Bulletin of Mathematical Biology Vol. 52, No. 3, pp. 455–475, 1990.
Printed in Great Britain.

0092-8240/90$3.00+0.00
Pergamon Press plc

THE EXPECTED FRACTION OF CLONABLE GENOMIC DNA*

■ BETTY TANG and MICHAEL S. WATERMAN
Department of Mathematics,
University of Southern California
Los Angeles, CA 90089-1113, U.S.A.

Random clone mapping of genomic DNA is a subject of great interest in molecular biology. *E. coli* has just been mapped and work is progressing on some human chromosomes. In this paper we give estimates of the fraction of genomic DNA which is not clonable by partial digest with a restriction enzyme.

1. Introduction. A revolution in biology has resulted from the recent advances in manipulating and reading DNA from the genomes of various organisms. All 48 502 base pairs (nucleotides) of λ, a bacteriophage that infects *E. coli*, are known, as are all 172 282 base pairs of Epstein–Barr virus that infects humans. In spite of all the progress of the last decade, the task of sequencing all 4.7×10^6 base pairs of *E. coli* (Daniels and Blattner, 1986) is still formidable, although it should be accomplished in the next few years. Recently there has been much talk of sequencing the human genome of 3×10^9 base pairs (Roberts, 1987). After much discussion, the work is proceeding by first mapping the genome. One approach to mapping is by overlapping clones (see below) along the linear sequence. *E. coli* has recently been mapped by overlapping clones and is the largest genome that has been characterized in this fashion (Kohara *et al.*, 1987). Two other large mapping projects have recently been reported: yeast (Olson *et al.*, 1986) and nematode (Coulson *et al.*, 1986).

Construction of *recombinant partial digest DNA libraries* is a standard technique in genetic engineering (Dahl *et al.*, 1981), and is, among other things, preliminary to mapping. The technique was developed to overcome the difficulty of handling a long DNA molecule by obtaining various DNA fragments short enough to be more easily studied in the laboratory. (The length of a fragment is the number of nucleotides contained in it. The unit is normally 1 kbp $\equiv$ 1000 nucleotides.) The procedure involves essentially the following three steps (Dahl *et al.*, 1981; Lewin, 1985).

(1) Degradation of DNA molecules into shorter fragments by restriction enzyme digestion. This is carried out with the help of site-specific endonucleases which are

* Research supported by grants from the National Institutes of Health and the System Development Foundation.

enzymes that cleave the DNA molecules at specific short sequences. A complete digest allows all the recognition sites (or restriction sites) of a specific restriction enzyme to be cleaved. In contrast, a partial digest is a digest that is interrupted before completion, hence only a certain fraction of the restriction sites are cut.

(2) Cloning the DNA fragments. The DNA fragments are inserted into viral carriers which commonly are either the *bacteriophage λ vectors*, *plasmid vectors* or *cosmid vectors*. There are restrictions on the size range of the DNA fragments that can be accommodated by the various vectors.

(3) Selection (usually random) of the recombinant DNA molecules to form a library.

One of the biologists' major concerns about a recombinant library is how representative it is, i.e. whether the library contains all the genetic information of the genome. It is known that certain sequence patterns will cause any fragment containing them to be unclonable in a λ vector. For mathematical purposes, we take the point of view that all clonable fragments are equally clonable. We identify two main causes for some nucleotides to be not represented in the library. Since a vector can accommodate DNA fragments within a certain size range, some nucleotides will not be *clonable*—they can only be contained in fragments of lengths unsuitable to be incorporated in the vectors. Furthermore, some nucleotides, even though clonable, could be *unselected*—they are contained only in fragments left out from the selection process and therefore missing from the library. (Our terminology differs slightly from the standard one used by molecular biologists.) Ideally only unclonable nucleotides are missing from the library.

The fraction of nucleotides that are unclonable after a complete digest can be determined easily by consideration of the distribution of restriction sites and is well known to be smaller than that by partial digest. The estimates of the fraction of nucleotides that are not clonable by partial digest and the fraction unselected have been carried out in two pioneering papers (Seed, 1982; Seed *et al.*, 1982). In work published by Seed in the journal *Biopolymers* (Vol. 21, 1982), the fraction is derived by considering whether a nucleotide is located between any two restriction sites spaced some suitable length apart and is expressed as a product of recursively defined conditional probabilities. The solution by recursion is computationally expensive and does not give any analytical insight about the dependence of the fraction on the biological parameters.

It is the purpose of this paper to obtain an estimate of the fraction of unclonable nucleotides in simple form for parameter ranges within practical regimes. A more practical consideration, considered by Seed *et al.* (1982) but not in this paper, is the coverage of the genome by a finite sampling of cloned DNA since all clonable fragments are not equally clonable. Our approach in this paper is of a combinatorial nature. The central idea is to describe the different patterns of the distribution of restriction sites with respect to an arbitrary nucleotide such that it is not clonable. It turns out that in some cases only four sites, two on each side of the nucleotide, are sufficient to characterize

these distributions, and we are able to express the fraction in closed form. In other cases where the four sites do not provide sufficient information, we obtain upper and lower bounds. Our study shows that the fraction unclonable depends mainly on the maximum clonable length $(L+r)$, and in a minor way on the minimum clonable length (L). The main conclusion of this paper is that for most practical purposes, a reasonably good estimate of the fraction of clonable nucleotides is:

$$1-e^{-p(L+r)}(1+p(L+r))-\tfrac{1}{6}p^2L^2e^{-p(L+2r)}(pL+3),$$

here p is the frequency of occurrence of restriction enzyme restriction sites.

This paper is organized as follows. Various assumptions and definitions used in our calculations are given in Section 2. For better understanding of later sections the estimation of the fraction unclonable by complete digest is presented in Section 3, even though the expression has been derived in a somewhat different context (Kuhn, 1930; Montroll and Simha, 1940). The main results are stated in Section 4 [equations (11) and (12)] and are explained in detail in Sections 5 and 6. Section 7 contains a calculation to show that all clonable fragments are present in typical experiments, although we do not study the properties of small samples (a few genome equivalents) from this distribution. Numerical studies of the derived estimates and concluding remarks are contained in Section 8.

2. Assumptions and Definitions. We basically adhere to the assumptions and notation of Seed (1982) with additional ones pertinent to our calculations. All the genetic information of an organism is assumed to be contained in one single large DNA molecule which can be taken as a random sequence of the four nucleotides. The fraction of unclonable nucleotides and the probability that an arbitrary nucleotide is unclonable are the same and will be referred to interchangeably. The total number of nucleotides in a genome is N and the number of (identical) DNA molecules in a sample prepared for complete or partial digest is denoted by K.

The restriction sites of a particular restriction enzyme are, to first approximation, randomly distributed throughout the DNA molecule (Hamer and Thomas, 1975). Let p be the frequency of occurence of such sites. In practice $N \gg 1/p$. To be precise, p is the probability that an internucleotide bond can be broken by the restriction enzyme. For example, the sequence GAATTC is the restriction site for the restriction enzyme Eco R1 which cuts the site by breaking the bond between G and A. In this case p can be viewed as the probability that any bond in the DNA molecule is between G and A in the sequence GAATTC. Henceforth a restriction site refers to the bond that can be broken by the restriction enzyme. It is also assumed that all the bonds in the DNA molecule have equal probability p of being restriction sites, even though

two such bonds must be at least some distance apart (6 nucleotides in the case of Eco R1). As we will see in the next section, when p is small enough the average distance between two restriction sites is large, so the error introduced with this assumption will be negligible.

For an arbitrary nucleotide b^* in the DNA molecule, we will label the i^{th} bond to the left and to the right of b^* as positions $-i$ and i respectively. If there are m restriction sites to the left of b^* at positions $x_1, x_2, \ldots, x_m$ where $-1 \geqq x_1 > x_2 > \ldots > x_m$, and n restriction sites to the right of b^* at positions $y_1, y_2, \ldots, y_n$ where $1 \leqq y_1 < y_2 < \ldots < y_n$, the location of restriction sites (relative to b^*) is denoted by:

$$(x; y)_{b^*} = (x_m, \ldots, x_2, x_1; y_1, y_2, \ldots, y_n).$$

We will call $(x; y)_{b^*}$ the *restriction sites configuration* (rsc) with respect to b^*. (The case of no restriction sites on either side of b^* is not "generic" when $N \gg 1/p$ and we assume in general there are restriction sites on both sides of b^*.) The subscript b^* will be conveniently dropped so $(x; y)$ refers to the rsc with respect to the nucleotide under consideration.

The size range of DNA fragments that can be incorporated in a particular type of cloning vectors is between $L+1$ and $L+r$. For many cloning vectors, $r \sim 10$ kbp and L varies between 0 and $2r$. Moreover, both L and r are relatively small compared with N. When the difference is insignificant, L, instead of $L+1$, is referred to as the minimum clonable length. In a partial digest where the restriction sites of one DNA molecule at x_i, $1 \leqq i \leqq m$, and at y_j, $1 \leqq j \leqq n$, are cut but those at $x_{i-1}, \ldots, x_1, y_1, \ldots, y_{j-1}$ are intact, the resulting DNA fragment containing b^* is denoted by $[x_i, y_j]$. The latter is called a *clonable fragment* if its length $l(x_i, y_j)$, given by:

$$l(x_i, y_j) = y_j - x_i - 1,$$

satisfies $L < l(x_i, y_j) \leqq L+r$. We will say b^* has a *clonable configuration* if its rsc $(x; y)$ is such that $[x_i, y_j]$ for some i and some j, $1 \leqq i \leqq m$, $1 \leqq j \leqq n$, is a clonable fragment when the appropriate restriction sites are cut.

A consecutive sequence of bonds located from position k_1 to k_2 (with respect to b^*) is denoted by the set:

$$\{k_1, k_2\} = \{k \mid k_1 \leqq k \leqq k_2\}.$$

Note the difference of this from the fragment $[k_1, k_2]$ where the bonds at k_1 and k_2 must be restriction sites. The bond at position k, where $k_1 \leqq k \leqq k_2$, is loosely referred to as the bond in the set $\{k_1, k_2\}$. The number of bonds in a set A is denoted by $|A|$. In particular,

$$|\{k_1, k_2\}| = \begin{cases} k_2 - k_1 + 1, & \text{if } \operatorname{sgn}(k_1) = \operatorname{sgn}(k_2), \\ k_2 - k_1, & \text{otherwise.} \end{cases}$$

Examples of some typical values of the various parameters are given in the appendix.

3. Fractions of Nucleotides not Clonable by Complete Digest. Since all restriction sites are broken in a complete digest, all of the K DNA molecules are broken, always at the same bonds, and it suffices to focus our attention on any one of them. Moreover, we can consider the DNA molecule as being broken up randomly with each bond having equal probability p of being broken. Theoretical study of such processes have been undertaken (Kuhn, 1930; Montroll and Simha, 1940) and we give a brief description according to our notation.

Let $f(l)$ be the probability that an arbitrary nucleotide b^* is contained in a fragment of length l after complete digest.

PROPOSITION 1. $f(l)=lp^2(1-p)^{l-1}$.

Proof. Observe that after complete digest b^* can only be contained in the fragment $[x_1, y_1]$, therefore $f(l)$ is also the probability that $(x; y)$ is such that $l(x_1, y_1)=y_1-x_1-1=l$. Since $y_1 \geqq 1$, the value of x_1 can only vary between -1 and $-l$. For each fixed value of x_1, there are $l-1$ bonds between x_1 and y_1 which must not be restriction sites and the probability for that to be the case is given by the geometric distribution $p^2(1-p)^{l-1}$. Summing over all the possible values of x_1 and we obtain the expression for $f(l)$. ■

Remark. Technically the above expression for $f(l)$ is not true if there are less than l bonds on either side of b^*. The correction term is however insignificant when $N \gg 1/p$.

Henceforth we will refer to the two restriction sites at x_1 and y_1 as the *flanking* restriction sites of b^*. Also when p is sufficiently small we can express $f(l)$ as:

$$f(l)=lp^2e^{-p(l-1)}. \tag{1}$$

It is easily shown that $f(l)$ is maximum when $l=1/p$ and is practically zero when l is very large. The shape of the curve of $f(l)$ is shown in Fig. 1.

Let F_0^+ and F_0^- be respectively the fraction of nucleotides which are clonable and unclonable by complete digest. Obviously:

$$F_0^+ = \sum_{l=L+1}^{L+r} f(l). \tag{2}$$

When $L+r$ is large enough, the summation in equation (2) can be approximated by an integral:

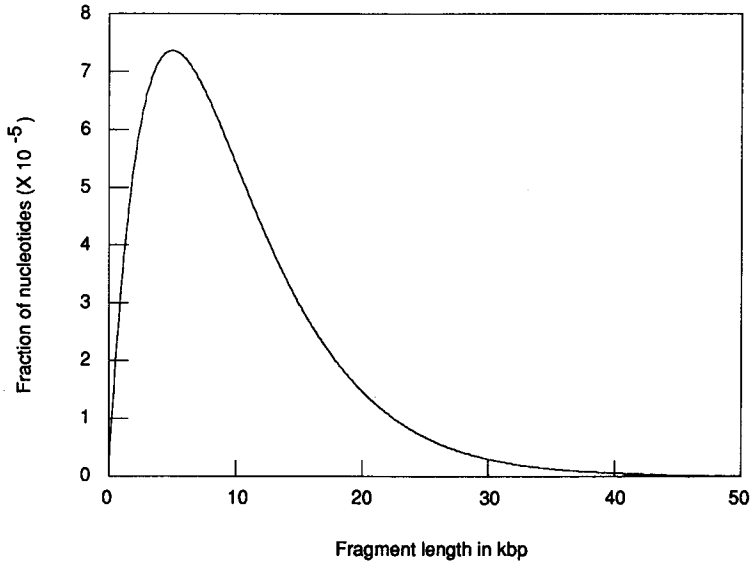

Figure 1. Length distribution of fragment lengths after complete digest where $p \approx 1/5\ \text{kbp}^{-1}$.

$$F_0^+ \approx \int_L^{L+r} f(l)\, dl. \tag{3}$$

Expressing $f(l)$ as in equation (1) and evaluating the integral, we obtain:

$$F_0^+ \approx (Lp+1)e^{-pL} - ((L+r)p+1)e^{-p(L+r)}. \tag{4}$$

Notice that F_0^+ is approximately the area under the graph of $f(l)$ between $l=L$ and $l=L+r$. If $L \ll 1/p$ and $r \gg 1/p$, as in the case for most bacteriophages, F_0^+ corresponds to the area of a large region and the fraction of unclonable nucleotides is not very large. But if $L \gg 1/p$, as is the case for most cosmid vectors, F_0^+ is the area of a small region so that a large part of the genome is not clonable. As an example, suppose the *E. coli* DNA molecule is digested with the enzyme Eco R1 where the corresponding restriction sites occur at $p \approx 1/5000$. If the DNA fragments are cloned by the phage λgt WES ($L=2.4$ kbp, $r=15$ kbp), $F_0^- \approx 22\%$. If the cosmid vector pJC74 ($L=19$ kbp, $r=17$ kbp) is used, $F_0^- \approx 90\%$. We will see later how much F_0^- can be reduced if the DNA molecules undergo partial digestion.

4. Fraction of Nucleotides not Clonable by Partial Digest. Let F^- be the fraction of nucleotides not clonable by partial digest. If the flanking restriction sites of an arbitrary nucleotide b^* are too far apart ($l(x_1, y_1) > L+r$), b^* will

never be contained in any clonable fragment. On the other hand, if they are too close together ($l(x_1, y_1) \leqq L$), b^* might still be clonable if it has a clonable configuration and at least one of the K DNA molecules is cut by the restriction enzyme at appropriate restriction sites to yield a clonable fragment containing b^*. The minimum clonable length is therefore not a major constraint. This is different from complete digest where the maximum and minimum clonable lengths are equally important constraints. It is thus clear that $F^- < F_0^-$.

Let S_+ and S_- be respectively the set of all possible forms of $(x; y)$ such that b^* has and has not a clonable configuration. We distinguish two cases or events such that b^* is not clonable:

(A1) $(x; y) \in S_-$;
(A2) $(x; y) \in S_+$ but none of the K DNA molecules is cut by the restriction enzyme in an appropriate fashion so that b^* is contained in a clonable fragment.

Let $P(\text{A1})$ and $P(\text{A2})$ be the probability of the two events respectively, then:

$$F^- = P(\text{A1}) + P(\text{A2}). \tag{5}$$

For K large enough, it can be shown that $P(\text{A2})$ is effectively zero (see Section 7), therefore:

$$F^- \approx P(\text{A1}). \tag{6}$$

Note that event A1 depends only on the distribution of restriction sites so F^- as expressed in (6) does not depend on the fraction of restriction sites cut in a partial digest.

In order to calculate $P(\text{A1})$, we first partition S_- into two subsets S_-^1 and S_-^2 where:

$$S_-^1 = \{(x; y) \in S_- \,|\, l(x_1, y_1) > L + r\},$$

the set of all rsc's such that after partial digest b^* can only be contained in fragments too long to be clonable, and:

$$S_-^2 = \{(x; y) \in S_- \,|\, l(x_1, y_1) \leqq L\},$$

the set of all those rsc's such that b^* can only be contained in fragments which are too short and/or too long to be clonable. Then $P(\text{A1})$ can be expressed as:

$$P(\text{A1}) = P((x; y) \in S_-^1) + P((x; y) \in S_-^2). \tag{7}$$

Henceforth we will denote $P((x; y) \in \Omega)$ by $P(\Omega)$ for any subset Ω of S_-^1 or S_-^2.

Since any $(x; y)$ in S_-^1 is characterized by the flanking restriction sites of b^* being too far apart, we can easily express $P(S_-^1)$ in terms of $f(l)$:

$$P(S_{-}^{1}) = \sum_{l=L+r+1}^{N} f(l),$$

$$\approx 1 - \sum_{l=1}^{L+r} f(l),$$

$$\approx 1 - \int_{0}^{L+r} lp^2 e^{-pl}\, dl,$$

where approximation is taken for very large N and $L+r$. On simplification we get:

$$P(S_{-}^{1}) \approx (p(L+r)+1)e^{-p(L+r)}. \tag{8}$$

The expression for $P(S_{-}^{2})$ is more complicated and we treat the case of $L \leqq r+1$ and of $L > r+1$ separately. (Most bacteriophages satisfy $L < r$ and most cosmid vectors satisfy $r < L < 2r$.)

THEOREM 1. *For* $L \leqq r+1$:

$$P(S_{-}^{2}) = p^2(1-p)^{2(L+r)} \sum_{t=1}^{L} (1-p)^{-(t+1)} g(t), \tag{9}$$

where:

$$g(t) = p^2t^3/6 + p(1-p/2)t^2 + (p^2/3 - p + 1)t.$$

See Section 5 for the proof of the theorem. Approximating the sum by an integral:

$$P(S_{-}^{2}) \approx G(L, r, p), \tag{10}$$

where:

$$G(L, r, p) = \tfrac{1}{6}p^2L^2e^{-p(L+2r)}(pL+3).$$

Note that $P(A1) = P(S_{-}^{1})$ if $L = 0$.

Substituting (8) and (10) into (7), an estimation for F^{-} for $L \leqq r+1$ can be obtained in closed form:

$$F^{-} \approx e^{-p(L+r)}(1+p(L+r)) + G(L, r, p). \tag{11}$$

Note that when r is large $G(L, r, p)$ is small compared with the first term. This is consistent with our earlier remark that the maximum clonable length is the

major constraint in a partial digest. Numerical studies of this expression is carried out in Section 8.

THEOREM 2. *For* $L>r+1$:
(a) $P(S_-^2)\geqq G(L, r, p)$;
(*b*) $P(S_-^2)<c(L)G(L, r, p)$ *where*:

$$c(L)=\begin{cases}1+e^{-pr}p(L-r)[1+p(L-r)/2], & L\leqq 2r,\\ 1+e^{p(L-2r)}, & 2r<L\leqq 4r,\\ [1+e^{p(L-3r)}]^2, & 4r<L.\end{cases}$$

See Section 6 for the proof. The corresponding estimate of F^- is:

$$F^- \sim e^{-p(L+r)}(1+p(L+r))+c'G(L, r, p), \tag{12}$$

where c' varies between 1 and $c(L)$. Note that when r is large, the upper bound and lower bound for $L\leqq 2r$ are very close to each other. For other values of L, the upper bound is only a very loose one. The comparison of the upper and lower bounds is carried out in Section 8.

5. Estimation of F^- for $L\leqq r+1$. For each $(x; y)\in S_-^2$, define:

$$t=\max_{\substack{i,j,\\ [x_i,y_j]\leqq L}} l(x_i, y_j).$$

In other words, t is the maximum length of all possible fragments in which b^* could be found after partial digest but are too short to be clonable. The two restriction sites which give rise to the fragment (containing b^*) of length exactly t are labelled x_a and y_d respectively.

LEMMA 1. *For each* $(x; y)\in S_-^2$, x_a *and* y_d *as described above can be identified and are unique.*

Proof. For each $(x; y)\in S_-^2$, $l(x_1, y_1)\leqq L$, so t is well defined and the labelling of x_a and y_d is always possible. Suppose x_a and y_d are not unique and there exist $\bar{x}_a$ and $\bar{y}_d$ such that $l(\bar{x}_a, \bar{y}_d)$ is also t. Without loss of generality assume $x_a<\bar{x}_a$ (hence $y_d<\bar{y}_d$) and note that $-t\leqq x_a\leqq -1$. Since $(x; y)$ does not have a clonable configuration, either $l(x_a, \bar{y}_d)\leqq L$ or $l(x_a, \bar{y}_d)>L+r$. If the former case holds, $[x_a, \bar{y}_d]$ is also too short to be clonable but longer than t, a contradiction. If the latter case holds, there are more than $t-1$ nucleotides between x_a and $\bar{x}_a$, again a contradiction. ■

For each t, $1\leqq t\leqq L$, let C_t be the set of all $(x; y)$ in S_-^2 such that the longest fragment too short to be clonable (namely $[x_a, y_d]$) is of length t.

THEOREM 3. *The family of sets* $\{C_t\}_{1 \leqq t \leqq L}$ *forms a partition of* S^2_-, *i.e.*

$$\bigcup_{t=1}^{L} C_t = S^2_- \text{ and } C_t \cap C_s = 0 \text{ for all } t \neq s.$$

COROLLARY 1. $P(S^2_-) = \sum_{t=1}^{L} P(C_t)$.

LEMMA 2. *If* $(x; y) \in C_t$, *any bond in* $\{y_d+1, x_1+(L+r)+1\} \cup \{y_1-(L+r)-1, x_a-1\}$ *is not a restriction site.*

Proof. If $(x; y) \in C_t$, x_a and y_d must satisfy $-t \leqq x_a \leqq -1$ and $y_d = x_a + t + 1$. Since all fragments of the form $[x_1, y_j]$ are not clonable, we must have $y_j < x_1 + L + 2$ or $y_j > x_1 + (L+r) + 1$ for any y_j. Similarly, we must have $y_j < x_a + L + 2$ or $y_j > x_a + (L+r) + 1$ when considering fragments of the form $[x_a, y_j]$. Hence all bonds in $\{x_1+L+2, x_1+(L+r)+1\} \cup \{x_a+L+2, x_a+(L+r)+1\}$ are not restriction sites. Observe that for $L \leqq r+1$:

$$x_1 + L + 2 - (x_a + (L+r) + 1) \leqq 1,$$

which means $\{x_1+L+2,\ x_1+(L+r)+1\} \cup \{x_a+L+2,\ x_a+(L+r)+1\} = \{x_a+L+2, x_1+(L+r)+1\}$. Next we consider bonds in $\{y_d+1, x_a+L+1\}$. If any one of these is a restriction site, there will be a fragment $[x_a, y_j]$ longer than t and still too short to be clonable, i.e. $(x; y) \in C_s$ where $t < s \leqq L$, which is a contradiction. Analogous arguments show that all the bonds in $\{y_1-(L+r)-1, x_a-1\}$ to the left of b^* also cannot be restriction sites. ■

As a result of the last lemma, all $(x; y) \in C_t$ must be of the form indicated in Fig. 2. Note that it is not necessary to specify the restriction sites in $\{x_a+1, x_1-1\} \cup \{y_1+1, y_d-1\}$, i.e. b^* is not clonable whether the bonds in these sets of $(x; y) \in C_t$ each of which characterized by the triple $v = (x_a, x_1, y_1)$ instead of regarding all rsc's in C_t as being distinct. (The restriction site y_d is omitted—it is

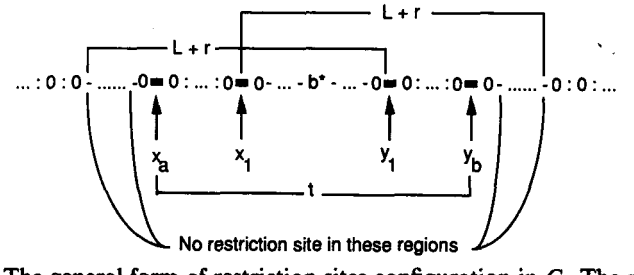

Figure 2. The general form of restriction sites configuration in C_t. The nucleotide under consideration is "b^*", while the other nucleotides are represented by "0". An internucleotide bond is indicated by "■" if it is a restriction site, by "-" if it is not, and are restriction sites or not. It is therefore natural to consider equivalence classes by ":" if it can be either.

determined by x_a.) Let I_t be the set of all possible choices of x_a, x_1 and y_1:

$$I_t=\{v=(x_a, x_1, y_1)\,|\,-t\leqq x_a\leqq -1,\ x_a\leqq x_1\leqq -1,\ 1\leqq y_1\leqq y_d=x_a+t+1\}.$$

For each triple $v\in I_t$, let $E_t(v)$ be the corresponding equivalence class of $(x; y)$. Then:

$$P(C_t)=\sum_{v\in I_t} P(E_t(v)).$$

THEOREM 4.

$$P(C_t)=p^2(1-p)^{2(L+r)-(t+1)}\times$$

$$\sum_{x_a=-t}^{-1}[-p^2x_a^2+p^2(1+t)x_a+(1-p)(1+tp)].$$

Proof. All the rsc's in each equivalence class $E_t(v)$ are such that the bonds in:

$$\{y_1-(L+r),\ x_a-1\}\cup\{x_1+1,\ -1\}\cup\{1,\ y_1-1\}\cup\{y_d+1,\ x_1+(L+r)+1\},$$

must not be restriction sites. Direct calculation shows there are $M=2(L+r)-(t+1)$ bonds in the above set. The total number of bonds which must be restriction sites, on the other hand, can be 2, 3 or 4 depending on whether $x_a=x_1$ or $y_d=y_1$. Let $n(v)$ be the number of restriction sites determined by v. When x_a is fixed let n_2, n_3 and n_4 be corrspondingly the number of choices of x_1 and y_1 such that $n(v)=2$, 3 or 4. Clearly, $n_2=1$. There are $-(x_a+1)$ choices of x_1 such that $x_a\neq x_1$, and $y_d-1=x_a+t$ choices of y_1 such that $y_1\neq y_d$, therefore $n_3=t-1$ and $n_4=-(x_a+1)(x_a+t)$. Consequently:

$$P(E_t(v))=p^{n(v)}(1-p)^M,$$

and:

$$P(C_t)=\sum_{v\in I_t} P(E_t(v)),$$

$$=p^2(1-p)^{2(L+r)-(t+1)}\sum_{x_a=-t}^{-1}[n_2+n_3p+n_4p^2]$$

$$=p^2(1-p)^{2(L+r)-(t+1)}\sum_{x_a=-t}^{-1}[1+(t-1)p-(x_a+1)(x_a+t)p^2],$$

which on rearrangement of terms gives the expression for $P(C_t)$ in the statement of the lemma. ■

The summation in Theorem 4 involves a quadratic polynomial in x_a and can be evaluated explicitly to give:

$$P(C_t)=p^2(1-p)^{2(L+r)-(t+1)}g(t), \tag{13}$$

where:

$$g(t)=p^2t^3/6+p(1-p/2)t^2+(p^2/3-p+1)t.$$

Hence by Corollary 1 we obtain:

$$P(S_-^2)=p^2(1-p)^{2(L+r)}\sum_{t=1}^{L}(1-p)^{-(t+1)}g(t).$$

The last summation can be approximated by the integral:

$$\int_0^L e^{pt}g(t)\,dt,$$

which upon evaluation gives:

$$\tfrac{1}{6}p^2e^{-p(L+2r)}\{pL^3+3(1-p)L^2+2pL+2e^{-pL}-2\},$$
$$\approx\tfrac{1}{6}p^2L^2e^{-p(L+2r)}(pL+3),$$

in (10).

Remark. Our description of nonclonable rsc's of b^*, in particular Lemma 2, assumes that there are at least $L+r$ bonds on both sides of b^*. If b^* is located at a distance less than $L+r$ from either end of the molecule, it is clearly more likely to be unclonable. In practice, $L+r$ is relatively small compared with N and our description does not hold for only a negligible portion of the nucleotides.

6. Estimation of F^- for $L>r+1$. When $L>r+1$ Lemmas 1 and 2 do not hold in general. We illustrate this with the following example.

Example 6.1. Suppose $L=8, r=2$ and $t=7$. Let $(x; y)\in E_7(-6, -2, 1)$. Then the bonds at -9, -8, -7, 3, 4, 5, 8, 9, 10 cannot be restriction sites. If the bonds in $\{-5, -2\}$ are not specified, there is no definite conclusion about the bonds at 6 and 7. This creates a "hole" to the right of b^* (at $\{6, 7\}$) instead of a consecutive string of bonds which must not be restriction sites. If there is a restriction site at 7 but not at -2, -3, -4, -5 or 6, $(x; y)$ is still a rsc in $E_7(-6, -2, 1)$. If there are restriction sites at -2 and 6, $(x; y)$ is also a rsc in E_7 $(-2, -1, 6)$. If the bonds at -3 and 6 are restriction sites, $(x; y)$ is in C_8. Lastly, if the bonds at -4 and 7 are restriction sites, $(x; y)$ is a clonable configuration.

Let $(x; y)\in C_t$ where there is a restriction site at k. Define:

$$k' = \begin{cases} k+t+2, \text{ if } k \leqq -1, \\ k-t-2, \text{ if } k \geqq 1, \end{cases}$$

and:

$$k'' = \begin{cases} k+(L+r)+1, \text{ if } k \leqq -1, \\ k-(L+r)-k \geqq 1. \end{cases}$$

In order that there is no fragment (containing b^*) of the form $[k, \cdot]$ ($[\cdot, k]$) which is longer than t and shorter than $L+r$, all bonds in $\{k', k''\}$ ($\{k'', k'\}$) must not be restriction sites. A hole is created to the right of b^* if:

$$x'_1 > x''_a + 1,$$

and to the left of b^* if:

$$y'_1 < y''_d - 1.$$

This is expressed more explicitly in the following lemma.

LEMMA 3. *Let* $(x; y) \in C_t$. *There is exactly one hole if and only if* $L \geqq r+2$, $(L+r)/2+1 \leqq t \leqq 2(L+r+1)/3$ *and either*:

$$(a)\ -t \leqq x_a \leqq -(L+r-t+2) \text{ and } L+r+x_a+1-t \leqq x_1 \leqq -1,$$

or:

$$(b)\ L+r-2t+1 \leqq x_a \leqq -1 \text{ and } 1 \leqq y_1 \leqq x_a+2t-(L+r).$$

The hole is to the right of b^* if (a) is satisfied and to the left if (b) is satisfied. There is simultaneously one hole on each side of b^* if and only if $L \geqq 2r+3$, $t > 2(L+r+1)/3$ and both (a) and (b) are satisfied.

For convenience sake, let $I_t = I_{t0} \cup I_{t1} \cup I_{t2}$ where I_{t0} is the set of those triples $v = (x_a, x_1, y_1)$ such that the corresponding rsc contains no hole(s), and I_{t1} and I_{t2} are respectively the set of those satisfying (a) and (b) in Lemma 3. Also the hole to the right (left) of b^*, i.e. the set of bonds in $\{x''_a+1, x'_1-1\}$ ($\{y''_1+1, y''_d-1\}$), will be denoted by $X(W)$, and the set of bonds between x_a and x_1 (between y_1 and y_b), by $Y(V)$.

When $L > r+1$, the expression for $P(C_t)$ given by (13) still holds if $t < (L+r)/2+1$. On the other hand, an accurate derivation for $P(C_t)$ requires more information on the location of restriction sites other than x_a, x_1, and y_1 when $t \geqq (L+r)/2+1$. Since it is not feasible to keep track of all the possible ways the restriction sites can be located in X, Y, W and V such that b^* does not have a clonable configuration, we will determine upper and lower bounds—in closed form—for $P(C_t)$ and consequently for $P(S^2_-)$.

Let $v \in I_{t1} \cup I_{t2}$. If $(x; y)$ is a rsc in $E_t(v)$ where there are some restriction sites in X and/or W, in some cases $(x; y)$ is also a rsc in $E_t(v')$ where $v \neq v'$, as seen in Example 6.1. Therefore even though:

$$\bigcup_{v \in I_t} E_t(v) = C_t,$$

the family of sets:

$$\{E_t(v)\}_{v \in I_t},$$

does not constitute a partition of C_t anymore. This gives an upper bound for $P(S_-^2)$:

$$P(S_-^2) = \sum_{t=1}^{L} P(C_t) < \sum_{t=1}^{L} \sum_{v \in I_t} P(E_t(v)). \tag{14}$$

This expression will be estimated by obtaining upper bounds of $P(E_t(v))$ for different values of t and v. We first establish a lower bound for $P(S_-^2)$.

If $v \in I_{t0}$, $E_t(v) \cap E_t(v') = 0$ for any $v' \neq v$, hence:

$$P(C_t) = \sum_{v \in I_{t0}} P(E_t(v)) + P\left(\bigcup_{v \in I_{t1} \cup I_{t2}} E_t(v) \right). \tag{15}$$

Consider $E_t(v)$ where $v \in I_{t1} \cup I_{t2}$. Let $E_t^0(v) \subset E_t(v)$ be the set of those $(x; y)$ such that all bonds in X and W are not restriction sites. In that case the bonds in Y and V can be unspecified. Such forms of $(x; y)$ are contained in some $E_t^0(v)$ for exactly one triple v, so:

$$P\left(\bigcup_{v \in I_{t1} \cup I_{t2}} E_t(v) \right) > \sum_{v \in I_{t1} \cup I_{t2}} P(E_t^0(v)),$$

and hence:

$$P(C_t) > \sum_{v \in I_{t0}} P(E_t(v)) + \sum_{v \in I_{t1} \cup I_{t2}} P(E_t^0(v)).$$

The forms of all $(x; y)$ included in the last expression are exactly the same as those described in the last section, so:

$$h(t) = p^2(1-p)^{2(L+r)-(t+1)} g(t), \tag{16}$$

the expression on the right hand side of (13), is a lower bound for $P(C_t)$. Direct substitution of (16) into $P(S_-^2)$ proves part (a) of Theorem 2.

The next lemma gives the main idea to obtain upper bounds of $P(E_t(v))$ for different values of t and v.

LEMMA 4. *Let* $v\in I_{t1}$ ($v\in I_{t2}$) *and* $(x; y)\in E_t(v)\backslash E_t^0(v)$ *such that there are* n *restriction sites in* X (*in* W). *Denote the location of the restriction sites in* X (*in* W) *by* $z(n)=(z_1, \ldots, z_n)$ *where* $z_1<z_2<\ldots<z_n$. *Then the total number of bonds in* $X\cup Y(W\cup V)$ *which cannot be restriction sites is* $|X|-n+M'(|W|-n+M')$ *where*:

$$M'=M'(z(n))=L+r-t+\sum_{i=1}^{n-1} min(L+r-t, z_{i+1}-z_i).$$

Proof. We prove the case of $v\in I_{t1}$. The total number of bonds in X which are not restriction sites is obviously $|X|-n$. Consider the bonds in Y. Let $A_i\subset Y$ be the set $\{z_i'', z_i'\}$, $1\leqq i\leqq n$. Note that $|A_i|=L+r-t$ for all i. Clearly the bonds in:

$$A=\bigcup_{i=1}^{n} A_i,$$

cannot be restriction sites, and the bonds in $Y\backslash A$ can be unspecified. If $z_{i+1}-z_i\leqq L+r-t$, $A_i\cup A_{i+1}=\{z_i'', z_{i+1}'\}$ and thus $|A_i\cup A_{i+1}|=|A_i|+z_{i+1}-z_i=L+r-t+z_{i+1}-z_i$. If $z_{i+1}-z_i>L+r-t$, A_i and A_{i+1} are disjoint and hence $|A_i\cup A_{i+1}|=2(L+r-t)$. Using induction on i, we get $|A|=M'$ and the proof is complete. ■

COROLLARY 2. *If* $v\in I_{t1}\backslash I_{t2}$:

$$P(E_t(v))=p^{n(v)}(1-p)^M\left\{1+\sum_{n=1}^{|X|}\sum_{z(n)} p^n(1-p)^{M'-n}\right\}, \tag{17}$$

where the last sum is over all z(n) *satisfying* $1\leqq z_1<z_2<\ldots<z_n\leqq|X|$.

THEOREM 5. *If* $r+2\leqq L\leqq 2r+2$, $(L+r)/2+1\leqq t\leqq L$ *and* $v\in I_{t1}\cup I_{t2}$, *then*:

$$P(E_t(v))<p^{n(v)}(1-p)^M\{1+e^{-pr}p(L-r)[1+p(L-r)/2]\}.$$

Proof. Note that v cannot be an element of both I_{t1} and I_{t2}. We show the case of $v\in I_{t1}$; the same proof also applies when $v\in I_{t2}$. Since $|X|=x_1-x_a+t-(L+r)\leqq 2t-(L+r)-1\leqq L+r-t+1$ in this case, any $z(n)$ satisfies $z_{i+1}-z_i\leqq L+r-t$ for all $1\leqq i\leqq n-1$, hence $M'(z(n))=z_n-z_1+L+r-t$ for all $z(n)$. Therefore:

$$\sum_{z(1)}(1-p)^{M'}=|X|(1-p)^{L+r-t},$$

and for $n\geqq 2$:

$$\sum_{z(n)} (1-p)^{M'}$$

$$= (1-p)^{L+r-t} \sum_{d=n-1}^{|X|-1} (|X|-d) \binom{d-1}{n-2} (1-p)^d.$$

Consequently:

$$\sum_{n=2}^{|X|} p^n \sum_{z(n)} (1-p)^{M'-n}$$

$$= (1-p)^{L+r-t} \sum_{n=2}^{|X|} p^n \sum_{d=n-1}^{|X|-1} (|X|-d) \binom{d-1}{n-2} (1-p)^{d-n}$$

$$= (1-p)^{L+r-t} \sum_{d=1}^{|X|} (|X|-d)(1-p)^d \sum_{n=2}^{d+1} \binom{d-1}{n-2} \left(\frac{p}{1-p}\right)^n$$

$$= (1-p)^{L+r-t-1} p^2 \sum_{d=1}^{|X|-1} (|X|-d)$$

$$= (1-p)^{L+r-t-1} p^2 |X|(|X|-1)/2$$

$$\approx (1-p)^{L+r-t-1} p^2 |X|^2/2,$$

which together with the term for $n=1$ gives:

$$|X| p (1-p)^{L+r-t-1} (1+p|X|/2).$$

Since $|X| \leqq 2t-(L+r)-1$ and $t \leqq L$, the last expression is bounded above by:

$$p(L-r-1)(1-p)^{r-1}(1+p(L-r-1)/2)$$

$$\approx e^{-pr} p(L-r)[1+p(L-r)/2]. \quad \blacksquare$$

It is not easy to estimate $M'(z(n))$ for any $z(n)$ when $L \geqq 2r+3$ and we only calculate a very loose upper bound.

THEOREM 6. *Suppose* $L \geqq 2r+3$.
(*a*) *If* $t \geqq (L+r)/2+1$ *and* $v \in I_{t1} \backslash I_{t2}$ *or* $v \in I_{t2} \backslash I_{t1}$,

$$P(E_t(v)) < p^{n(v)}(1-p)^M\{1+e^{(L-2r)p}\};$$

(*b*) *if* $t > 2(L+r+1)/3$ *and* $v \in I_{t1} \cap I_{t2}$,

$$P(E_t(v)) < p^{n(v)}(1-p)^M\{1+e^{(L-3r)p}\}^2.$$

Proof. (a) Since $M'(z(n)) \geqq L+r-t+n-1$ for all $z(n)$:

$$\begin{aligned} P(E_t(v)) &< p^{n(v)}(1-p)^M\{1+\sum_{n=1}^{|X|}\binom{|X|}{n}p^n(1-p)^{L+r-t-1}\} \\ &= p^{n(v)}(1-p)^M\{1+(1-p)^{L+r-t-1}[(1+p)^{|X|}-1]\} \\ &< p^{n(v)}(1-p)^M\{1+(1-p)^{r-1}[(1+p)^{L-r-1}-1]\} \\ &\approx p^{n(v)}(1-p)^M\{e^{(L-2r)p}+1\}. \end{aligned}$$

(b) If $v \in I_{t1} \cap I_{t2}$, either X, or W, or both can contain restriction sites, therefore:

$$P(E_t(v)) < p^{n(v)}(1-p)^M\left\{1+\sum_{n=1}^{|X|}\binom{|X|}{n}p^n(1-p)^{L+r-t-1}\right\}\times \left\{1+\sum_{n=1}^{|W|}\binom{|X|}{n}p^n(1-p)^{L+r-t-1}\right\}.$$

The proof is complete by noting that both $|X|$ and $|W|$ cannot exceed $3t-2(L+r)-2$. ∎

COROLLARY 3. (a) *If* $2r+3 \leqq L \leqq 3r+p^{-1}ln(e^{rp}-2)$:

$$P(E_t(v)) < p^{n(v)}(1-p)^M\{1+e^{(L-2r)p}\}$$

for all $v \in I_{t1} \cup I_{t2}$ *and for all* $t \geqq (L+r)/2+1$;
(b) *if* $L > 3r+p^{-1}\,ln(e^{rp}-2)$:

$$P(E_t(v)) < p^{n(v)}(1-p)^M\{1+e^{(L-3r)p}\}^2,$$

for all $v \in I_{t1} \cup I_{t2}$ *and for all* $t \geqq (L+r)/2+1$. (*When* $r \gg p$, $p^{-1}\,ln(e^{rp-2})$ *can be replaced by* r.)

In summary, for all $1 \leqq t \leqq L$ and $v \in I_t$:

$$P(E_t(v)) < \begin{cases} p^{n(v)}(1-p)^M\{1+e^{-pr}[1+p(L-r)/2]\}, & r < L \leqq 2r, \\ p^{n(v)}(1-p)^M\{1+e^{(L-2r)p}\}, & 2r < L \leqq 4r, \\ p^{n(v)}(1-p)^M\{1+e^{(L-3r)p}\}^2, & 4r < L. \end{cases}$$

Note that since r is very large in practice, terms like $r+1$ and $2r+3$ are replaced by r and $2r$ to give simpler relation between L and r. Direct substitution of the above upper bound on $P(E_t(v))$ into (14) proves Theorem 2(b).

Remark. Our estimate of the upper bound on $P(E_t(v))$ is very generous, especially when $L > 2r$, and we expect the actual value of F^- to be closer to the lower bound than to the upper bound. This is corroborated in our simulation in Section 8.

7. Clonable Configuration and Clonability. In this section we show that $P(\text{A2})$ is practically zero if K is large enough. In other words, all nucleotides with clonable configurations can be considered as clonable. The argument is suggested by Louis Gordon.

Since clonable range is between $L+1$ and $L+r$, the expected number of restriction sites between the two ends of a clonable fragment is between Lp and $(L+r)p$. Let μ be the fraction of restriction sites that are cut in a partial digest. The lower bound of the probability that any clonable fragment is obtained from one DNA molecule (one genome) is approximately:

$$\mu^2(1-\mu)^{(L+r)p}.$$

Therefore, any clonable fragment will be obtained on the average once in not more than

$$\frac{1}{\mu^2(1-\mu)^{(L+r)p}},$$

DNA molecules.

Suppose every restriction site is the left end of some clonable fragment(s). There are approximately Np restriction sites and each is approximately the left end of rp distinct clonable fragments, so there are altogether Nrp^2 clonable fragments. A very generous overestimate of the number of molecules required to yield all these fragments is:

$$\frac{Nrp^2}{\mu^2(1-\mu)^{(L+r)p}}. \tag{18}$$

As an illustration, suppose *E. coli* is being digested with Eco R1 with $\mu=0.5$, and the cloning vector is pJC74 where $L=19\times10^3$ and $r=17\times10^3$. The number computed from equation (18) is approximately 1.8×10^6, which is well within the limit of a normal sample size of about 2×10^9 molecules.

8. Numerical Studies. We first see how the fraction unclonable by partial digest compares with that by complete digest. For the same example at the end of Section 3, the fraction unclonable in a partial digest is about 14% and 0.7% respectively when λgt WES and pJC74 are used as cloning vectors. The improvement for pJC74 is remarkable. Since the maximum clonable length is the major constraint in a partial digest, most cosmid vectors, which can accommodate longer DNA fragments, are much better cloning vectors than bacteriophages in this respect.

When r is large, the estimates given in (11) and (12) are very close to each other for $L\leqq 2r$, and are also approximately the same as $P(S_-^1)\approx e^{-p(L+r)}(1+p(L+r))$ (8). This can be seen in Fig. 3 where the fraction of nucleotides not

clonable corresponding to a restriction site frequency $p = 1/5\ \text{kbp}^{-1}$ and $r = 15$ kbp is plotted for L varying between 0 and $2r$. This agrees with our earlier remark that the maximum clonable length is the major constraint.

To test the validity of (11) and (12), a random number generator is used to assign restriction sites ($p = 1/5\ \text{kbp}^{-1}$) on a molecule of the size of the genome of *E. coli*. The fraction of unclonable nucleotides in that molecule is then obtained for $r = 5$ kbp and L varies between 0 and $4r$. (While typically $L < 2r$ in most λ-vectors, in practice shorter fragments are discarded and the "usual" insert size in a λ library is in the range of 15–20 kbp, i.e. $r = 5$ kbp, $L \approx 3$r.) The process is repeated 10 000 times and the mean as well as the 99% confidence interval of F^- are calculated and compared with the lower bound given by (11). The comparison, shown in Fig. 4, indicates very good agreement between the two.

We conclude that a nucleotide is unclonable mostly because its flanking restriction sites are too far apart and for practical purposes $e^{-p(L+r)}(1 + p(L + r)) + \frac{1}{6}p^2L^2e^{-p(L+2r)}(pL+3)$ is a reasonably good estimate of the fraction of unclonable nucleotides.

APPENDIX

For *E. coli*, $N \approx 4.7 \times 10^6$. The restriction sites for the enzyme Eco R1 occur at probability $p \approx 1/5000$, and for Hae it is about 1/3500 (Hamer and Thomas, 1975). The weight of a typical sample for digestion is about 10 μg, which means $K \approx 2 \times 10^9$ for *E. coli*. The size restriction of different cloning vectors is given in Table 1.

Table 1. Cloning capacity of some vectors recommended for cloning DNA fragments digested with Eco R1. (Adapted from Dahl *et al.*, 1981.)

Phage	Phage λ cloning vectors L (kbp)	r (kbp)
λgt WES	2.4	15
Charon 3	0	8.6
Charon 4	7.3	12
Charon 21A	0	8.2
NM 641	0	12.5
Cosmid	**Cosmid vectors L (kbp)**	**r (kbp)**
pJC74	19	17
pJC75-58	23	17
pHC	29	15
pJB-8	30	17
MUA	31	17

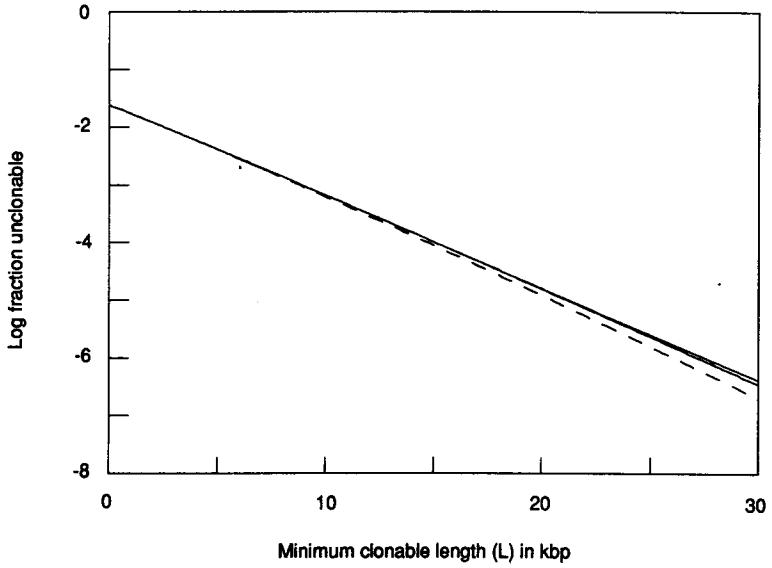

Figure 3. Comparison of the different estimates of F^- for $p=1/5$ kbp^{-1}, $r=15$ kbp and L varying between 0 and $2r$. The broken line corresponds to $P(S_-^1)$ and the upper and lower solid lines correspond respectively to upper and lower bounds for F^-. Note that the two solid lines are not that much different from each other.

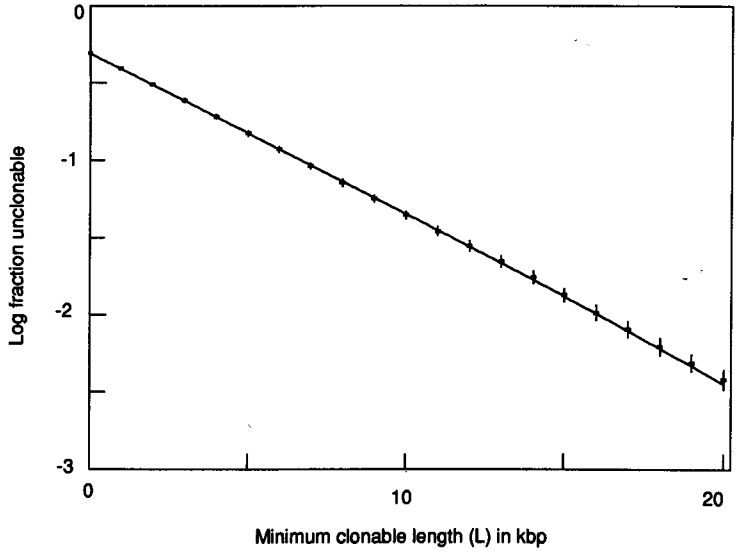

Figure 4. The mean and 99% confidence interval of F^- calculated from 10 000 randomly generated DNA molecules for $p=1/5$ kbp^{-1}, $r=5$ kbp and different values of L is compared with the lower bound given by (11). When L is not too large the confidence interval is virtually of no width.

LITERATURE

Coulson, A., J. Sulston, S. Brenner and J. Karn. 1986. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. natl Acad. Sci. U.S.A.* **83**, 7821–7825.

Dahl, H. H., R. A. Favell and F. G. Grosveld. 1981. The use of genomic libraries for the isolation and study of eukaryotic genes. In *Genetic Engineering*, Vol. 2. R. Williamson (Ed.). London: Academic Press.

Daniels, D. L. and F. R. Blattner. 1987. Mapping using gene encyclopedias. *Nature* **325**, 831–832.

Hamer, D. H. and C. A. Thomas, 1975. The cleavage of *Drosophila melanogaster* DNA by restriction endonucleases. *Chromosoma* **49**, 243–267.

Kohara, Y., K. Akiyama and K. Isono. 1987. The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**, 495–508.

Kuhn, W. 1930. Uber die Kinetik des Abbaues hochmolekularer Ketten. *Ber. dtsch. chem. Ges.* **63**, 1503–1508.

Lewin, B. 1985. *Genes*, New York: Wiley.

Montroll, E. W. and R. Simha. 1940. Theory of depolymerization of long chain molecules. *J. chem. Phys.* **8**, 721–727.

Olson, M. V., J. E. Dutchik, M. Y. Graham, G. M. Brodeur, C. Helms, M. Frank, M. MacCollin, R. Sheinman and T. Frank. 1986. Random-clone strategy for genomic restriction mapping in yeast. *Proc. natl Acad. Sci. U.S.A.* **83**, 7826–7830.

Roberts, L. 1987. Who owns the human genome? *Science* **237**, 358–361.

Seed, B. 1982. Theoretical study of the fraction of a long-chain DNA that can be incorporated in a recombinant DNA partial-digest library. *Biopolymers* **21**, 1793–1810.

Seed, B., R. C. Parker and N. Davidson. 1982. Representation of DNA sequences in recombinant DNA libraries prepared by restriction enzyme partial digestion. *Gene* **19**, 201–209.

Received 6 February 1988