A LIMIT THEOREM FOR THE REPLICATION TIME OF A DNA MOLECULE

by

Richard Cowan*, S.N. Chiu* and Lars Holst**

Download version of paper which appeared in *J. Appl. Prob.* **32**, 296-303 (1995).

Abstract: The DNA of higher animals replicates by an interesting mechanism. Enzymes recognise specific sites randomly scattered on the molecule and establish a bi-directional process of unwinding and replication from these sites. We investigate the limiting distribution of the completion time for this process by considering related coverage problems investigated by Janson (1983) and Hall (1988).

Keywords: DNA replication, limit theorem, asymptotics, coverage.
AMS Subject Classification: Primary 92C40
Secondary 60F05, 60K40

Introduction

The DNA molecule in higher animals is topologically linear. It is replicated by an interesting process which commences at specific sites called 'origins of replication'. These origins are randomly scattered along the linear molecule. Due to random evolution of the extremely long DNA molecule, these specific sites can be modelled well by a Poisson process on the real line.

Each site is recognised, after some random delay time, by an enzyme-complex which then binds to the site. The complex immediately initiates, via its influence on other enzymes, a bi-directional movement along the DNA. At this moving 'frontier', the helical structure unwinds and separates into two single strands. Replication of these single-stranded regions then takes place (see Kornberg (1980) for biological details).

In this paper we pose a stochastic model of this process. The Poisson process of origins has intensity γ. These sites must wait an iid exponentially distributed time with mean $1/\mu$ until the approach of the enzyme complex, the times being independent of the Poisson process. If a site has not yet been passed over by a moving frontier initiated elsewhere, molecular binding takes place and the two bidirectional frontiers commence at the site. The frontiers move at constant speed r in each direction. When frontiers meet, they stop and the enzymes involved in unwinding fall off the DNA. Let T_L be the time when an interval of length L, say $[0, L]$, is completely unwound. In this paper we find a useful asymptotic result for the distribution of T_L as $L \to \infty$. Our methods draw upon results by Janson (1983) and Hall (1988) on coverage processes.

Our work relates to a model studied by Vanderbei and Shepp (1988) and, in a context unrelated to DNA, by Quine and Robinson (1990, 1992). In this related model, there are no specified sites of initiation; instead initiation may occur randomly and indiscriminantly at any location that has not yet been 'passed over' by a frontier, that is, according to a Poisson process which is stationary in space-time. This model is not appropriate for DNA replication, because origins of replication are defined by the specific local stereo-chemistry of the DNA; the approach of the enzyme-complex is not indiscriminant. So the model of Vanderbei and Shepp, though interesting, is not applicable in DNA studies.

Nevertheless, we also apply our method to this related model. As a result, we are able to prove the results mentioned by Vanderbei and Shepp, without the 'gaps' in proof that are acknowledged in their paper.

The analysis of our model

Janson (1983) and Hall (1988) have considered a coverage problem which relates to our model. Line-segments of random iid lengths are placed with left-ends (or centres) at the points of a stationary Poisson process. Janson and Hall considered the probability that these (possibly overlapping) segments cover the interval $[0, y]$.

Our situation fits into this framework. Consider the Poisson point process of all origins which have been *approached* by an enzyme-complex by time t. This includes those origins which were already 'passed over' at the time of enzyme approach, so this Poisson process has intensity $\gamma(1 - e^{-\mu t})$. If at the epoch of approach to each origin a line-segment starts to grow at rate $v = 2r$, then by time t each point has acquired a line-segment with generic length U such that

$$P\{U \leq x\} = P\{v(t - V) \leq x | V \leq t\}, \quad x > 0, \tag{1}$$

where

$$P\{V \leq x\} = 1 - e^{-\mu x}, \quad x > 0. \tag{2}$$

In the real process, passed-over origins do not initiate the growth of a line-segment, but the event that the interval $[0, y]$ is covered has the same probability whether or not we allow such initiations.

Janson (1983, Lemma 2) and Hall (1988, Theorem 2.5) each prove a limit theorem of the following character. If their Poisson process intensity, λ say, tends to infinity and the mean segment-length a tends to 0 such that

$$\frac{e^{a\lambda}}{\lambda} \to e^{u} \tag{3}$$

where $-\infty < u < \infty$, and if a number of other conditions are imposed on the *distribution* of segment-length and its mode of change as a changes, then

$$P\{[0, y] \text{ is covered}\} \to e^{-ye^{-u}}.$$

These theorems do not have an immediate application to our problem, because it turns out that one of the 'other conditions' of Janson or Hall does not hold in our case. Nevertheless their work has helped us prove the following theorem concerning the time T_L until the interval $[0, L]$ is covered.

THEOREM 1. For each real number u,

$$P\left\{v\gamma T_L - \log(\gamma L) - \frac{v\gamma}{\mu} \le u\right\} \to e^{-e^{-u}} \quad \text{as} \quad L \to \infty.$$

Our result involves a limit as $L \to \infty$. At each stage in this limiting process we may rescale the real axis by a factor L, so that $[0, L]$ becomes $[0, 1]$. This means that at time t, the mean rescaled segment length, denoted by a, is

$$a = \frac{v}{L}\left(\frac{t}{1 - e^{-\mu t}} - \frac{1}{\mu}\right),$$

whilst the initiated origins form a Poisson process with intensity

$$\lambda = \frac{v\gamma}{a}\left(1 - e^{-\mu t}\right)\left(\frac{t}{1 - e^{-\mu t}} - \frac{1}{\mu}\right).$$

As $L \to \infty$, it is clear that $a \to 0$ and $\lambda \to \infty$, but for fixed t there does not exist a number $u \in R$ such that (3) holds. It is necessary that t changes with L. Suppose, for $c_1 > 0$,

$$t = c_0 + c_1 \log L.$$

Then (3) holds, provided $c_1 = 1/\gamma v$ and $c_0 = 1/\mu + (u + \log\gamma)/\gamma v$. Suppose therefore that for some $u \in R$

$$t = \frac{\log(\gamma L) + u}{\gamma v} + \frac{1}{\mu}. \tag{4}$$

Expressing a and λ as functions of t (eliminating L) we have

$$a_t = \frac{f(t)}{t}\left(\frac{t}{1 - e^{-\mu t}} - \frac{1}{\mu}\right) \overset{t\to\infty}{\longrightarrow} 0,$$
$$\lambda_t = \left(1 - e^{-\mu t}\right)\frac{\gamma v t}{f(t)} \overset{t\to\infty}{\longrightarrow} \infty,$$

where

$$f(t) = \gamma v t \exp\left(u + \frac{\gamma v}{\mu} - \gamma v t\right) \overset{t\to\infty}{\longrightarrow} 0. \tag{5}$$

With relationship (4) between t and L, (3) holds as $L \to \infty$ and hence as $t \to \infty$. This suggests the application of the theorems of Janson or Hall, yet there is a difficulty. From (1), (2) and (4) the segment-lengths after rescaling have distribution function H_t given by

$$\begin{aligned} H_t(x) =&\ 1 - \frac{1 - \exp[-\mu t\{1 - x/f(t)\}]}{1 - e^{-\mu t}} && x < f(t) \\ &= 1 && x \geq f(t). \end{aligned} \tag{6}$$

$H_t(x)$ changes with t in a different manner, however, from that which is required in Janson or Hall. They require that H_t satisfies $H_t(x) = H_1(a_1 x/a_t)$ for all $t > 0$. Let $p_t(y)$ be the probability that $[0, y]$ is covered at time t. From Janson (1983, Lemma 1) or Hall (1988, Theorem 2.6) we have that (in our notation)

$$\begin{aligned} \pi_t(s) &\equiv \int_0^\infty e^{-sy} p_t(y) dy \\ &= \frac{1}{s} - \frac{1}{s^2 \int_0^\infty \exp\left[-sy + \lambda_t \int_y^\infty \{1 - H_t(x)\} dx\right] dy} \\ &= \frac{1}{s} - \frac{1}{s^2 \int_0^\infty \exp\left[-sy + I_t(y)\right] dy} \qquad \text{say.} \end{aligned} \tag{7}$$

In our context, the essence of Janson/Hall limit theory is a proof that if $H_t(x) = H_1(a_1 x/a_t)$ for $t > 0$, then as $t \to \infty$, (and hence $\lambda_t \to \infty$, $a_t \to 0$ with $\exp(a_t \lambda_t)/\lambda_t \to e^u$),

$$\int_0^\infty \exp\left[-sy + I_t(y)\right] dy \to \frac{1 + s e^u}{s}. \tag{8}$$

Our approach, with an H_t not satisfying the regularity condition, is firstly to introduce a family of distribution functions G_t having mean a_t such that $G_t(x) = G_1(a_1 x/a_t)$. Then

$$\begin{aligned} I_t(y) &= \lambda_t \int_y^\infty \{1 - G_t(x)\} dx + \lambda_t \int_y^\infty \{G_t(x) - H_t(x)\} dx \\ &= J_t(y) + K_t(y) \qquad \text{say.} \end{aligned}$$

Let us (fairly arbitrarily) choose $G_t(x) = 1 - \exp(-x/a_t)$ for $x \geq 0$. Elementary calculations yield

$$\begin{aligned} K_t(y) &= \gamma v t \left[1 - \frac{y}{f(t)} - \frac{1}{\mu t}\{1 - \exp[-\mu t(1 - y/f(t))]\}\right] \\ &\quad -\lambda_t a_t \exp(-y/a_t), && y < f(t), \\ &= -\lambda_t a_t \exp(-y/a_t), && y \geq f(t). \end{aligned}$$

For fixed t, $K_t(y) < 0$ when $y \geq f(t)$ and we shall now show that $K_t(y) \leq 0$ for $y < f(t)$. Note that, for fixed t, $K_t(0) = 0$. K_t is also continuous with continuous derivative for all y, including at $y = f(t)$. The derivatives in the range $0 \le y \le f(t)$ are of the form

$$K_t'(y) = \lambda_t\big[B(e^{by} - 1) - (1 - e^{-cy})\big],$$
$$K_t''(y) = \lambda_t\big[Bbe^{by} - ce^{-cy}\big],$$

where the constants here are unimportant except that $b > 0,\ c > 0,\ B > 0$. Moreover it can be simply shown that $Bb < c$, so $K_t''(0) < 0$. Since $K_t'(0) = 0$, we see that K_t has a turning-point maximum at $y = 0$. Since $K_t(f(t)) < 0$, K_t cannot ever be positive if there are less than two roots of $K_t'(y) = 0$ in the range $0 < y < f(t)$. A plot of both $B(e^{by} - 1)$ and $(1 - e^{-cy})$ against y is helpful, revealing precisely one root of $K_t'(y) = 0$ within that range. Thus $K_t(y) \le 0$ for all $y \ge 0$. So for fixed t,

$$\int_0^\infty \exp\big[-sy + I_t(y)\big]\,dy < \int_0^\infty \exp\big[-sy + J_t(y)\big]\,dy. \tag{9}$$

Thus, this choice of G_t satisfying the Janson/Hall conditions enables us to find an upper bound in (9) which tends to $(1 + se^u)/s$ as $t \to \infty$. We now find a lower bound which has the same limit. Due to the form of (6) we split the range of integration in (8) into $[0, f(t))$ and $[f(t), \infty)$ yielding terms T_1 and T_2, where $T_2 = e^{-sf(t)}/s$ and

$$\begin{aligned}
T_1 &= \int_0^{f(t)} \exp\Big[-sy + \frac{\gamma v}{\mu}\Big\{\mu t - \frac{\mu t y}{f(t)} - 1 + \exp\Big[-\mu t\Big(1 - \frac{y}{f(t)}\Big)\Big]\Big\}\Big]\,dy \\
&> \int_0^{f(t)} \exp\Big[\frac{\gamma v}{\mu} e^{-\mu t}\Big] \exp\Big[-sy + \frac{\gamma v}{\mu}\Big\{\mu t - \frac{\mu t y}{f(t)} - 1\Big\}\Big]\,dy \\
&= \exp\Big[\frac{\gamma v}{\mu} e^{-\mu t}\Big] \exp\Big[\frac{\gamma v}{\mu}(\mu t - 1)\Big] \int_0^{f(t)} \exp\Big[-y\Big(s + \frac{\gamma v t}{f(t)}\Big)\Big]\,dy \\
&= \frac{\exp\big[\frac{\gamma v}{\mu} e^{-\mu t}\big] \exp\big[\frac{\gamma v}{\mu}(\mu t - 1)\big]\big\{1 - \exp[-(sf(t) + \gamma v t)]\big\}}{s + \gamma v t / f(t)} \\
&= T_1^* \qquad \text{say.}
\end{aligned}$$

It is a simple matter using (5) to show that, as $t \to \infty$, $T_2 \to 1/s$ and $T_1^* \to e^u$. Thus

$$T_1^* + T_2 < \int_0^\infty \exp[-sy + I_t(y)]dy,$$

where this lower bound tends to $(1 + se^u)/s$ as $t \to \infty$. Therefore (8) holds, despite the absence of one of the Janson/Hall conditions. So, from (7),

$$\begin{aligned}
\pi_t(s) &\to \frac{1}{s} - \frac{1}{s(1 + se^u)} \\
&= \frac{e^u}{1 + se^u}.
\end{aligned}$$

Thus $p_t(y) \to e^{-ye^{-u}}$. Now our original interval $[0, L]$ has been scaled to be $[0, 1]$, so as $L \to \infty$,

$$P\{v\gamma T_L - \log(\gamma L) - \frac{v\gamma}{\mu} \le u\} = P\{T_L \le t\} = p_t(1) \to e^{-e^{-u}},$$

which proves the assertion.

REMARK 1. A consequence of the result above is that

$$\frac{v\gamma T_L}{\log(\gamma L)} \to 1 \quad \text{in probability as} \quad L \to \infty.$$

REMARK 2. We have taken care not to refer to L as the length of the DNA molecule, merely a long part of the molecule. The DNA is modelled by the whole real line. So our model is not totally accurate since DNA has finite length, albeit very long relative to $1/\lambda$. Arguably we should treat L as the DNA length and study a process where origins outside $[0, L]$ cannot be initiated. Let T_L^* be the completion time for this new process. Usually $T_L^* = T_L$, but their difference (if any) is governed by the proportion, in our studied process, of $[0, L]$ covered by frontiers which are initiated outside $[0, L]$. This proportion at time t is bounded above by vt/L, a term which goes to zero as t and L become large in the manner of (4).

Analysis of the indiscriminant model

Consider a Poisson point process of rate γ in the (x, t)-plane (space-time). Each point in the process initiates on the x-axis a line-segment growing with speed v. As before, T_L is the time when the interval $[0, L]$ is completely covered by such segments.

THEOREM 2. For each real number u,

$$P\left\{\sqrt{\gamma v \log \frac{\gamma L^2}{v}}\, T_L - \log \frac{\gamma L^2}{v} - \tfrac{1}{2} \log\log \frac{\gamma L^2}{v} \le u\right\} \to e^{-e^{-u}} \quad \text{as} \quad L \to \infty.$$

PROOF. At a fixed time t the x-coordinates of the initiated points form a Poisson process on the real line with intensity γt. The line-segments given by this process have lengths which are iid random variables with distribution like vtU where U is uniformly distributed on the unit interval $[0, 1]$. Their mean length is $vt/2$.

Rescale the real axis so that the interval $[0, L]$ becomes $[0, 1]$. Thus the rescaled mean segment length a equals $vt/2L$ and the initiated points form a Poisson process with rescaled intensity $\lambda = \gamma vt^2/2a$. In order to find the appropriate relationship between t and L, we write

$$\begin{aligned}\frac{e^{a\lambda}}{\lambda} &= \exp\left[\frac{\gamma v t^2}{2} - \log(\gamma L t)\right] \\ &= \exp\left[\frac{\gamma v t^2}{2} - \tfrac{1}{2}\log\left(\gamma v t^2 \cdot \frac{\gamma L^2}{v}\right)\right] \\ &= \exp\left[\tfrac{1}{2}\{b - c - \log b\}\right]\end{aligned}$$

where $b = \gamma v t^2$ and $c = \log(\gamma L^2/v)$. We must find, for given $u > 0$, a relationship between b and c such that $b - c - \log b \to 2u$. For reasons, which will become clear it is preferable to express b as a function of c. The relationship

$$b = \left(\sqrt{c} + \frac{\log\sqrt{c}}{\sqrt{c}} + \frac{u}{\sqrt{c}}\right)^2 \tag{10}$$

has the appropriate properties, as we now demonstrate.

$$\begin{aligned}b - c - \log b = {} & \left(\frac{\log\sqrt{c}}{\sqrt{c}}\right)^2 + \frac{u^2}{c} + 2\log\sqrt{c} + 2u + \frac{2u\log\sqrt{c}}{c} \\ & - 2\log\left(\sqrt{c} + \frac{\log\sqrt{c}}{\sqrt{c}} + \frac{u}{\sqrt{c}}\right) \\ \stackrel{c\to\infty}{\longrightarrow} {} & 2u.\end{aligned}$$

So (10) establishes the appropriate relationship between t and L, one which guarantees that $a \to 0$, $\lambda \to \infty$, with $e^{a\lambda}/\lambda \to e^u$. In principle, a and λ can be written as a function of t only (by eliminating L). Also

$$\begin{aligned}H_t(x) &= x/2a_t \qquad && 0 \le x \le 2a_t \\ &= 1 && x > 2a_t.\end{aligned}$$

Clearly, H_t satisfies the Janson/Hall condition $H_t(x) = H_1(a_1 x/a_t)$. So, as $t \to \infty$, $p_t(1) \to e^{-e^{-u}}$. Since

$$\begin{aligned}p_t(1) &= P\{T_L \le t\} \\ &= P\left\{T_L \le \frac{\sqrt{b}}{\sqrt{\gamma v}}\right\} \\ &= P\left\{\sqrt{\gamma v}T_L \le \sqrt{c} + \frac{\log\sqrt{c}}{\sqrt{c}} + \frac{u}{\sqrt{c}}\right\} \\ &= P\left\{\sqrt{\gamma v c}T_L - c - \tfrac{1}{2}\log c \le u\right\},\end{aligned}$$

the theorem is proved.

REMARK 1. A consequence of the result above is that

$$\frac{\sqrt{\gamma v}T_L}{\sqrt{\log\frac{L^2\gamma}{v}}} \to 1 \quad \text{in probability as} \quad L \to \infty.$$

REMARK 2. Vanderbei and Shepp introduced the function ψ defined by the equation $\psi(x) = \log\psi(x) + x$ for $x \geq 1$, and gave the limit distribution using this function. It is easily seen that $\psi(x) = x + \log x + o(1)$ as $x \to \infty$. Using this, it follows after a simple calculation that their limiting asymptotic formula for the distribution of T_L is equivalent to ours. Their method is completely different and their 'proof' has some gaps.

References

Hall, P. (1988). *Introduction to the Theory of Coverage Process.* Wiley, New York.

Janson, S. (1983). Random coverings of the circle with arcs of random lengths. in *Essays in honour of Carl-Gustav Esseen.* ed. A. Gut & L. Holst. Dept. Math. Uppsala Univ.

Kornberg, A. (1980). *DNA Replication.* Freeman, San Francisco.

Quine, M.P. and Robinson, J. (1990). A linear random growth model. *J. Appl. Prob.* **27,** 499–509.

Quine, M.P. and Robinson, J. (1992). Estimation for a linear growth model. *Statistics & Probability Letters* **15,** 293–297.

Vanderbei, R.J. and Shepp, L.A. (1988). A probabilistic model for the time to unravel a strand of DNA. *Stochastic Models* **4,** 299–314.