

MOMENTS OF THE DISTRIBUTION OF OKAZAKI FRAGMENTS

KRZYSZTOF BARTOSZEK¹ AND JUSTYNA SIGNERSKA²

ABSTRACT. This paper is a continuation of [1] which provides formulae for the probability distributions of the number of Okazaki fragments at time t during the process of DNA replication. Given the expressions for the moments of the probability distribution of the number of Okazaki fragments at time t in the recursive form, we evaluated formulae for the third and fourth moments, using *Mathematica*, and obtained results in explicit form. Having done this, we calculated the distribution's skewness and kurtosis.

Okazaki fragments are small fragments of DNA remaining after the process of DNA replication. Denote the number of Okazaki fragments at the time $t \geq 0$ by $N_t(\omega)$. Let $g_i(t)$ denote the probability that at time t there are exactly i Okazaki fragments, $g_i(t) = P(N_t = i)$. In [2], [3], [4] it is shown that g_i , $i = 0, 1, \dots$ satisfy the following system of equations,

$$\begin{aligned} g_0(t) &= e^{-\lambda t} + \int_0^{at} g_0(t-y)\lambda e^{-\lambda y} dy \\ g_i(t) &= h_i(t) + \int_0^{at} g_i(t-y)\lambda e^{-\lambda y} dy, \end{aligned} \tag{1}$$

where $0 < a < 1$, $\lambda > 0$ and

$$h_i(t) = \begin{cases} e^{-\lambda t} & i = 0 \\ \int_0^{at} g_{i-1}(t-y)\lambda e^{-\lambda y} dy & i = 1, 2, 3, \dots \end{cases} \tag{2}$$

Denote $\mathbf{g} = \{g_i\}_{i=1}^{\infty}$. In [2] it was proved that \mathbf{g} defines a probability distribution on positive integers and the recursive formula for its moments was derived. Let $n_k(t) = E(N_t^k)$. We consider the following formula for $n'_k(t)$ taken from [1],

$$\begin{aligned} n'_k(t) &= \lambda^2 \sum_{j=0}^{k-1} \binom{k}{j} \int_{at}^t n_j(t-s)e^{-\lambda s} ds + \\ &\lambda e^{-\lambda t} + \lambda \sum_{j=0}^{k-1} \binom{k}{j} \left[\int_{at}^t n'_j(t-s)e^{-\lambda s} ds - a n_j(bt)e^{-\lambda at} \right], \end{aligned} \tag{3}$$

in order to evaluate the explicit expression for the k -th moment, $n_k(t)$. We specifically aimed at calculating $n_3(t)$ and $n_4(t)$ ($n_1(t)$ and $n_2(t)$ were already calculated in [1]). For this

¹ Student of the Department of Mathematics, Gdańsk University of Technology, ul. Narutowicza 11/12, 80-952 Gdańsk, Poland, *E-mail*: kbart@sphere.pl.

² Student of the Department of Mathematics, Gdańsk University of Technology, ul. Narutowicza 11/12, 80-952 Gdańsk, Poland, *E-mail*: jussig@wp.pl.

```

In[1]:=
nderiv0[t_] := 0
ptemp0[t_] := 0
n0[t_] := 1
nderiv1[t_] := L * (1 - a) * e-L*a*t
ptemp1[t_] :=  $\frac{1-a}{a} * (e^{-L*a*t})$ 
n1[t_] :=  $\frac{1-a}{a} * (1 - e^{-L*a*t})$ 
nderivk[t_] := nderivk[t] =
  L2 *  $\sum_{j=0}^{k-1} \left( \text{Binomial}[k, j] * \int_{a*t}^t n_j[t-s] * e^{-L*s} ds \right) +$ 
  L * e-L*t + L *  $\sum_{j=0}^{k-1} \left( \text{Binomial}[k, j] * \left( \int_{a*t}^t \text{nderiv}_j[t-s] * e^{-L*s} ds -$ 
  a * nj[(1-a)*t] * e-L*a*t \right) \right)
ptempm[t_] := ptempm[t] =  $\int \text{nderiv}_m[t] dt$ 
nm[t_] := nm[t] = ptempm[t] - ptempm[0]

r = 4
Do[t1 = TimeUsed[]; nderivm[t];
  ptempm[t_] =  $\int \text{nderiv}_m[t] dt$ ; nm[t_] = ptempm[t] - ptempm[0];
  t2 = TimeUsed[]; diff = t2 - t1; Print[m];
  Print[nm[t]]; Print[nm[0]]; Print[diff],
  {m, 2, r, 1}]

Mean[t_] := n1[t]
Mean[t]
Varian[t_] := n2[t] - n1[t] * n1[t]
Varian[t]
Skewness[t_] :=  $\frac{n_3[t] - 3 * n_1[t] * n_2[t] + 3 * n_1[t]^3 - n_1[t]^3}{\sqrt{\text{Varian}[t]^3}}$ 
Skewness[t]
Kurtosis[t_] :=  $\frac{n_4[t] - 4 * n_3[t] * n_1[t] + 6 * n_2[t] * n_1[t]^2 - 4 * n_1[t]^4 + n_1[t]^4}{\text{Varian}[t]^2} - 3$ 
Kurtosis[t]

```

FIGURE 1. *Mathematica* program

purpose we created a short program in *Mathematica* (Figure 1), which also calculates their limits as t goes to infinity and the skewness and kurtosis of g .

The computational complexity of this formula is $T(n) = \sum_{j=0}^{n-1} T(j) + 1$ whose solution is $T(n) = 3^n$. Therefore the complexity is $\Theta(3^n)$ assuming that all *Mathematica* operations are done in $O(1)$ time. Exponential complexity is a result of the form of expression (3). Times required to compute consecutive moments on a Celeron 1.80GHz *Windows XP Home Edition Mathematica 5* are shown in Figure 2.

Using the program from Figure 1 we evaluated the formulae for $n_2(t)$, $n_3(t)$, and $n_4(t)$. The formulae are very long, n_3 takes up 1 page while n_4 takes up 8, therefore they are not included. Their graphs with parameters $a = 0.4$ and $\lambda = 1$ are shown in Figure 3. We computed for $\lambda = 1$ for simplicity of calculation and also because a change of the value of λ merely changes the time scale. Special biological significance is given for values of a near 0.4 and near 0.006 [4]. We chose $a = 0.4$. Calculating the limits of $n_1(t)$, $n_2(t)$, $n_3(t)$,

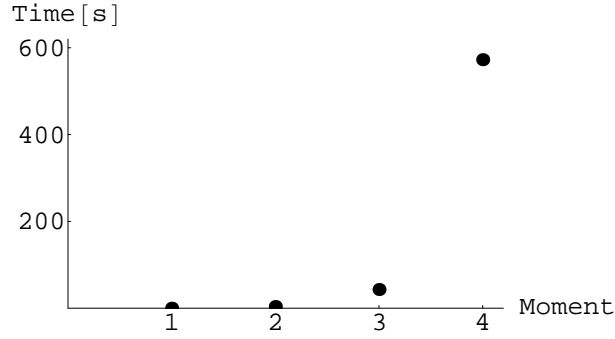


FIGURE 2. Time of computation of moments.

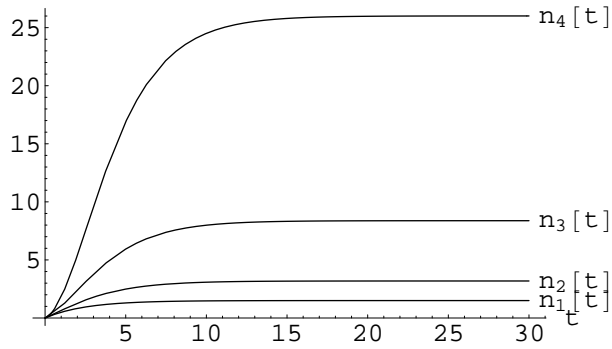


FIGURE 3. The first, second, third, fourth moments.

and $n_4(t)$ as t goes to infinity with the same parameters λ and a gave the following results.

$$\begin{aligned} \lim_{t \rightarrow \infty} n_1(t) &= 1.5000 & \lim_{t \rightarrow \infty} n_2(t) &= 3.1875 \\ \lim_{t \rightarrow \infty} n_3(t) &= 8.3771 & \lim_{t \rightarrow \infty} n_4(t) &= 26.008 \end{aligned} \quad (4)$$

It was found in [2] (see also [1]) that $Var(\mathbf{g}) = \frac{1-a}{1-(1-a)^2}$. We further investigated other features of the distribution \mathbf{g} , particularly the skewness, $S(t)$, and kurtosis, $K(t)$.

$$S(t) = \frac{\mathbb{E}((X - \mathbb{E}X)^3)}{\sqrt{\mathbb{E}((X - \mathbb{E}X)^2)^3}} \quad K(t) = \frac{\mathbb{E}((X - \mathbb{E}X)^4)}{\mathbb{E}((X - \mathbb{E}X)^2)^2} - 3 \quad (5)$$

Again the formulae are very long so we present only their graphs with the graphs of the mean value and variance (as before, with $\lambda = 1$ and $a = 0.4$) in Figure 4.

The limits of the mean, the variance, the skewness and the kurtosis as t goes to infinity are the following,

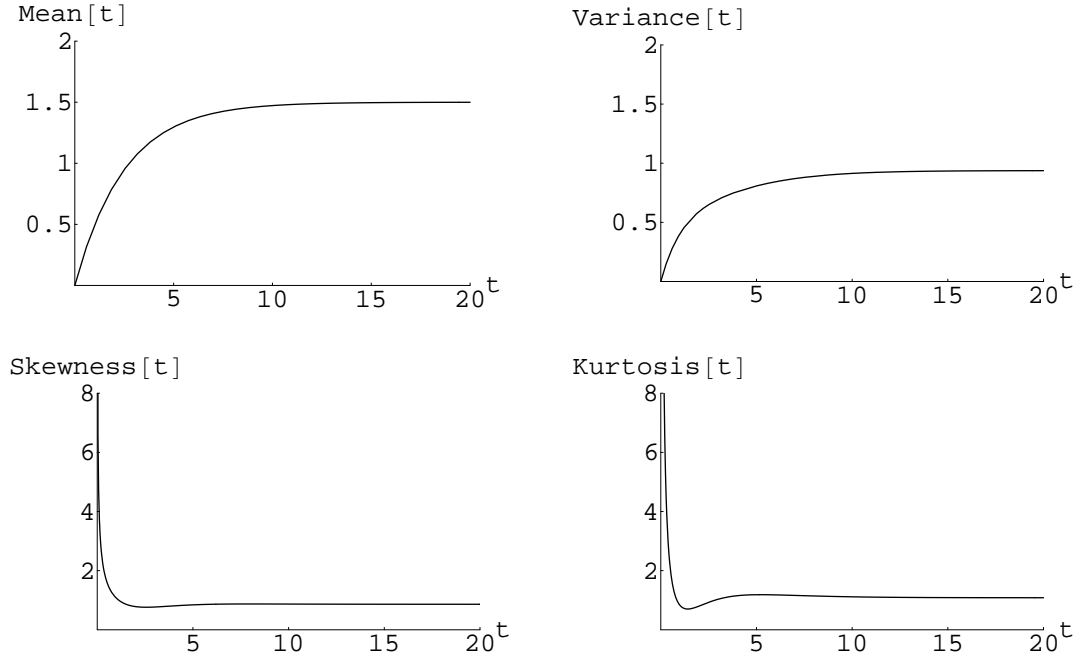


FIGURE 4. The mean, variance, skewness and kurtosis values of \mathbf{g} .

$$\begin{aligned}
 \mathbb{E} &= \lim_{t \rightarrow \infty} n_1(t) = 1.5 & \text{Var} &= \lim_{t \rightarrow \infty} n_2(t) - n_1(t)^2 = 0.9375 \\
 S &= \lim_{t \rightarrow \infty} S(t) = 0.86299 & K &= \lim_{t \rightarrow \infty} K(t) = 1.08352.
 \end{aligned} \tag{6}$$

In [1] the following formula was proved,

$$g_i = \prod_{j=1}^{\infty} (1 - b^j) \cdot \sum_{m=i}^{\infty} \frac{b^m}{1 - b^m} \Psi_{i,m}(b), \tag{7}$$

where for all $r \geq 1$ we put $\Psi_{1,r}(b) \equiv 1$ and for $s \geq i + 1$,

$$\Psi_{i+1,s}(b) = \sum_{r=i}^{s-1} \frac{b^r}{1 - b^r} \Psi_{i,r}(b). \tag{8}$$

Using this representation for g_i the authors of [1] computationally obtained the approximate values for g_i , $i = 0, 1, 2, \dots, 10$, where $g_i = \lim_{t \rightarrow \infty} g_i(t)$. We present these results in Figure 5.

The kurtosis, which says about the degree of peakedness, is about 1 (6) (for comparison, the kurtosis of the normal distribution is 0). The fact that the skewness is 0.86299 (6) indicates that the discussed distribution is positively skewed; i.e., if the distribution was "extended" to the interval $(-\infty, \infty)$, the right tail would be more pronounced than the left tail.

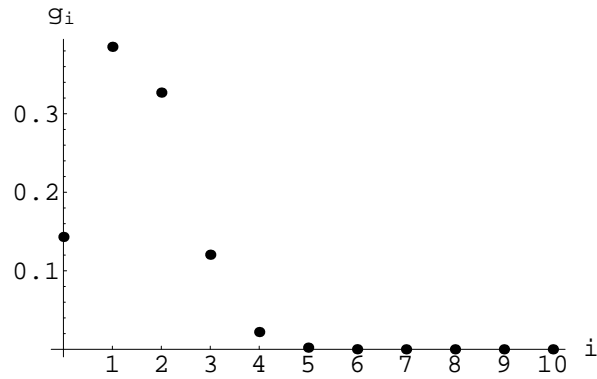


FIGURE 5. The distribution \mathbf{g} : values of g_i for $i = 0, 1, 2, \dots, 10$.

Further research to find an effective method of calculating the moments is necessary due to the exponential complexity of formula (3). Also it should be investigated what effect different values of a have on the moments.

REFERENCES

- [1] K. Bartoszek and W. Bartoszek. On the time behaviour of Okazaki fragments. *Journal of Applied Probability*, 43:500–509, 2006.
- [2] R. Cowan. A new discrete distribution arising in a model of DNA replication. *Journal of Applied Probability*, 38:754–760, 2001.
- [3] R. Cowan. Stochastic models for DNA replication. In C.R. Rao and D.N. Shanbbang, editors, *Handbook of Statistics*, Volume 20, pages 137–166. North-Holland, Amsterdam, 2001.
- [4] D. Piau. Quasi-renewal estimates. *Journal of Applied Probability*, 37:269–275, 2000.