A new discrete distribution arising in a model of DNA replication

Richard Cowan
School of Mathematics and Statistics,
University of Sydney, NSW 2006, Australia.

August 31, 2000

Abstract

During DNA replication, small fragments of DNA are formed. These have been observed experimentally and the mechanism of their formation modelled mathematically. Using the stochastic model of Cowan and Chiu (1992, 1994), we find the probability distribution of the number of fragments. A new discrete distribution arises. The work has interest as an application of the recent theory on quasi-renewal equations in Piau (2000).

Keywords: DNA Replication, Okazaki fragment, quasi renewal equations, discrete distributions.

AMS 1991 Subject Classification: Primary 60E05. Secondary 60K05, 60K40, 92C40, 92C45.

1 Finding water for the town

We firstly phrase the problem of this paper in a simple non-biological setting. Consider an exploration team which starts at a town located on the x-axis at the origin O. The team's job is to find water for the residents of the town. The team commences at O and moves along the positive x-axis at a constant speed r. When a water spring is seen, a member of the team immediately starts to build a pipeline back to the town. He builds at a rate c. The rest of the team continues at speed r to look for more springs. Whenever the next spring is seen, another man starts to build a pipeline toward the town. He will eventually reach the earlier pipeline which started at the first spring and the pipes are instantaneously joined. This process continues; explorers move away from the town at speed r and pipe-builders construct pipes toward the town at speed c. Pipes are joined whenever constructed segments of pipe meet. The team of explorers is assumed infinite and the location of springs is assumed to follow a Poisson process on the x-axis. For large t, what is the distribution of the number of pipe-builders at work at time t (construction by a builder on any segment having ceased when it joins another segment)?

In this paper we answer this problem, but in the biological setting which motivated it, namely the formation of Okazaki fragments during DNA replication. A model for this process was first presented in Cowan and Chiu (1992,1994) and Cowan (2001), where links to the relevant biological literature are given.

2 The pipe segments and the DNA analogy

DNA is a double-stranded molecule. When it commences its replication, the process starts at sites known as origins. At an origin, the two strands separate. Under enzymic control the separation extends bidirectionally thereby creating a structure which is called an *eye* or *replicon* (see Fig. 1a where the position of the origin O is marked on both strands). The advancing frontiers of strand separation (called *replication forks*) move at speeds r. Within the eye, strands are now single-stranded, ready for replication.

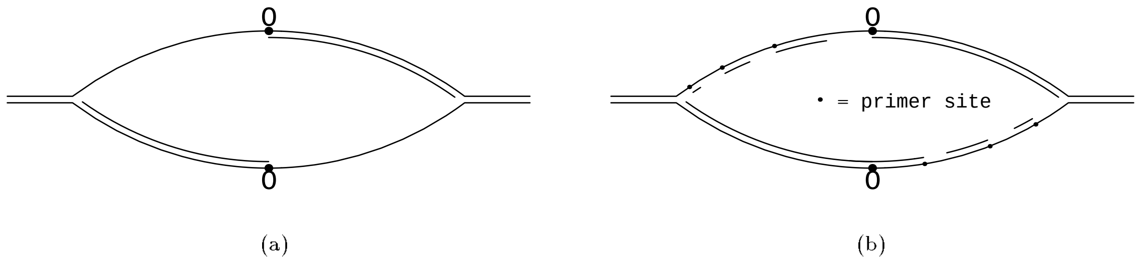

Figure 1: (a) An eye and the effect on the leading strand of replication by the same enzyme that separates the two strands. (b) The discontinuous replication on the lagging strand from primer sites, with $c/r = 0.6$.

Replication starts at each O. In the top strand a copying process takes place in a left-to-right direction following directly behind the moving fork. On the lower strand, a similar process takes place in a right-to-left direction. So half of the eye shown in Fig. 1 is copied almost immediately.

The DNA strand is highly directional and replication cannot take place in the opposite directions to those described above. So the cell needs another mechanism to copy the parts of the eye which remain single-stranded in Figure 1a.

There are sites randomly scattered on each strand known as *primer sites* (shown in Fig 1b). At these sites, copying can also commence as soon as the site is exposed by the moving fork. Copying starts at these sites in the direction allowed, but at a different speed than that associated with origin-initiated copying. This new speed is c (thought to be $< r$, though we do not use this constraint in the sequel). As Figure 1b shows, this mechanism leads to fragments of copied DNA along the DNA eye. The two strands at a fork are called *leading* and *lagging*; the copying is continuous on the leading strand and discontinuous on the lagging strand.

The discontinuous pieces, called *Okazaki fragments* after their discoverers (Okazaki *et al*, 1968), grow and coalesce with each other. Also, when the first fragment reaches O, it joins the copy created earlier by the other mechanism and so is no longer a fragment. Other fragments join onto the first and so on; all fragments which are created eventually join onto other fragments connected to O and hence cease to be fragments. Cowan and Chiu refer to the extant fragments as *islands* in contrast to the O-connected copies which they call the *mainland*. They model the positions of the primer sites by a Poisson process along the DNA, as do we. Thus the times at which sites are passed by the moving fork also form a Poisson process (with rate denoted by λ per unit time).

So, to the right of O, the primer sites are like the water springs and the moving fork is akin to the movement of the exploration team. The Okazaki fragments on the lower strand are analogous to the segments of water-pipes under construction.

We now proceed to find the probability mass function of N_t, defined as the number of extant fragments at a given fork at time t (or equivalently the number of busy pipe-builders). We focus on the equilibrium case of large t.

Before embarking an our analysis, we note a notational change from Cowan and Chiu. Our c was called b in that paper, but here we use the symbol b in another way, in conformity with a recent paper of Piau (2000) which we draw upon for some important analytical tools.

3 The distribution of N_t: the quasi-renewal equations.

Let $g_j(t) := \mathbb{P}\{N_t = j\}$ and $g_j := \lim_{t\to\infty} g_j(t)$. We find a system of integral equations for $g_j(\cdot)$ by exploiting a regenerative argument after conditioning on Y_1, the time until the first primer site is passed by the fork. Focussing on g_0 first, we can write (using $a := c/(c+r)$):

$$\begin{aligned}
g_0(t) &= \int_0^\infty g_0(t|y)\lambda e^{-\lambda y}dy \\
&= \int_0^{at} g_0(t-y)\lambda e^{-\lambda y}dy + \int_{at}^{t} 0\cdot\lambda e^{-\lambda y}dy + \int_t^\infty 1\cdot\lambda e^{-\lambda y}dy \\
&= e^{-\lambda t} + \int_0^{at} g_0(t-y)\lambda e^{-\lambda y}dy. \qquad (1)
\end{aligned}$$

This equation captures the fact that if $Y_1 > t$ there will certainly be no Okazaki fragments at t, whilst if $Y_1 < at$, the fragment initiated at the first primer-site will have had sufficient time to join the mainland by time t (hence ceasing to be a fragment by t). In the latter case, the only fragments extant at t will be those created in the remaining time $t-y$ (the regenerative argument).

A similar argument shows that, for $j \geq 1$,

$$g_j(t) = \int_0^{at} g_j(t-y)\lambda e^{-\lambda y}dy + \int_{at}^{t} g_{j-1}(t-y)\lambda e^{-\lambda y}dy. \qquad (2)$$

We can write (1) and (2) as one system for $j \geq 0$, namely

$$g_j(t) = h_j(t) + \int_0^{at} g_j(t-y)\lambda e^{-\lambda y}dy, \qquad (3)$$

where

$$\begin{aligned}
h_j(t) &= e^{-\lambda t}, & j = 0, \\
&= \int_{at}^{t} g_{j-1}(t-y)\lambda e^{-\lambda y}dy, & j \geq 1. \qquad (4)
\end{aligned}$$

The equations in (3) are examples of the quasi-renewal equation studied by Piau (2000). They are not renewal equations because $a < 1$. Piau studied the single equation

$$g(t) = h(t) + \int_0^{at} g(t-y)\lambda e^{-\lambda y}dy,$$

where $a \in (0,1]$. He showed that, if L is the Laplace transform operator, defined for $x \geq 0$ by

$$Lg(x) := \int_0^\infty e^{-xt}g(t)dt \qquad Lh(x) := \int_0^\infty e^{-xt}h(t)dt$$

and $b := 1 - a$, then

$$Lg(x) = \sum_{m \geq 0} (-1)^m Lh(\ A^m(x)\ )\big(1 + \frac{A^m(x)}{\lambda}\big) \prod_{k=0}^{m} \frac{\lambda}{A^k(x)} \tag{5}$$

where A is the affine function $A(x) := (x + \lambda a)/b$ and A^n is the n-th iterate of A, that is, $A^n(x) = -\lambda + (x + \lambda)/b^n$. Piau also showed that, as $t \to \infty$,

$$g(t) - h(t) \to\ \lambda \sum_{n \geq 0} (-1)^n Lh(x_n) \frac{P_n(b)}{b^n}, \tag{6}$$

where $x_n := A^n(0) = \lambda(b^{-n} - 1)$, $P_0(b) := 1$ and, for $n \geq 1$,

$$P_n(b) := \prod_{k=1}^{n} \frac{b^k}{1 - b^k} = \prod_{k=1}^{n} \frac{\lambda}{x_k}.$$

The application of Piau's results to our system of quasi-renewal equations follows. We note from (4) that

$$\begin{aligned} Lh_j(x) &= \frac{1}{\lambda + x}, && j = 0, \\ &= \frac{\lambda}{\lambda + x} Lg_{j-1}(\ A(x)\ ), && j \geq 1. \end{aligned} \tag{7}$$

Note that from (4), $h_j(t) \to 0$ as $t \to \infty$ for all j. This is obvious for $j = 0$. For $j \geq 1$, $g_j(t) := \mathbb{P}\{N_t = j\} \leq 1$ for all t, so

$$h_j(t) \leq \int_{at}^{t} 1 \cdot \lambda e^{-\lambda y} dy \to 0.$$

So the RHS of (6) is the limit of $g(t)$.

Equation (5) in our context takes a shorter form when (7) is substituted.

$$Lg_j(x) = \frac{1}{\lambda} \sum_{m \geq 0} (-1)^m \prod_{k=0}^{m} \frac{\lambda}{A^k(x)}, \qquad j = 0, \tag{8}$$

$$= \sum_{m \geq 0} (-1)^m Lg_{j-1}(\ A^{m+1}(x)\ ) \prod_{k=0}^{m} \frac{\lambda}{A^k(x)}, \qquad j \geq 1. \tag{9}$$

Consider (3) when $j = 0$. From (6)

$$\begin{aligned} g_0(t) \to g_0 &= \lambda \sum_{n \geq 0} (-1)^n Lh_0(x_n)\ \frac{P_n(b)}{b^n} \\ &= \lambda \sum_{n \geq 0} (-1)^n \frac{1}{\lambda + x_n}\ \frac{P_n(b)}{b^n} \\ &= \sum_{n \geq 0} (-1)^n P_n(b) \qquad \text{since } \lambda + x_n = \lambda / b^n \\ &= \prod_{n \geq 1} (1 - b^n). \end{aligned} \tag{10}$$

The last step follows from an identity originally proved by Euler in 1748 and generalised by Heine in the 19th century (see formula 2.2.6 of Andrews (1976), or formula (1.195) of Johnson, Kotz and Kemp (1993).) The identity is

$$\sum_{n\geq 0} z^n P_n(b) = \prod_{n\geq 1}(1+zb^n) \tag{11}$$

valid for $|b|<1$ and all z. We have used $z=-1$ to derive (10).

For $j\geq 1$, commencing with (6) and using (7), we have

$$\begin{aligned} g_j(t) \to g_j &= \lambda \sum_{n\geq 0}(-1)^n Lh_j(x_n)\frac{P_n(b)}{b^n} \\ &= \lambda \sum_{n\geq 0}(-1)^n \frac{\lambda}{\lambda+x_n} Lg_{j-1}(\ A(x_n)\ )\frac{P_n(b)}{b^n} \\ &= \lambda \sum_{n\geq 0}(-1)^n P_n(b) Lg_{j-1}(x_{n+1}), \end{aligned} \tag{12}$$

where we have used the simple identity

$$A^k(x_n) := A^k(A^n(0)) = A^{k+n}(0) =: x_{k+n}. \tag{13}$$

Equation (12) with $j=1$, in combination with (8) and (13), provides us with a result for g_1:

$$\begin{aligned} g_1 &= \lambda \sum_{n_1\geq 0} P_{n_1}(b)\frac{1}{\lambda}\sum_{m\geq 0}(-1)^{n_1+m}\prod_{k=0}^{m}\frac{\lambda}{x_{k+n_1+1}} \\ &= \sum_{n_1\geq 0}\sum_{m\geq 0}(-1)^{n_1+m}P_{n_1+m+1}(b) \quad \text{since} \prod_{k=0}^{m}\frac{\lambda}{x_{k+n}} = \frac{P_{n+m}(b)}{P_{n-1}(b)} \\ &= \sum_{n_1\geq 0}\sum_{n\geq n_1}(-1)^n P_{n+1}(b) \quad \text{where } n := n_1+m \\ &= \sum_{n\geq 0}(-1)^n P_{n+1}(b)\sum_{n_1=0}^{n} 1 \\ &= \sum_{n\geq 0}(-1)^n (n+1) P_{n+1}(b) \\ &= \sum_{n\geq 1}(-1)^{n-1} n P_n(b) \\ &= g_0\ \theta_1 \qquad \text{where } \theta_k := \sum_{n\geq 1}\left(\frac{b^n}{1-b^n}\right)^k. \end{aligned} \tag{14}$$

The last step uses an identity found through the differentiation of (11) with respect to z then letting $z=-1$.

The analysis of (12) can be taken further when $j \geq 2$, using (9):

$$\begin{aligned}
g_j &= \lambda \sum_{n_1 \geq 0} (-1)^{n_1} P_{n_1}(b) \sum_{n_2 \geq 0} (-1)^{n_2} Lg_{j-2}(\ A^{n_2+1}(x_{n_1+1})\ ) \prod_{k=0}^{n_2} \frac{\lambda}{A^k(x_{n_1+1})} \\
&= \lambda \sum_{n_1 \geq 0} P_{n_1}(b) \sum_{n_2 \geq 0} (-1)^{n_1+n_2} Lg_{j-2}(x_{n_1+n_2+2}) \ \prod_{k=0}^{n_2} \frac{\lambda}{x_{k+n_1+1}} \\
&= \lambda \sum_{n_1 \geq 0} P_{n_1}(b) \sum_{n_2 \geq 0} (-1)^{n_1+n_2} Lg_{j-2}(x_{n_1+n_2+2}) \ \frac{P_{n_1+n_2+1}(b)}{P_{n_1}(b)} \\
&= \lambda \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} (-1)^{n_1+n_2} Lg_{j-2}(x_{n_1+n_2+2}) \ P_{n_1+n_2+1}(b). \qquad (15)
\end{aligned}$$

Another application of (9) to (15) for $j \geq 3$ yields

$$\begin{aligned}
g_j &= \lambda \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} P_{n_1+n_2+1}(b) \sum_{n_3 \geq 0} (-1)^{n_1+n_2+n_3} Lg_{j-3}(\ A^{n_3+1}(x_{n_1+n_2+2})\ ) \prod_{k=0}^{n_3} \frac{\lambda}{A^k(x_{n_1+n_2+2})} \\
&= \lambda \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} P_{n_1+n_2+1}(b) \sum_{n_3 \geq 0} (-1)^{n_1+n_2+n_3} Lg_{j-3}(x_{n_1+n_2+n_3+3}) \ \frac{P_{n_1+n_2+n_3+2}(b)}{P_{n_1+n_2+1}(b)} \\
&= \lambda \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} \sum_{n_3 \geq 0} (-1)^{n_1+n_2+n_3} Lg_{j-3}(x_{n_1+n_2+n_3+3}) \ P_{n_1+n_2+n_3+2}(b) \qquad (16)
\end{aligned}$$

and continued iteration shows that

$$g_j = \ \lambda \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} \cdots \sum_{n_j \geq 0} (-1)^{n_1+\ldots+n_j} Lg_0(x_{n_1+n_2+\ldots+n_j+j}) P_{n_1+\ldots+n_j+j-1}(b). \qquad (17)$$

Finally we apply (8) to get, writing $n[j] := \sum_{i=1}^{j} n_i$,

$$\begin{aligned}
g_j &= \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} \cdots \sum_{n_j \geq 0} P_{n[j]+j-1}(b) \sum_{m \geq 0} (-1)^{n[j]+m} \prod_{k=0}^{m} \frac{\lambda}{x_{k+n[j]+j}} \\
&= \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} \cdots \sum_{n_j \geq 0} \sum_{m \geq 0} (-1)^{n[j]+m} P_{n[j]+m+j}(b) \\
&= \sum_{n_1 \geq 0} \sum_{n_2 \geq 0} \cdots \sum_{n_j \geq 0} \sum_{n \geq n[j]} (-1)^{n} P_{n+j}(b) \quad \text{where } n := n[j] + m \\
&= \sum_{n \geq 0} (-1)^{n} P_{n+j}(b) \sum_{n_1=0}^{n} \sum_{n_2=0}^{n-n_1} \cdots \sum_{n_j=0}^{n-n[j-1]} 1 \\
&= \sum_{n \geq 0} (-1)^{n} \frac{(n+1)(n+2)\ldots(n+j)}{j!} P_{n+j}(b) \\
&= \sum_{n \geq j} (-1)^{n-j} \binom{n}{j} P_n(b) \qquad (18)
\end{aligned}$$

arriving at a formula which holds for all $j \geq 0$, since it agrees with (10) and (14) found earlier.

The probability generating function is

$$\begin{aligned}\phi(s) := \sum_{j\geq 0} g_j s^j = \sum_{j\geq 0}\sum_{n\geq j}(-1)^{n-j}\binom{n}{j}P_n(b)s^j &= \sum_{n\geq 0}P_n(b)\sum_{j=0}^{n}\binom{n}{j}s^j(-1)^{n-j}\\ &= \sum_{n\geq 0}P_n(b)(s-1)^n\\ &= \prod_{n\geq 1}(1+(s-1)b^n), \end{aligned} \tag{19}$$

so one sees that $\phi(1) = 1$ as required for a probability distribution. Differentiating (19) and substituting $s = 1$, we obtain the first two moments:

$$\phi'(s) = \phi(s)\sum_{n\geq 1}\frac{b^n}{1+(s-1)b^n} \quad \Rightarrow \quad \mu = \frac{b}{1-b} = \frac{r}{c};$$

$$\phi''(s) = \phi'(s)\sum_{n\geq 1}\frac{b^n}{1+(s-1)b^n} - \phi(s)\sum_{n\geq 1}\frac{(b^n)^2}{(1+(s-1)b^n)^2} \quad \Rightarrow \quad \sigma^2 = \frac{b}{1-b^2} = \frac{r(c+r)}{c(c+2r)}.$$

These results for the first two moments agree with those given in Cowan and Chiu (1994).

It does not seem possible to find a simpler expression than (18) for g_j, though alternative expressions for g_0 and g_1 have been presented in (10) and (14). We are able, however, to give a recurrence relationship involving the sequence of functions θ_k which we defined in (14).

$$g_j = -\frac{1}{j}\sum_{i=0}^{j-1}(-1)^{j-i}g_i\theta_{j-i}(b). \tag{20}$$

4 Acknowledgment

I thank Didier Piau for keeping me in touch with his work and for his comments on an earlier draft.

References

Andrews, G. (1976) The Theory of Partitions. *Encyclopedia of Mathematics and Its Applications. Volume 2: Number Theory.* Addison-Wesley.

Cowan, R. (2001) Stochastic models for DNA replication. In *Handbook of Statistics* **Vol. 20**. Eds. C. R. Rao and D. N. Shanbhag. Elsevier.

Cowan, R. and Chiu, S. N. (1992) Mathematics of DNA replicating forks. Research Report **28**, University of Hong Kong.

Cowan, R. and Chiu, S. N. (1994) A stochastic model of fragment formation when DNA replicates. **31**, 301–308.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1993) *Univariate Discrete Distrubutions.* Wiley.

Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K., Kainuma, R., Sugimo, A. and Iwatsuki, N. (1968) *In vivo* mechanism of DNA chain growth. *Cold Spring Harbour Symposium on Quantitative Biology*, **33**, 129–143.

Piau, D. (2000) Quasi-renewal estimates. *J. Appl. Prob.* **37**, 269–275.