distribution of Initiation Sites for Mammalian DNA Replication: A Statistical Analysis

Joel E. Cohen, Barbara R. Jasny and Igor Tamm

The Rockefeller University, 1230 York Avenue, New York
N.Y. 10021, U.S.A.

(Received 28 June 1978, and in revised form 30 October 1978)

We have examined DNA replication in three mammalian cell lines (L, muntjac and MDBK) by statistical analyses of light microscopic DNA fiber autoradiograms in order to determine whether the sites where replication is initiated are spatially random or organized. A quantitative model was developed to predict the properties of a random spatial arrangement of initiation sites and was tested against experimental data. A description of the actual spatial organization of activated initiation sites is proposed.

The number of sites per strand and the mean distance between sites depend heavily on the length of the strands measured, according to both observation and theory. However, in all three cell lines, the observed relationship between those variables and strand length differed from the relationship predicted by the random model. The modal inter-initiation distance was nearly the same in all three cell lines (5 to 15 μm). Three methods were used to provide estimates of, or lower bounds on, the mean inter-initiation distances on the unbroken DNA fiber. The ranges of estimates were 8 to 23 μm for muntjac cells, 22 to 45 μm for MDBK cells, and 14 to 63 μm for L cells.

Because inter-initiation distance depends on strand length and strands break with excess probability near hot-labeled regions, alternatives to the testing of pooled samples were developed. Statistical tests of the randomness of inter-initiation distances were applied to individual strands containing at least nine initiation distances. A test of exponentiality against alternative Weibull distributions demonstrated that the probability of the distribution of inter-initiation distances being random was less than one in 10,000. The Kolmogorov–Smirnov and Keiding tests indicated differences in organization between the three cell lines. Long exposure to 5-fluoro-2′-deoxyuridine (5×10^{-7} M) shifted the organization of initiation sites in MDBK cell DNA toward randomness. Neither the mean nor the modal inter-initiation distance detected these changes in organization. The tests for exponentiality employed here provide more sensitive tools for studying the organization of DNA replication.

1. Introduction

In chromosomes, the activity of DNA replication is organized spatially and temporally. In a variety of cell lines, each chromosome has a characteristic reprodncible regional pattern of replication during DNA synthesis (S) phase (Taylor, 1958; Gavosto *et al.*, 1968; Ganner & Evans, 1971; Latt, 1973). Multiple sites of initiation of replication have been distinguished on the double-stranded fiber by means of light microscope DNA fiber autoradiography (Cairns, 1966; Huberman & Riggs, 1968; Hand & Tamm, 1972,1973; Callan, 1972) and electron microscopy (Wolstenholme, 1973; Kriegstein & Hogness, 1974; Newlon *et al.*, 1974). The size of a "replication" unit

0022–2836/79/060219–27 $02.00/0

defined as a segment of DNA in which replication starts at one initiation site or origin and proceeds in opposite directions at two replication forks, has been measured in mammalian and non-mammalian cells. Differences in replication unit size have been correlated with different developmental stages (Callan, 1972; Blumenthal *et al.*, 1973) and with species differences (Hand & Tamm, 1974; Edenberg & Huberman, 1975).

Qualitative analyses of grain density distributions in DNA fiber autoradiograms after pulse-labeling at high and low specific activities have suggested that initiation sites are not randomly distributed in mammalian cells (Huberman & Riggs, 1968; Hand, 1975). A non-random distribution of initiation sites would be consistent with the demonstrated organization of DNA replication at the chromosome level and might relate to the distribution of gene sequences.

In a previous study of L cells (Jasny *et al.*, 1978), we demonstrated a non-random interaction between the location of breaks in the DNA fiber when cells were spread for light microscope autoradiography and the location of activated sites of initiation of DNA replication. We also detected a suggestive pattern of deviations from randomness in a preliminary analysis of the distribution of distances between activated initiation sites. We now return to this suggestion with more refined techniques.

In this paper we describe the initiation site distributions of three mammalian cell lines. We find that estimates of replication unit size and frequency depend markedly on strand length, which previous studies have not adequately considered. We offer a quantitatively precise formulation of the properties of a spatially random distribution of initiation sites. From a variety of statistical analyses, we conclude that the spatial organization of initiation sites is not random. We propose a tentative quantitative description of the pattern of deviation from randomness.

2. Materials and Methods

(a) *Cell culture*

L929 cells, a continuous line of mouse fibroblasts, were grown in monolayer cultures in Eagle's minimal essential medium (Eagle, 1959) supplemented with 5% fetal calf serum. The MDBK cell line is epithelioid and was derived from a kidney of a male steer (Madin & Darby, 1958). The muntjac cell line is fibroblastic. It was derived from a skin biopsy of an adult male muntjac (*Muntiacus muntjac*) and is especially noted for its low chromosome number (Wurster & Benirschke, 1970). MDBK and muntjac cells were grown in reinforced Eagle's medium (Bablanian *et al.*, 1965) supplemented with 10% fetal calf serum. L cells were obtained from Microbiological Associates, Bethesda, Md. MDBK and muntjac cell lines were originally obtained from the American Type Culture Collection, Rockville, Md.

Asynchronous populations were employed for the experiments to be described. Cells were passaged when they were less than confluent into 28 cm^2 Petri dishes at 3×10^4 to 1×10^5 cells/dish and allowed to double 2 to 3 times before use.

(b) *Radioisotopes and chemicals*

[^{3}H]thymidine (50 to 60 Ci/mmol, 1 mCi/ml) was obtained from New England Nuclear, Boston, Mass. 5-Fluoro-2′-deoxyuridine was a gift from Hoffman-La Roche, Inc. NTB2 emulsion was obtained from the Eastman Kodak Company, Rochester, N.Y.

(c) *Radioisotope labeling and DNA fiber autoradiography*

Exponentially growing cells were treated for 0·5 h with 2×10^{-6} M-FdUrd† to deplete the thymidylate pool. The cells were then pulse-labeled for 10 min with high specific

† Abbreviations used: FdUrd, 5-fluoro-2′-deoxyuridine; dThd, thymidine.

activity [^{3}H]dThd (50 to 60 Ci/mmol), followed by 3 h of low specific activity (5 to 6 Ci/mmol) labeling. The cells were lysed on glass slides by placing 25 μl of cells next to 25 μl of a solution of 1% sodium dodecyl sulfate + 0·01 M-EDTA in phosphate buffered saline, and mixing the drops by placing a glass rod across the slide. The mixture was then gently pulled down the slide with the rod. After drying, the DNA was precipitated in 5% trichloroacetic acid, dehydrated in ethanol at increasing concentrations, and the slides were dipped in Kodak NTB2 emulsion. After exposure for approximately 5 months, the slides were developed.

(d) *Terminology*

We refer to each activated initiation site for DNA replication, detected by our protocol, as a site. We use fiber to refer to the unbroken DNA molecule. An end of a fragment of a fiber is called free if labeling ends clearly and unambiguously rather than, for example, in a mass of fragments. Only fragments with 2 free ends have a well-defined length. An observed end may be either the physical end of a fragment or the point where labeling stopped, with the adjacent DNA unlabeled. Because the cells were not synchronized and were growing exponentially, the labeled DNA is probably representative of the total DNA.

An inter-initiation distance is defined as the measured distance (in μm of DNA; 1 μm = 3000 base-pairs, approx.) between 2 adjacent sites. The location of a site is taken as the center of a pre-pulse or post-pulse figure (Jasny *et al.*, 1978).

The exponential distribution is the predicted distribution of lengths or distances between events which occur according to the models of randomness described in the Appendix. A test of exponentiality is a statistical test as to whether a sample of observed lengths or distances conforms as well as might be expected by chance to an exponential distribution. The 0-truncated geometric distribution, or geometric distribution for short, is the predicted distribution of the number of sites per strand, assuming randomness (see Appendix).

(e) *Sampling procedures*

The samples of fragments used in the following analyses were constructed in one of two ways. Initially, in "complete" samples, a strand was defined as a fragment with 2 free ends and at least 1 site (Jasny *et al.*, 1978). Fragments without sites were omitted from complete samples because of the difficulty of counting and measuring them reliably. All strands (as just defined) on each slide examined were recorded. This procedure minimizes subjectivity and systematic errors in the choice of strands. It also eliminates the effects of the variation which we noted in strand length and in the number of sites/strand among different regions of an individual slide. Complete samples of strands were examined repeatedly by the same and different observers (Jasny, 1978). Among strands at least 50 μm long, measurements of strand length and counts of sites/strand can be made reliably.

In samples of "many-site" strands, a strand was defined as a fragment with at least 9 inter-initiation distances but not necessarily with 2 free ends. Since the 2 ends of a many-site strand were not determined by termination of label, its length was defined by measuring as much of the strand as was visible. When a strand entered a mass of nuclear material its measured length stopped where the strand was no longer clearly identifiable. At least 19 strands from at least 2 experiments were measured for each cell type and treatment. As in Jasny *et al.* (1978), only inter-initiation distances bounded by "internal" sites, which were therefore inter-initiation distances that could be measured unambiguously, counted among the 9 required for each strand.

(f) *Tests of exponentiality*

To test the exponentiality of lengths of all fragments without sites, we used a conventional χ^2-test of goodness of fit.

In samples of many-site strands, the exponentiality of the inter-initiation distances on each individual strand was tested by each of 5 methods.

Durbin's (1975) adaptation of the Kolmogorov–Smirnov D_n statistic allows the minimum possible inter-initiation distance (α in his notation) to be fixed at 0 or to be estimated from the data for each strand. We analyzed each strand both ways. We obtained values of P, the probability that a worse fit to the exponential distribution would have occurred by chance, by linear interpolation in Durbin's Table 3 (Durbin, 1975). This P value is 2-tailed, since it measures deviation of any form from the exponential distribution.

The test of 2-parameter exponentiality against 3-parameter Weibull alternatives, due to Engelhardt & Bain (1975), estimates the minimum inter-initiation distance (θ in their notation) for each strand and subtracts it from the inter-initiation distances on that strand. The test compares an estimated shape parameter (β in their notation) of a Weibull distribution fitted to the inter-initiation distances minus the minimum, with the theoretical value of 1, which would occur if the data were exponential. If the shape parameter exceeds 1 significantly, then inter-initiation distances are more clustered around the mean than expected from the exponential distribution. If the shape parameter is significantly less than 1, then more extremely short and more extremely long inter-initiation distances are observed than expected from an exponential distribution. Calculation of P for this test relied on the parameters in Table 3 of Engelhardt & Bain (1973). Since we rejected the exponential distribution if the shape parameter was either too high or too low, each 1-tailed P_1 obtained from the procedure of Engelhardt & Bain (1975) was converted to a 2-tailed P_2 by the formula $P_2 = 2(0{\cdot}5 - |P_1 - 0{\cdot}5|)$. A simpler Weibull *versus* exponential test appeared after our calculations were carried out (Engelhardt & Bain, 1977).

Keiding's (1977) test of exponentiality rests on the relation between the exponential and gamma distributions (Johnson & Kotz, 1970). If $\bar{x}$ is the mean of n independent observations from an exponential distribution with a minimum (inter-initiation distance, in our case) of 0, then $K(0) = 2n\bar{x}/s$, where s is the sample standard deviation of these observations, has approximately the distribution of a χ^2 variate with $2n$ degrees of freedom. If the minimum is α, then the modified Keiding statistic $K(\alpha) = K(0) - 2n\alpha/s$ should have approximately a χ^2 distribution with $2n$ degrees of freedom. As in the Kolmogorov–Smirnov and Weibull tests, we estimated the minimum α as the sample minimum and retained $2n$ degrees of freedom as conservative. For each strand we calculated $K(0)$ and $K(\alpha)$. The 1-tailed probability obtained from standard χ^2 tables using this test was converted to a 2-tailed probability as before, since an exponential distribution was rejected by extremely large or extremely small values of K.

Thus the five methods used are: the Kolmogorov–Smirnov test with $\alpha = 0$; the same with α estimated; the Weibull test; the Keiding test with $\alpha = 0$; the same with α estimated.

Five tests of exponentiality were used because each gave different information about deviations from randomness. The Kolmogorov–Smirnov test was an omnibus test for any deviation from an exponential distribution, but was probably less sensitive than the Weibull test to deviations from randomness which arise from a Weibull distribution. The Weibull test assumed that if data were not exponential, then they were Weibull distributed. It gave a quantitative description, under that assumption, of the pattern of deviation from exponentiality. The Keiding test rested on the identity between the mean and standard deviation of a standard exponential distribution; it accepted as exponential any distribution that approximately satisfied that identity and rejected any distribution that did not. We know of no systematic comparisons of the sensitivities of these tests.

(g) *Pooling of* P *values*

Suppose a series of n independent tests of some hypothesis (for example, tests of exponentiality on n strands) gave a set $P_{21}, P_{22}, \ldots, P_{2n}$ of 2-tailed probabilities. To estimate the overall significance of deviations from the hypothesis we compared $-2\ \Sigma_i \ln P_{2i}$ to a χ^2 variate with $2n$ degrees of freedom (Fisher, 1970).

3. Results

(a) *Randomness of strand length*

To test whether breaks in the fiber occur randomly, we compared the distribution of length of strands at least 50 μm long in complete samples with equation (1a) (see Appendix) (Fig. 1; see caption for details of fitting). For all three cell lines, as judged by a χ^2-test of goodness of fit, the agreement between the experimental and theoretical curves was either barely acceptable or was not quite acceptable at the 0·01 probability level. In each cell line, there were more strands observed than predicted in the range of length 50 to 150 μm, and fewer in the range 150 to 250 μm. The same systematic pattern of deviation was observed when an exponential distribution truncated below 50 μm was fitted to these data, and a χ^2-test of goodness of fit indicated even more significant deviations from observation. Thus the observed distribution of strand lengths suggested a discrepancy between the random model and the process which generates strands, but did not specify where the discrepancy lies.

The average length of strands from muntjac cells (93 μm) was significantly less than the average strand length from L cells (122 μm), which was significantly less than the

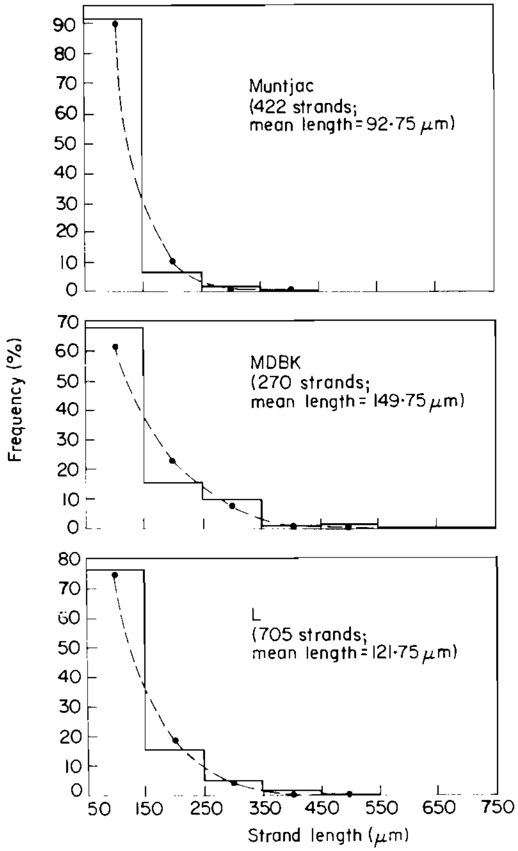

Fig. 1. Frequency distribution of length of strands from complete samples of 3 experiments for muntjac and L cells, and 2 experiments for MDBK cells. Experimental data, grouped into 100 μm-long categories (————); theoretical distribution (– – – – –) was calculated according to $F_{50}(h)$ in Appendix, section (c) by choosing λ and ν to minimize the χ^2 statistic: for muntjac, $\lambda = 0{\cdot}0222$, $\nu = 0{\cdot}1901$; for MDBK, $\lambda = 0{\cdot}0095$, $\nu = 0{\cdot}1667$; for L, $\lambda = 0{\cdot}0132$, $\nu = 0{\cdot}1493$. For all 3 cell lines, the fit of the theoretical distribution was barely or not quite acceptable at the 1% level.

average strand length from MDBK cells (150 μm). These differences may have been due to variation in the spreading procedure or to differences between cell lines.

For L cells, the distribution of lengths of fragments of DNA which contained no sites was compatible with an exponential distribution, according to a sample of fragments with two free ends from three different slides in two experiments (Table 1). This suggested that the presence of sites in a region of DNA may affect the randomness of breakage in that region. As the location of sites rather than the location of breaks was of primary interest, comparable samples of fragments with no sites were not developed for the other cell lines.

TABLE 1

Frequency distribution of length of L cell DNA fragments with no initiation sites, compared to an exponential distribution

Length (μm)	Observed† Number	Observed† Relative frequency	Predicted‡ Number	Predicted‡ Relative frequency
50–100	95	0·841	95·4	0·844
100–150	14	0·124	14·7	0·130
150–200	3	0·027	2·5	0·022
$\geq$200	1	0·009	0·3	0·003
Total	113	1·001	112·9	0·999

† Mean length = 76·9 μm, standard deviation = 26·3 μm.

‡ From Appendix, equation (1b). Goodness of fit χ^2 = 1·43 with 2 degrees of freedom, $P > 0{\cdot}4$.

(b) *Number of sites per strand*

In complete samples of strands (which were defined as containing at least 1 site) from two experiments with L cells, the observed frequency distribution of the number of sites per strand (Table 2) agreed very well with the predicted geometric distribution (Appendix, section (d)). These samples included all strands of length at least 1·71 μm, which was the smallest fragment recorded as a strand.

The frequency distribution of sites per strand reported here was not sensitive to the dependence between fiber breakage and the location of sites which we detected in strand lengths (section (a), above) and previously (Jasny *et al.*, 1978) in the distribution of external segments. The following more refined indicators of the site distribution revealed deviations from the random model.

(c) *Relationship between number of sites and strand length*

In complete samples of strands at least 50 μm long, the mean number of sites per strand increased with the length of the strand in all three cell lines (Fig. 2).

If sites occurred at random on strands of any length, the predicted relationship between strand length and the average number of sites would be given by equation (4), in which ν is the average number of sites per μm of DNA fiber. On the unbroken fiber, ν equals both 1/(mean inter-initiation distance) and (total number of sites)/(total fiber length).

We estimated ν in three ways for each cell line, using all strands at least 50 μm long in complete samples: (A) as (total number of sites on the strands)/(total of

TABLE 2

Frequency distribution of number of initiation sites per strand of L cell DNA, compared to a truncated geometric distribution

Number of sites per strand	Experiment 5		Experiment 6	
	Number of strands			
	Observed	Predicted	Observed	Predicted
1	96	96·7	315	323·2
2	15	13·9	103	90·0
3	2	2·3	20	25·1
≥ 4	0		10	9·7
Total	113	112·9	448	448·0
	$\chi^2 = 0\cdot14$ with 1 degree of freedom $P = 0\cdot71$		$\chi^2 = 3\cdot11$ with 2 degrees of freedom $P = 0\cdot22$	

lengths of all strands); (B) as 1/(mean inter-initiation distance), weighting each inter-initiation distance equally; (C) as the reciprocals of the high and low estimates obtained in Appendix, section (i). These methods are examined further in section (g), below and in the Discussion.

For each of these methods of estimation and each cell line, Table 3 gives $1/\nu$, the estimated mean distance between sites on the unbroken DNA fiber. Methods (A), (B), and (C-high) gave estimates which were consistent within a factor of two.

Returning now to the mean number of sites per strand in complete samples, for each cell line we calculated a predicted curve using each estimate of ν. Invariably estimate (A) gave the closest agreement to the observations. Only the predicted curves based on estimate (A) are plotted in Figure 2.

Method (A) was superior to method (B) probably because only internal sites were used in measurements of inter-initiation distances. The ordinate in Figure 2 is the total number of sites, not the number of internal sites, and the abscissa is total strand length, not the portion of the strand which fell between two internal sites.

TABLE 3

Mean distance (μm) between initiation sites on the unbroken DNA fiber of 3 cell lines, estimated by 3 methods

Cell line	Method			
	A	B	C-High	C-Low
Muntjac	23	21	20	8
MDBK	45	40	32	22
Mouse L	63	37	28	14

A Total length of all strands/total number of initiation sites in complete samples of strands ≥ 50 μm long.

B Mean inter-initiation distance, weighting each inter-initiation distance equally, in complete samples of strands ≥ 50 μm long.

C High estimate, Appendix, section (i).

C Low estimate, Appendix, section (i).

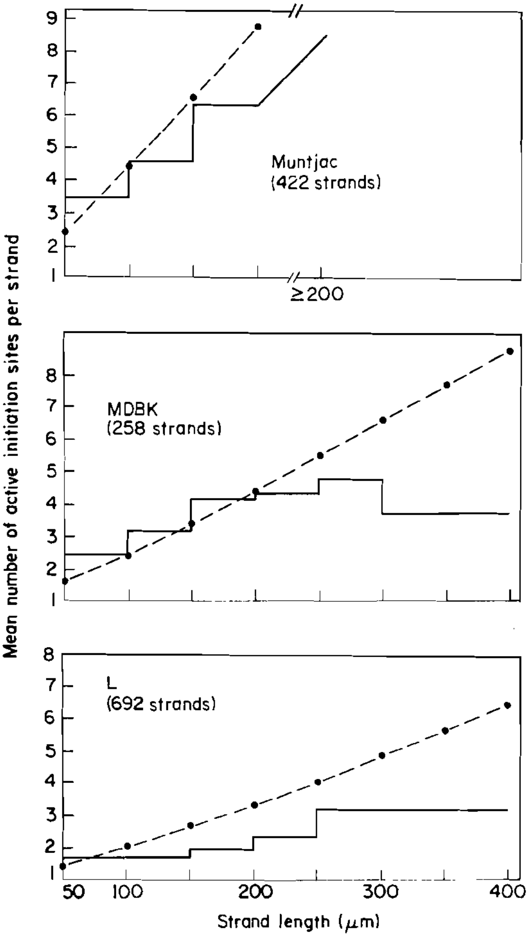

FIG. 2. Relationship between the mean number of initiation sites per strand and strand length in complete samples. The experimental curve (———————) represents strands 50 to 400 μm long from among strands used in Fig. 1. Each horizontal bar represents the mean number of initiation sites on at least 15 strands. The theoretical curve (– – – – –) was calculated by Appendix equation (4), with $\nu = 0{\cdot}044/\mu$m for muntjac, $0{\cdot}022/\mu$m for MDBK, $0{\cdot}016/\mu$m for L cells. These values are the reciprocals of the distances in column A of Table 3.

The predicted curves in Figure 2 agreed reasonably well with the observations for strands up to 200 μm long. For longer strands, especially in MDBK and L cells, systematically more sites were predicted than observed.

(d) *Relationship between inter-initiation distances and strand length*

In complete samples of strands from all three cell types, the mean inter-initiation distance increased with the length of strands (Fig. 3). For MDBK and muntjac cells, complete samples of slides from three experiments per cell line provided enough strands for each length category. For L cells, in addition to complete samples of strands from three experiments, strands at least 154 μm long from a slide of a fourth experiment were included without altering the effects observed.

For strands of a given length, muntjac cells had the shortest mean inter-initiation distances, while those of MDBK and L cells were very similar.

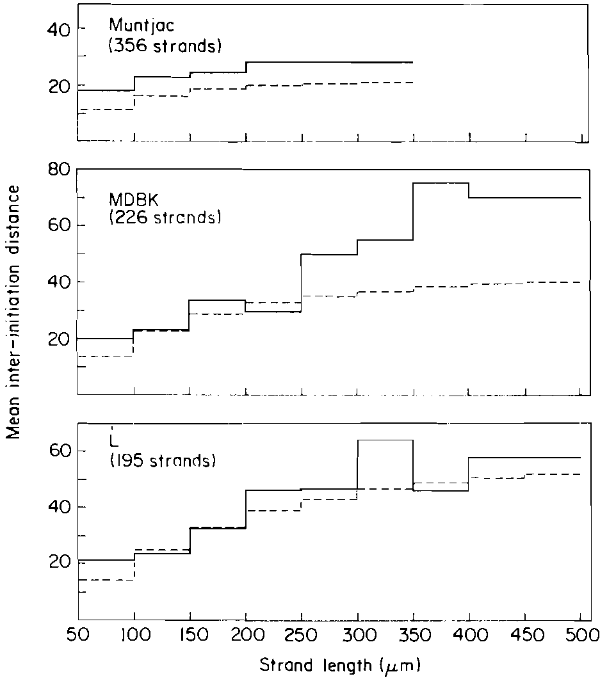

Fig. 3. Relationship between inter-initiation distance and strand length in complete samples. Each inter-initiation distance was weighted equally. At least 3 experiments are represented for each cell line. Each experimental (———) horizontal bar represents the mean of 47 to 522 inter-initiation distances for muntjac; of 24 to 147 distances for MDBK; and of 16 to 98 distances for L cells. The theoretical curve (-----) was calculated by Appendix equation (5), with ν as in Fig. 2.

If sites occurred at random on strands of any length, the predicted relationship between strand length and the average inter-initiation distance would be as given by equation (5) in the Appendix. When the predicted curves were calculated using each estimate of ν from Table 3, method (A) gave the closest agreement between predictions and observations. The predicted mean inter-initiation distance appeared to be systematically too low for the longer strands from MDBK and L cells, and too low for all strands from muntjac cells.

The distributions of inter-initiation distances within strand length categories (Figs 4 to 6) showed that the major difference among the length categories was in the appearance of long inter-initiation distances as strand length increased. This was to be expected as no inter-initiation distance could be longer than the strand on which it was observed. The modal inter-initiation distances (5 to 15 μm) remained nearly the same over a wide range of strand lengths and in all three cell lines. The differences in the mean inter-initiation distances in relation to strand length observed among muntjac, MDBK and L cells (Fig. 3) were evident in the distributions of inter-initiation distances. For example, on strands 250 to 350 μm long, 11·1% of the inter-initiation distances in muntjac cells were greater than 50 μm. In L cells and MDBK cells, 34·6% and 35·0% of the inter-initiation distances, respectively, were greater than 50 μm.

Thus the mean inter-initiation distance in strands of a given length changes with that length. It follows that the pattern of inter-initiation distances pooled from a

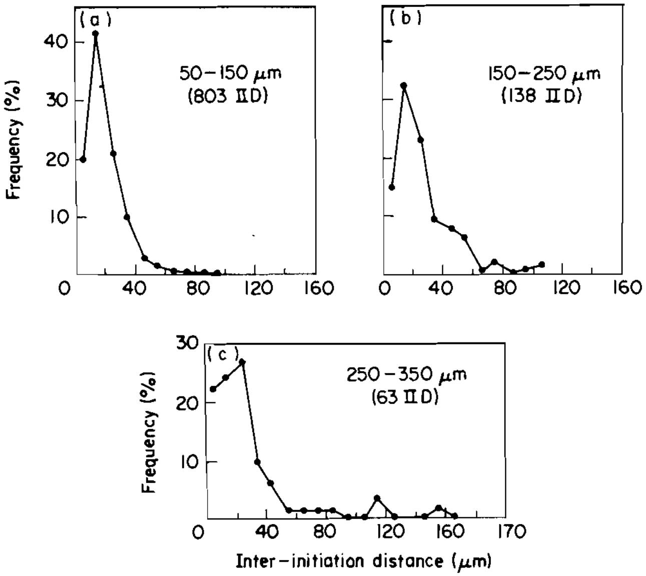

Fig. 4. Frequency distribution of inter-initiation distances in 3 strand-length categories in complete samples of muntjac cells. Frequencies are plotted at the midpoints of 10-μm intervals. IID, Inter-initiation distances.

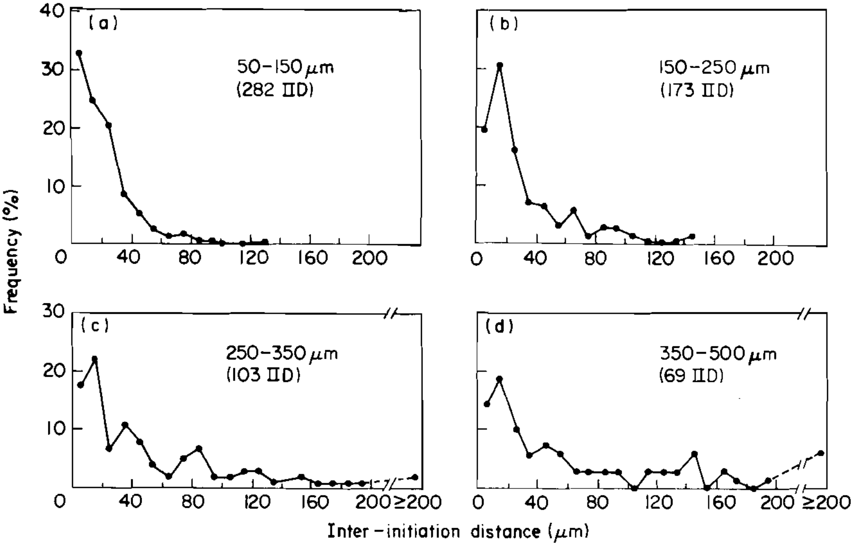

Fig. 5. Frequency distribution of inter-initiation distances in 4 strand-length categories in complete samples of MDBK cells. Frequencies are plotted at the midpoints of 10-μm intervals. IID, Inter-initiation distances.

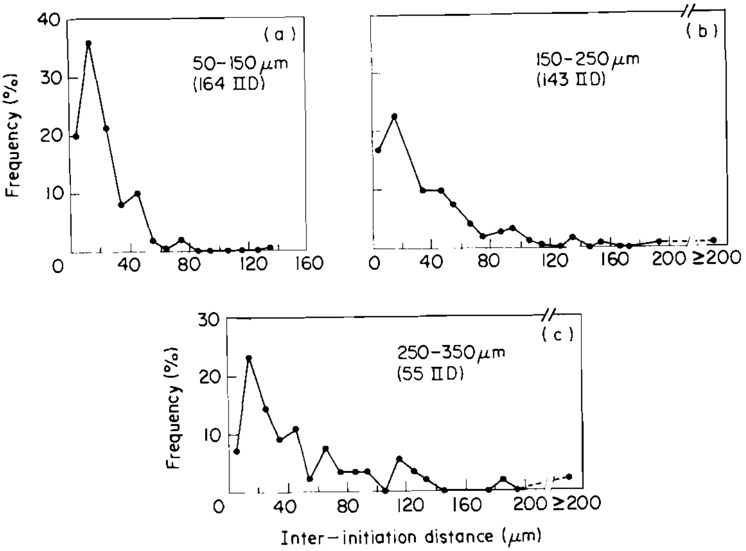

FIG. 6. Frequency distribution of inter-initiation distances in 3 strand-length categories in complete samples of L cells. Frequencies are plotted at the midpoint of 10-μm intervals. IID, Inter-initiation distances.

collection of strands of varying length would strongly depend on the distribution of length of those strands.

To determine whether sites were randomly placed on the fiber, regardless of other deviations from the random model, a different kind of information was required.

(e) *Spatial organization*

In order to study as nearly as possible the pattern of inter-initiation distances on the unbroken fiber, strands with many sites were selected. The rationale for this choice was that the pattern, random or non-random, of distances between sites within a strand could not have been altered by breakage of the fiber. Strands were sought with enough inter-initiation distances to offer a reasonable possibility of detecting a deviation from randomness. The choice of at least nine inter-initiation distances, or at least ten internal sites, is explained in the Appendix, section (g).

Tests of exponentiality were applied to each strand individually. Within each cell line and for each test, the *P* values were pooled (as described in Materials and Methods, section (g)) to obtain an overall *P* value for each cell line and each test (Table 4).

In the calculations reported here, statistically significant gaps in labeling (excluding those in prepulse figures) were not included in measurements of inter-initiation distances. For example, if a lightly labeled region of 20 μm abutted a blank region of 8 μm, which in turn abutted a colinear lightly labeled region of 30 μm, the length of DNA was taken as 20 + 30 = 50 μm, on the assumption that the 20 μm and 30 μm portions belonged to a single fragment and were separated by the spreading procedure. Only portions of DNA that were in perfect visual alignment were added in this way. A parallel calculation, not reported here (Jasny, 1978), has been carried out on the

TABLE 4

Distribution of inter-initiation distances in 3 cell lines (FdUrd, 2×10^{-6} M, 0·5 h). Two-tailed tests of exponentiality. Gaps excluded

Cells	A. Muntjac	B. MDBK	C. L
Number of strands	22	24	20
Strand length (μm)			
Mean	504·7	588·5	512·9
Maximum	1114·1	1191·0	1319·3
Minimum	115·4	119·7	136·8
Number of IID† per strand			
Mean	19	15	16
Maximum	38	27	52
Minimum	9	9	9
Mean IID per strand (μm):			
Grand mean (weighting each strand equally)	24·6	33·5	29·1
Maximum	46·7	85·5	92·5
Minimum	10·1	9·6	11·8
Standard deviation of IID per strand:			
Mean over strands	22·6	40·1	24·6
Maximum	64·6	144·5	138·3
Minimum	5·5	5·1	5·6
Minimum IID per strand (μm)			
Mean over strands	4·6	5·0	5·4
Maximum	9·4	9·1	10·3
Minimum	3·4	3·4	3·4
Durbin Kolmogorov–Smirnov test			
$P\ \alpha = 0$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
$P\ \alpha$ est.	$0·01 < P < 0·025$	$<10^{-4}$	$0·2 < P < 0·3$
Weibull test			
P	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Shape parameter β			
Maximum	2·87	3·05	3·83
Minimum	0·99	0·56	0·97
Median IID (mean over strands) (μm)	25·0	26·8	29·8
Keiding test			
$P\ \alpha = 0$	$0·0001 < P < 0·0005$	$<10^{-4}$	$<10^{-4}$
$P\ \alpha$ est.	$P = 0·24$	$P = 0·06$	$P = 0·39$

† IID, Inter-initiation distance.
est., estimated.

opposite assumption that the statistically significant gaps represented stretches of DNA. In the example, the length of DNA would have been taken as $20 + 8 + 30 = 58$ μm. In no case did the exclusion or inclusion of gaps affect the conclusions, which were as follows.

The Kolmogorov–Smirnov test ($\alpha = 0$), the Weibull test which estimated α, and the Keiding test ($\alpha = 0$) all clearly indicated a non-random distribution of inter-initiation distances in all three cell lines. The probability that the observed distances were randomly distributed was between 1 in 10,000 and 5 in 10,000 as measured by the Keiding test ($\alpha = 0$) in muntjac cells and less than 1 in 10,000 for the other cell types. The probability of randomness was less than 1 in 10,000 by the Kolmogorov–Smirnov test ($\alpha = 0$) and by the Weibull test for all three cell types.

The shortest inter-initiation distance on each strand was taken as the location parameter α. Table 4 gives the maximum, the minimum, and the average of α over all many-site strands in each cell line.

The Kolmogorov–Smirnov test when α was estimated did not reveal a significant deviation from randomness (exponentiality) in the muntjac or L cell lines. Similarly, the Keiding test with α estimated did not detect randomness in any of the three cell lines. Thus these tests were unable to detect non-randomness in the excess of inter-initiation distances over the minimum, on the order of 5 to 10 μm.

The failure of the Kolmogorov–Smirnov and Keiding tests to detect non-randomness when α was estimated did not suggest, however, that inter-intiation distances would be entirely random except for the shortest distances. The Weibull test also estimated α and gave a clear indication of the deviation from randomness in all three cell lines. The difference between the Weibull and the other two tests with α estimated probably lay in the greater sensitivity of the Weibull test to a specific *form* of deviation from randomness.

The shape parameter (β) of the Weibull distribution fitted to each strand ranged from less than one, or approximately equal to one, to greater than one, for all cell lines. When inter-initiation distances were clustered (β greater than 1), they clustered around a value approximately equal to the median. The means of the median inter-initiation distances for the many-site strands were 27 μm for MDBK cells, 30 μm for L cells, and 25 μm for muntjac cells (gaps excluded).

The range of the shape parameter suggested two types of organization (Fig. 7). On strands with high β values, the inter-initiation distances were short and evenly spaced. The strands with low β values had some very long inter-initiation distances, sometimes seeming to separate small clusters of shorter inter-initiation distances. Most of the strands observed appeared to be a mixture of the two types.

To determine whether short inter-initiation distances were significantly clustered together, the ordering of short and long inter-initiation distances on individual many-site MDBK strands was tested for randomness by the runs test (Sokal & Rohlf, 1969) and by a test of O'Brien (1976) and the resulting P values were pooled as in

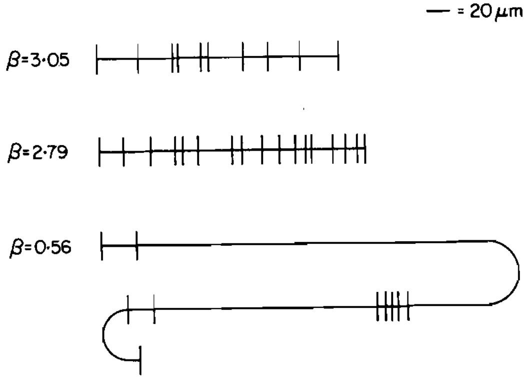

FIG. 7. Examples of initiation site distributions on MDBK strands with high or low Weibull shape parameters (β). The vertical lines represent the sites of initiation along the strand.

Materials and Methods, section (g). Perhaps as a result of the small number of inter-initiation distances on each strand, no statistically significant indication of non-randomness appeared. Consequently the tests were not extended to the muntjac and L cell lines.

(f) *Effect of FdUrd on initiation events in MDBK cells*

Previous work (Painter & Schaefer, 1971; Amaldi *et al.*, 1972; Taylor, 1977) indicated that synchronization of cells by FdUrd might markedly change the organization of sites. FdUrd inhibits thymidylate synthetase. It has been suggested that potential initiations are accumulated in the presence of FdUrd. MDBK cells were chosen for analysis of this suggestion.

Before attempting to synchronize cells with FdUrd, it was desirable to re-evaluate the possible effect of the 0·5-hour exposure to 2×10^{-6} M-FdUrd, which was generally used before labeling to deplete the thymidylate pool and to give clear autoradiograms. Although it was more difficult to distinguish hot-pulse labeled figures without prior depletion of the thymidylate pool, a sample of many-site strands was examined from MDBK cells that had never been exposed to FdUrd (Table 5A). The strand length, mean inter-initiation distance, and number of inter-initiation distances per strand were comparable to those of the DNA strands from FdUrd-pretreated cells (Table 5B). The Durbin Kolmogorov–Smirnov tests and the Weibull test indicated that there was less than a 10^{-4} probability of the distribution of inter-initiation distances being random. The Keiding tests detected no significant non-randomness, since the mean inter-initiation distance was close to the standard deviation. However, the clear results of the other tests confirmed that the inter-initiation distances in MDBK cells were distributed non-randomly along the fiber, regardless of a brief treatment with FdUrd.

Previous studies have indicated that when exponentially growing MDBK cells were treated with 5×10^{-7} M-FdUrd for 16 hours, 80 to 90% of the cells were in S phase upon release.

When we synchronized MDBK cells by exposure to 5×10^{-7} M-FdUrd for 16 hours, the strand length, the mean inter-initiation distance, and the mean number of inter-initiation distances per strand were similar to those seen in the untreated or briefly treated cells (Table 5C). However, the Kolmogorov–Smirnov test ($\alpha = 0$) indicated that the probability of randomness had increased more than tenfold. The probability of the distribution of inter-initiation distances being random increased 100-fold in the Weibull test. According to the Keiding test with α estimated, the apparent randomness of inter-initiation distances increased as the duration of exposure to FdUrd increased from none to 0·5 hour to 16 hours. No similar trends appeared in the Kolmogorov–Smirnov test with α estimated or in the Keiding test with $\alpha = 0$. It can be concluded that while the inter-initiation distances of synchronized cells still may have been non-randomly distributed, the distribution was markedly closer to random than in untreated MDBK cells.

This effect of synchronization was not observable in the frequency distribution of the inter-initiation distances from these many-site strands (Fig. 8). The overall distribution of inter-initiation distances was the same whether FdUrd was used to synchronize, used briefly to deplete the thymidylate pool, or not used at all. It was the pattern within the regions of DNA fiber represented by many-site strands that was changing, and the overall mean or modal inter-initiation distances were not sensitive to certain alterations in the process of replication.

TABLE 5

Effect of FdUrd on inter-initiation distances in MDBK cells. Two-tailed tests of exponentiality. Gaps excluded

FdUrd	A None	B 2×10^{-6} M 0·5 h	C 5×10^{-7} M 16 h
Number of strands	27	24	19
Strand length (μm):			
Mean	655·9	588·5	663·0
Maximum	1289·3	1191·0	1376·5
Minimum	148·8	119·7	140·2
Number of IID† per strand:			
Mean	15	15	15
Maximum	27	27	42
Minimum	9	9	9
Mean IID per strand (μm):			
Grand mean (weighting each strand equally)	41·7	33·5	39·6
Maximum	134·4	85·5	81·2
Minimum	11·1	9·6	13·5
Standard deviation of IID per strand:			
Mean over strands	42·1	40·1	37·8
Maximum	143·5	144·5	74·7
Minimum	7·2	5·1	8·0
Minimum IID per strand (μm):			
Mean over strands	6·3	5·0	6·5
Maximum	11·6	9·1	15·9
Minimum	3·4	3·4	3·4
Durbin Kolmogorov–Smirnov test			
P $\alpha = 0$	$<10^{-4}$	$<10^{-4}$	$0·001 < P < 0·005$
P α est.	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Weibull test			
P	$<10^{-4}$	$<10^{-4}$	$0·005 < P < 0·01$
Shape parameter β			
Maximum	2·68	3·05	2·02
Minimum	0·79	0·56	0·75
Median IID (mean over strands) (μm)	37·7	26·8	35·7
Keiding test			
P $\alpha = 0$	$0·4 < P < 0·5$	$<10^{-4}$	$0·1 < P < 0·2$
P α est.	$P = 0·03$	$P = 0·06$	$P = 0·43$

† IID, Inter-initiation distance.

(g) *Mean distance between sites on the DNA fiber*

In Table 3, estimates (A) and (B) of the mean distance between sites on the DNA fiber neglected strands less than 50 μm long and fragments which contained no sites.

To consider the range of possible effects of strands less than 50 μm long on the estimated mean distance between sites (Appendix, section (i)), the breakage of the

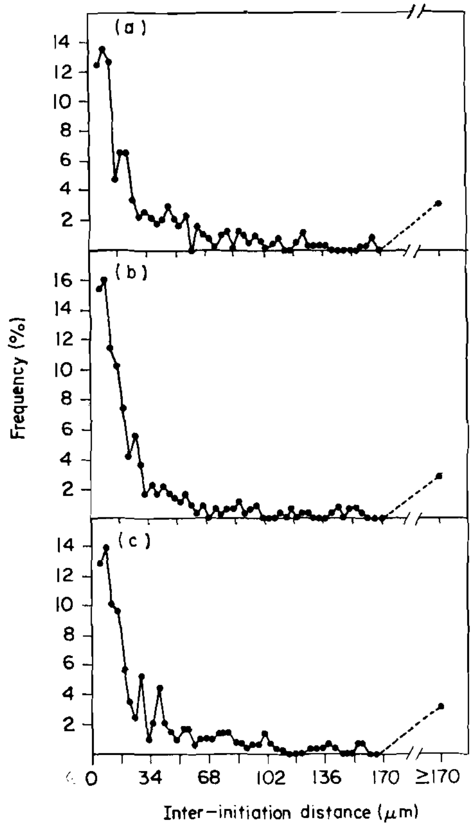

FIG. 8. The effect of FdUrd on the frequency distribution of inter-initiation distances in MDBK cells. Strands containing 9 or more inter-initiation distances were examined. Frequencies are plotted at the midpoints of the intervals. (a) No FdUrd, 386 inter-initiation distances; (b) 2×10^{-6} M-FdUrd, 0·5 h, 364 inter-initiation distances; (c) 5×10^{-7} M-FdUrd, 16 h, 289 inter-initiation distances.

fiber into strands was taken as approximately exponential. For strands at least 50 μm long, the mean inter-initiation distance for strands of a given length was taken to be the mean calculated from the data (Fig. 3). In method (C-high), the mean inter-initiation distance on strands less than 50 μm long was taken as the smallest mean inter-initiation distance in any observed length category; in method (C-low) it was taken as zero. Mean inter-initiation distances were estimated by approximating the distribution of distances on the fiber.

The estimates in Table 3 were subject to two countervailing biases. The exclusion of fragments of DNA fiber without initiation sites from "complete" samples tended to lower the estimates below the true value in each line. The potential fusion of sites very close together as a result of the 10 minutes of hot labeling eliminated very small inter-initiation distances and may have raised the estimates above the true value. The range of estimates in Table 3 indicated, at a minimum, the uncertainty which should be attached to these numbers.

The method leading to equation (1c) in the Appendix escaped both of these biases.

It depends on the mean length of siteless fragments and of strands (with at least one site), rather than on measured distances between sites. Since length measurements were least reliable for very short strands or fragments, the strands and siteless fragments could be chosen from the range where length measurements were reliable. However, this method does use all the assumptions of randomness in the model.

For L cells, the mean length of siteless fragments (from Table 1) was $H_0 = 76{\cdot}9\ \mu$m. The mean length of the strands included in Figure 1 was $H_1 = 121{\cdot}8\ \mu$m. For both siteless fragments and strands the minimum lengths included in these samples were $\theta_0 = \theta_1 = 50\ \mu$m.

Appendix equation (1c) leads to $\nu = (76{\cdot}9—50)^{-1} — (121{\cdot}8—50)^{-1} = 0{\cdot}0233$ per μm or an estimated mean inter-initiation distance between sites on the fiber of 43 μm. This estimate falls in the middle of the range of estimates for L cells in Table 3, suggesting that, at least for L cells, the two biases above roughly cancelled.

4. Discussion

A diploid set of chromosomes of a typical mammalian cell contains 1·5 to 2 m of DNA. A pair of replication forks travelling bidirectionally from a single origin of replication, or activated initiation site, can replicate on average 1 μm of DNA per minute, or approximately 0·5 mm of DNA in the eight hours of a typical synthesis (S) phase of a mammalian cell cycle. Replication of the entire genome therefore requires the concurrent activation of thousands of initiation sites distributed throughout the genome. Conceivably these sites could be placed at regular intervals along the DNA fiber, distributed at random, or organized in some other way, e.g. in clusters.

From statistical analyses of light microscopic DNA fiber autoradiograms in three mammalian cell lines (mouse L cells, muntjac, and bovine kidney cells), we infer that the initiation sites observable by our protocol are neither regularly nor randomly placed, but may be spatially organized into co-ordinately activated clusters. Temporal aspects of this co-ordination will be reported elsewhere (Jasny & Tamm, manuscript in preparation).

The spatial pattern which we propose as typical of initiation sites activated in any short interval of time is suggested by the three MDBK strands in Figure 7. This pattern consists of clusters of up to 40 sites with up to 25 μm between sites. The large clusters observed would be comparable in size to those observed by Willard (1977) as independently replicating subregions of chromosome bands. The longest inter-initiation distance we observed was 432 μm, close to the maximum distance two replicating forks moving in opposite directions could travel in eight hours. An inter-cluster region might be replicated at a different time, and possibly with a different organization than the clusters at its borders.

This proposed pattern permits the following interpretation of site distributions such as those in Figure 7. When two breaks defining the ends of a strand fell within a single cluster, sites appeared to be more regularly spaced than expected assuming randomness. Inter-initiation distances were distributed more closely around the median than expected from an exponential distribution and were better fitted by a Weibull distribution with shape parameter β significantly greater than one. When, on the contrary, two breaks defining the ends of a strand fell in different clusters, at least one inter-cluster region of DNA without sites appeared in the strand. There were more extremely short inter-initiation distances and more extremely long inter-initiation distances observed than expected from an exponential distribution with a mean equal to the mean inter-initiation distance on that strand. In this case, inter-initiation distances were better fitted by a Weibull distribution with β significantly less than one.

The observational, theoretical and statistical procedures reported here were required to demonstrate that the pattern suggested by Figure 7 represented the organization of sites on the DNA fiber and was not the result of selecting a few striking or atypical fragments. A theoretical model was developed to predict the properties of the spatial arrangement of initiation sites assuming randomness. These predictions were tested against experimental data.

The theoretical model assumed that (1) breaks in the DNA fiber were determined by a Poisson process; (2) sites were placed on the strand by another Poisson process; (3) these two Poisson processes were independent.

As predicted, the distribution of length of fragments without sites was approximately exponential (Table 1). This exponential distribution of length, based on direct optical measurements, differed markedly from the breakage into fragments of equal length inferred to occur under defined experimental conditions by Hershey & Burgi (1960).

By contrast with siteless fragments, the distribution of length of strands (containing at least 1 site) was not very well described by the distribution derived from the random model (Fig. 1). The presence of sites appeared to be associated with deviations from the random model.

We have no evidence that the significant difference between cell lines in mean strand length (93 μm for muntjac, 122 μm for L, 150 μm for MDBK) were due to differences in the spreading procedure. Several comparisons of strand length revealed no difference between slides within cell lines. The rank order of the three cell lines according to mean strand length was the same as their rank order by one estimate of mean inter-initiation distance (Table 3B). This suggestion of an association between inter-initiation distances and mean strand length was consistent with the earlier demonstration (Jasny *et al.*, 1978) that there were far more breaks adjacent to hot-labeled regions of the fiber than predicted by the random model. These breaks may be due to spreading, to radiation damage, or to prepulse figures in which large gaps preclude recognition of both halves.

The observed geometric distribution of the number of sites per strand (Table 2) was predicted (Appendix, section (d)) by assuming that sites were random and independent of breaks. This distribution was evidently not sensitive to the deviations from the assumptions of the random model revealed by other measures.

The mean number of sites per strand and the mean inter-initiation distance increased with the length of the strands measured, according to both observation (Figs 2 and 3) and theory (Appendix, sections (e) and (f)). Thus the sample mean inter-initiation distance given without reference to the distribution of strand length, which has been used in the past as an index of differences between cell types, does not adequately characterize mean replication unit size.

In all three cell lines, the observed increase of mean sites per strand and of mean inter-initiation distance with increasing strand length differed from the relationships predicted by the random model. The long strands, especially in MDBK and L cells, had markedly fewer sites on average than predicted, even though the short strands agreed fairly well with the random theory. This is compatible with the possibility that the long strands were more likely to contain a long region without sites. Such a region might correspond to the proposed inter-cluster region. Similarly, the mean inter-initiation distance on strands of given length appeared to rise with increasing strand length more rapidly than predicted by the random model. This discrepancy may again be interpreted as evidence of the appearance in long strands of inter-cluster regions without sites.

On strands of a given length, muntjac cells appeared to have more sites and lower mean inter-initiation distances than L and MDBK cells. The rate of progression of muntjac replication forks was approximately half that of MDBK and L cells, and the S phases of the three cell lines were comparable (Jasny, 1978).

In complete samples of strands, the modal inter-initiation distance was similar (5 to 15 μm) in all three cell lines over a wide range of strand lengths (Figs 4 to 6). Some underlying feature of chromatin or DNA structure may favor placement of initiation sites at 10-μm intervals, or there may be fusion of smaller replication units during the 10-minute hot pulse.

The tests reviewed so far of the assumption that inter-initiation distances on the fiber are random were subject to influence by the interaction between the location of sites and the location of breaks. As the pattern of inter-initiation distances within a strand could not have been altered by breakage of the fiber, strands with at least nine inter-initiation distances were selected to study as nearly as possible the pattern of inter-initiation distances on the unbroken fiber. Tests of exponentiality applied to each strand individually gave two-tailed P values describing the probability that a worse fit to an exponential distribution would have occurred by chance. Treating each strand as an independent test of exponentiality, we combined these P values to estimate the overall significance of deviations from exponentiality for each of five tests and for each cell line.

Three tests clearly indicated a non-random distribution of inter-initiation distances in all three cell lines ($P < 0{\cdot}0005$). Two other tests were unable to detect non-randomness in the excess of inter-initiation distances over the minimum, which was of the order of 5 to 10 μm. This minimum is approximately the distance that a pair of replication forks moving in opposite directions would travel during the 10-minute period of hot-pulse labeling. Any activated sites less than 10 μm distant could conceivably have been fused as a result of the labeling protocol. However, one of the three tests which clearly indicated non-randomness was based entirely on the excess of inter-initiation distances over the minimum estimated for each strand. This test was probably the most sensitive to deviations from exponentiality of specific (Weibull) form. The failure of two tests to detect non-randomness in the excess of inter-initiation distances over the minimum is interpreted as due to the relative insensitivity of those tests to a specific pattern of deviation from exponentiality.

When MDBK cells were exposed for 16 hours to 5×10^{-7} M-FdUrd, the frequency distribution of inter-initiation distances shifted toward exponentiality according to four of the five tests and registered no change according to the remaining test. Yet the minimum, mode and average inter-initiation distances on the many-site strands treated with FdUrd did not appear to change when compared to the same variables on many-site strands treated briefly or not at all with FdUrd. By contrast, Taylor (1977) observed a striking decrease in inter-initiation distances in CHO cells using a much higher concentration of FdUrd.

The sensitivity to perturbations offered by tests of the exponentiality of inter-initiation distances may be useful in future efforts to detect subtle changes in the pattern of organization of initiation sites that may occur in transformed cells.

As a result of fiber breakage, no strands observed in these studies were more than 1·4 mm long. Inter-initiation distances longer than 1·4 mm could not have been detected by our experimental procedure. Our results describe those portions of the fiber where there was at least some activity of initiation sites during our labeling period. Because the activation is observed of only those replication units which can be detected with a 10-minute hot pulse, the mean measured inter-initiation distance probably exceeds the mean distance between adjacent origins of replication which would be obtained by following a given strand throughout S phase.

Subject to those limitations, we have estimated the mean distance between sites by several methods (Table 3). One method (C) is based on the harmonic mean of inter-initiation distances obtained from specific strand-length categories, weighted by a length-weighted exponential distribution. This method is explicitly recognized as likely to provide a lower bound on the true mean inter-initiation distance on the fiber, and does in fact give estimates which are lower than those obtained by the other methods. The difference between the other two methods (A and B) lies essentially in the treatment of sites and regions of DNA at the ends of strands. These sites and regions are taken into consideration in method (A) and are excluded when there is ambiguity in the estimation of an inter-initiation distance in method (B). Both methods (A) and (B) rely essentially on the ratio of length of DNA to number of sites occurring in that length. Both give estimates of the mean inter-initiation distance on the fiber which differ by less than a factor of two within any cell line. For L cells, these estimates are approximated by yet another method of estimation (Appendix, section (c)) which does, however, rest on the assumptions of the random model.

For the purpose of estimating the mean inter-initiation distance on the fiber, we would consider it prudent not to claim more precision for our observational and analytical methods than is suggested by the range of estimates for each cell line in Table 3. This same uncertainty does not attach to our conclusions concerning the non-random patterning of inter-initiation distances on the fiber.

APPENDIX

This Appendix describes the assumptions, and derives some predictions, of a model for the spatially random distribution of initiation sites. We use the terminology defined in Materials and Methods, section (d), in the main text.

We treated length as a continuous variable, rather than measuring it discretely in numbers of nucleotides, for two reasons. First, the actual measurements were made in units of length. Second, the mathematical theory was much easier when length was treated as a continuous variable. Taking length as a continuous variable was valid, because there are 3000 base-pairs per μm and the lower limit of resolution of our optical observations exceeded 1 μm.

(a) *The model*

As in Jasny *et al.* (1978), our model assumes that the average length of a strand is much less than the length of the fiber. Our data show that average strand length is several orders of magnitude smaller than the estimated length of DNA in the genome or in a single chromosome. If the points at which the fiber breaks to produce strands are determined by a purely random process known as a Poisson process (Parzen, 1962), then a break is as likely to occur at any point in the fiber as it is at any other, regardless of the distance to the nearest break. Let the mean distance between breaks in the fiber be $1/\lambda$. The dimension of λ is length^{-1}. Then the probability density function of the distance between breaks is exponential with parameter λ.

We specify the location of each site by a single point. The length of the prepulse or postpulse figure which identifies a site is irrelevant. Only the location of the origin of replication is considered. The model of randomness assumes that the sites are distributed on the fiber by a Poisson process with parameter ν (dimension: length^{-1}). Consequently, on the fiber prior to breakage into fragments, the probability density function of the distance between sites is exponential with parameter ν; the number of sites per unit length of the fiber is Poisson distributed with parameter ν.

The model further assumes that the random process which determines the distribution of sites is independent of the random process which determines the distribution of breaks.

To summarize, the model supposes that (1) breaks are located on a fiber of effectively infinite length by a Poisson process with parameter λ; (2) sites are located on the fiber by a Poisson process with parameter ν; (3) the two Poisson processes are independent. Finally, as described in Materials and Methods, section (e), fragments with no hot-pulse labeled regions were not considered strands. Thus (4) we observed all strands with one or more sites. The effects of including only such strands have been considered in each derivation.

(b) *Notation*

Let L be a random variable describing the length of a strand, $L \geq 0$; U be a random variable describing the number of sites (or units) on a strand, $U = 1, 2, \ldots,$ and I be a random variable describing an inter-initiation distance, that is a distance between two adjacent sites on a strand, $I \geq 0$. We shall use the conventional abbreviation *pdf* for probability density function, and $P(\quad)$ for the probability of the event described in parenthesis. Thus $P(L \geq h)$ is the probability according to the model that a strand (with at least 1 site) is greater than or equal to h μm long; $P(U = k|L = h)$ is the conditional probability, given that a strand is h μm long, that exactly k sites are present on the strand, $k = 1, 2, \ldots$. We shall use the notation $E(\quad)$ for the mean or average or expected value of the random variable enclosed in parenthesis.

(c) *Length of strands*

The probability density function (*pdf*) of the distance h between adjacent breaks in the fiber is $\lambda e^{-\lambda h}$. The distance between two adjacent breaks corresponds to a strand length $L = h$ if and only if there is at least one site between the breaks. There is at least one site in an interval of length h on the fiber with probability $1 - e^{-\nu h}$ according to the Poisson distribution. Since the location of breaks and the location of sites are independent, the *pdf* of $L = h$ is proportional to the product $\lambda e^{-\lambda h}(1 - e^{-\nu h})$. Since the integral of the *pdf* from $L = 0$ to infinity is 1, we have $pdf(L = h) = \lambda(1 + \lambda/\nu)e^{-\lambda h}(1 - e^{-\nu h})$. Therefore the cumulative distribution function (*cdf*) $F(h)$ of strand length, which is the probability that $L \leq h$, is given by

$$1 - F(h) = (1 + \lambda/\nu)e^{-\lambda h} - (\lambda/\nu)e^{-(\nu+\lambda)h}, \; h \geq 0. \tag{1a}$$

If only strands greater than or equal to some threshold $\theta > 0$ are considered, then the conditional *cdf* $F_\theta(h)$ of those strands is given by $1 - F_\theta(h) = (1 - \mathrm{F}(h))/(1 - \mathrm{F}(\theta))$, $h \geq \theta$.

Equation (1a) is not an exponential distribution because two breaks which are close together on the fiber are less likely to have a site between them, and hence less likely to create a strand in our samples. However, $F(h)$ becomes very nearly exponential when short strands are excluded.

Since the *pdf* of length h of a fragment (with or without sites) is $\lambda \mathrm{e}^{-\lambda h}$ and the probability of no site occurring on a fragment of length h is $\mathrm{e}^{-\nu h}$, the *pdf* of length of siteless fragments is

$$g(h) = (\nu + \lambda)\mathrm{e}^{-(\nu+\lambda)h}, \; h \geq 0, \tag{1b}$$

which is identical to the *pdf* of external segment length given by Jasny *et al.* (1978).

Suppose that only strands (with at least one site) of length greater than or equal to some threshold $\theta_1 \gg 1/\nu$ are observed, so that by equation (1a) their length distribution is approximately exponential with parameter λ. Let H_1 be the mean length of these strands; approximately, $H_1 - \theta_1 = 1/\lambda$. Suppose also that a sample of siteless fragments longer than some threshold $\theta_0 \geq 0$ is observed, with sample mean length H_0. Then by equation (1b), $H_0 - \theta_0 = 1/(\nu + \lambda)$. Combining these two equations gives

$$\nu = 1/(H_0 - \theta_0) - 1/(H_1 - \theta_1). \tag{1c}$$

Thus without measurement of any inter-initiation distances, equation (1c) gives a way of estimating the mean distance between sites on the unbroken fiber; the information required is the mean length of siteless fragments and the mean length of strands which are long enough ($\geq \theta_1$) to be approximately exponential in length.

(d) *Sites per strand*

On the fiber, the number of sites in an interval of length h is a Poisson-distributed random variable. Any interval of length h containing no sites is not observed as a strand, so the number of sites on a strand of length h has the 0-truncated Poisson distribution. That is, the probability that a strand contains exactly k sites, given that the strand is of length h, is

$$P(U = k|L = h) = \mathrm{e}^{-h\nu}(h\nu)^k/[k!(1 - \mathrm{e}^{-h\nu})]. \tag{2}$$

Consequently, the number of sites per strand in the entire population of strands has the *pdf*

$$P(U = k) = \int_{h=0}^{\infty} P(U = k|L = h) pdf(L = h)dh = (\lambda/\nu)(\nu/[\nu + \lambda])^k. \tag{3}$$

This is a 0-truncated geometric distribution. The mean number of sites per strand is

$$E(U) = \sum_{k=1}^{\infty} kP(U = k) = 1 + \nu/\lambda.$$

In a sample of strands, if the observed mean number of sites per strand is $\bar{U}$, a natural estimate of ν/λ is obtained by replacing $E(U)$ by $\bar{U}$. The result may then be substituted into equation (3) to generate the relative frequencies predicted by the 0-truncated geometric distribution for a complete sample of strands which include all lengths greater than or equal to 0.

The significance of this calculation is that even though the number of sites on strands of fixed length $L = h$ is 0-truncated Poisson distributed, the number of sites per strand after allowing for the random variation in strand length is 0-truncated geometrically distributed.

(e) *Mean number of sites on strands of given length*

From equation (2), the average number of sites on strands of length $L = h$ is

$$E(U|L = h) = \sum_{k=1}^{\infty} kP(U = k|L = h) = h\nu/(1 - e^{-h\nu}). \tag{4}$$

Thus the expected number of sites on strands of a given length is not strictly proportional to length, because only strands with one or more sites are observed. For strands of length significantly longer than $1/\nu$, $e^{-h\nu}$ approaches 0 and the expected number of sites becomes nearly proportional to length.

(f) *Mean inter-initiation distance on strands of given length*

If there are $k = 1, 2, \ldots$ sites on a strand of length $L = h$, then on that strand there are $k - 1$ inter-initiation distances between the adjacent sites. The expected length of each of these inter-initiation distances is $h/(k + 1)$. Therefore, weighting each inter-initiation distance equally, the mean inter-initiation distance on strands of length $L = h$ is

$$\begin{aligned} E(I|L = h) &= \sum_{k=1}^{\infty} (k - 1)(h/[k + 1])P(U = k|L = h) \\ &\quad \div \sum_{k=1}^{\infty} (k - 1)P(U = k|L = h) \\ &= (h\nu[1 + e^{-h\nu}] - 2[1 - e^{-h\nu}])/[\nu(h\nu - 1 + e^{-h\nu})]. \end{aligned} \tag{5}$$

As expected, $E(I|L = h) \rightarrow 1/\nu$ as $h \rightarrow \infty$.

(g) *Distribution of inter-initiation distances*

On a strand of length $L = h$ with exactly $k \geq 2$ sites, we wish to determine the distribution of the random variable I, which measures the distance between two adjacent sites. Let us measure distances along the strand starting from the left end as zero and proceeding up to h at the right end. Let $X_{(i)}$ be the position (i.e. distance from 0) of the ith site on the strand, $i = 1, \ldots, k$. With probability 1, the k sites divide the strand into $k + 1$ segments of positive length. As in Jasny *et al.* (1978), we refer to the first segment, from 0 to $X_{(1)}$, as the left external segment, and the last segment, from $X_{(k)}$ to h, as the right external segment. If $k > 1$, the strand also contains $k - 1$ inter-initiation distances of length $X_{(i+1)} - X_{(i)}$, $i = 1, \ldots, k - 1$.

According to our model, the locations of breaks and the locations of sites on the fiber are determined by independent Poisson processes. It may be shown that positions $X_{(i)}$ have exactly the distribution of the k order statistics of k independent and identically distributed uniform random variables on the interval $[0,h]$. If we define $X_{(0)} = 0$ and $X_{(k+1)} = h$, then for $i = 0, 1, \ldots, k$, and for any number t such that $0 \leq t \leq k$,

$$P(k(X_{(i+1)} - X_{(i)})/h > t) = P(I > ht/k) = (1 - t/k)^k.$$

Although this exact distribution of inter-initiation distances is not an exponential distribution, in the limit as the number k of sites on a strand gets large, for any fixed length $L = h$,

$$\lim_{k \to \infty} P(I > ht/k) = e^{-t}. \qquad (6)$$

Thus in the limit the lengths of the external segments and the inter-initiation distances are exponentially distributed with average length h/k. Moreover, these distances are also asymptotically independent. A derivation of these results is given by Feller (1966).

For practical applications it is essential to know how large k must be for the exponential distribution in equation (6) to provide a good approximation to the exact distribution. For nine inter-initiation distances, or $k = 10$ sites, the agreement between the exact distribution and the exponential limit is excellent (calculations not shown). Hence tests for exponentiality of inter-initiation distances were applied only to strands with at least nine inter-initiation distances.

(h) *Shortcomings of the model*

To investigate whether inter-initiation distances are random, as supposed in assumption (2) above, we introduced other assumptions in order to adapt to the way in which the data were collected. Since the analysis of data led us to reject the null hypothesis of randomness, it is important to consider whether our conclusions were affected by aspects of the manner of data collection which the model has not recognized.

One shortcoming of the model is that, contrary to assumption, sites are not points of length 0 but are the centers of hot-labeled regions at least 1·7 μm long. Some pre-pulse figures are more than an order of magnitude larger than the minimum. Suppose, as an average, that a single replication fork travels 0·5 μm per minute on the fiber. Suppose two sites were within 5 μm of each other, and both began to replicate at the beginning of the 10-minute hot pulse. By the end of the hot pulse, the replicated regions originating at the two sites would have fused. These regions would have been recorded, in our protocol, as a single site located between the two actual sites. The same fusion of replicated figures is possible with pre-pulse figures. Our model does not recognize the inability of the observational protocol to resolve sites which are close together.

We would expect to observe fewer very short inter-initiation distances than predicted by our model. However, this shortcoming of the model in relation to the observational protocol is not sufficient to explain why the distribution of inter-initiation distances is non-random for some cell lines, even when α is estimated, nor to explain why, as the Weibull and Keiding tests suggest, FdUrd treatment changes the non-randomness of inter-initiation distances longer than the estimated α in MDBK cells.

It would be highly desirable to remove this shortcoming of the model in the future, taking account of available information on the synchrony of initiation of adjacent sites, but we conclude that the non-randomness of sites which we find is not an artifact due to this shortcoming.

At a more fundamental level, if an origin of replication is an identifiable physical structure where, for example, a specific protein interacts with the fiber, the origin is likely to be several and perhaps hundreds of bases long. Such lengths are too short to affect our tests.

(i) *Estimating a lower bound for the mean distance between sites on the fiber*

The purpose of the following analysis is to gain information about the average distance between sites on the fiber. It is not possible to exploit the simple assumptions of our initial model to estimate the mean for several reasons. First, the locations of breaks are not independent of the locations of sites. We have no good theory of the relation between sites and break points. Second, we have excluded from observation fragments of the fiber with no sites. Third, it is not possible to estimate inter-initiation distances on those strands with only one site. Fourth, our data on inter-initiation distances are clearly not consistent with an exponential distribution.

Therefore we have attempted to construct an estimate which is as free as possible of unverified assumptions. As an empirical approximation, we take the distribution of length of strands at least 50 μm long as exponential. The result of our analysis is a lower bound on the mean inter-initiation distance on the fiber (that is, a number lower than the true mean) rather than an estimate of the mean itself.

If we ignore end effects, the mean inter-initiation distance on the fiber equals the total length of the fiber divided by the total number of inter-initiation distances on the fiber. We wish to estimate both the numerator and denominator. Given a sample of strands (from which fragments of the fiber with no sites are excluded), it seems likely that our estimate of the numerator, which excludes all external segments, will underestimate total fiber length by a larger proportion than our estimate of the denominator will underestimate the total number of inter-initiation distances. Hence our resulting ratio will probably underestimate the mean inter-initiation distance on the fiber.

Let the *pdf* of strand length L be $pdf(L = h) = f(h)$. The shortest fragment of the fiber which counts as a strand is 1·71 μm long. We take $f(h) = 0$ for $h < 1{\cdot}71$. Let $g(h) = E(I|L = h)$ be the mean inter-initiation distance on strands of exact length $L = h$. The shortest measurable inter-initiation distance is 3·42 μm. We assume $g(h) > 0$ for all $h \geq 3{\cdot}42$ μm, $g(h) = 0$ for $h < 3{\cdot}42$ μm and that the variance of inter-initiation distances on the fiber is finite. Then the expected number of inter-initiation distances on a strand of length h is approximately $h/g(h)$ (Cox, 1962). Consequently, in a large sample of n strands, the number of strands of length h is approximately $nf(h)$, and the total number of inter-initiation distances measured would be approximately

$$n\int_{h=3{\cdot}42}^{\infty} f(h)(h/g(h))dh. \tag{7}$$

In a large sample of n strands, the total length of strands of length h is approximately $nhf(h)$. Hence the total length of strands in the sample is approximately

$$n\int_{h=1{\cdot}71}^{\infty} hf(h)dh. \tag{8}$$

The ratio I^* of equation (8) divided by equation (7) is the harmonic mean of inter-initiation distances obtained from specific strand-length categories, weighted by a length-weighted *pdf* $f(h)$ of strand length. I^* is independent of sample size n.

We now describe how we have converted I^* into a practical computing formula based on the available data for a given cell line. Let H be the mean length of strands which are $\geq$ 50 μm long. Our data give H and show that the sample *pdf* of length $L = h$ (in μm) for these strands is of the order of magnitude of $\gamma e^{-\gamma(h-50)}$, where

$\gamma = (H - 50)^{-1}$. In order to estimate the proportion of strands between 1·71 μm and 50 μm long, we assume that in the entire population of strands $pdf(L = h) = \gamma e^{-\gamma(h-1\cdot71)}$, $h \geq 1\cdot71$.

To compute the average inter-initiation distance on strands as a function of length, the strands have been sorted into length categories, starting with the interval from 50 μm to 100 μm, 100 μm to 150 μm, and so on. We shall denote the boundaries of these length categories by $a_1 = 50$ μm, $a_2 = 100$ μm and so on up to $a_{c+1} = \infty$, where c is the number of categories. We introduce this notation because both boundaries and the number of the length categories varied from one cell line to another, depending on the available number of strands. The boundaries of the ith category are from a_i to a_{i+1}, including the lower boundary and excluding the upper. Let $a_0 = 1\cdot71$ and let g_i be the average inter-initiation distance on all strands in the ith category of length from a_i to a_{i+1}. Then g_0 is unknown. g_i, $i = 1, 2, \ldots, c$ are obtained from the data. Dropping n, and using the exponential $pdf(L = h)$ in place of $f(h)$ in equation (8), we find analytically that the numerator of I^* is $1\cdot71 + 1/\gamma = H - (50 - 1\cdot71)$. Again dropping n, we approximate the denominator equation (7) of I^* by

$$\int_{h=3\cdot42}^{\infty} (hf(h)/g(h))dh = \sum_{i=0}^{c} g_i^{-1} \int_{h=a_i}^{a_{i+1}} h\gamma\, e^{-\gamma(h-1\cdot71)}dh$$
$$= \sum_{i=0}^{c} g_i^{-1} [(a_i + 1/\gamma)e^{-\gamma(a_i-1\cdot71)} - (a_{i+1} + 1/\gamma)e^{-\gamma(a_{i+1}-1\cdot71)}]. \quad (9)$$

Given H, a_i, and g_i, $i = 1, 2, \ldots, c$, all the quantities required to compute I^* approximately are known except g_0, the mean inter-initiation distance on strands between 1·71 μm and 50 μm long. We observe that as i increases, g_i increases; that is, longer strands have longer inter-initiation distances. We infer that g_0 lies somewhere between 0 and g_1. Therefore we compute two values I_0^* and I_1^* of I^*, where I_0^* is obtained by taking $g_0 = 0$ and I_1^* by taking $g_0 = g_1$. Because the resulting numerical values (in Table 3) of I_0^* (C-low) and I_1^* (C-high) differ by at most 14 μm in the three cell lines, we infer that an exact knowledge of g_0 would make little difference to our estimate of I^* as a lower bound on the mean inter-initiation distance in the unbroken fiber.

This work was supported by Research grant CA-18608 and by Program Project grant CA-18213 awarded by the National Cancer Institute, and by National Science Foundation grant DEB74-13276. One of the authors (B. R. J.) was a predoctoral trainee under the Institutional National Research Service award CA-09256 from the National Cancer Institute.

REFERENCES

Amaldi, F., Carnevali, F., Leoni, L. & Mariotti, D. (1972). *Expt. Cell Res.* **74**, 367–374.

Bablanian, R., Eggers, H. J. & Tamm, I. (1965). *Virology*, **26**, 100–113.

Blumenthal, A. B., Kriegstein, H. J. & Hogness, D. S. (1973). *Cold Spring Harbor Symp. Quant. Biol.* **38**, 205–224.

Cairns, J. (1966). *J. Mol. Biol.* **15**, 372–373.

Callan, H. G. (1972). *Proc. Roy. Soc. ser. B*, **181**, 19–41.

Cox, D. R. (1962). In *Renewal Theory*, p. 49, Methuen & Co., London.

Durbin, J. (1975). *Biometrika*, **62**, 5–22.

Eagle, H. (1959). *Science*, **130**, 432–437.

Edenberg, H. J. & Huberman, J. A. (1975). *Annu. Rev. Genet.* **9**, 245–284.

Engelhardt, M. & Bain, L. J. (1973). *Technometrics*, **15**, 541–549.

Engelhardt, M. & Bain, L. J. (1975). *Technometrics*, **17**, 353–356.

Engelhardt, M. & Bain, L. J. (1977). *Technometrics*, **19**, 323–331.

Feller, W. (1966). In *An Introduction to Probability Theory and Its Applications*, vol. 2, p. 23, John Wiley & Sons, New York.

Fisher, R. A. (1970). *Statistical Methods for Research Workers*, 14th edit., p. 99, Oliver & Boyd, Edinburgh.

Ganner, E. & Evans, H. J. (1971). *Chromosoma (Berlin)*, **35**, 326–341.

Gavosto, F., Pegararo, L., Masera, P. & Rovera, G. (1968). *Expt. Cell Res.* **49**, 340–358.

Hand, R. (1975). *J. Cell Biol.* **64**, 89–97.

Hand, R. & Tamm, I. (1972). *Virology*, **47**, 331–337.

Hand, R. & Tamm, I. (1973). *J. Cell Biol.* **58**, 410–418.

Hand, R. & Tamm, I. (1974). In *Cell Cycle Controls* (Padilla, G. M., Cameron, J. L. & Zimmerman, A. M., eds), pp. 273–288, Academic Press, New York.

Hershey, A. D. & Burgi, E. (1960). *J. Mol. Biol.* **2**, 143–152.

Huberman, J. & Riggs, A. D. (1968). *J. Mol. Biol.* **32**, 327–341.

Jasny, B. R. (1978). Ph.D. thesis, The Rockefeller University, New York.

Jasny, B. R., Cohen, J. E. & Tamm, I. (1978). In *DNA Synthesis: Present and Future* (Kohiyama, M. & Molineaux, I., eds), pp. 175–183, Plenum Press, New York.

Johnson, N. L. & Kotz, S. (1970). In *Continuous Univariate Distributions*, vol. 1, p.222, Houghton Mifflin, Boston.

Keiding, N. (1977). *Am. Nat.* **111**, 1211–1219.

Kriegstein, H. J. & Hogness, D. S. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 135–139.

Latt, S. A. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 3395–3399.

Madin, S. H. & Darby, N. B. Jr (1958). *Proc. Soc. Expt. Biol. Med.* **98**, 574–576.

Newlon, C. S., Petes, T. D., Hereford, L. M. & Fangman, W. L. (1974). *Nature (London)*, **247**, 32–35.

O'Brien, P. C. (1976). *Biometrics*, **32**, 391–401.

Painter, R. B. & Schaefer, A. W. (1971). *J. Mol. Biol.* **58**, 289–295.

Parzen, E. (1962). *Stochastic Processes*, Holden-Day, San Francisco.

Sokal, R. R. & Rohlf, F. J. (1969). In *Biometry*, pp. 624–627, W. H. Freeman & Co., San Francisco.

Taylor, J. H. (1958). *Expt. Cell Res.* **15**, 350–357.

Taylor, J. H. (1977). *Chromosoma (Berlin)*, **62**, 291–300.

Willard, H. F. (1977). *Chromosoma (Berlin)*, **61**, 61–73.

Wolstenholme, D. R. (1973). *Chromosoma (Berlin)*, **43**, 1–18.

Wurster, D. H. & Benirschke, K. (1970). *Science*, **168**, 1364–1366.