

# ON CLONING AND CLONE LIBRARIES FOR FINITE AND INFINITE LENGTH GENOMES

YI FANG AND SUSAN R. WILSON

## 1. INTRODUCTION

In view of the recent success of the International Human Genome Sequencing Consortium [1] in obtaining an initial ‘clone-based’ physical map of the human genome, and of Venter et al [5] (2001), in sequencing (most of) the human and other genomes based on good libraries, it is apposite to provide some further details of the underlying mathematical theory on cloning and clone libraries.

A clone (or DNA) library is a collection of clones (restriction endonuclease-generated fragments) containing DNA fragments from the genome. A cloning vector is generally a plasmid, phage or eukaryotic virus-derived linear or circular DNA that is capable of reproduction (most commonly in bacteria or yeast; see [2]). Cloning vectors can only accept inserts within a certain size range dependent on the choice of vector. So the length of DNA that can be put into a vector is limited, and hence so is the DNA that can be contained in a library.

The mathematical theory that has been developed to determine the representativeness of clone libraries is implicitly based on the underlying simplifying assumption that the length of the genome is (effectively) infinite (Chapter 5 of [6]). In the following this model is referred to as the **infinite model**. Here the aim is to develop corresponding theory for models where the genome can be finite. In particular two models are considered. One is the **loop model** for genomes of length  $G$  base pairs (bps) which are assumed to be connected by  $G$  bonds in a loop. The other is the **interval model** which has a start and an end base pair, and so it is of length  $G$  bps with  $G - 1$  bonds.

To form the fragments, it is assumed that a cut site (restriction site) occurs between any two base pairs with probability  $p$ , where  $p$  is small. In the following it is assumed that the cuts are such as to be independently and identically distributed (iid). This assumption was made in [6] where it was noted that, in principle, this assumption would allow two adjacent bonds to be cut, and as sites are four or more base pairs in length, this is not realistic. However, when  $p$  is small, such as  $p = 1/5000$ , then under the simple iid model, there is only a vanishingly small probability that the cuts are adjacent.

In the next section the theory for the finite loop and the interval models, as well as for the infinite model, is developed, and the means and variances of fragment lengths, and of a randomly selected fragment, are determined. In the third section, the probability of the event that a particular base pair is clonable is calculated for all models. The final section gives results for partial digestion, namely for stopping the digest before all sites are cut.

## 2. THE LOOP, THE INTERVAL, AND THE INFINITE MODELS

For a fixed base pair  $b$ , computation of the probability  $\mathbb{P}(b(l))$ , that  $b$  is in a fragment of length  $l$  is required, with  $1 \leq l \leq G$  for the loop and the interval models, and  $1 \leq l < \infty$  for the infinite model, and such that  $\sum \mathbb{P}(b(l)) = 1$ .

**Lemma 1.** *Assume restriction sites are distributed along a genome of  $G$  bps according to a Bernoulli process with  $p = \mathbb{P}(\text{restriction site})$ .*

*Then under the loop model, for a fixed base pair  $b$ ,*

$$(1) \quad \mathbb{P}(b(l)) = lp^2(1-p)^{l-1}, \quad 1 \leq l \leq G-1; \quad \mathbb{P}(b(G)) = Gp(1-p)^{G-1} + (1-p)^G.$$

*Under the interval model, the probability of inclusion of a base pair depends also on its position, so the base pairs are labelled as  $b_i$ ,  $1 \leq i \leq G$ , and the event  $(i, l)$  is that  $b_i$  belongs to a fragment of length  $l$ ,  $1 \leq l \leq G$ . Then*

$$(2) \quad \mathbb{P}((1, l)) = \begin{cases} p(1-p)^{l-1} & 1 \leq l \leq G-1; \\ (1-p)^{G-1} & l = G. \end{cases}$$

*For  $2 \leq i \leq [G/2]$ ,*

$$(3) \quad \mathbb{P}((i, l)) = \begin{cases} lp^2(1-p)^{l-1} & 1 \leq l \leq i-1; \\ p(1-p)^{l-1} + (i-1)p^2(1-p)^{l-1} & i \leq l \leq G-i; \\ 2p(1-p)^{l-1} + (G-l-1)p^2(1-p)^{l-1} & G-i < l < G; \\ (1-p)^{G-1} & l = G. \end{cases}$$

*By symmetry,  $\mathbb{P}((i, l)) = \mathbb{P}((G-i+1, l))$ , for  $i > [G/2]$ . When  $G$  is even, all the cases have been covered. The only remaining case is when  $G = 2n+1$  and  $i = n+1$ , for which*

$$(4) \quad \mathbb{P}((n+1, l)) = \begin{cases} lp^2(1-p)^{l-1} & 1 \leq l \leq n; \\ 2p(1-p)^{l-1} + (G-l-1)p^2(1-p)^{l-1} & n < l < 2n+1; \\ (1-p)^{2n} & l = 2n+1. \end{cases}$$

*For an infinite model (i.e. assuming that the genome has an infinite number of bps without either a beginning or an ending)*

$$\mathbb{P}(b(l)) = lp^2(1-p)^{l-1}, \quad 1 \leq l < \infty.$$

*Proof.* First consider the loop model. Since the cuts are independent, the probability of the event that there are  $k$  cuts is a binomial distribution with parameters  $(G, p)$ . Note that fragments of length  $G$  are produced by either just one cut, or no cut at all, so  $\mathbb{P}(b(G)) = Gp(1-p)^{G-1} + (1-p)^G$ .

Note that since the genome is in a loop, if  $b$  belongs to a fragment of length  $l$ ,  $1 \leq l < G$ , then there is more than one fragment and  $b$  belongs to only one of them. In particular, there are at least two cuts. Thus in this situation, fragments of length  $l$ ,  $1 \leq l \leq G-1$ , are produced by the configuration cut- $\{\text{no cut}\}^{l-1}$ -cut. Since  $b$  could be at any of the  $l$  positions of the configuration,  $\mathbb{P}(b(l)) = lp^2(1-p)^{l-1}$ .

Similar and simpler arguments apply to the infinite model.

In the interval model, it is necessary to distinguish the cases where there is either one or two cuts to a sequence, and it is also necessary to consider the position of the base pair  $b_i$  within that specific sequence. This depends both on the position  $i$  and the length  $l$ , and all the cases need to be considered.

In the interval model, the cuts are distributed as a binomial distribution with parameters  $(G - 1, p)$ .

The bonds are labelled by  $c_i$ ,  $1 \leq i \leq G - 1$ , such that  $b_1c_1b_2c_2 \cdots b_{G-1}c_{G-1}b_G$  is the sequence. If a sequence has two cuts at  $c_j$  and  $c_k$ ,  $c_j$  is called the left cut if  $j < k$ .

$\mathbb{P}((1, l)) = p(1-p)^{l-1}$  is obvious, since there is only one way to include  $b_1$  in a fragment of any length  $1 \leq l \leq G-1$ , i.e., one cuts at  $c_l$ . For any  $1 \leq i \leq G$ ,  $\mathbb{P}((i, G)) = (1-p)^{G-1}$  for all  $i$ , because there would have been no cuts at all.

Consider positions such that  $2 \leq i \leq [G/2]$ . If  $1 \leq l \leq i - 1$ , then any fragment of length  $l$  and containing  $b_1$  or  $b_G$  will not contain  $b_i$ . Thus for a length  $l$  fragment containing  $b_i$ , there are exactly two cuts. Also since  $b_i$  could be at any of the  $l$  positions in the fragment, it follows that  $\mathbb{P}((i, l)) = lp^2(1-p)^{l-1}$ .

If  $i \leq l \leq G - i$ , there are two subevents. Let  $E_1$  be the subevent that  $b_i$  is contained in a fragment of length  $l$  which also contains  $b_1$ . Then  $\mathbb{P}(E_1) = p(1-p)^{l-1}$ . Let  $E_2$  be the subevent that  $b_i$  is contained in a fragment of length  $l$  which does not contain  $b_1$ . Since  $l \leq G - i$  and  $i \leq [G/2]$ , a fragment of length  $l$  that contains  $b_i$  cannot also contain  $b_G$ . Thus if the fragment does not contain  $b_1$ , it must have two cuts. One of the cuts is  $c_k$ ,  $k < i$ . Since any  $c_j$ ,  $1 \leq j \leq i - 1$  could be a cut, it follows that  $\mathbb{P}(E_2) = (i - 1)p^2(1-p)^{l-1}$ . Thus

$$\mathbb{P}((i, l)) = \mathbb{P}(E_1 \cup E_2) = p(1-p)^{l-1} + (i-1)p^2(1-p)^{l-1}, \quad i \leq l \leq G - i.$$

If  $G - i < l < G$ , again there are two subevents. Let  $E_1$  be the subevent that a fragment of length  $l$  contains  $b_i$  and also contains either  $b_1$  or  $b_G$ . Then obviously  $\mathbb{P}(E_1) = 2p(1-p)^{l-1}$ . Let  $E_2$  be the subevent that a fragment containing  $b_i$  contains neither  $b_1$  nor  $b_G$ , then such a fragment has two cuts. If  $c_j$  is the left cut, then the right cut is at  $c_k$ , where  $k = j + l \leq G - 1$ . Thus it follows that that  $j \leq G - l - 1$ . Since any of  $c_j$ ,  $1 \leq j \leq G - l - 1$  could be the left cut, it follows that  $\mathbb{P}(E_2) = (G - l - 1)p^2(1-p)^{l-1}$ . Therefore,

$$\mathbb{P}((i, l)) = \mathbb{P}(E_1 \cup E_2) = 2p(1-p)^{l-1} + (G - l - 1)p^2(1-p)^{l-1}, \quad G - i < l < G.$$

Analogous arguments to the above can be applied to obtain (4).  
Using the formula

$$\sum_{l=k}^n r^l = \frac{r^k(1 - r^{n-k+1})}{1 - r} = \frac{r^k - r^{n+1}}{1 - r},$$

it can easily be checked that for the loop model,

$$\begin{aligned}
\sum_{l=1}^G \mathbb{P}(l) &= Gp(1-p)^{G-1} + (1-p)^G + \sum_{l=1}^{G-1} lp^2(1-p)^{l-1} \\
&= Gp(1-p)^{G-1} + (1-p)^G - p^2 \frac{d}{dp} \sum_{l=1}^{G-1} (1-p)^l \\
&= Gp(1-p)^{G-1} + (1-p)^G - p^2 \frac{d}{dp} \frac{1-p - (1-p)^G}{p} \\
&= Gp(1-p)^{G-1} + (1-p)^G - p^2 \left( -\frac{1}{p^2} + \frac{(1-p)^G}{p^2} + \frac{G(1-p)^{G-1}}{p} \right) \\
&= Gp(1-p)^{G-1} + (1-p)^G + 1 - (1-p)^G - Gp(1-p)^{G-1} \\
&= 1.
\end{aligned}$$

For the infinite model,

$$\sum_{l=1}^{\infty} \mathbb{P}(l) = p^2 \sum_{l=1}^{\infty} lp^2(1-p)^{l-1} = -p^2 \frac{d}{dp} \sum_{l=1}^{\infty} (1-p)^l = -p^2 \frac{d}{dp} \frac{1-p}{p} = 1.$$

For the interval model,

$$\begin{aligned}
\sum_{l=1}^G \mathbb{P}((1, l)) &= p \sum_{l=1}^{G-1} (1-p)^{l-1} + (1-p)^{G-1} \\
&= p \frac{1 - (1-p)^{G-1}}{p} + (1-p)^{G-1} = 1.
\end{aligned}$$

For  $2 \leq i \leq [G/2]$ ,

$$\begin{aligned}
\sum_{l=1}^G \mathbb{P}((i, l)) &= p^2 \sum_{l=1}^{i-1} l(1-p)^{l-1} + p[1 + (i-1)p] \sum_{l=i}^{G-i} (1-p)^{l-1} + (1-p)^{G-1} \\
&\quad + p[2 + (G-1)p] \sum_{l=G-i+1}^{G-1} (1-p)^{l-1} - p^2 \sum_{l=G-i+1}^{G-1} l(1-p)^{l-1} \\
&= -p^2 \frac{d}{dp} \sum_{l=1}^{i-1} (1-p)^l + [1 + (i-1)p] [(1-p)^{i-1} - (1-p)^{G-i}] + (1-p)^{G-1} \\
&\quad + [2 + (G-1)p] [(1-p)^{G-i} - (1-p)^{G-1}] + p^2 \frac{d}{dp} \sum_{l=G-i+1}^{G-1} (1-p)^l \\
&= -p^2 \frac{d}{dp} \frac{1-p - (1-p)^i}{p} + [1 + (i-1)p] [(1-p)^{i-1} - (1-p)^{G-i}] \\
&\quad + [2 + (G-1)p] [(1-p)^{G-i} - (1-p)^{G-1}] + (1-p)^{G-1} \\
&\quad + p^2 \frac{d}{dp} \frac{(1-p)^{G-i+1} - (1-p)^G}{p} \\
&= -p^2 \left[ \frac{-1 + (1-p)^i}{p^2} + \frac{i(1-p)^{i-1}}{p} \right] + [1 + (i-1)p] [(1-p)^{i-1} - (1-p)^{G-i}] \\
&\quad + [2 + (G-1)p] [(1-p)^{G-i} - (1-p)^{G-1}] + (1-p)^{G-1} \\
&\quad + p^2 \left[ \frac{(1-p)^{G-i+1} - (1-p)^G}{-p^2} - \frac{(G-i+1)(1-p)^{G-i} - G(1-p)^{G-1}}{p} \right] \\
&= 1 - (1-p)^i - pi(1-p)^{i-1} + [1 + (i-1)p] [(1-p)^{i-1} - (1-p)^{G-i}] \\
&\quad + [2 + (G-1)p] [(1-p)^{G-i} - (1-p)^{G-1}] + (1-p)^{G-1} \\
&\quad - (1-p)^{G-i+1} + (1-p)^G - p(G-i+1)(1-p)^{G-i} + pG(1-p)^{G-1} \\
&= 1 + (1-p)(1-p)^{i-1} - (1-p)^i + (1-p)(1-p)^{G-i} - (1-p)^{G-i+1} \\
&\quad - (1-p)(1-p)^{G-1} + (1-p)^G \\
&= 1.
\end{aligned}$$

□

**Theorem 1.** *Let  $X$  be a random variable representing the fragment length for the loop model or the infinite model. Let  $X(i)$  be a random variable representing the length of a fragment containing  $b_i$  for the interval model. Let  $\tilde{l}$  be the average length of a fragment. Then under the loop model*

$$\begin{aligned}
\tilde{l} &= \mathbb{E}(X) = 2p^{-1}[(1-p)^2 - (1-p)^{G+1}] + 3 - 2p - (G+1)(1-p)^G \\
(5) \quad &= 2 \sum_{l=2}^G (1-p)^l + 3 - 2p - (G+1)(1-p)^G
\end{aligned}$$

Under the interval model, the average length of a fragment including  $b_i$  depends not only on  $p$ , but also on the position of the base pair. For  $i = 1$ ,

$$(6) \quad \tilde{l}(1) = \mathbb{E}(X(1)) = p^{-1}[1 - (1-p)^G] = \sum_{l=0}^{G-1} (1-p)^l.$$

For  $2 \leq i \leq [G/2]$ ,

$$(7) \quad \begin{aligned} \tilde{l}(i) &= \mathbb{E}(X(i)) = p^{-1}[2(1-p)^2 - (1-p)^i - (1-p)^{G-i+1}] + 3 - 2p \\ &= \sum_{l=2}^{i-1} (1-p)^l + \sum_{l=2}^{G-i} (1-p)^l + 3 - 2p. \end{aligned}$$

Under the infinite model,

$$(8) \quad \tilde{l} = \mathbb{E}(X) = 2p^{-1}(1-p)^2 + 3 - 2p = 2 \sum_{l=2}^{\infty} (1-p)^l + 3 - 2p.$$

When  $G$  is large and  $p$  is small, in both the loop and the infinite models the approximation

$$(9) \quad \tilde{l} \cong 2p^{-1}(1-p)^2 + 3 = 2 \sum_{l=1}^{\infty} (1-p)^l + 3.$$

can be used. Thus in this case the average length of a fragment only depends on  $p$ .

For the interval model, assuming that  $i$  is near the middle of the sequence where  $G$  is large, the approximation (9) can be used.

The variances of these random variables are as follows.

For the loop model,

$$(10) \quad \begin{aligned} \text{Var}(X) &= 2p^{-2}[(1+2p)(1-p)^3 + (1-pG)(1-p)^{G+1} - 2(1+pG)(1-p)^{2G+1}] \\ &+ 2(1-p)(3-2p) - (G+1)(2G-1)(1-p)^G - (G+1)^2(1-p)^{2G}. \end{aligned}$$

For the interval model, the variance of  $X(1)$  is

$$(11) \quad \text{Var}(X(1)) = \text{Var}(X(G)) = p^{-2}[1 - (1-p)^{2G}] - p^{-1}[1 + (2G-3)(1-p)^G].$$

For  $i \neq 1, G$ , the variance of  $X(i)$  is

$$(12) \quad \begin{aligned} \text{Var}(X(i)) &= 2p^{-2}(1+2p)(1-p)^3 + 2(1-p)(3-2p) \\ &- p^{-1}[(2i-1)(1-p)^i + (2G-2i+1)(1-p)^{G-i+1}]. \end{aligned}$$

Under the infinite model,

$$(13) \quad \text{Var}(X) = 2p^{-2}(1+2p)(1-p)^3 + 2(1-p)(3-2p).$$

If  $p$  is small and  $G$  is large, in almost all cases the approximation

$$(14) \quad \text{Var}(X) \cong 2p^{-2}(1+2p)(1-p)^3 + 2(1-p)(3-2p)$$

can be used.

**Remark 1.** The significance of the expressions for  $\mathbb{E}(X)$  is that if the average length of the fragments is known, then this can be used to determine  $p$ , the rate at which an enzyme is cutting the genome.

**Remark 2.** The equations (7) and (12) show that  $\mathbb{E}(X(i))$  and  $\text{Var}(X(i))$  are symmetric, i.e.,  $\mathbb{E}(X(i)) = \mathbb{E}(X(G-i+1))$ ,  $\text{Var}(X(i)) = \text{Var}(X(G-i+1))$  for  $1 \leq i \leq G$ . This is consistent with the probabilities  $\mathbb{P}((i, l))$  being symmetric. Thus when  $G = 2n + 1$ , and  $i = n + 1$ ,

$$\begin{aligned} \tilde{l}(n+1) &= \mathbb{E}(X(n+1)) = 2p^{-1}[(1-p)^2 - (1-p)^{n+1}] + 3 - 2p \\ (15) \quad &= 2 \sum_{l=2}^n (1-p)^l + 3 - 2p. \end{aligned}$$

$$\begin{aligned} \text{Var}(X(n+1)) &= 2p^{-2}(1+2p)(1-p)^3 + 2(1-p)(3-2p) \\ (16) \quad &\quad - 2(2n+1)p^{-1}(1-p)^{n+1}. \end{aligned}$$

*Proof.* For the loop model,

$$\begin{aligned} \mathbb{E}(X) &= \sum_{l=1}^G l\mathbb{P}(l) = p^2 \sum_{l=1}^{G-1} l^2(1-p)^{l-1} + pG^2(1-p)^{G-1} + G(1-p)^G \\ &= p^2 \sum_{l=1}^{G-1} (l+1)l(1-p)^{l-1} - p^2 \sum_{l=1}^{G-1} l(1-p)^{l-1} + pG^2(1-p)^{G-1} + G(1-p)^G \\ &= p^2 \frac{d^2}{dp^2} \sum_{l=1}^{G-1} (1-p)^{l+1} - 1 + pG(G+1)(1-p)^{G-1} + (G+1)(1-p)^G \\ &= p^2 \frac{d^2}{dp^2} \frac{(1-p)^2 - (1-p)^{G+1}}{p} - 1 + pG(G+1)(1-p)^{G-1} + (G+1)(1-p)^G \\ &= p^2 \left[ \frac{2(1-p)^2 - 2(1-p)^{G+1}}{p^3} + \frac{4(1-p) - 2(G+1)(1-p)^G}{p^2} + \frac{2 - G(G+1)(1-p)^{G-1}}{p} \right] \\ &\quad + pG(G+1)(1-p)^{G-1} + (G+1)(1-p)^G - 1 \\ &= 2p^{-1}(1-p)^2 - 2p^{-1}(1-p)^{G+1} + 3 - 2p - (G+1)(1-p)^G \\ &= 2p^{-1}[(1-p)^2 - (1-p)^{G+1}] + 3 - 2p - (G+1)(1-p)^G \\ &= 2 \sum_{l=2}^G (1-p)^l + 3 - 2p - (G+1)(1-p)^G. \end{aligned}$$

For the interval model, when  $i = 1$ ,

$$\begin{aligned}
\mathbb{E}(X(1)) &= \sum_{l=1}^G l\mathbb{P}((1, l)) = p \sum_{l=1}^{G-1} l(1-p)^{l-1} + G(1-p)^{G-1} \\
&= G(1-p)^{G-1} - p \frac{d}{dp} \frac{1-p - (1-p)^G}{p} \\
&= -p \left( \frac{-1}{p^2} + \frac{G(1-p)^{G-1}}{p} + \frac{(1-p)^G}{p^2} \right) + G(1-p)^{G-1} = p^{-1}[1 - (1-p)^G] = \sum_{l=0}^{G-1} (1-p)^l.
\end{aligned}$$

For  $2 \leq i \leq [G/2]$ ,

$$\begin{aligned}
\mathbb{E}(X(i)) &= \sum_{l=1}^G l\mathbb{P}((i, l)) = p^2 \sum_{l=1}^{i-1} l^2(1-p)^{l-1} + [p + (i-1)p^2] \sum_{l=i}^{G-i} l(1-p)^{l-1} \\
&\quad + (2p + Gp^2) \sum_{l=G-i+1}^{G-1} l(1-p)^{l-1} - p^2 \sum_{l=G-i+1}^{G-1} (l+1)l(1-p)^{l-1} + G(1-p)^{G-1} \\
&= p^2 \frac{d^2}{dp^2} \sum_{l=1}^{i-1} (1-p)^{l+1} + p^2 \frac{d}{dp} \sum_{l=1}^{G-i} (1-p)^l - (p + ip^2) \frac{d}{dp} \sum_{l=i}^{G-i} (1-p)^l \\
&\quad - (2p + Gp^2) \frac{d}{dp} \sum_{l=G-i+1}^{G-1} (1-p)^l - p^2 \frac{d^2}{dp^2} \sum_{l=G-i+1}^{G-1} (1-p)^{l+1} + G(1-p)^{G-1} \\
&= p^2 \frac{d^2}{dp^2} \frac{(1-p)^2 - (1-p)^{i+1}}{p} + p^2 \frac{d}{dp} \frac{1-p - (1-p)^{G-i+1}}{p} \\
&\quad - (p + ip^2) \frac{d}{dp} \frac{(1-p)^i - (1-p)^{G-i+1}}{p} - (2p + Gp^2) \frac{d}{dp} \frac{(1-p)^{G-i+1} - (1-p)^G}{p} \\
&\quad - p^2 \frac{d^2}{dp^2} \frac{(1-p)^{G-i+2} - (1-p)^{G+1}}{p} + G(1-p)^{G-1} \\
&= p^{-1}[2(1-p)^2 - (1-p)^i - (1-p)^{G-i+1}] + 3 - 2p \\
&= \sum_{l=2}^{i-1} (1-p)^l + \sum_{l=2}^{G-i} (1-p)^l + 3 - 2p.
\end{aligned}$$

Under the infinite model,

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{l=1}^G l\mathbb{P}(l) = p^2 \sum_{l=1}^{\infty} l^2(1-p)^{l-1} = p^2 \frac{d^2}{dp^2} \sum_{l=1}^{\infty} (1-p)^{l+1} - 1 \\
&= p^2 \frac{d^2}{dp^2} \frac{(1-p)^2}{p} - 1 = 2p^{-1}(1-p)^2 + 3 - 2p.
\end{aligned}$$

Now the variance in the infinite model is calculated.

$$\begin{aligned}
E(X^2) &= \sum_{l=1}^{\infty} p^2 l^3 (1-p)^{l-1} = \sum_{l=1}^{\infty} p^2 (l+2)(l+1)l (1-p)^{l-1} - \sum_{l=1}^{\infty} p^2 (3l^2 + 2l) (1-p)^{l-1} \\
&= -p^2 \frac{d^3}{dp^3} \sum_{l=1}^{\infty} (1-p)^{l+2} - 3E(X) - 2 = -p^2 \frac{d^3}{dp^3} \frac{(1-p)^3}{p} - 3E(X) - 2 \\
&= 6p^{-2}(1-p)^3 + 12p^{-1}(1-p)^2 + 7 - 6p.
\end{aligned}$$

Thus

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = 2p^{-2}(1+2p)(1-p)^3 + 2(1-p)(3-2p).$$

For the interval model with  $i = 1$ ,

$$\begin{aligned}
\mathbb{E}(X^2(1)) &= \sum_{l=1}^G l^2 \mathbb{P}((1, l)) = p \sum_{l=1}^{G-1} l^2 (1-p)^{l-1} + G^2 (1-p)^{G-1} \\
&= p \sum_{l=1}^{G-1} (l+1)l (1-p)^{l-1} - p \sum_{l=1}^{G-1} l (1-p)^{l-1} + G^2 (1-p)^{G-1} \\
&= p \frac{d^2}{dp^2} \frac{(1-p)^2 - (1-p)^{G+1}}{p} - \mathbb{E}(X(1)) + G(G+1)(1-p)^{G-1} \\
&= p \left( \frac{2(1-p)^2 - 2(1-p)^{G+1}}{p^3} + \frac{4(1-p) - 2(G+1)(1-p)^G}{p^2} + \frac{2 - G(G+1)(1-p)^{G-1}}{p} \right) \\
&\quad - \mathbb{E}(X(1)) + G(G+1)(1-p)^{G-1} \\
&= 2p^{-2}[(1-p)^2 - (1-p)^{G+1}] + 3p^{-1} - 2 - (2G-1)p^{-1}(1-p)^G,
\end{aligned}$$

thus

$$\begin{aligned}
\text{Var}(X(1)) &= \mathbb{E}(X^2(1)) - [\mathbb{E}(X(1))]^2 \\
&= p^{-2}[1 - (1-p)^{2G}] - p^{-1}[1 + (2G-3)(1-p)^G].
\end{aligned}$$

Similarly the variance of  $X$  or  $X(i)$ ,  $i > 1$ , under the loop and the interval model can be calculated. The tedious calculation is omitted.  $\square$

Consider the loop, the interval and the infinite models. Interest lies in the average length of a randomly selected fragment under these models.

**Theorem 2.** *Let  $X$  be the random variable representing the length of a randomly selected fragment. Then under the infinite model*

$$(17) \quad \mathbb{E}(X) = p^{-1}.$$

*Under the interval model,*

$$(18) \quad \mathbb{E}(X) = p^{-1}[1 - (1-p)^G] = \sum_{l=0}^{G-1} (1-p)^l.$$

Under the loop model,

$$\begin{aligned}
 \mathbb{E}(X) &= G(1-p)^G + \sum_{k=1}^G \frac{G}{k} \binom{G}{k} p^k (1-p)^{G-k} \\
 (19) \qquad &= G(1-p)^G + pG \sum_{k=0}^{G-1} \frac{1}{G-k} \sum_{i=k}^{G-1} (1-p)^i.
 \end{aligned}$$

For the loop model, simple lower and upper bounds are

$$(20) \qquad p^{-1}[1 - (1-p)^G] \leq \mathbb{E}(X) \leq 2p^{-1}[1 - (1-p)^G] - G(1-p)^G.$$

**Remark 3.** If  $p < G^{-1}$ , then  $p^{-1} > G$ , and so equation (17) is obviously absurd in this situation. Thus, although the infinite model is a good approximation, it is best to be aware that it is only an approximation.

In any case, the intuition that the average length of a typical fragment should be  $pG$  is unfounded.

The interval model gives a simple formula for  $\mathbb{E}(X)$ , and it is always less than  $G$  as can be easily seen from equation (18).

Nevertheless, using either equation (17) or (18) and the strong law of large numbers, and assuming that  $n$  is large,  $p$  can be estimated by the formula

$$(21) \qquad p \cong \left( \frac{\sum_{i=1}^n l_i}{n} \right)^{-1},$$

where  $l_i$  is the length of the  $i$ -th fragment.

*Proof.* Under the infinite model, the probability of a fragment having length  $l$  is  $p(1-p)^{l-1}$ , and the mean of this geometric distribution is  $p^{-1}$ , see for example, [3]. So (17) is true.

For the interval and the loop models, the cuts are distributed as a binomial distribution with parameters  $(G-1, p)$  and  $(G, p)$  respectively. Let  $Y$  be the random variable indicating the number of cuts. Then

$$(22) \qquad \mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)).$$

For the interval model,

$$(23) \qquad \mathbb{E}(\mathbb{E}(X|Y)) = \sum_{k=0}^{G-1} \mathbb{E}(X|Y=k) \mathbb{P}(Y=k).$$

For the loop model,

$$(24) \qquad \mathbb{E}(\mathbb{E}(X|Y)) = \sum_{k=0}^G \mathbb{E}(X|Y=k) \mathbb{P}(Y=k).$$

Let  $(i_1, \dots, i_k)$ ,  $1 \leq i_1 < \dots < i_k \leq G-1$  be the event that the  $i_1$ -th,  $\dots$ ,  $i_k$ -th bonds are cut. Let  $l_j$ ,  $1 \leq j \leq k+1$  be the length of the fragments under the cuts  $(i_1, \dots, i_k)$ ,

then any of the  $k + 1$  fragments has equal probability of  $1/(k + 1)$  of being selected. Thus, since  $\sum_{j=1}^{G-1} l_j = G$ ,

$$\mathbb{E}(X|\{Y = k, (i_1, \dots, i_k)\}) = \sum_{j=1}^{k+1} l_j \frac{1}{k+1} = \frac{G}{k+1}.$$

Since each combination of  $k$  bonds  $(i_1, \dots, i_k)$  has the same probability, it follows that

$$\mathbb{E}(X|Y = k) = \sum_{1 \leq i_1 < \dots < i_k \leq G-1} \mathbb{E}(X|\{Y = k, (i_1, \dots, i_k)\}) \mathbb{P}((i_1, \dots, i_k)) = \frac{G}{k+1}.$$

From (23)

$$\begin{aligned} \mathbb{E}(X) &= \sum_{k=0}^{G-1} \frac{G}{k+1} \binom{G-1}{k} p^k (1-p)^{G-1-k} = \sum_{k=0}^{G-1} \binom{G}{k+1} p^k (1-p)^{G-1-k} \\ &= p^{-1} \sum_{k=1}^G \binom{G}{k} p^k (1-p)^{G-k} = p^{-1} [1 - (1-p)^G]. \end{aligned}$$

A similar argument for the loop model gives  $\mathbb{E}(X|Y = 0) = G$  and

$$\mathbb{E}(X|Y = k) = \frac{G}{k}, \quad 1 \leq k \leq G.$$

Thus by (24) it follows that

$$\begin{aligned}
\mathbb{E}(X) &= G(1-p)^G + \sum_{k=1}^G \frac{G}{k} \binom{G}{k} p^k (1-p)^{G-k} \\
&= G(1-p)^G + G \sum_{k=1}^G \binom{G}{k} (1-p)^{G-k} \int_0^p s^{k-1} ds \\
&= G(1-p)^G + G \int_0^p \left( \sum_{k=1}^G \binom{G}{k} s^{k-1} (1-p)^{G-k} \right) ds \\
&= G(1-p)^G + G \int_0^p \left( s^{-1} \sum_{k=1}^G \binom{G}{k} s^k (1-p)^{G-k} \right) ds \\
&= G(1-p)^G + G \int_0^p s^{-1} [(s+1-p)^G - (1-p)^G] ds \\
&= G(1-p)^G + G \int_0^p \sum_{k=0}^{G-1} (1-p)^k (s+1-p)^{G-1-k} ds \\
&= G(1-p)^G + G \sum_{k=0}^{G-1} \frac{1}{G-k} (1-p)^k (s+1-p)^{G-k} \Big|_0^p \\
&= G(1-p)^G + G \sum_{k=0}^{G-1} \frac{1}{G-k} (1-p)^k [1 - (1-p)^{G-k}] \\
&= G(1-p)^G + pG \sum_{k=0}^{G-1} \frac{1}{G-k} \sum_{i=k}^{G-1} (1-p)^i.
\end{aligned}$$

To obtain the lower bound, consider

$$\begin{aligned}
\sum_{k=1}^G \frac{G}{k} \binom{G}{k} p^k (1-p)^{G-k} &\geq \sum_{k=1}^G \frac{G+1}{k+1} \binom{G}{k} p^k (1-p)^{G-k} \\
&= \sum_{k=1}^G \binom{G+1}{k+1} p^k (1-p)^{G-k} = p^{-1} \sum_{k=1}^G \binom{G+1}{k+1} p^{k+1} (1-p)^{G+1-(k+1)} \\
&= p^{-1} \sum_{k=2}^{G+1} \binom{G+1}{k} p^k (1-p)^{G+1-k} = p^{-1} [1 - p(G+1)(1-p)^G - (1-p)^{G+1}].
\end{aligned}$$

This gives the lower bound at (20).

For the upper bound, note that  $G/k \leq 2(G+1)/(k+1)$ , so

$$\begin{aligned} \sum_{k=1}^G \frac{G}{k} \binom{G}{k} p^k (1-p)^{G-k} &\leq 2 \sum_{k=1}^G \frac{G+1}{k+1} \binom{G}{k} p^k (1-p)^{G-k} \\ &= 2 \sum_{k=1}^G \binom{G+1}{k+1} p^k (1-p)^{G-k} = 2p^{-1} \sum_{k=1}^G \binom{G+1}{k+1} p^{k+1} (1-p)^{G+1-(k+1)} \\ &= 2p^{-1} \sum_{k=2}^{G+1} \binom{G+1}{k} p^k (1-p)^{G+1-k} = 2p^{-1} [1 - p(G+1)(1-p)^G - (1-p)^{G+1}]. \end{aligned}$$

Thus

$$\mathbb{E}(X) \leq 2p^{-1} [1 - (1-p)^G] - G(1-p)^G = 2 \sum_{l=0}^{G-1} (1-p)^l - G(1-p)^G.$$

□

### 3. THE PROBABILITY OF A BASE PAIR BEING CLONABLE

Let  $L$  be the lower size limit of clonable fragments, and  $U$  the upper limit, so only fragments of length  $l$ ,  $0 < L \leq l \leq U < G$ , are clonable. The following theorem gives the probability that a fixed base pair is clonable under the three models.

**Theorem 3.** *Assume restriction sites are distributed along a genome of  $G$  bps according to a Bernoulli process with  $p = \mathbb{P}(\text{restriction site})$ . Let  $C(b)$  be the event that a fixed base pair  $b$  is clonable. Then for both the finite loop model and the infinite model in Lemma 1,*

$$(25) \quad \mathbb{P}(C(b)) = \sum_{l=L}^U \mathbb{P}(l) = [p(L-1) + 1](1-p)^{L-1} - (pU + 1)(1-p)^U,$$

$$(26) \quad \mathbb{P}(\widetilde{C}(b)) = 1 - [p(L-1) + 1](1-p)^{L-1} + (pU + 1)(1-p)^U,$$

where  $\widetilde{C}(b)$  is the complement event of  $C(b)$ , i.e.,  $b$  is not clonable, or equivalently,  $b$  belongs to a fragment of length  $l$ , either  $l < L$ , or  $l > U$ .

For the interval model, the position of the base pair  $b_i$  has to be considered, and so some examples are given. The two most important cases are  $L < i < U \leq \lfloor G/2 \rfloor$  and  $L < U < i < G - U$ . The first case has a larger probability compared with the analogous probabilities for other two models, while the second case has the same probability as for the loop and infinite models. Let  $C(i)$  be the event that the  $i$ -th base pair  $b_i$  is clonable.

$$(27) \quad \mathbb{P}(C(1)) = (1-p)^{L-1} - (1-p)^U.$$

For  $2 \leq i < L < U < [G/2]$ ,

$$\begin{aligned} \mathbb{P}(C(i)) &= [p(i-1) + 1](1-p)^{L-1} - [p(i-1) + 1](1-p)^U \\ (28) \quad &= p[p(i-1) + 1] \sum_{l=L-1}^{U-1} (1-p)^l. \end{aligned}$$

For  $2 \leq i \leq L \leq G - i \leq U$ ,

$$\begin{aligned} \mathbb{P}(C(i)) &= [p(i-1) + 1](1-p)^{L-1} - (pi + 1)(1-p)^U \\ (29) \quad &+ (U - G + i + 1)p(1-p)^{U-G+i}. \end{aligned}$$

For  $L < U \leq i < G - U$ ,

$$(30) \quad \mathbb{P}(C(i)) = [p(L-1) + 1](1-p)^{L-1} - (pU + 1)(1-p)^U.$$

For  $L \leq i \leq U \leq [G/2]$ ,

$$(31) \quad \mathbb{P}(C(i)) = [p(L-1) + 1](1-p)^{L-1} - [p(i+1) + 1](1-p)^U.$$

*Proof.* For the loop and the infinite models, and when  $L < U < i < G - U$  in the interval model, the proof is very simple. By Lemma 1, summing the probabilities that the fragment containing  $b$  has length  $l$ ,  $0 < L \leq l \leq U < G$ ,

$$\begin{aligned} \mathbb{P}(C(b)) &= \sum_{l=L}^U lp^2(1-p)^{l-1} = -p^2 \frac{d}{dp} \sum_{l=L}^U (1-p)^l = -p^2 \frac{d}{dp} \left( \frac{(1-p)^L}{p} - \frac{(1-p)^{U+1}}{p} \right) \\ &= [p(L-1) + 1](1-p)^{L-1} - (pU + 1)(1-p)^U, \end{aligned}$$

and

$$\mathbb{P}(\widetilde{C}(b)) = 1 - \mathbb{P}(C(b)) = 1 - [p(L-1) + 1](1-p)^{L-1} + (pU + 1)(1-p)^U.$$

For the other cases given for the interval model, the proof is similar but tedious, and so it is omitted.  $\square$

**Remark 4.** With the necessary simplification of iid, all the results given here are exact. The formulae in Theorem 3 are true for any  $0 < L < U$ , and it is not necessary to assume that  $L$  and  $U$  are both large, or  $p$  is small (although this is necessary for adopting the iid model, as explained above). When  $p$  is really small, then from the approximation

$$(1-p)^N = \exp(N \log(1-p)) = \exp\left(-N \sum_{k=1}^{\infty} \frac{1}{k} p^k\right) \cong e^{-pN}$$

it follows that

$$\mathbb{P}(C(b)) \cong [p(L-1) + 1]e^{-p(L-1)} - (pU + 1)e^{-pU}.$$

Moreover,  $\mathbb{P}(\widetilde{C}(b))$  can be calculated accurately; this is one of the main advantages of working with a clearly established model. Compare our probability,  $\mathbb{P}(\widetilde{C}(b))$ , with the lower bound estimate from Theorem 5.2 of [6], which states that

$$\mathbb{P}(\widetilde{C}(b)) \geq \exp(pLe^{-p(U-L)} - (1 - e^{-p(U-L)}) - pU) = \exp[-1 - pU + (pL + 1)e^{-p(U-L)}].$$

**Remark 5.** Note that in the probability of  $\widetilde{C}(b)$ , the first term,

$$1 - [p(L-1) + 1](1-p)^{L-1} = p^2 \sum_{l=1}^{L-1} l(1-p)^{l-1} = \sum_{l=1}^{L-1} \mathbb{P}(l)$$

is the probability of the event that site is in a fragment that is too short to be clonable; the second term,

$$(pU + 1)(1-p)^U = \sum_{l>U} \mathbb{P}(l),$$

is the probability of the event that the site is in a fragment that is too long to be clonable.

Dependent on the values of  $p$ ,  $L$ , and  $U$ , either of the two terms could be larger than the other, even significantly. For  $\lambda$  cloning vectors  $L = 2k$  bps (or  $= 10k$  bps) and  $U = 20k$  bps, while for cosmid vectors  $L = 20k$  bps and  $U = 45k$  bps ([6]). Taking  $p = 1/5000$ , the probability that the site is in a fragment that is too short to be clonable is 0.062 (0.594) for  $\lambda$  cloning vectors and 0.908 for cosmid vectors, while the probability that the site is in a fragment that is too long to be clonable is 0.092 for  $\lambda$  cloning vectors and 0.001 for cosmid vectors. Thus the conclusion in [4], that “a nucleotide is unclonable mostly because its flanking restriction sites are too apart” is not tenable.

**Theorem 4.** *In both the loop and the infinite models, and in the case of  $L < U < i$  in the interval model, the average length of clonable fragments is*

$$(32) \quad \mathbb{E}(X | L \leq X \leq U) = 2p^{-1}[(1-p)^{L+1} - (1-p)^{U+2}] + (2L+1)(1-p)^L - (2U+3)(1-p)^{U+1} + pL^2(1-p)^{L-1} - p(U+1)^2(1-p)^U.$$

If  $p$  is small and  $L$  and  $U$  are both large, then

$$(33) \quad \mathbb{E}(X | L \leq X \leq U) \cong 2p^{-1}[(1-p)^{L+1} - (1-p)^{U+2}] = 2 \sum_{l=L+1}^{U+1} (1-p)^l.$$

*Proof.*

$$\begin{aligned} \mathbb{E}(X | L \leq X \leq U) &= \sum_{l=L}^U p^2 l^2 (1-p)^{l-1} \\ &= \sum_{l=L}^U p^2 (l+1)l(1-p)^{l-1} - p^2 \sum_{l=L}^U l(1-p)^{l-1} \\ &= p^2 \frac{d^2}{dp^2} \sum_{l=L}^U (1-p)^{l+1} + p^2 \frac{d}{dp} \sum_{l=L}^U (1-p)^l \\ &= p^2 \frac{d^2}{dp^2} \frac{(1-p)^{L+1} - (1-p)^{U+2}}{p} + p^2 \frac{d}{dp} \frac{(1-p)^L - (1-p)^{U+1}}{p} \\ &= 2p^{-1}[(1-p)^{L+1} - (1-p)^{U+2}] + (2L+1)(1-p)^L \\ &\quad - (2U+3)(1-p)^{U+1} + pL^2(1-p)^{L-1} - p(U+1)^2(1-p)U. \end{aligned}$$

□

**Remark 6.** Equation (33) shows that the average length of a clonable fragment may be quite different from the intuitive guess  $(U + L)/2$ .

For the interval model, it follows from Theorem 3, that if it was possible to observe that the frequency of a certain base pair appearing in a clone was much smaller than the average frequency, then by Lemma 1, it could be inferred that  $b$  was near the ends (either beginning or ending) of the genome. If very large fragment sizes were possible, then the following result would be useful.

**Theorem 5.** *Assume restriction sites are distributed along a genome of  $G$  bps according to a Bernoulli process with  $p = \mathbb{P}(\text{restriction site})$ . Assume that  $U - L$  is large, then under the interval model, the relative frequency of the base pair  $j$  and base pair  $i$ ,  $1 \leq j < L$ ,  $U < i \leq \lfloor G/2 \rfloor$ , is*

$$\frac{\mathbb{P}(C(i))}{\mathbb{P}(C(j))} = \frac{[p(L-1) + 1](1-p)^{L-1} - (pU+1)(1-p)^U}{[p(j-1) + 1](1-p)^{L-1} - [p(j-1) + 1](1-p)^U} \cong \frac{p(L-1) + 1}{p(j-1) + 1}.$$

*Proof.* When  $U - L$  is large,

$$1 - (1-p)^{U-L+1} \cong 1$$

and

$$1 - \frac{pU+1}{p(L-1)+1}(1-p)^{U-L+1} \cong 1.$$

□

#### 4. THE PARTIAL DIGEST LIBRARIES

As described in [6], partial restriction digests are performed by stopping the digest before all sites are cut. Let  $\mu$  be the fraction of sites that are cut. Here it helps to think in terms of there being two independent Bernoulli processes; first an attempt is made to cut the site with probability  $\mu$ , then second this cut is successful with probability  $p$ . So the same models as above can be used with  $\mu p$  in stead of  $p$ . Therefore the probability that  $b$  (or  $b_i$ ) is clonable under partial digestion is obtained by substituting  $p$  by  $\mu p$  in the above formulae. In particular it is noted that a partial digest may increase the probability that the site is clonable.

Two theorems relating specifically to partial digests follow.

**Theorem 6.** *Under the loop and the infinite model, or under the interval model and  $L < U < i < G - U$ , for any  $0 < p < 1$ , define*

$$(34) \quad \mu_0(p, L, U) = p^{-1} \left[ 1 - \left( \frac{L(L-1)}{U(U+1)} \right)^{\frac{1}{U-L+1}} \right] > 0.$$

*If  $\mu_0 < 1$ , then*

$$\mathbb{P}_{\mu_0}(C(b)) > \mathbb{P}_{\mu}(C(b)), \quad \forall \mu \in [0, 1], \quad \mu \neq \mu_0;$$

otherwise

$$\mathbb{P}_1(C(b)) = \mathbb{P}(C(b)) > \mathbb{P}_\mu(C(b)), \quad \forall \mu \in [0, 1).$$

*Proof.* Define

$$f(\mu) = \mathbb{P}_\mu(C(b)) = [\mu p(L-1) + 1](1 - \mu p)^{L-1} - (\mu p U + 1)(1 - \mu p)^U,$$

then  $f(0) = 0$ .

$$f'(\mu) = \mu p^2(1 - \mu p)^{L-2}[U(U+1)(1 - \mu p)^{U-L+1} - L(L-1)].$$

$$\begin{aligned} f''(\mu) &= p^2(1 - \mu p)^{L-2}[U(U+1)(1 - \mu p)^{U-L+1} - L(L-1)] \\ &\quad - (L-2)\mu p^3(1 - \mu p)^{L-3}[U(U+1)(1 - \mu p)^{U-L+1} - L(L-1)] \\ &\quad - \mu p^3 U(U+1)(U-L+1)(1 - \mu p)^{U-2}. \end{aligned}$$

Since  $\mu p^2(1 - \mu p)^{L-2} > 0$  for  $\mu \in (0, 1)$ ,  $f$  can have at most one critical point in  $(0, 1)$ . If  $f$  does have a critical point in  $(0, 1)$ , it must be  $\mu_0 \in (0, 1)$ . In this case,  $f''(\mu_0) = -\mu_0 p^3 U(U+1)(U-L+1)(1 - \mu_0 p)^{U-2} < 0$ . So  $f(\mu_0)$  is a local maximum value. Since  $f$  does not have any other critical points in  $(0, 1)$ ,  $f(\mu_0)$  is also a maximum value in  $[0, 1]$ .

If  $\mu_0 > 1$ , then  $f$  does not have critical points in  $(0, 1)$ , so  $f$  is either strictly increasing, or strictly decreasing in  $[0, 1]$ . Since  $f(1) > 0$  and  $f(0) = 0$ ,  $f$  must be strictly increasing. Hence  $f(\mu) < f(1)$ ,  $0 \leq \mu < 1$ .  $\square$

Next, estimation of the conditional probability that a fixed base pair  $b$  is clonable but that in the partial digest library it is not clonable is determined.

**Theorem 7.** *Let  $E(\mu)$  be the event that  $b$  is not clonable under the partial digest,  $C(b)$  be the event that  $b$  is clonable. Then under both the loop model and the infinite model,*

$$(35) \quad \mathbb{P}(E(\mu)|C(b)) \leq \frac{(1 - \mu)(\mu p U + 1)(1 - \mu p)^{U-1}}{\mathbb{P}(C(b))}.$$

*Of course corresponding formulae for the other cases of the interval model could be given; the only difference is the formulae for  $\mathbb{P}_\mu((i, k))$ ,  $k > U$  in the various cases.*

*Proof.* Let  $C(l)$  be the event that  $b$  belongs to a length  $l$  fragment, and  $C_\mu(l)$  be the event that  $b$  belongs to a length  $l$  fragment under partial digestion. Then

$$E(\mu) = (\cup_{l=1}^{L-1} C_\mu(l)) \cup (\cup_{l \geq U+1} C_\mu(l)) = E_1(\mu) \cup E_2(\mu).$$

Since  $b$  is clonable but is not clonable under the partial digestion, then  $b$  belongs to a fragment of length  $l$ ,  $L < l < U$ , but under the partial digestion  $b$  belongs to a fragment of length  $l > U$ . Thus

$$\begin{aligned} C(b) &= \cup_{l=L}^U C(l), \quad E(\mu) \cap C(b) = E_2(\mu) \cap C(b). \\ \mathbb{P}(E(\mu)|C(b)) &= \frac{\mathbb{P}(E(\mu) \cap C(b))}{\mathbb{P}(C(b))} = \frac{\mathbb{P}(E_2(\mu) \cap C(b))}{\mathbb{P}(C(b))}. \end{aligned}$$

$$\mathbb{P}(E_2(\mu) \cap C(b)) = \sum_{l=L}^U \sum_{k>U} \mathbb{P}(C_\mu(k) \cap C(l)).$$

If  $b$  belongs to a fragment of length  $l$ , but under the partial digestion  $b$  belongs to a fragment of length  $k > U$ , it must be that a cut was not made because the position was not chosen. So

$$\mathbb{P}(C_\mu(k) \cap C(l)) \leq \frac{1-\mu}{1-\mu p} \mathbb{P}_\mu(C_\mu(k)).$$

But for both the loop model and the infinite model,

$$\sum_{k>U} \mathbb{P}(C_\mu(k)) = (\mu p U + 1)(1 - \mu p)^U.$$

Thus,

$$\begin{aligned} \mathbb{P}(E_2(\mu) \cap C(b)) &\leq \frac{1-\mu}{1-\mu p} (\mu p U + 1)(1 - \mu p)^U \\ &= (1-\mu)(\mu p U + 1)(1 - \mu p)^{U-1} \end{aligned}$$

□

**Remark 7.** Since

$$\mathbb{P}(E(\mu)|C(b)) \leq \frac{(1-\mu)(\mu p U + 1)(1 - \mu p)^{U-1}}{\mathbb{P}(C(b))} \cong \frac{(1-\mu)(\mu p U + 1)e^{-\mu p(U-1)}}{\mathbb{P}(C(b))},$$

if  $\mu p$  is small, it is found that  $\mathbb{P}(E(\mu)|C(b))$  is almost zero when  $U$  is large.

#### REFERENCES

- [1] International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409: 860-921, 2001.
- [2] G. P. Rédei (1998). *Genetics Manual*. World Scientific, Singapore
- [3] S. M. Ross (1993). *Introduction to Probability Models*, 5th Edition. Academic Press, Boston
- [4] B. Tang and M. S. Waterman (1990). The expected fraction of clonable genomic DNA. *Bulletin of Mathematical Biology*, 52(3):455-75, 1990.
- [5] J. C. Venter, et al (2001). The sequence of the human genome. *Science*, 291:1304-1351, 2001.
- [6] M. S. Waterman (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall, London

CENTRE FOR MATHEMATICS AND ITS APPLICATIONS, SCHOOL OF MATHEMATICAL SCIENCES,  
AUSTRALIAN NATIONAL UNIVERSITY, CANBERRA, ACT 0200, AUSTRALIA

*E-mail address:* yi@maths.anu.edu.au

CENTRE FOR BIOINFORMATION SCIENCE, SCHOOL OF MATHEMATICAL SCIENCES, AUSTRALIAN  
NATIONAL UNIVERSITY, CANBERRA, ACT 0200, AUSTRALIA

*E-mail address:* Sue.Wilson@anu.edu.au