Stochastic models for DNA replication

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

October 26, 2000

1 Variation in biology

A statistician or probabilist entering into the biological domain may do so at a variety of levels. Most statisticians, when consulted for help by biologists, find themselves dealing at the *animal or plant level* – the animal (or plant) is the experimental unit. Since measurements usually display considerable variability from animal to animal, the consulting statistician finds ready application for the well-developed repertoire of statistical analyses – analysis of variance, regression, survival analysis, and so on – in interpreting the experiment. Quite often in our experiences at this level, the statistician is not greatly involved with the biology, beyond what the data tell him/her about these experimental units.

At *the organ or tissue level*, where a part of the animal or plant is studied, there are often similar considerations. For example, a sample of leaves from a single plant may be subjected to experiment. Once again, the standard statistical techniques might be applied, with the leaf being the experimental unit.

One might wish to study the animal or plant at its *cellular level*. Or, one may be working with an organism, such as a bacterium or yeast, which is unicellular and so the study of the "animal" necessarily means observation of the "cell". If we are to take observations on a single cell, either a microscope or a "flow cytometer" is typically involved. For example, we might observe a culture of cells on a petri dish and, by meticulous observation under the microscope, record the individual generation time of each cell (Powell, 1953, 1958, Cowan and Staudte, 1986) or the cell's size at division (Staudte *et al*, 1984). Or we might perhaps observe the cells individually as they flow, in a dilute solution, past some electronic sensor which measures (say) the amount of fluorescence emitted by each cell, the cell population having been previously labelled by a fluorescent dye which intercollates with the cell's DNA. This standard technique of "flow cytometry" is able to stage the cell's progression through its division cycle by effectively measuring the amount of DNA it currently has.

Such observations of individual cells can show considerable variability; for example, the generation times of the bacterial cells in Powell's culture had a coefficient of variation in the range 15–25%.

So it is possible to make the sort of observations on individual cells which would draw the statistician into biology at the cellular level. Very often, however, measurements are not done individually but are instead made on a large population of cells. For example, a biologist may measure the optical density of the millions of cells in a flask by using an appropriate scanner and may have the readings for 20 flasks, say. The flask now becomes the experimental unit and so the statistician's exposure to cellular biology may turn out again to be minimal. Certainly the inter-cellular variation is not being measured and is lost in the aggregate nature of the observation.

A great deal of the literature in biology is now at the *molecular level* and, at this level, observation of *the individual* is a rarity indeed. Generally, our powers of observation do not allow us to look at the individual molecule. We know from physical chemistry that these individuals display a great deal of randomness – in their movement, their speed, their orientation and energy levels, their electron status, and, for the larger molecules, their isomeric form, shape and configuration. Yet we don't see as a rule, except in the occasional electron micrograph, direct observation of this variation for individual molecules.

Our observations usually involve aggregates, where the impact of *Avogadro's number* and various *laws of large numbers* average out molecular variation. Hence one might think that there is little of stochastic interest in biology at the molecular level and little need for the skills of the statistician or probabilist. In this article, by focussing on a range of problems associated with the replication of DNA, I aim to show otherwise.

2 Stochastic chemistry

Firstly, however, I shall review the small literature which deals with these aggregation effects in systems with a large number of reacting molecules. This is the literature (surveyed by McQuarrie, 1967) which treats chemical reactions as stochastic processes.

Consider the reversible chemical reaction $A+B \rightleftharpoons AB$. This has been modelled as a Markov process with a state described by the triple (a_t, b_t, n_t). Here a_t and b_t denote the numbers of free A and B molecules respectively at time t and n_t is the number of AB molecules at t. Indeed the state can be described merely by n_t since, due to the conservation laws for the A and B molecules, $a_t + n_t = a_0 + n_0$ and $b_t + n_t = b_0 + n_0$. Introducing $p_n(t) := \mathbb{P}\{n_t = n\}$ and using the standard methods of Markovian process analysis, combined with *first-order chemical kinetics* for the breakdown of AB and *law of mass action* assumptions for the combining of free A with free B, we may write a family of differential equations. For $0 < n < n_0 + \min(a_0, b_0)$,

$$\begin{aligned} p_n'(t) =& \rho(n+1)p_{n+1}(t) - [\rho n + \theta(a_0 + n_0 - n)(b_0 + n_0 - n)]p_n(t) \\ &+ \theta(a_0 + n_0 - n + 1)(b_0 + n_0 - n + 1)p_{n-1}(t) \end{aligned} \tag{1}$$

where ρ is the *infinitessimal rate* of breakdown of an AB molecule and θ is the infinitessimal rate of aggregation of every pair of free A and B. When $n = n_0 + \min(a_0, b_0)$ or $n = 0$, the first or third term respectively is omitted from (1).

The model leading to (1) is not unique from a chemical viewpoint; for example, the first-order-kinetics assumption for breakdown of AB into its products is one which attributes the breakdown solely to energy received from collision of an AB with either the side of the flask or the H_2O molecules which form the solution and not to the energy of collisions with other AB, A or B molecules. If these latter collisions were important, then the equation should contain breakdown rates which involve products of various molecular abundances. The model has also been criticised for the glib way it handles the combining events $A + B \rightarrow AB$, focussing as it does merely on abundance of the free A and B without regard for their spatial distribution (Clifford, Green & Pilling, 1987).

The stochastic approach leading to (1) is nevertheless a considerable refinement over the deterministic approach which is prevalent among chemists. The chemists' approach writes equations in terms of *chemical concentrations*, $[A]_t, [B]_t$ and $[AB]_t$ assumed to be continuous and differentiable functions of t. Whilst this assumption is strictly incorrect, since concentration is the number of molecules (an integer at all times) divided by the volume, its adoption leads to the differential equation

$$\begin{aligned}\frac{d[AB]_t}{dt} &= -\rho[AB]_t + \theta[A]_t[B]_t \\ &= -\rho[AB]_t + \theta([A]_0 + [AB]_0 - [AB]_t)([B]_0 + [AB]_0 - [AB]_t).\end{aligned} \tag{2}$$

If required, the transient solution of this Riccati equation (2) is readily found, though in practice this does not seem to interest chemists greatly. Rather, they focus on the equilibrium state trivially found by equating the time derivative to zero.

By contrast, the transient solution of the stochastic model (1) is not easy to find except in the irreversible case where $\rho = 0$ (McQuarrie, 1967; Ishida 1964). Darvey, Ninham and Staff (1966) have, however given explicit results for the equilibrium probability distribution when the reaction is reversible (that is, $\rho > 0$) in terms of Hypergeometric functions.

Ishida's calculations for the irreversible case have revealed that $E(n_t)$ from the stochastic theory is virtually the same as $[AB]_t$ from the deterministic theory, except in the final stages of the reaction when nearly all of the less abundant of A or B has been converted. A minor, but physically insignificant, difference exists at this late stage. His formulae also show that, in the canonical case where $n_0 = 0$ and $a_0 = b_0$, the actual process n_t has very low variation about its mean, until this later stage (when a_t and b_t have become small).

The attitude expressed in the probability literature of that time (McQuarrie, 1967) was one of disappointment. Was the stochastic approach worthwhile, given its complexity and the identical practical conclusions to that of the simpler deterministic theory? Moreover Avagadro's number suggested that most reactions of interest to chemists involve massive numbers of molecules. Similar discouragement arose from the study of other reactions, for example, the stochastic study of the simple conversion of one substance, S say, to another, P say, through the action of an enzymic catalyst, E. This 'Michaelis-Menten' reaction, $S + E \rightleftharpoons ES \rightarrow P + E$, was studied stochastically by Heyde and Heyde (1971). With large numbers of molecules involved, stochastic variation is low and unobservable, though it might be reasoned that the time for the last few S to be converted to P could have sizeable stochastic variation relative to its mean (albeit unobservable in a practical sense).

Some change of thinking by mathematicians on the role of probability in chemical reactions came with the work of Hall (1983) and Dunstan and Reynolds (1981). Hall considered the $A + B \rightleftharpoons AB$ reaction when the ratio θ/ρ is either very small or very large. If the ratio is very small, then there will be very few AB molecules at equilibrium. If it is very large, then the less abundant of A and B will be a rarity at equilibrium. He showed that either the *Poisson distribution* or a so-called *modified Bessel distribution* approximates well the number at equilbrium of this less abundant free molecule (which of the two distributions being determined by subtleties on the numbers of the *more* abundant of A or B). Hall did not consider small numbers of molecules; rather his results were asymptotic as the numbers of molecules involved became infinite. The interesting results came because the ratio θ/ρ of reaction rates simultaneously became extreme as molecular abundance became infinite.

Hall (1983) also made precise an interesting observation made by Dunstan and Reynolds (1981), who had demonstrated earlier that a *normal* distribution could approximate the complicated equilibrium distribution of Darvey, Ninham and Staff (1966) for the reaction $A+B \rightleftharpoons AB$. Dunstan and Reynolds have one numerical example, whereby $a_0 = b_0 = 10^9$, $n_0 = 0$ and $\theta/\rho = 6^{-1} \times 10^{-14}$. They calculate that, at large times t, $n_t \sim N(1667, 1667)$ approximately, in this example. Noting a coefficient of variation of about 2.5%, they warn that a deterministic analysis may be somewhat inaccurate for molecular numbers below 10^9.

I believe that the warning of Dunstan and Reynolds needs qualification. The non-negligible coefficient of variation is a reflection, as in Hall's work, of reaction imbalance; their ratio θ/ρ is very small mathematically – only about 1667 molecules of bound AB exist at equilibrium compared to about 2×10^9 of free molecules. Had they chosen a ratio equal to 1, they would have concluded that $n_t \sim N(0.999968 \times 10^9, 126^2)$ approximately; the coefficient of variation would then have been truly negligible. Thus the warning based merely on initial numbers of A and B is unsound – the reaction imbalance is a major factor. In their defence, reversible reactions in practice are often energetically disposed to be rather one-directional and, in those cases, their warning may be apposite.

Hall neatly ties up the relationship between the modified Bessel, Poisson and Normal distributions by showing that the modified Bessel family is approximated by the Poisson or Normal distributions for some of its parameter values. So the theories for the population size of the rarer molecules of a reaction are unified by Hall.

Thus the work of Hall, Dunstan and Reynolds adds considerable insight to the earlier theories of stochastic chemistry. We might summarise that non-negligible stochastic effects for molecular-abundance quantities like n_t occur in these simple chemical reactions:

- whenever numbers of all molecular types are small;
- whenever we look at the rarer molecules in reversible reactions which are close to being irreversible;
- toward the completion of irreversible reactions when numbers of some molecules are becoming small.

3 Exponentially distributed waiting times

Later in this article, we need, for our model building, assumptions about the distribution of waiting time for one particular site on the DNA molecule to bind an enzyme (from a pool of identical enzymes which are attracted to that site). We have found that the exponential distribution is a convenient choice because of its mathematical simplicity. Is this an appropriate choice, however? Let us explore the issue of waiting times with reference to our learning example $A + B \rightleftharpoons AB$.

Models in chemistry of a Markovian nature, such as the one discussed above for the reversible reaction $A + B \rightleftharpoons AB$, have inbuilt consequences about the exponential distribution. Firstly, the time that a given AB molecule stays in a bound form is exponentially distributed with pdf $\rho e^{-\rho x}$. Secondly, the time from t until the next event (either a joining of a free A and B or a disintegration of an AB) is, given n_t, exponentially distributed by model assumption. Unconditional upon n_t, this time is distributed as a mixture of exponentials, having pdf

$$\sum_{n=0}^{\infty} \mathbb{P}\{n_t = n\} \lambda_n e^{-\lambda_n x},$$

where $\lambda_n = \rho n + \theta(a_0 + n_0 - n)(b_0 + n_0 - n)$. It must be noted, however, that unless a_0, b_0 and n_0 are *all* quite small, the *random variables* n_t and λ_{n_t} will have negligible coefficient of variation and so the mixed exponential will differ little in practical terms from a pure exponential distribution as a model for the *next event* waiting time.

Of more interest though, for some of the problems we shall be considering later, is the time T until a given free A molecule first binds to a B molecule. Here envisage the particular A as having a label which identifies it, but any B suffices.

Once again there are heuristic arguments which show that T is approximately exponentially distributed provided there are a reasonably large number of B molecules *and* provided that the chemical reaction is reversible. Our arguments draw on ideas from the theory of stationary point processes.

To see how the exponential arises in the simplest setting, consider the stationary point process formed by the epochs when our labelled A commences to bind with the i-th B molecule in the system. This A will repeatably bind with this particular B when the reaction is reversible, staying bound for a while, and after many liaisons with other B's will find the i-th B again. We assert that the expected number of "binding commencements" with that particular B per unit time, say λ, is finite and positive.

If b is the total number of B molecules in the system, either free or bound, then there will be b such point processes. These are not independent, but any dependence is weak and gets weaker as b is made larger.

Thus, heuristically at least, we see that the point process formed by the superposition of these b point processes will, for large b, be approximately a Poisson point process. This process is none other than the point process of all "binding commencements" of the labelled A molecule. Thus the times between such binding epochs will be exponentially distributed approximately.

4 The biological cell

Stochastic interest in molecular biology can be high, because many important molecules in the cell are there in small numbers. DNA is the prime example. An *E. Coli* bacterial cell [1] has just one DNA molecule[2]. On this single DNA molecule, there is just one site which is receptive to the molecular complexes which initiate the replication of the DNA. This site is called *the origin* of replication. The initiating complexes themselves are in greater abundance, a few hundred perhaps (though there is no good data on the number); they are manufactured by the cell for a short period of each generation time.

One speculates that there may be some random delay as this unique site waits for a complex to bind and that this is part of the explanation as to why this early phase of the cell generation time, prior to the commencement of DNA replication, is variable in duration (Smith and Martin, 1973). The overall generation times of bacterial cells are quite variable, with coefficients of variation in the range 15-25% (Powell, 1953, 1958), and it has been found that the greatest source of the variation is in the waiting period until this critical binding event which initiates DNA relication.

[1] growing at normal temperatures and with normal nutrient, so that its generation time is about one hour on average (Mitchison, 1971, p.102)

[2] We ignore here the DNA contained within the plasmids and phages, biological entities which reside within the bacterium.

There are, of course, some molecules in the cell which are present in massive numbers, perhaps even in Avogadrian abundance. Amongst these are the *nucleotides*, *amino acids* and *fatty acids*, the basic building blocks for the polymers which the cell manufactures. These polymers are the DNA itself, numerous RNA's, the proteins/enzymes and the cellular membranes.

There also exist some less abundant macro-molecules which possess multiple binding sites for some enzyme or protein. So the abundance of these sites arises not from an abundance of the macro-molecules, but from the multiplicity of the sites. For example, there is but one membrane (a geometrically–regular polymer of fatty acids) which encloses the cell, yet it has a vast number of receptors which attract other molecules (as part of the import and export of matter between the cell and its external environment). So these receptors form an abundant class of binding sites.

A second, and perhaps greater element of interest for stochastic modellers, is that the binding sites on some of the macro-molecules are arranged spatially in a quite random way. DNA is a prime example. A strand of this "double-stranded" polymer is a seemingly random sequence of its building blocks, the nucleotide "bases" *adenine, cytosine, guanine* and *thiamadine*, called A, C, G and T – for example, ... T-G-T-C-A-A-G-A-T-G-C- .... (The other strand is determined from the first by simple matching rules, A opposite T and C opposite G, and the two strands are held together in the famous Crick-Watson helical structure.) Other molecules, enzymes usually, can be attracted to specific "letter" patterns on one strand; these patterns obviously occur at variable distances from each other.

For example, there is an enzyme (called Hpa I) which has a chemical affinity with any site on the DNA where a sequence G-T-T-A-A-C occurs. The local geometry and electron densities caused by these "letters" in sequence create the affinity. Yet, because of the seemingly random sequence of letters along the whole DNA molecule, the locations of the specific pattern G-T-T-A-A-C appear to be random. For patterns such as this, with sufficient rarity due to its 6 "bases", any map of the sites (known as a *restriction map*) gives the impression of a Poisson point process. Indeed Poisson limit theorems of the type discussed in Barbour *et al* (1992) and by Chryssaphinou and Papastavridis (1988) provide some mathematical substance to a claim that, despite the discrete nature of the problem, the location of Hpa I sites form a Poisson point process on a the continuum which is the linear DNA molecule.

Some biologist may raise a few objections to the statement that DNA is a random molecule, at least in the regions of that molecule which have a definite role to play in directing and regulating the nature of the animal or plant. Yet, a few hours spent reading Chapter 1 of the standard text, Alberts *et al* (1983), shows a molecular evolution of DNA driven by random polymerisation and random mutation, with a gradual emergence of regions of some functional importance to the organism. These "functional regions", genes or regulatory sequences, tended to stop mutating at some point in evolution. Or rather, the random changes to the DNA polymer still occurred but produced an organism which did not survive to reproduce itself.

Even then, some random change was (and is) still allowed in these "functional" zones. For example, a gene can accept considerable mutation (at the third position of each *codon*) without affecting in any way the blueprint of life contained in the gene [3].

Thus, the view that the DNA molecule is a random sequence of bases (along one of its two strands) prevails, provided we recognise that some regions are still randomly evolving and other

[3] A codon is a unit, comprising three bases, which provides information to the cell. The translation of that information for the cell's benefit is based on a genetic code that has many redundancies. Some random mutation of the DNA is allowed without changing the information that the cell receives.

regions are now fixed, with the result of past random change now "frozen" in time [4].

5 A glib mathematical abstraction

Consider an interval $[0, L]$ as a section of our DNA molecule, with a point process of sites (which are perhaps of some importance because they attract some enzyme). We do not assume that the point process is Poisson, yet it soon becomes one under the following rather-glib model of mutation.

Because of mutation the sites may die (ie. change into a form which has lost its role). This occurs randomly and is like a "thinning" of the point process. There is, however, the genesis of new sites through mutation. These are equally likely to occur anywhere in $[0, L]$ and independently so. Thus this spatial birth and death process quickly gives rise to a Poisson point process.

We assume throughout the sequel of our paper that sites which interest us form a Poisson process along the DNA. Our first example of this is with sites known as "replication origins".

6 The spatial pattern of replication origins

We have mentioned above that a bacterial cell has only one *origin of replication* on its sole DNA molecule. On the DNA of higher organisms, however, there are a large number of "origins" on each of the many DNA molecules that the cell possesses. These origins are randomly scattered along the linear molecule. Due to random evolution of the extremely long DNA molecule, perhaps along the lines of our glib model above, these specific sites can be modelled well by a Poisson process along the DNA molecule.

The replication of DNA is an interesting process (Kornberg, 1980). Each origin 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 both moving 'frontiers', the double-stranded structure unwinds from its helical structure and separates into two single strands. Because of the strand separation, the frontiers are called *forks* and the region of separated DNA between the pair of forks is known as an "eye" or "replicon". Replication of the single-strands in the "eye" then takes place (see later for details of the actual replication mechanism).

When a fork meets a fork from another eye, both forks stop and the enzymes involved in unwinding/separating fall off the DNA. The two eyes simply merge into one larger eye.

Beautifully clear photographs of this process, taken under the electron microscope, appear in Blumenthal, Kriegstein and Hogness (1974) and Kriegstein and Hogness (1973). We draw a typical picture below (Fig. 1). One can see various segments of DNA within the photographic field and numerous fully-visible *eyes* where the DNA has separated into two strands. The authors used the fruit fly, *Drosophila melanogaster*, in the period prior to the so-called *mid-blastula transition*, for their studies. Generation times are extremely short, around 10 minutes, during the first 12 generations of the embryo before the "transition" and DNA replication is complete within about 3-4 minutes. (Later in the fly's development, generation times of cells lengthen markedly, as does the time taken for DNA replication; it is thought that a much smaller proportion of origins are activated in this later period of development.)

[4] to borrow from Francis Crick who described the genetic code itself as a *frozen accident*.

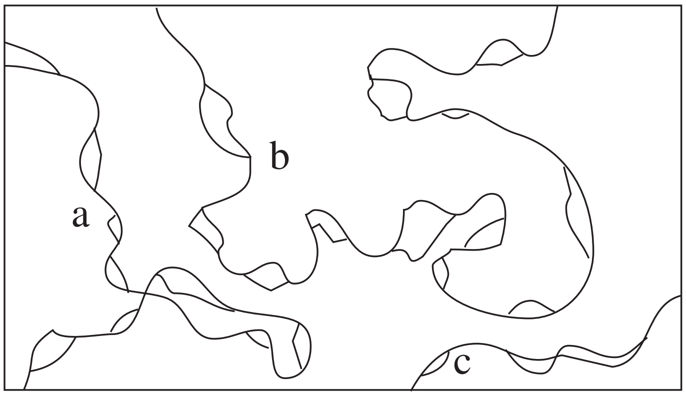


Figure 1: The type of micrograph which shows the *eyes* of replicating DNA. Three DNA segments are seen in the rectangular field: a has 8 fully-visible eyes and one eye truncated by the boundary; b has 12 fully-visible eyes; c has two, one of which crosses over itself.

The lengths of the eyes extant at the time the photo is taken are quite variable, as are the distances between eyes. These are evidenced by histograms compiled from Blumenthal *et al* (1974).

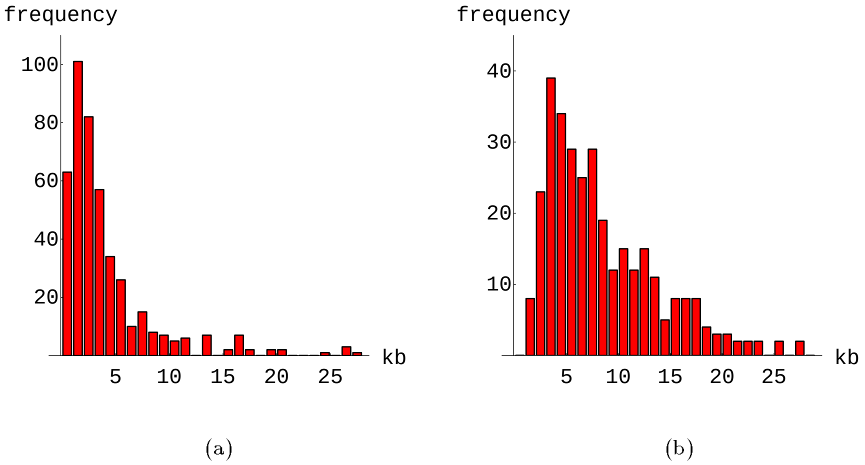


Figure 2: (a) The histogram of eye length for 439 fully visible eyes; (b) the histogram of eye-to-eye distance as measured along the DNA from centre to centre of fully-visible adjacent eyes.

There are some difficulties in the interpretation of these histograms. One problem, of which the authors were well aware, was that the eyes were observed from molecules at different stages of the process. To combat this problem, they also published data on the eye-to-eye distances for molecular segments grouped according to the proportion of the molecule which has separated into two strands (a good indication of the stage in the process). These also show great variability.

There are other difficulties of interpretation due to the sampling protocols used and calculations employed by the authors (for example, in the treatment of eyes truncated by the edge of the photo and in the use of the full eyes and eye-to-eye distances only from segments with at least two full eyes). The authors do not adequately deal with the truncation and size-biassing effects encountered. The main message we draw now, however, is simply that the entities such as eye length and eye-to-eye distance show considerable variability; this provides the stochastic interest for the probabilist.

In Cowan, Chiu and Holst (1995), a stochastic model of this "strand-separation" process was posed. The Poisson process of origins was given an 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. There is a reasonable sized pool of enzyme complexes so, as we discussed above, an exponential distribution makes a plausible assumption.

If, in the waiting for an enzyme complex, a site has not been passed over by a moving fork initiated elsewhere, molecular binding of the complex takes place and the two bidirectional forks commence at the site. The forks move at constant speed r in each direction. The assumption of constant speed is reasonable and accords with experimental evidence. In reality, one might expect that the time to unwind/separate the two strands a distance of just *one* base might indeed be a random variable, but thousands of these small steps are involved in rapid succession, so variability is not observed in the total time to unwind (say) a thousand bases, variability being nullified by the strong law of large numbers. For *Drosophila*, Blumenthal *et al* estimate fork speeds of 2,600 bases per minute.

Let T_L be the time when a section of length L, represented by $[0, L]$, is completely separated into single strands. In the paper, the authors found a useful asymptotic result for the distribution of T_L as $L \to \infty$. Their methods drew upon mathematical results by Janson (1983) and Hall (1988) on coverage processes.

We now look at the Cowan/Chiu/Holst model in some detail.

7 The time to separation of a long DNA molecule

Janson (1983) and Hall (1988) considered a coverage problem which relates to the model of Cowan, Chiu and Holst. 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 approached but 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

$$\begin{aligned} \mathbb{P}\{U \leq x\} &= \mathbb{P}\{v(t - V) \leq x | V \leq t\} \\ &= \frac{e^{\mu(t - x/v)} - e^{-\mu t}}{1 - e^{-\mu t}} \quad 0 < x \leq vt, \end{aligned} \tag{3}$$

since

$$\mathbb{P}\{V \leq x\} = 1 - e^{-\mu x}, \quad x > 0. \tag{4}$$

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 (and it is convenient to allow it in this analysis so that we can draw on the ideas of Janson and Hall).

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

$$\frac{e^{a\lambda}}{\lambda} \to e^u \tag{5}$$

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

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

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

Theorem 1. For each real number u,

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

The 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 $u \in \mathsf{R}$ such that (5) holds. It is necessary that t changes with L. Suppose, for $c_1 > 0$,

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

Then (5) holds, provided $c_1 = 1/\gamma v$ and $c_0 = 1/\mu + (u + \log\gamma)/\gamma v$. Suppose therefore that for some $u \in \mathsf{R}$

$$t = \frac{\log(\gamma L) + u}{\gamma v} + \frac{1}{\mu}. \tag{6}$$

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

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

With relationship (6) between t and L, (5) 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 (3), (4) and (6) the segment-lengths after rescaling have distribution function H_t given by

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

$H_t(x)$ 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 the current 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 $\exp(a_t \lambda_t)/\lambda_t \to e^u$),

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

Cowan, Chui and Holst then prove that (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 $L \to \infty$,

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

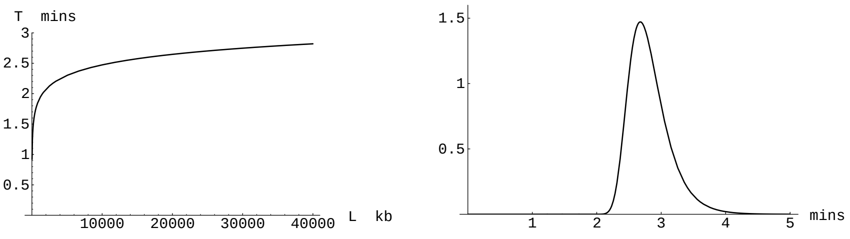
which proves the theorem. Note that the distribution function on the RHS has mean 0.577 (Euler's constant) and variance $\pi^2/6$.

Remark 1. A consequence of the result above is that

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

Thus for extremely large L, T_L is near to the constant $\log(\gamma L)/v\gamma$. Interestingly, for all the stochasticity in the model, the result is close to a constant for very large L. $\mathbb{E}T_L$ grows slowly like $\log L$ (see Fig. 3(a)) but $\mathbb{V}\mathrm{ar}T_L$ settles down to $\pi^2/(6v^2\gamma^2)$.

The longest chromosome in *Drosophila melanogaster* is 62,000 kb. Allowing 40,000 for the longer of the two arms on either side of its centromere, we see below in Figure 3(b) a sample plot of the approximate pdf of T_L. Note that the variance is still appreciable for $L = 40,000$ kb and the chosen parameter values.



(a) (b)

Figure 3: (a) The slow logarithmic growth of the mean of T_L for lengths of DNA (in kb) up to the longest chromosomal arm of the fruit fly; (b) the approximate pdf of T_L for $L = 40,000$ kb. In both plots, $v = 8$ kb/min, $1/\gamma = 2$ kb and $1/\mu = 12$ secs .

Remark 2. Cowan, Chiu and Holst took care not to refer to L as the length of the DNA molecule, merely that $[0, L]$ was a long section of the molecule. The DNA was modelled by the whole real line. So their model is not totally accurate since DNA has finite length, albeit very long relative to $1/\gamma$. Arguably, as they remark, 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 (6).

Remark 3. The work of Cowan/Chiu/Holst relates to "a DNA model" studied by Vanderbei and Shepp (1988) and, in a context unrelated to DNA, by Quine and Robinson (1990, 1992). In the former model of Vanderbei/Shepp, 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 fork frontier, that is, according to a Poisson process which is stationary in space-time. This model is not appropriate for DNA replication, because (as we have seen) 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, Cowan, Chiu and Holst also applied their method to this related model. As a result, they were able to prove rigorously the results suggested by Vanderbei and Schepp.)

Remark 4. Holst, Quine and Robinson (1996) have addressed the distribution of T_L for finite L. They have found the Laplace transform of a related quantity and, in principle, this is invertible and the exact distribution found. They have not done the inversion, but as this chapter was going to press, Chiu and Yin (2000) have found the inverse Laplace transform. They present an explicit, though complicated, expression for the distribution of T_L. See these papers for details and generalisations.

8 The proportion of origins initiated

It is clear that some origins are not initiated during the process of strand separation described above. In the Cowan/Chiu/Holst model, it is possible to find, at least for large L, the probability that a given origin O is initiated and not passed over by the fork of an eye initiated elsewhere.

Let us condition on the time S until O is first approached by an enzyme complex. Suppose $S = s$. Let $q(s)$ be the probability that a fork from the right of O passes over O before time s. Note that the only origins lying to the right of O which can initiate moving forks in time to reach O by time s, are those within distance rs. Suppose there are $k > 0$ of these. Given k, the locations $U_1, U_2, \ldots U_k$ of these k origins are independently distributed, uniformly on $[0, rs]$.

The times $S_1, S_2, \ldots S_k$ at which these k origins are approached by enzyme are independent and identically distributed with pdf $\gamma e^{-\gamma t}$. Thus, for the ith such origin, the chance that it is not approached earlier enough for a frontier to reach O by time s is

$$\begin{aligned}\mathbb{P}\{S_i > s - \frac{U_i}{r}\} &= \int_0^{rs} e^{\gamma(s-u/r)} \frac{du}{rs} \\ &= \frac{1}{\gamma s}(1 - e^{-\gamma s}).\end{aligned}$$

The origin O is not passed over from the right by time s if and only if all k such sites are approached too late, that is with probability (given k and s) of

$$\left[\frac{1}{\gamma s}(1 - e^{-\gamma s})\right]^k.$$

Therefore, after incorporating the case $k = 0$ as well,

$$q(s) = \sum_{k=0}^{\infty} \frac{(\mu rs)^k e^{\mu rs}}{k!} \frac{(1 - e^{\gamma s})^k}{(\gamma s)^k} = \exp\left[\frac{\mu r}{\gamma}(1 - \gamma s - e^{-\gamma s})\right].$$

Clearly, the chance that O is not passed over by time s from either the right or left is $[q(s)]^2$. Therefore, unconditional upon s, the probability that origin O is not passed over, and so initiates its own eye, is, with v replacing $2r$,

$$\int_0^{\infty} \gamma e^{-\gamma s}[q(s)]^2 ds = \int_0^{\infty} e^{-w} \exp\left[\frac{\mu v}{\gamma}(1 - w - e^{-w})\right] dw. \tag{9}$$

This result can be found in Chiu (1992) and, in a quite unrelated context, in the work of McPeek and Speed (1995). There is also a generalised formula for the case where the space-time positions of "enzyme approaches to sites" form a more general space-time Poisson process (Chiu and Quine, 1997).

9 Mean eye lengths and eye–to–eye distances

Blumenthal *et al* collect statistics on mean eye lengths, c say, and mean eye-to-eye distances, d say (see their Figure 6). We do not reproduce their numbers, because, as mentioned above, we question their statistical protocol. We shall, however, attempt to emulate the broad character of their Figure 6, if not an exact fit.

Their figure plots the two means against the proportion of the molecule which has separated into eyes. We shall therefore try to relate these entities theoretically in the Cowan/Chiu/Holst model. We are able to do so at various times t, assuming a very long DNA molecule (indeed infinite, as we model the DNA molecule by the real line).

Consider again the preamble to equation (3). At time t, the origins which have been approached by an enzyme complex form a point process of intensity $\lambda_t = \gamma(1 - e^{-\mu t})$. The distribution function of "line segment" which has grown at these sites is given by (3). The mean length is

$$\begin{aligned} a_t = &\int_0^{vt} \left[1 - \frac{e^{\mu(t-x/v)} - e^{-\mu t}}{1 - e^{-\mu t}}\right] dx \\ = &\frac{vt}{1 - e^{-\mu t}} - \frac{v}{\mu}. \end{aligned} \tag{10}$$

The union of all these line segments is a realisation of the Boolean process of intervals on the line. Coverage results for such a process are summarised in Hall (1988, p.89) in terms of the mean value a of line-segment lengths and the intensity λ of nuclei for these segments (Hall uses α instead of a). We have shown these entities above as a function of t.

Thus we can cite Hall's results. Let p_t be the proportion of the line covered by line segments at time t. This is given [Hall's (3.4)] as

$$p_t = 1 - e^{-a_t \lambda_t}. \tag{11}$$

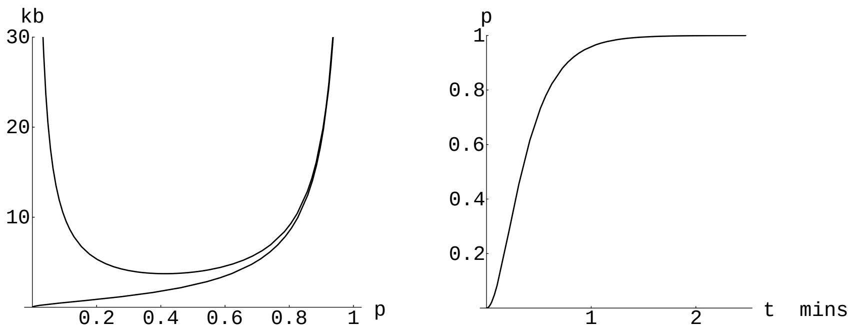
The mean eye length (Hall's clump size, Thm. 2.2) has mean

$$c_t = \frac{e^{a_t \lambda_t} - 1}{\lambda_t} \tag{12}$$

whilst the mean gap of DNA separating two eyes can easily be shown to have mean $1/(a_t \lambda_t)$. Thus the mean eye–to–eye distance is

$$d_t = \frac{1}{a_t \lambda_t} + \frac{e^{a_t \lambda_t} - 1}{\lambda_t}. \tag{13}$$

We seek to relate both mean eye length c and mean eye-to-eye distance d to p. This is achieved by eliminating t from the equations above, since it is the common dependence on t which creates the sought relationship. Elimination of t is not possible by simple algebraic manipulation.



(a) (b)

Figure 4: (a) The upper curve shows the mean eye–to–eye distance d as a function of p, the proportion of a long DNA molecule which has separated. The lower curve shows the mean eye length c as a function of p. (b) p as a function of t. In both plots, $v = 8$ kb/min, $1/\gamma = 2$ kb and $1/\mu = 12 secs$.

We can, however, prepare "parametric plots" using a package such as *Mathematica*; typical parametric plots are shown in Figure 4(a). Figure 4(b) plots p_t against t. The parameter values chosen for Figure 4 mean that fork speed is 4 kb/min, origins are on average 2 kb apart and the mean time until an origin is first approached by an enzyme complex is 12 seconds. From (9), I have calculated that 13.19% of origins are actually initiated when these parameter values are used.

My Figure 4(a) is of the same general character as Figure 6 in Blumenthal *et al.* There is not an exact numerical fit between the two figures and I have not attempted to find one. Some explorations suggest that the model will not capture the flatness of the plateau for d evident in their figure. I am not in a position to know if this is due to a failure of the Cowan/Chiu/Holst model or to inappropriate statistical calculations in the paper of Blumenthal *et al* or to inapplicability of plots such as my Figure 4(b) when the DNA segments analysed are not extremely long.

I leave these as *open* issues and move on to other problems associated with DNA replication.

10 What is happening inside the eyes?

We have indicated that the separation of the DNA into 2 single strands allows these strands to be copied. The single strand acts as a template for the building of the copy; for example, an A on the template attracts a T to bind opposite, and vice versa, whilst C and G have similar complementarity. Thus, due to this base-pairing phenomenon, the copy formed is identical to the strand that was formerly bound to the template before the separation into the "eye" form.

This copying does not, however, rely on the attraction of A to T and C to G by chemical diffusion alone. It is enzyme mediated; an enzyme binds to the DNA and rapidly finds, from the neighbouring solution, the right match to place directly opposite each base in the template. The enzyme acts as a catalyst and makes the binding quick and efficient.

How does this start? In the most popular theory of DNA replication (Kornberg, 1980; Huberman and Horwitz, 1973; Denhardt and Faust,1986), two distinct enzymic mechanisms are postulated. The first involves the same enzyme which separates the DNA. As this enzyme moves along the double-stranded molecule, separating the DNA and pushing the fork further away from the eye's "origin", its rear portion remains in contact with one of the two single strands behind it. This rear portion attracts the right base to match each base on this particular single strand, immediately converting the strand back into a double-stranded molecule. If the eye's origin is called O, then one strand is immediately replicated between O and the right-moving fork whilst the other strand is immediately replicated between O and the left-moving fork (there being a location O on both strands after separation occurs). Figure 5(a) shows this process.

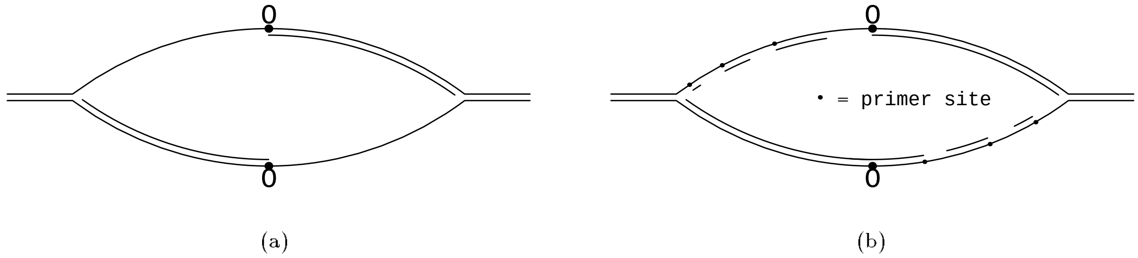


Figure 5: (a) An eye and the effect on the leading strand of replication by the same enzyme that separates the two strands. (b) The discontinuous replication on the lagging strand from primer sites, with $b/r = 0.6$.

Note, however, that half of the DNA is not replicated by this enzymic method, since the rear portion of the enzyme retains contact with only one strand, the so-called *leading* strand of that replicating fork. The strand that it loses contact with, the so-called *lagging* strand, does not enjoy the replicating assistance of this "separating" enzyme (and so *lags* somewhat in its replication).

Nature has devised another enzymic mechanism to replicate the lagging strand. Scattered randomly along a single strand of DNA are very short sequences of "letters" – specific letters but the sequence is not currently known. These letter patterns attract short fragments of RNA known as *primers*; thus the sequence is called a *primer site*. When the primer lodges on the primer site, by random diffusion and chemical attraction, another enzyme known as *DNA polymerase* also lodges and starts the process of building the DNA copy on the single strand (by acting as the catalyst for base pair attraction and moving along the DNA, creating the correct base pairing as it goes).

Noteworthy, however, is that this enzyme can only function in the "moving catalyst" role in a given direction, namely away from the fork toward O. Thus, if we could look at the eye centred at O (before any of the complications that arise when eyes merge), we would see a pattern as in Figure 5(b). (Note that the magnification in micrographs such as Figure 1 is insufficient to see this detail.)

As Figure 5(b) shows, there are newly-copied 'islands' on the lagging strand adjacent to the fork (i.e. fragments of newly-created double-stranded DNA set on the single-stranded material). The first such fragment formed on (say) the right-hand side of O in the figure grows leftwards and soon reaches O, where it joins the large section of double-stranded DNA already in place to the left of O. Eventually each fragment on the right-hand side of O joins to the fragment on its left, and so becomes connected to the origin O in due course. In contrast to our 'islands', we refer to the collection of joined fragments connected to 0 as the 'mainland'. Any fragment which becomes connected to the mainland is no longer called a 'fragment'.

If the process is interrupted at time t, and the DNA heated so that all strands separate, short fragments of single-stranded DNA can be observed (our islands); these have been named Okazaki fragments after Okazaki *et al.* (1968) who observed these fragments in prokaryotic cells. (Huberman and Horwitz (1973) later established the existence of such fragments in eukaryotic cells.)

Of course the mean length of Okazaki fragments does not equal the mean inter-primer-site distance. Some Okazaki fragments will comprise just one inter-site spacing which is not completely copied and others will be the amalgam of numerous copied and partly copied inter-site spacings. The investigation of the relationship between mean length of Okazaki fragments, our 'islands', and the mean inter-site distance, μ say, was the major aim of a paper by Cowan and Chiu (1994). They formulated a stochastic model of fragment formation and, using data on mean lengths of Okazaki fragments, showed that the unknown μ can be estimated.

I give an account of their model, together with some useful recent analysis by Piau (2000).

11 The Cowan-Chiu model of fragment formation

We focus on the lagging strand to the right of O in Figure 5(b). Measured along this strand, let X_1 be the distance between O and the first primer site and X_i the distance between the $(i-1)$st and the ith sites $(i = 2, 3, \cdots)$. It is readily shown that for almost any plausible model of DNA structure, these inter-site distances are approximately exponentially distributed, so we model the sites by a stationary Poisson process of intensity $1/\mu$.

Denote the fork speed by r and define $Y_i = X_i/r$. Thus $\{\sum_{i=1}^{n} Y_i\}_{n=1}^{\infty}$ is the Poisson process of times at which the right-moving fork passes the primer sites of the lagging strand. Site i becomes exposed to primers at time Y_i and, if there is no delay[5], the fragment at this ith site starts immediately and grows leftward at rate b. This temporal Poisson process of primer sites "emerging" from the fork has intensity $\lambda = r/\mu$.

We assume that, when the fragment grows to reach the previous primer site, there is a time δ for the complete joining of the fragment to its neighbouring fragment (which may by this time have joined the mainland). Let D_t denote the distance between the fork and the nearest end of the mainland at time t, whilst P_t denotes the cumulative length of all copying up until time t. Also define L_t as the total length of extant Okazaki fragments at time t, whilst N_t is the number of such fragments at time t. Note that N_t decreases by one whenever two fragments join or when a fragment joins to the mainland, but increases when a primer site emerges from the fork. Since the length of the mainland equals $P_t - L_t$ and also equals $rt - D_t$, we have the identity

$$L_t = P_t - (rt - D_t). \tag{14}$$

The mean Okazaki-fragment length is effectively equal to $\mathbb{E}(L_t)/\mathbb{E}(N_t)$. To see this, note that experiments which try to estimate the mean length of Okazaki fragments are performed with cultures containing a massive number of cells. So there are large numbers of active forks at any time in the experiment (especially so in the eukaryote case where each cell has thousands of active forks). If fork i has variates L_t^i and N_t^i, the observed mean is $\sum_i L_t^i / \sum N_t^i$. The forks are independent and their number large, so the strong law of large numbers ensures that this ratio effectively equals $\mathbb{E}(L_t)/\mathbb{E}(N_t)$.

12 The expectations of N_t and P_t: renewal equations

We find $\mathbb{E}(N_t)$ first, using an argument which conditions on Y_1 and exploits the regenerative situation at time Y_1. Denoting $\mathbb{E}(N_t)$ by $n(t)$, $\mathbb{E}(N_t|Y_1 = y)$ by $n(t|y)$ and for brevity $b/(b+r)$ by a, one finds that $n(t|y) = 0$ for $y \geq t$. So for $t \geq 0$,

$$n(t) = \int_0^t n(t|y)\lambda e^{-\lambda y} dy.$$

When $t \geq \delta$, we have the breakdown

$$\begin{aligned} n(t) &= \int_0^{a(t-\delta)} n(t-y)\lambda e^{-\lambda y} dy + \int_{a(t-\delta)}^t (1 + n(t-y))\lambda e^{-\lambda y} dy \\ &= e^{-\lambda a(t-\delta)} - e^{-\lambda t} + \int_0^t n(t-y)\lambda e^{-\lambda y} dy, \end{aligned}$$

[5]We assume no delay. This is for mathematical convenience but also fairly realistic. If there is a delay it should be random in duration, presumably exponentially distributed, but with a very small mean due to the obvious abundance of the primers.

whilst for $0 \leq t < \delta$,

$$
\begin{aligned}
n(t) &= \int_0^t (1 + n(t-y))\lambda e^{-\lambda y} dy \\
&= 1 - e^{-\lambda t} + \int_0^t n(t-y)\lambda e^{-\lambda y} dy.
\end{aligned}
$$

Therefore, for all $t \geq 0$

$$
n(t) = m(t) + \int_0^t n(t-y)\lambda e^{-\lambda y} dy,
$$

where $m(t) = \exp(-\lambda a[t-\delta]_+) - e^{-\lambda t}$ and $[x]_+ := \max(x, 0)$. Applying standard methods for this typical renewal equation, we find the solution

$$
n(t) = m(t) + \lambda \int_0^t m(u) du. \tag{15}
$$

So

$$
\begin{aligned}
n(t) &= \lambda t & t < \delta \\
&= \lambda\delta + \frac{r}{b}\big(1 - e^{-\lambda a(t-\delta)}\big) & t \geq \delta.
\end{aligned} \tag{16}
$$

Therefore

$$
n := \lim_{t\to\infty} \mathbb{E}(N_t) = \lambda\delta + \frac{r}{b}. \tag{17}
$$

This result can also be derived by Little's formula, under quite general point-process models for the primer sites. Consider "births" of Okazaki fragments, at rate λ, when primer sites emerge form the fork and "deaths" when the fragments join to the previous fragment. Since μ equals r/λ, ν, the mean time between birth and death, is $\delta + r/(\lambda b)$. So by Little's formula, the expected number alive at equilibrium is $\nu\lambda$, or $\lambda\delta + r/b$.

As an aside, we note that similar regenerative arguments can be used to find higher moments of N_t. We merely note for the record that

$$
\mathbb{E}(N_t^2) \to \frac{(b + 2r + 2b\lambda\delta)e^{\lambda a\delta}}{ab} - \frac{2(b+r)^3 e^{\lambda(2-a)a\delta}}{b^2(b+2r)} + 1, \qquad t \to \infty.
$$

A similar regenerative argument also yields an integral equation for $\mathbb{E}(P_t)$. Condition on $Y_1 = y$. If $0 < y \leq at$, the copying at the first primer site has progressed ry. If $at < y \leq t$, the copying has progressed $b(t-y)$. Defining $p(t) := \mathbb{E}(P_t)$ and $p(t|y) := \mathbb{E}(P_t|Y_1 = y)$, we have that for $t \geq 0$,

$$
\begin{aligned}
p(t) &= \int_0^t p(t|y)\lambda e^{-\lambda y} dy \\
&= \int_0^{at} (ry + p(t-y))\lambda e^{-\lambda y} dy + \int_{at}^t (b(t-y) + p(t-y))\lambda e^{-\lambda y} dy \\
&= q(t) + \int_0^t p(t-y)\lambda e^{-\lambda y} dy
\end{aligned}
$$

where $q(t) = \big[r + be^{-\lambda t} - (r+b)e^{-\lambda at}\big]/\lambda$. It follows from the form of (15) that for $t \geq 0$,

$$p(t) = rt - \frac{r}{\lambda a}\big(1 - e^{-\lambda at}\big).$$

As $t \to \infty$, $rt - \mathbb{E}(P_t) \to r/\lambda a$. Thus if $\ell := \lim_{t\to\infty} \mathbb{E}(L_t)$ and $d := \lim_{t\to\infty} \mathbb{E}(D_t)$, then from (14),

$$\ell = d - r/\lambda a. \tag{18}$$

13 The expectation of D_t: the quasi-renewal equation.

To find an equation for $\mathbb{E}(D_t)$, again condition on $Y_1 = y$. Consider $t \geq \delta$. If $y > a(t-\delta)$, then the copying started at the first primer site has not reached the mainland, so the distance between the fork and the mainland is rt. If $0 < y \leq a(t-\delta)$, it has reached the mainland and so the said expected distance is equal to $\mathbb{E}(D_{t-y})$. Define $d(t) := \mathbb{E}(D_t)$ and $d(t|y) := \mathbb{E}(D_t|Y_1 = y)$. Clearly $d(t) = rt$ for $t < \delta$, whilst for $t \geq \delta$

$$\begin{aligned} d(t) &= \int_0^\infty d(t|y)\lambda e^{-\lambda y} dy \\ &= \int_0^{a(t-\delta)} d(t-y)\lambda e^{-\lambda y} dy + \int_{a(t-\delta)}^\infty rt\lambda e^{-\lambda y} dy \\ &= rte^{-\lambda a(t-\delta)} + \int_0^{a(t-\delta)} d(t-y)\lambda e^{-\lambda y} dy. \end{aligned}$$

Define $d_\delta(t) := d(\delta + t)$ for $t \geq 0$ and $\delta \geq 0$. Then for $t \geq 0$, $\delta \geq 0$

$$d_\delta(t) = r(t+\delta)e^{-\lambda at} + \int_0^{at} d_\delta(t-y)\lambda e^{-\lambda y} dy. \tag{19}$$

This is not a renewal equation since $a < 1$. For reasons which will become clear, we define a new function g_δ by $g_\delta(t) := d_\delta(t)/r$. When (19) is suitably modified, to yield (20), we note that g_δ depends on r and b only via a, which itself depends on r and b solely via the ratio b/r.

$$g_\delta(t) = (t+\delta)e^{-\lambda at} + \int_0^{at} g_\delta(t-y)\lambda e^{-\lambda y} dy. \tag{20}$$

It is a simple matter to show, from (20), that

$$g_\delta(t) = g_0(t) + \delta.$$

Thus we may focus our attention on the case $\delta = 0$, whereby (20) looks like

$$g_0(t) = te^{-\lambda at} + \int_0^{at} g_0(t-y)\lambda e^{-\lambda y} dy. \tag{21}$$

Let us call equations of this type, *quasi-renewal equations*.

Cowan and Chiu were unable to find an explicit analytic solution to equation (21) and adopted a numerical approach. They presented graphs of g_0 versus t for a range of b/r, with λ set at 1.

They remark that the limit of g_0 as $t \to \infty$ is important to the theory, since $\mathbb{E}(D_t)/r \to d/r = d_\delta(\infty)/r = g_\delta(\infty) = g_0(\infty) + \delta$. They were, however, only able to give numerical calculations for $g_0(\infty)$.

Recently, Piau (2000) has made a study of the quasi-renewal equation, deriving explicit expressions for the limiting value of the solution as $t \to \infty$. Applying his results to (20), he derived the satisfying analytic result

$$g_0(\infty) = \frac{1}{\lambda(1-a)} \sum_{n \geq 1} \left[1 - \prod_{k \geq n} [1 + (1-a)^k] \right]. \tag{22}$$

Piau's calculations with (22) give numerical answers close to those of Cowan and Chiu, albeit slightly higher.

14 Relationship between fragment length and primer-site spacing

A major concern is the ratio at equilibrium, R say, of mean Okazaki-fragment length to mean inter-primer-site distance, that is,

$$R = \frac{\ell/n}{r/\lambda} = \frac{\lambda g_0(\infty) + \lambda\delta - 1/a}{\lambda\delta + r/b} \tag{23}$$

from (17) and (18). Note that R depends on b and r solely via b/r. Figure 6 shows a plot of R for a range of b/r, with λ set at 1 and $\delta = 0$.

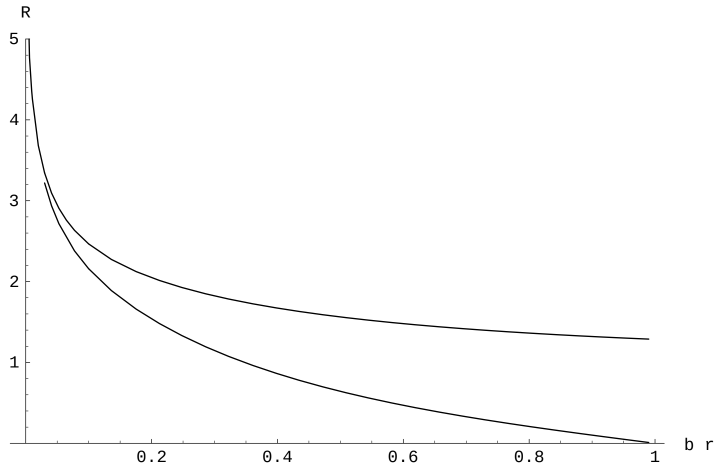


Figure 6: The upper curve shows R, the ratio of mean Okazaki-fragment length to mean inter-primer-site distance, plotted against b/r. The lower curve shows R^*, an entity which arises in the competing model (see sections 16–19). The two curves merge for very small values of b/r, so, in order to avoid clutter, the R^* plot is not shown completely.

When $\lambda \neq 1$, Cowan and Chiu show that, if $\alpha := b/r$, then $n(t, \alpha, \lambda, \delta) = n(\lambda t, \alpha, 1, \lambda\delta)$, where the expression is now written as a function of four variables. Likewise, they show that $\lambda g_\delta(t, \alpha, \lambda) = g_{\lambda\delta}(\lambda t, \alpha, 1)$ for any $\delta \geq 0$. Equation (23) shows that R is a function of α, δ and λ, namely

$$R(\alpha, \delta, \lambda) = \frac{\lambda g_0(\infty, \alpha, \lambda) + \lambda\delta - (1+\alpha)/\alpha}{\lambda\delta + 1/\alpha}.$$

So one can clearly see that, for all λ,

$$R(\alpha, \delta, \lambda) = R(\alpha, \lambda\delta, 1).$$

Thus Figure 6, based on $\delta = 0$ and $\lambda = 1$, has applicability for general λ since $R(\alpha, 0, \lambda) = R(\alpha, 0, 1)$. For $\delta \neq 0$, the use of (23) combined with Figure 6 easily yields the answers needed.

15 Estimated spacing between primer sites

The pattern of 'bases' which form a primer site is not known. It would be useful to know the expected distance between primer sites as this would give a guide to the number of invariant 'bases' in the primer site. For example, if the site had 4 invariant bases (e.g. $GCTA$ or $GXAXGT$ where X = anything), then the approximate average distance between sites is $4^4 = 256$ bases.

The Cowan/Chiu theory, combined with data on the average length of Okazaki fragments, provides an estimate of the mean inter-primer-site distance and hence some information on the primer site pattern. Experimental data give mean Okazaki fragment length of 150 bases in eukaryotes and 1500 bases in prokaryotes (Kornberg, 1980).

Good data on the parameters b, r and δ are not available, but it is known roughly that b/r is about 0.6 for eukaryotes and perhaps 100 times less than this for prokaryotes. From (23) and Piau's formula (22), assuming $\delta = 0$ in the absence of better experimental knowledge, R is about 1.48 for eukaryotes and 4.77 for prokaryotes. Thus the mean inter-primer site distance is respectively $150/1.48 = 101$ and $1500/4.77 = 314$.

The data on which these calculations are based are very rough, but this calculation already suggests that, if these species have a common primer-site structure, then it might have 4 invariant bases and be spaced on average every 256 bases. The near contenders, 3 or 5 invariant bases, would lead to spacings of about 64 or 1024 bases respectively.

16 The competing theory

The discussion of Okazaki fragments cannot be closed without reference to an alternative theory considered by some biologists. We have stated, in building the model above, that the DNA copying on the *leading* strand is continuous due to the role of the unwinding/separating enzyme as a catalyst in the copying of the leading strand.

There have been suggestions, however, that replication of the *leading* strand is also done in discontinuous pieces and that the unwinding/separating enzyme plays no role at all in copying. Some experiments point to this (Callan, 1973; Kowalski and Denhardt, 1982). The putative mechanism is essentially similar to that described for the lagging strand, namely the lodging of primers at specific primer sites on the DNA and the subsequent copying by the moving *polymerase* enzyme. The only difference is that, because of the strict directionality of the polymerase movement along the DNA, the polymerase on the leading strand is moving toward the fork (in contrast to its movement away from the fork on the lagging strand). There would be a constraint, of course, that the speed b of the polymerase *must* be no greater than the fork speed r or else the polymerase would be bumping into the fork enzyme. Figure 7 illustrates this competing theory.

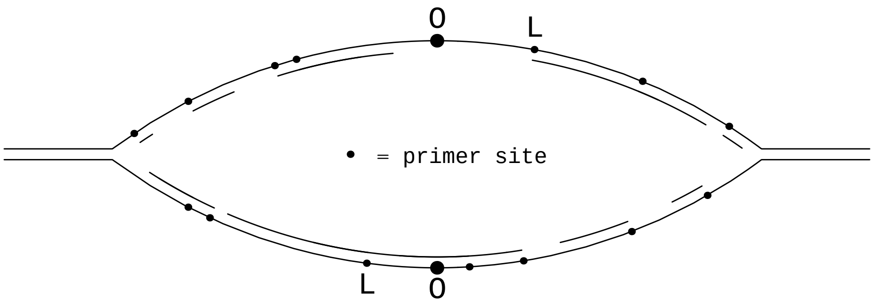


Figure 7: The discontinuous-replication model for both the leading and lagging strands. Here $b/r = 0.6$. There are 3 "islands" and one "mainland" on the lower strand, the mainland being the piece which covers the "leading site" marked L. On the upper-strand there are 4 "islands" and the "mainland" which (at this early stage in the eye's development) is the piece with the "leading site " L at its left-hand end.

17 Notations for the competing theory

We now present an analysis of the Cowan/Chiu model, modified to take into account this competing theory. We follow the unpublished report of Cowan and Chiu (1992) in calling the new model, "Model B", and the earlier model, "Model A".

Let us focus on the lower strand of Figure 7. Okazaki fragments are present in both the leading strand (of the fork to the left of O) and the lagging strand (of the fork to the right of O). In model A, the 'mainland' commences at time zero because it is assumed that copying action starts immediately on the leading strand at O. In model B, we need to define what is meant by 'mainland', since polymerase activity is not initiated on the leading strand until after the leading-strand fork passes the first primer site on the leading strand (call this the *leading site*).

Let $X_1,\ X_2,\ \cdots$ be defined as before for the lagging strand and define $X_1^*,\ X_2^*,\ \cdots$ as the equivalent quantities on the leading strand. Thus the leading site is distance X_1^* from O. A polymerase lodges on this site at time Y_1^*, where $Y_i^* := X_i^*/r$. Prior to this time, we define the mainland to be solely the leading site itself. At later times, the mainland is the contiguous section of double-stranded DNA which either *covers* the leading site or has one end at the leading site. Thus D_t equals $rt + X_1^*$ when $t < Y_1^*$ and D_t^*, the distance from the leading-strand fork to the mainland, is $rt - X_1^*$ for this same range of t.

Since one polymerase which lodges on the lagging strand can pass through O and travel on the leading strand, we need to be precise in our definitions of variables such as N_t, L_t and P_t. An Okazaki fragment is defined to be a *lagging*-strand fragment (resp. *leading*-strand fragment) if its polymerase lodged on the lagging strand (resp. leading strand). So, for example, P_t is the cumulative distance travelled by polymerases which lodge on the lagging strand. P_t^* is similarly defined for the leading strand, as are N_t^*, L_t^* and the functions n^*, p^*, $\cdots$.

18 Analysis of the lagging strand for Model B

Lagging-strand properties are changed slightly from Model A. The methodology is simple and illustrated by $\mathrm{E}(N_t)$. Condition on $\{Y_1^* = z\}$ and let $n(t|z)$ be $\mathrm{E}(N_t|Y_1^* = z)$, different in meaning to that ascribed to $n(t|y)$ in Section 12. Then

$$n(t) = \int_0^\infty n(t|z)\lambda e^{-\lambda z}\, dz \tag{24}$$

and, by further conditioning on $\{Y_1 = y\}$,

$$\begin{aligned} n(t|z) &= \int_0^t \big[1 + n(t-y|0)\big]\lambda e^{-\lambda y}\, dy \qquad t < \delta + z/\alpha \\ &= \int_0^{a(t-\delta-z/\alpha)} n(t-y|0)\lambda e^{-\lambda y} dy + \int_{a(t-\delta-z/\alpha)}^t \big[1 + n(t-y|0)\big]\lambda e^{-\lambda y}\, dy \end{aligned}$$

when $t \geq \delta + z/\alpha$. Recall $\alpha = b/r$. Note that $n(\cdot|0)$ is known, from Model A, as the n in (16). So $n(t|z)$ can be calculated directly, and (24) used to show that, for Model B,

$$\begin{aligned} n(t) &= \lambda t & 0 \leq t < \delta \\ &= \lambda\delta + \frac{1}{\alpha}\big(1 - e^{-\lambda\alpha(t-\delta)}\big) & t \geq \delta, \\ \mathbb{E}(N) &= \lambda\delta + \frac{1}{\alpha}. \end{aligned} \tag{25}$$

For finite t, this result differs from the equivalent result for Model A, but asymptotically, there is no difference. A similar methodology shows that

$$\mathbb{E}(P_t) = rt - \frac{r}{\lambda\alpha}(1 - e^{-\lambda\alpha t}) \qquad t \geq 0.$$

As $t \to \infty$, $rt - \mathbb{E}(P_t) \to r/(\lambda\alpha)$. It can also be readily calculated that, although $d(t)$ differs between Models A and B for finite t, $d(\infty)$ is the same in both cases (essentially the stochastic process D_t is equal under both models except for a short transient difference).

Due to our convention that the leading site is the separation point between lagging and leading strands, (14) no longer holds. It is replaced by

$$L_t = P_t - rt + D_t - X_1^*.$$

Yet, as $t \to \infty$, we find that (18) still holds. So

$$\mathbb{E}(L) = rg_0(\infty) + r\delta - \frac{r}{\lambda a} \tag{26}$$

where $g_0(\infty)$ is the same as in Model A. The average length of lagging-strand Okazaki fragments is $\mathbb{E}(L)/\mathbb{E}(N)$, easily calculated from (26) and (25).

19 Analysis of the leading strand for Model B

The properties of the leading strand can be derived using a similar methodology. We firstly suppose $\{Y_1^* = 0\}$ and analyse the leading strand, using regenerative arguments. Then we condition on $\{Y_1^* = z\}$, find simple linkages with the results under $\{Y_1^* = 0\}$, and finally integrate over z. Details are omitted. We merely report that

$$\begin{aligned} n^*(t) &= n(t) - 1 + e^{-\lambda t} & t \geq 0, \\ \mathbb{E}(N^*) &:= \lim_{t\to\infty} \mathbb{E}(N_t^*) = \mathbb{E}(N) - 1, \\ p^*(t) &= p(t). \end{aligned} \tag{27}$$

Thus, as $t \to \infty$, $rt - \mathbb{E}(P_t^*) \to r/(\lambda\alpha)$, the same as on the lagging strand. We can also show that, if $g_0^*(t)$ is defined as $\mathbb{E}(D_t^*)/r$ for $t \geq 0$ when $\delta = 0$ and if $g_0^*(t|z)$ denotes $\mathbb{E}(D_t^*|Y_1^* = z)$, then $g_0^*(\cdot|0)$ satisfies

$$g_0^*(t|0) = (1-\alpha)te^{-\lambda\alpha t} + \int_0^{\alpha t} g_0^*(t-u|0)\lambda e^{-\lambda u}\, du. \tag{28}$$

Since $g_0^*(t|z) = g_0^*(t-z|0)$ for $t \geq z$, the solution of (28) also provides the solution for $g_0^*(\cdot|z)$. Recall $\alpha := b/r \leq 1$, so (28) is not a renewal equation except in one special case, $\alpha = 1$, where the solution of (28) is trivially $g_0^*(t|0) = 0$. For $\alpha < 1$, (28) is a quasi-renewal equation which can be analysed using Piau's methods. As before, it is only the limit $g_0^*(\infty|0)$ which is important, and moreover, $g_0^*(\infty|0) = g_0^*(\infty)$. Piau's method's show that

$$g_0^*(\infty) = \sum_{n\geq 1}\left[1 - \prod_{k\geq n}[1+(1-\alpha)^k]\right].$$

For the leading strand, it is easy to see that

$$L_t^* = P_t^* - rt + D_t^* + X_1^*.$$

So, as $t \to \infty$,

$$\begin{aligned}\mathbb{E}(L^*) &= \mathbb{E}(D^*) - \frac{r}{\lambda\alpha} + \frac{r}{\lambda} \\ &= rg_0^*(\infty) + r\delta - \frac{r}{\lambda}\big(e^{-\lambda\alpha\delta} - \frac{1}{\alpha}\big).\end{aligned} \tag{29}$$

The average length, R^* say, of leading-strand Okazaki fragment is $\mathbb{E}(L^*)/\mathbb{E}(N^*)$, easily derived from (29), (27) and (25) as:

$$R^* = \frac{\lambda g_0^*(\infty) + \lambda\delta + e^{-\lambda\alpha\delta} - 1/\alpha}{\lambda\delta + 1/\alpha - 1}.$$

In Figure 6, R^* is plotted against $\alpha := b/r$ for $\delta = 0$ and $\lambda = 1$. The plot captures the intuitive result that, as $b \to r$, the mean length of the Okazaki fragments on the leading strand tends to zero. It is easy to show that, in an analogous manner to the lagging-strand, $R^*(\alpha, \delta, \lambda) = R^*(\alpha, \lambda\delta, 1)$.

Finally, experiments which measure average length of extant Okazaki fragments involve averaging over all such fragments, be they on the leading or lagging strand. It is a simple matter to calculate this average and reach an overall R figure:

$$R_{\text{overall}} = \frac{\lambda[E(L) + E(L^*)]}{r[E(N) + E(N^*)]}.$$

I find model B to be extremely plausible since nature needs only to devise one method, namely the one involving the DNA polymerase and primers in order to achieve replication. It is not necessary to endow the unwinding/separating enzyme with the added role of copying the leading strand and one wonders why nature would evolve a redundancy of this sort. Further biological research will, no doubt, point to the correct model, A or B.

20 Concluding remarks

In this paper I have considered a model for the formation of *eyes* in a molecule of DNA when it is undergoing replication. The sites where these eyes initiate, called *replication origins* are modelled by a Poisson process along the DNA and the times at which certain enzyme complexes first approach a given site have the exponentail distribution (in line with our general discussion on the role of exponential distributions in chemical reactions). These two stochastic ingredients to the model provide the interest for the probabilist. Results on the time to replicate a long section of DNA and on the proportion of origins actually initiated are presented.

Secondly, I considered a model for the actual replication within the *eye*, identifying different mechanisms for the so-called *leading* and *lagging* strands. The model for the lagging strand succeeds in capturing the formation of Okazaki fragments. We are able to present results on their mean length. The same results are presented for an alternative model favoured by some biologists.

In presenting the material, I have been concious of drawing on the known biological facts and the snippets of relevant data to assist the estimation of our models' parameters. These data are not precise but their informal use enables us to cast some light on a few biological questions.

For the probabilist, my aim has been to describe the way that stochastic effects arise in molecular biology, in part due to the small numbers of some molecules (or molecular sites) – leading to greater stochastic variation in the chemistry – but mainly due to the random spatial organisation of binding sites on the long DNA molecule.

I hope that interest is aroused in this branch of *applied probability modelling.*

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1983) *Molecular biology of the cell.* Garland, New York.

Barbour, A. D., Holst, L. and Janson, S. (1992) *Poisson Approximation*, Oxford University Press, Oxford.

Blumenthal, A. B., Kriegstein, H. J. and Hogness, D. S. (1974) The Units of DNA Replication in *Drosophila melanogaster* Chromosomes. *Cold Spring Harbor Symposium on Quantitative Biology.* **38**, 205–213.

Callan, H. G. (1973) DNA replication in the chromosomes of eukaryotes. *Cold Spring Harbor Symposium on Quantitative Biology.* **38**, 195–203.

Chiu, S. N. (1992) M.Phil. Thesis, University of Hong Kong.

Chiu, S. N. and Quine, M. P. (1997) Central limit theory for the number of seeds in a growth model in R^d with inhomogeneous Poisson arrivals. *Ann. Appl. Prob.* **7**, 802–814.

Chiu, S. N. and Yin, C. C. (2000) The time of completion of a linear birth-growth model. Preprint received. Dec. 1999.

Chryssaphinou, O. and Papastavridis, S. (1988) A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials. *J. Appl. Prob.* **25**, 428–431.

Clifford, P., Green, N. J. B. and Pilling, M. J. (1987) Statistical Models of Chemical Kinetics in Liquids. *J. Roy. Stat. Soc, Series B* **49**, 266–300.

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.

Cowan, R., Chiu, S. N. and Holst, L. (1995) A limit theorem for the replication time of a DNA molecule. *J. Appl. Prob.* **32**, 296–303.

Cowan, R. and Staudte, R. (1986) The bifurcating autoregression model in cell lineage studies. *Biometrics* **42**, 769–783.

Darvey, I. G., Ninham, B. W. and Staff, P. J. (1966) Stochastic models for second-order chemical kinetics. The equilibrium state. *J. Chem. Phys.* **45**, 2145–2155.

Denhardt, D.T. and Faust, E. A. (1986) Eukaryotic DNA replication. *BioEssay*, **2**, 148–154.

Dunstan, F. D. J. and Reynolds, J. F. (1981) Normal approximation for distributions arising in the stochastic approach to chemical reaction kinetics. *J. Appl. Prob.* **18**, 263–267.

Hall, P. (1983) On the roles of the Bessel and Poisson distributions in chemical kinetics. *J. Appl. Prob.* **20** 585–599.

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

Heyde, C. C. and Heyde, E. (1971) Stochastic Fluctuations in a One Substrate One Product Enzyme System: Are they ever Relevant? **30**, 395–404.

Holst, L., Quine, M. P. and Robinson, J. (1996). A general stochastic model for nucleation and linear growth. *Ann. Appl. Probab.* **6**, 903–921.

Huberman, J A. and Horwitz, H. (1973) Discontinuous DNA synthesis in mammalian cells. *Cold Spring Harbour Symposium on Quantitative Biology*, **38**, 233–238

Ishida, K. (1964) Stochastic Model for Bimolecular Reaction. *J. Chem. Phys.* **41**, 2472–2478.

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.

Kowalski, J. and Denhardt, D. T. (1982) Adenovirus DNA replication *in vivo.* Properties of short DNA molecules extracted from infected cells. *Biochem. Biophys. Acta*, **698**, 260–270.

Kriegstein, H. J. and Hogness, D. S. (1973) The mechanism of DNA replication in *Drosophila* chromosomes: Structure of replication forks and evidence for bidirectionality. *Proc. Nat. Acad. Sci.* **71(1)**, 135–139.

McPeek, M. S. and Speed, T. P. (1995) Modeling Interference in Genetic Recombination. *Genetics* **139**, 1031–1044.

McQuarrie, D. A. (1967) Stochastic approach to chemical kinetics. *J. Appl. Prob.* **4,** 413–478.

Mitchison, J. M. (1971) *The Biology of the Cell Cycle.* Cambridge University Press, Cambridge.

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.

Powell, E. O. (1953) Some features of the generation times of individual bacteria. *Biometrika* **42**, 16–44.

Powell, E. O. (1958) An Outline of the Pattern of Bacterial Generation Times. *J.gen.Microbiol.* **18**, 382–417.

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.

Smith, J. A. and Martin, L. (1973) Do cells cycle? *Proc. Nat. Acad. Sci. USA* **70**, 1263–1267.

Staudte, R., Guiguet, M. and Collyn d'Hooghe, M. (1984) Additive models for dependent cell populations. *J.theor.Biol.* **109**, 127–146.

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.