GENOMICS 26, 84–100 (1995)

Genomic Mapping by End-Characterized Random Clones: A Mathematical Analysis

ETHAN PORT, FENGZHU SUN, DANIELA MARTIN, AND MICHAEL S. WATERMAN[1]

Department of Mathematics, DRB-155, University of Southern California, Los Angeles, California 90089-1113

Received July 6, 1994; revised November 17, 1994

Physical maps can be constructed by "fingerprinting" a large number of random clones and inferring overlap between clones when the fingerprints are sufficiently similar. E. Lander and M. Waterman (*Genomics* 2: 231–239, 1988) gave a mathematical analysis of such mapping strategies. The analysis is useful for comparing various fingerprinting methods. Recently it has been proposed that ends of clones rather than the entire clone be fingerprinted or characterized. Such fingerprints, which include sequenced clone ends, require a mathematical analysis deeper than that of Lander–Waterman. This paper studies clone islands, which can include uncharacterized regions, and also the islands that are formed entirely from the ends of clones.

1. INTRODUCTION

An increasing number of approaches to the physical mapping of genomes have been developed. A physical map consisting of overlapping clones that span the genome of an organism is the goal of these approaches. The map is then the basis for further genetic analyses such as gene location or sequencing of specific regions. Early physical maps were constructed for *Saccharomyces cerevisiae* (Olson *et al.*, 1986), *Caenorhabditis elegans* (Coulson *et al.*, 1986), and *Escherichia coli* (Kohara *et al.*, 1987), and these efforts have been extended to many other organisms.

While many variations have been developed, the basic principle of many physical mapping projects is first to fingerprint the clones and then to infer the overlap of clones when there is sufficient similarity of fingerprints. The three projects cited above all used clone restriction fragment information as a basis for the fingerprint. In other approaches STSs or anchors are used to overlap the clones containing a given anchor. All of these projects have a combinatorial aspect: a large number of clones are chosen at random from a library and fingerprint data are used to infer overlap. If there are 10,000 clones, there are approximately $\binom{10,000}{2} \approx 5 \times 10^7$ pairs of clones so that the true genomic overlaps must be located in a large number of potential overlaps.

Lander and Waterman (1988) gave the first mathematical analysis of genome mapping by fingerprinting random clones. For fixed-length clones they model the fingerprinting schemes by the parameter θ, the fraction of clone overlap needed to detect true genomic overlap. They give formulas for expected number of islands, contigs (islands of two or more clones), and expected island length. These formulas have proved to be useful in analyzing potential mapping strategies. Later, several papers described the mathematical properties of genome mapping by anchoring random clones (Arratia *et al.*, 1991; Barillot *et al.*, 1991; Ewens *et al.*, 1991; Torney, 1991; Marr *et al.*, 1992). These papers required a more mathematically sophisticated analysis. These problems fall under the general heading of coverage processes, and we recommend the excellent book by P. Hall (1988).

More recently another mapping strategy has been proposed in which the ends of clones are characterized but the central region of the clone is not. See Edwards and Caskey (1991) and Richards *et al.* (1994). The easiest model to picture is when the clone ends are sequenced, say 500 bp at each end. Of course, it is possible to apply any fingerprinting scheme to these clone ends. Clone overlap is then inferred from end overlaps. It is possible to have two clones overlap where the characterized end of one clone lies in the uncharacterized center of another, and therefore the overlap cannot be detected. Variations of this scheme using end sequencing are discussed in Chen *et al.* (1993) and Smith *et al.* (1994). Their results combine a physical map with the partial sequencing of the results from sequenced ends.

In this paper we describe the mathematical properties of genome mapping with end-characterized clones. While the overlap model of Lander and Waterman using the parameter θ is used, the straightforward analysis of Lander and Waterman cannot be directly carried over due to the more complex statistical dependence between clone ends. We are only able to handle fixed-length clones. In the course of our work we found two additional formulas for the Lander–Waterman setting, which we include in Section 2. In addition, we include a modification of these results for the case where there

[1] To whom correspondence should be addressed.

0888-7543/95 $6.00

is differential cloning efficiency in the genome. Then in Section 3.1 we study the properties of what we call *gapped islands,* where the islands consist of the overlapped clones. Note that these islands can contain uncharacterized regions, which we refer to as "gaps." We are able to obtain results only for islands that satisfy a special condition. Then in Section 3.2 we study *block islands,* which are islands of overlapped characterized ends. In Section 4 we present graphs of some relevant quantities for certain biological examples.

2. MAPPING BY CLONE OVERLAP

In Lander and Waterman (1988) physical mapping by fingerprinting random clones is addressed. Lander and Waterman used a discrete model in which each basepair (bp) is effectively an integer. In Arratia *et al.* (1991) a related analysis is conducted for physical mapping by random anchors, which involves aligning clones in a library where they share a short DNA sequence or marker unique in the genome called a *sequenced tagged site* (STS). We present here the Lander–Waterman model for physical mapping by random clones in a continuous setting. We essentially follow the setup of Arratia *et al.* (1991), which models clone locations by a homogeneous Poisson process. This allows us to present the results about progress in a physical mapping project using the Lander–Waterman model of clone inserts and overlap and the gapped clone model of Section 3 within the same framework. Additionally, in this section we give two new results, Theorem 1(v′) and (vi), for the Lander–Waterman model. These results for the expected length and genomic coverage by *islands* of at least two members (contigs) are a useful addition to the earlier results.

First we define the Lander–Waterman model and give some notation. For a given genome of length G, we assume a uniformly representative genome library with clone inserts of equal length L. It is convenient to rescale length by L, so that clone lengths are taken to be $L/L = 1$, and the genome corresponds to an interval of length $g = G/L$. Note that in this scaled metric each basepair corresponds to an interval of length $1/L$. We will use a continuous model.

Here we will use the following symbols:

G = genome length;
L = clone insert length;
N = number of clones with right ends in $(0, G)$;
c = LN/G, expected number of clones covering a random point;
T = length needed to detect overlap;
θ = T/L;
g = G/L;
σ = $1 - \theta$.

Assume that clones are placed on the real line $\mathbb{R}$ by a homogeneous Poisson process with rate $c = LN/G = N/g$. The genome corresponds to the interval $(0, g)$. The process of the location of right ends of clones can thus be modeled as a Poisson process $\{A_i : i \in \mathbb{Z}\}$ labeled by

$$\cdots A_{-2} < A_{-1} < A_0 \leq 0 < A_1 < \cdots < A_N < g \leq A_{N+1} < \cdots. \quad [1]$$

Note that N is now the random number of clones whose right ends belong to $(0, g)$ and that N has a Poisson distribution with mean cg. In formulas below, for convenience N appears as a constant. Those formulas can be translated into probability statements if desired.

There is a boundary effect that occurs since some clones with left ends before 0 might have right ends inside $(0, g)$, and some clones beginning in $(0, g)$ might have right ends greater than g. As we will see, these boundary effects become negligible as we take g going to infinity.

Throughout the paper almost sure convergence will be shown by the usual equal sign. By the ergodic theorem, e.g., we write

$$\lim_{g\to\infty} N/g = c,$$

while the left side is a random variable equal to c with probability 1.

Recall that for a Poisson process with rate c the interarrival times $A_i - A_{i-1}$ are independent, identically distributed exponential variates with mean $1/c$ and density ce^{-cx}, $x > 0$. A useful property of exponentials is the lack of memory property: If X is an exponential variate, then

$$\mathbb{P}(X > t + s \,|\, X > t) = \mathbb{P}(X > s).$$

Hence, the distribution of the set $\{A_i : i \in \mathbb{Z},\ i \neq 0\}$, conditioned upon the event $\{A_0 = 0\}$, is the same as the distribution of the set $\{A_i : i \in \mathbb{Z}\}$ before conditioning. We use this fact repeatedly in what follows, referring to a "given clone" when we are conditioning on having a clone at a given location. Finally we note that the choice of right clone ends is arbitrary, so that, for example, the process of left clone ends (or the process of centers of clones) is also a Poisson process with rate c.

Next we give the Lander–Waterman results for completeness and add the new formulas (v′) and (vi). While (vi) holds only for $\theta < \frac{1}{2}$, a general formula (not shown here) has been derived by David Torney at Los Alamos National Laboratory. The word "apparent" is used to emphasize the distinction between actual genome islands and those detected by the scientists.

THEOREM 1 (Lander–Waterman). *With the above notation and assumption*

(*i*) *The expected number of apparent islands is* $Ne^{-c\sigma}$.

(*ii*) *The expected number of apparent islands consisting of* j *clones,* $j = 1, 2, \ldots,$ *is*

$$Ne^{-2c\sigma}(1 - e^{-c\sigma})^{j-1}.$$

(*iii*) *The expected number of contigs is*

$$Ne^{-c\sigma} - Ne^{-2c\sigma}.$$

(*iv*) *The number of clones in an apparent island is geometrically distributed with mean* $e^{c\sigma}$.

(*v*) *The expected length of an apparent island is* $L\lambda$, *where*

$$\lambda = (e^{c\sigma} - 1)/c + 1 - \sigma.$$

(*v'*) *The expected length of a contig (non-singleton apparent island) is* $L\lambda'$, *where*

$$\lambda' = (1 - e^{-c\sigma})^{-1}\left(\frac{e^{c\sigma} - 1}{c} + 1 - \sigma - e^{-c\sigma}\right).$$

(*vi*) *If* $\theta < \frac{1}{2}$, *the expected fraction of the genome covered by contigs is*

$$1 + e^{-2c\sigma}[(c\theta - 1)(2 - e^{-c\sigma}) - c] - e^{-c}(1 - e^{-c\sigma})^2.$$

(*vii*) *The probability that an ocean of length at least* xL *occurs at the end of an apparent island is* $e^{-c(x+1-\sigma)}$. *In particular, taking* $x = 0$, *the probability that an apparent ocean is real (as opposed to an undetected overlap occurring) is* $e^{-c(1-\sigma)}$.

(*viii*) *The corresponding results for the actual islands that would result if all overlaps could be detected are obtained by setting* $\theta = 0$. *For example, the expected number of actual islands is* Ne^{-c}.

Proof. Part (v) is proved in Lander and Waterman (1988) using Wald's lemma (Hoel *et al.*, 1971). We present an alternative ergodic argument for the case $\theta < \frac{1}{2}$, because this gives us a simpler example of the ergodic argument that we apply in Section 3 in a case where Wald's lemma fails.

Just as in Arratia *et al.* (1991), we form the process of the right ends of apparent islands. Because clone length is constant, one apparent island cannot be completely contained in another apparent island. Therefore, we can order apparent islands, using either the right or the left end of islands. Hence we label apparent islands by their right ends $\{E_j : j \in \mathbb{Z}\}$ with

$$\cdots E_{-2} < E_{-1} < E_0 \leq 0 < E_1 < \cdots < E_K < g \leq E_{K+1} < \cdots,$$

so that K is the random number of apparent islands that have their right ends in the genome $(0, g)$. Clearly the set of right clone ends $\{A_i : i \in \mathbb{Z}\} \supseteq \{E_j : j \in \mathbb{Z}\}$. Although there is a positive probability that an apparent island with a right end in $(0, g)$ begins before 0 and that an apparent island with a right end greater than g begins inside $(0, g)$, these boundary effects will become negligible as g becomes large.

Let S_j be the length of the jth apparent island, and let X_t be the number of apparent islands containing $t \in (0, g)$. If $\theta < \frac{1}{2}$, then $X_t \leq 2$. If r_i is the probability that a point is covered by exactly i apparent islands, it follows that $r_0 + r_1 + r_2 = 1$ and

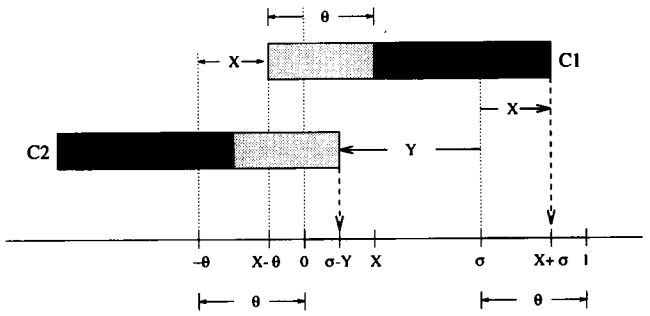

FIG. 1. Two apparent islands covering a fixed point $t = 0$.

$$\lim_{g\to\infty} \frac{1}{g} \sum_{j=1}^{K} S_j = \lim_{g\to\infty} \frac{1}{g} \int_0^g X_t dt = \mathbb{E}(X_1) = r_1 + 2r_2 = 1 - r_0 + r_2. \quad [2]$$

(If $\theta > \frac{1}{2}$ then we can have $X_t \geq 3$. Note that Lander and Waterman (1988) show that Theorem 1(v) holds for all $\theta \in [0, 1]$.)

We also utilize the fact that by stationarity, the labeling of the points in the genome $(0, g)$ was arbitrary, and we can relabel our coordinate system when it is convenient.

A point is not covered by a clone with probability $r_0 = e^{-c}$. To calculate r_2, fix a point $t \in (0, g)$. For t to be covered by two apparent islands there must be at least two clones covering t. Moreover, every clone covering t must have an end within θ of t, since any clone with both ends of distance greater than θ from t would overlap any other clone covering t. Referring to Fig. 1 with $t = 0$, we see that there must be at least one clone C_1 with a left end in $(t - \theta, t)$ and another clone C_2 with a right end in $(t, t + \theta)$ such that no other clones overlap both of these two clones by more than θ. The overlap between C_1 and C_2 must be less than θ. Now let X be the distance from $t - \theta$ to the first such clone C_1 with a left end in $(t - \theta, t)$. This corresponds to the first right clone end occurring in $(1 + t - \theta, 1 + t) = (\sigma + t, 1 + t)$. Let Y be the distance from $\sigma + t$ to the first right end occurring before $\sigma + t$, and denote this clone by C_2. Hence we have that $X \in (0, \theta)$, and given $X = x$, $Y \in (1 - \theta - X, 1 - \theta) = (\sigma - x, \sigma)$. Note that these conditions are independent of t, as is implied by stationarity of the $\{A_i\}$ process, so we may relabel t by 0. Let $D = \{(x, y) : x \in (0, \theta), y \in (\sigma - x, \sigma)\}$. This set characterizes the event that t is covered by two apparent islands. The random variables X and Y have independent exponential distributions with mean $1/c$.

$$\begin{aligned} r_2 &= \int\int_D f(x, y)dxdy \\ &= \int_0^\theta \int_{\sigma-x}^{\sigma} c^2 e^{-c(x+y)} dydx \\ &= [c(1 - \sigma) - 1]e^{-c\sigma} + e^{-c}. \end{aligned}$$

By Eq. [2],

$$\lim_{g\to\infty} \frac{1}{g} \sum_{j=1}^{K} S_j = 1 + (c\theta - 1)e^{-c\sigma}. \quad [3]$$

Also, by the ergodic theorem,

$$\lim_{g\to\infty} \frac{K}{g} = ce^{-c\sigma}, \quad [4]$$

since clones end with rate c, and a clone is the end of an island with probability $e^{-c\sigma}$. Thus Eqs. [3] and [4] imply that the expected length of an apparent island is

$$\lim_{g\to\infty} \frac{1}{K} \sum_{j=1}^{K} S_j = \frac{1 + (c\theta - 1)e^{-c\sigma}}{ce^{-c\sigma}} = 1 - \sigma + \frac{e^{c\sigma} - 1}{c}.$$

To prove (v′), we label the right ends of contigs by the process

$$\cdots E'_{-2} < E'_{-1} < E'_0 \leq 0 < E'_1 < \cdots < E'_{K'} < g \leq E'_{K'+1} < \cdots,$$

so that K' is the random number of contigs that have their right ends in the genome $(0, g)$.

Let S'_j be the length of the jth contig. Define

$$\overline{S'_g} = \frac{1}{K'} \sum_{j=1}^{K'} S'_j = \frac{g}{K'} \frac{1}{g} \sum_{j=1}^{K'} S'_j. \quad [5]$$

As in Eq. [4], the ergodic theorem yields

$$\lim_{g\to\infty} \frac{K'}{g} = ce^{-c\sigma}(1 - e^{-c\sigma}), \quad [6]$$

since c is the clone rate, $e^{-c\sigma}$ is the probability that a clone is the right end of an island, and $1 - e^{-c\sigma}$ is the probability that the clone is in a contig.

By similar reasoning,

$$\lim_{g\to\infty} \frac{1}{g} (\text{No. of singleton islands}) = ce^{-2c\sigma}.$$

Hence, using Eq. [3],

$$\lim_{g\to\infty} \frac{1}{g} \sum_{j=1}^{K'} S'_j = \lim_{g\to\infty} \left(\frac{1}{g} \sum_{j=1}^{K} S_j - \frac{1}{g} (\text{No. of singleton islands}) \right) = 1 + (c\theta - 1)e^{-c\sigma} - ce^{-2c\sigma}. \quad [7]$$

Thus, by Eqs. [5], [6], and [7], a contig has expected length

$$\lim_{g\to\infty} \overline{S'}_g = \frac{1 + (c\theta - 1)e^{-c\sigma} - ce^{-2c\sigma}}{ce^{-c\sigma}(1 - e^{-c\sigma})} = \frac{1}{1 - e^{-c\sigma}} \left(1 - \sigma + \frac{e^{c\sigma} - 1}{c} - e^{-c\sigma} \right).$$

To prove (vi), recall $\theta < \frac{1}{2}$ and we have each point of the genome in at most two islands. The fraction of the genome covered by contigs is

$$f_c = \lim_{g\to\infty} \left(\frac{1}{g} \sum_{j=1}^{K'} S'_j - \frac{1}{g} \{S^*\} \right), \quad [8]$$

where S^* = length of the genome covered by two contigs.

The first term of Eq. [8] is given by Eq. [7]. To calculate the second term in Eq. [8], fix $t \in (0, g)$ and define X, Y, C_1, and C_2 as in the proof of Eq. [3] and Fig. 1. Then t is covered by two contigs if and only if C_1 and C_2 are not singleton clones. Figure 1 shows then that there must be a clone C_3 with a right end in $(X + \sigma, X + 2\sigma)$ overlapping C_1, and a clone C_4 with a right end in $(-Y, -Y + \sigma)$ overlapping C_2. Since these two intervals are disjoint and both of length σ, this event has probability $(1 - e^{-c\sigma})^2$ of occurring, so the probability that t is covered by two contigs is

$$r'_2 = (1 - e^{-c\sigma})^2 r_2 = (1 - e^{-c\sigma})^2 \times ([c(1 - \sigma) - 1]e^{-c\sigma} + e^{-c}).$$

By the ergodic theorem,

$$\lim_{g\to\infty} \frac{1}{g} S^* = r'_2. \quad [9]$$

Equations [7], [8], and [9] then imply, as $g \to \infty$, that the proportion of the genome covered by contigs is

$$f_c = 1 + (c\theta - 1)e^{-c\sigma} - ce^{-2c\sigma} - (1 - e^{-c\sigma})^2 (c\theta e^{-c\sigma} - e^{-c\sigma} + e^{-c}) = 1 + e^{-2c\sigma}[(c\theta - 1)(2 - e^{-c\sigma}) - c] - e^{-c}(1 - e^{-c\sigma})^2. \quad \blacksquare$$

One assumption underlying Theorem 1 is that the rate of the Poisson process of right ends of clones is constant. This assumption was stated in the original paper as having a perfectly representative genome library. Bias in cloning efficiency will of course result in less rapid progress. We model this bias using an inhomogeneous Poisson process with rate $c(t)$ at point $t \in (0, g)$. The number of clone ends in (s_1, s_2) is Poisson with mean $\int_{s_1}^{s_2} c(t)dt$. When $c(t) \equiv c$, the mean is $c(s_2 -$

s_1) as in the earlier model. The total number of clones is now $\int_0^g c(t)dt$. The analog of Theorem 1 is given next. Implications of this theorem are discussed with a numerical example in Section 4.1.

THEOREM 1'. *With the above notation and assumption*

(*i'*) *The expected number of apparent islands is*

$$\int_0^g c(t)e^{-\int_t^{t+\sigma} c(s)ds}dt.$$

(*iii'*) *The expected number of contigs is*

$$\int_0^g c(t)e^{-\int_t^{t+\sigma} c(s)ds}(1 - e^{-\int_{t-\sigma}^{t} c(s)ds})dt.$$

(*vi'*) *The probability that a point t is covered by islands is*

$$1 - e^{-\int_t^{t+1} c(s)ds}$$

and, when $0 < \theta < \frac{1}{2}$, *t is covered by contigs with probability*

$$1 - e^{-\int_t^{t+1} c(s)ds} - \int_\sigma^1 c(t + 1 - x)e^{-\int_{t-x+\sigma}^{t+1} c(s)ds}dx$$
$$- \int_\theta^\sigma c(t + 1 - x)e^{-\int_{t-x+\theta}^{t+1-x+\sigma} c(s)ds}dx$$
$$- \int_0^\theta c(t + 1 - x)e^{-\int_{t+1-x}^{t+1-x+\sigma} c(s)ds}\left\{e^{-\int_t^{t+1-x} c(s)ds}\right.$$
$$\left. + \int_\sigma^{1-x} c(t + 1 - x - y)e^{-\int_{t+1-x-y}^{t+1-x-y-\sigma} c(s)ds}dy\right\}dx.$$

(*vii'*) *The probability that an ocean of length at least xL occurs at the end of an apparent island ending at t is* $exp(-\int_{t+\sigma}^{t+1+x} c(s)ds)$. *In particular, taking* $x = 0$, *the probability that an apparent ocean is real is* $exp(-\int_{t+\sigma}^{t+1} c(s)ds)$.

Lee (1992) gives results generalizing the work of Arratia *et al.* (1991) for mapping by anchoring random clones. Lee's work generalizes the Poisson process of anchor locations to general renewal processes. Karlin and Macken (1991) have formulas for expected number of contigs and expected coverage for the case of $\theta = 0$ with random clone length, and clones are located by an inhomogeneous Poisson process.

3. MAPPING BY GAPPED CLONES

As discussed in the introduction, there are several situations in which the ends of clones are characterized or fingerprinted. These characterized ends will be referred to as *blocks*. See Fig. 2. Then mapping proceeds by comparing the fingerprints of the blocks. When the blocks of two clones have enough fingerprint in common, the clones are overlapped as in the preceding section. The fact that there is an uncharacterized region or *gap* in the middle of clones makes the physical map more complex. There is one class of islands that result from block overlaps themselves that we will call *block islands*. These islands consist only of the characterized part of the clones. Obviously block islands have a dependence structure and are more complex than the islands in the previous section. Then there are the islands that result when the entire clones are taken together. These can have uncharacterized regions in them and we call these *gapped islands*. See Fig. 3.

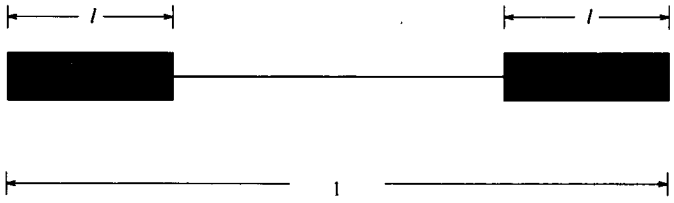

FIG. 2. A gapped clone with blocks of length l.

The notation from the previous section is maintained: g is the genome length, N is the number of clones, 1 is the clone length, and $c = N/g$. There is a new parameter l that is the (scaled) length of a block. See Fig. 2. For mathematical reasons, we require that $(2\sigma + 1)l < 1$, and therefore a block in one clone cannot simultaneously overlap both blocks of another clone. Two clones are declared to overlap if their blocks overlap by amount θl, where $0 \leq \theta \leq 1$. The case $\theta = 0$, closely corresponding to characterizing by sequencing the ends, is the easiest to establish results, but we include all θ in our theorems.

3.1. Apparent Greedy Islands and Contigs

This section presents results that are analogous to those in Theorem 1. The results that we give below are about "greedy islands" instead of gapped islands. To motivate this, recall that our method of proof is to compute the probability that a fixed clone is the right end of an island. Consider clones that prevent a fixed clone from being the right end of a gapped island. We divide these clones into two classes. Class 1 clones have left ends in the "No Clone" regions of Fig. 4. These include all clones that sufficiently intersect either end of our fixed clone and prevent an island end. Class 2 clones are all clones that extend the island to the right but establish connection with the given clone through other clones in the island. See Fig. 5 for illustration of class 1 and class 2 clones.

We count as greedy islands those that end by the "class 1" condition; that is, that satisfy the conditions of Fig. 4. The results for greedy islands (Theorem 2) give an overestimate for the number of gapped islands and an underestimate for the average length of gapped islands. Jared Roach pointed out to us that greedy islands

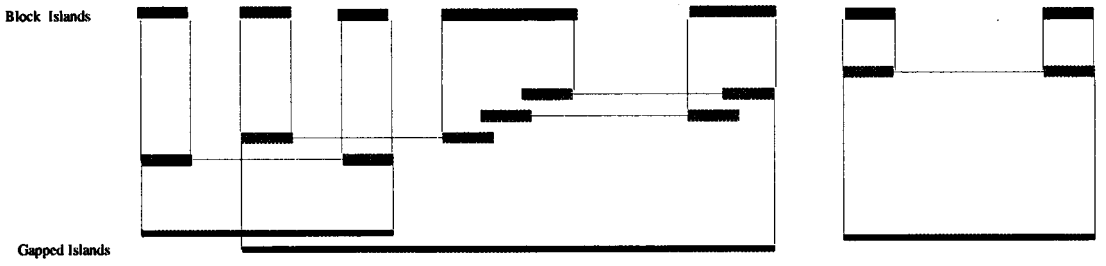

FIG. 3. Apparent block islands and greedy islands with $\theta = 0$.

and gapped islands are not equivalent. We include the results for greedy islands in the belief that it is the first step toward determining the corresponding theorem for gapped islands. We anticipate that the expected number of gapped islands is $Ne^{-\lambda c\sigma l}$, where $\lambda > 3$.

THEOREM 2 (apparent greedy islands). *With the same notation as above, assume that* $l < 1/(2\sigma + 1)$. *Then*

(i) $p = e^{-3c\sigma l}$ *is the probability that a right end of a clone is the right end of an apparent greedy island, and the expected number of apparent greedy islands is* Np.

(ii) The expected number of apparent greedy islands consisting of j *clones* ($j = 1, 2, \ldots$) *is*

$$Np^2(1-p)^{j-1}.$$

(iii) The expected number of clones in an apparent greedy island is $1/p$.

(iv) The expected length of an apparent greedy island is λL, *where*

$$\lambda = \frac{e^{3c\sigma l} - 1}{c} + \theta l + (1 - l - 2\sigma l)e^{2c\sigma l}.$$

(v) The proportion of the genome covered by apparent greedy islands is $1 - e^{-c}$.

Proof. Throughout this subsection, island refers to greedy island.

(i) Label the right end of a clone with coordinate 1. The right end of this clone is the right end of an island if and only if there are no clones with a left end in $(0, \sigma l)$ or in $(1 - l - \sigma l, 1 - l + \sigma l) = (1 - l - \sigma l, 1 - \theta l)$. See Fig. 4. This event occurs with probability $e^{-c\sigma l}e^{-2c\sigma l} = e^{-3c\sigma l} \equiv p$. Since the number of islands is equal to the number of times that we exit a clone without detecting overlap, the expected number of islands is Np.

(ii) The above reasoning shows that the number of clones J in an island follows the geometric distribution with mean $1/p$; that is, the probability that an island contains exactly j clones is

$$(1-p)^{j-1}p.$$

It follows that the expected number of apparent islands with j clones $= (Np)(p(1-p)^{j-1})$.

(iii) This follows immediately from (ii).

(iv) Consider an apparent island consisting of J clones, where J is geometrically distributed with mean $1/p$. We order the set of clones $\{C_j : j = 1, \ldots, J\}$ in the apparent island from right to left, so that C_1 is the rightmost clone, and C_J is the leftmost clone in the apparent island. Let A_j' be the right end of clone C_j, $j = 1, \ldots, J$. Notice that unlike the Lander–Waterman model of Section 2, there may be clones overlapping (A_j', A_1') that are *not* included in the apparent greedy island, so that $\{A_j', j = 1, \ldots, J\}$ may not be a contiguous

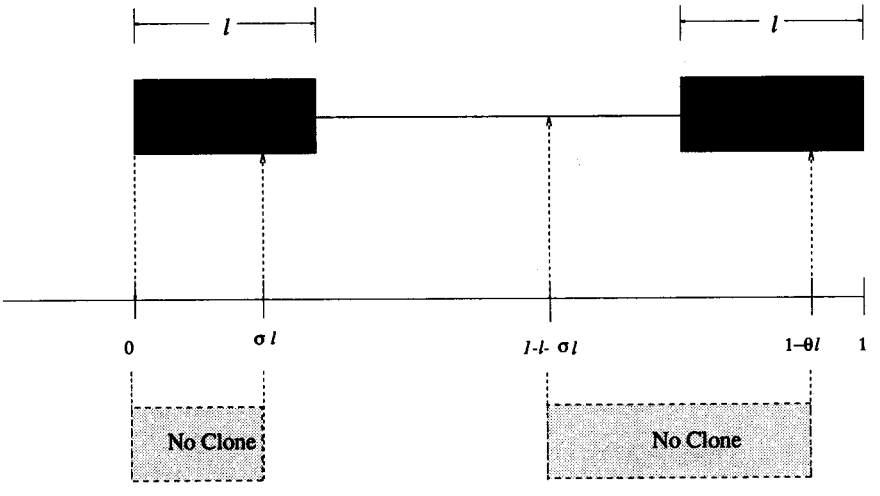

FIG. 4. The right end of an apparent greedy island, beginning at an arbitrary point 0.

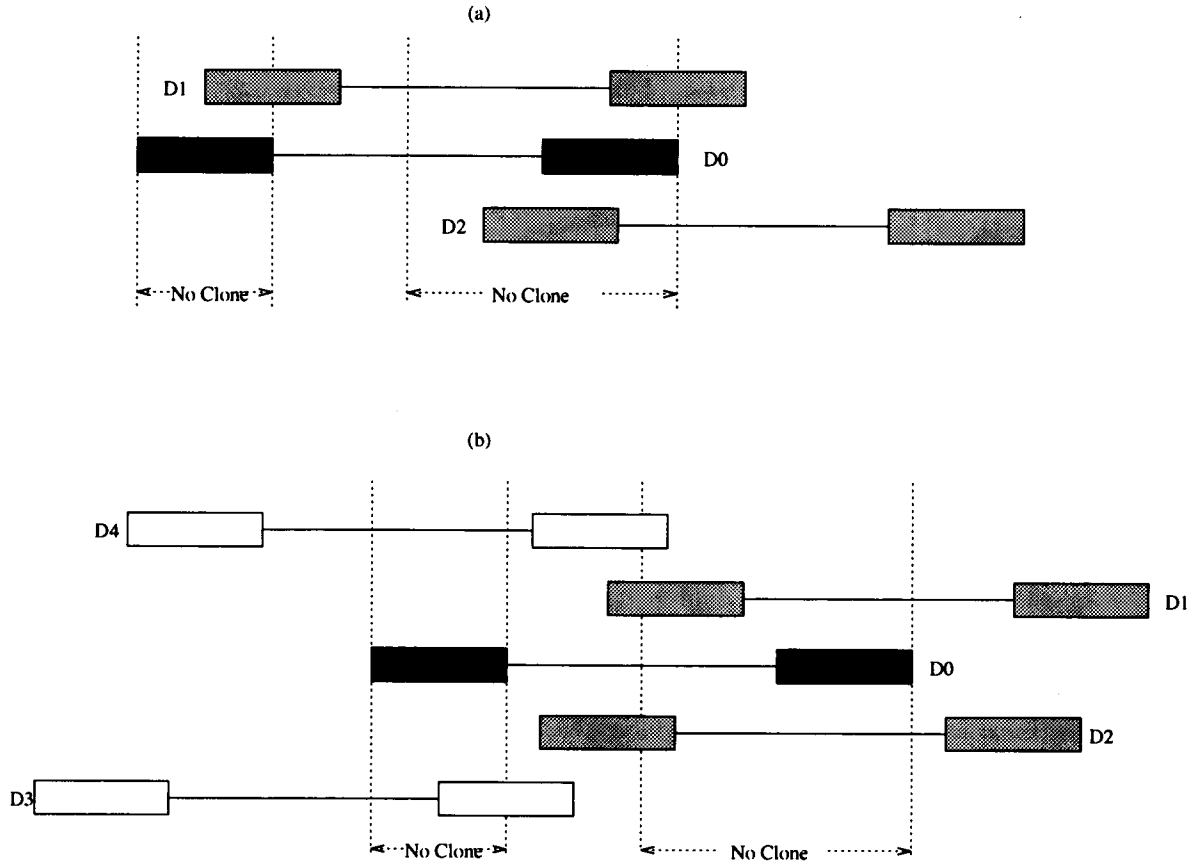

FIG. 5. Islands contain $D0$ ($\theta = 0$). (**a**) $D1$ and $D2$ are of class 1. (**b**) $D1$ and $D2$ are of class 2 since $D1$ is connected to $D0$ via $D3$ and $D4$, while $D2$ is connected to $D0$ via $D3$.

subset of $\{A_i : i \in \mathbb{Z}\}$. See for example Fig. 3. Define the coverage of clone C_j to be

$$X_j = A'_j - A'_{j+1}$$

for $j = 1, \ldots, J-1$ and $X_J = 1$. Let $G(x) = \mathbb{P}(X_1 > x)$. Using the lack of memory property, we next show that the X_i are identically distributed with

$$G(x) = \begin{cases} e^{-cx}, & 0 \leq x < \sigma l, \\ e^{-c\sigma l}, & \sigma l \leq x < 1 - (1+\sigma)l, \\ e^{-c(-1+l+2\sigma l)}e^{-cx}, & 1-(1+\sigma)l \leq x < 1-(1-\sigma)l, \\ e^{-3c\sigma l}, & 1-(1-\sigma)l \leq x < 1, \\ 0, & x \geq 1. \end{cases} \quad [10]$$

Equation [10] can be proved in the following way. We consider four cases:

(a) $0 \leq x < \sigma l$. Then $\mathbb{P}(X > x) = \mathbb{P}$ (no clones have a right end in $(0, x)$) $= e^{-cx}$.

(b) $\sigma l \leq x < 1 - (1+\sigma)l$. Then $\mathbb{P}(X > x) = \mathbb{P}$ (no clones have a right end in $(0, \sigma l)$) $= e^{-c\sigma l}$.

(c) $1 - (1+\sigma)l \leq x < 1 - (1-\sigma)l$. Then

$$\begin{aligned} \mathbb{P}(X > x) &= \mathbb{P} \text{ (no clones have right end in } (0, \sigma l)) \\ &\quad \text{or in } (1 - l - \sigma l, x)) \\ &= e^{-c\sigma l}e^{-c(x-(1-l-\sigma l))} = e^{-c(-1+l+2\sigma l)}e^{-cx}. \end{aligned}$$

(d) $1 - (1-\sigma)l \leq x < 1$. $\mathbb{P}(X > x) = p = e^{-3c\sigma l}$.

Hence we have shown Eq. [10]. Therefore

$$\begin{aligned} \mathbb{E}X &= \int G(x)dx \\ &= \frac{1 - e^{-3c\sigma l}}{c} + (1 - l - 2\sigma l)e^{-c\sigma l} + \theta l e^{-3c\sigma l}. \end{aligned}$$

$\{J \geq j\}$ is determined by $X_1, \ldots, X_{j-1}$, so J is a stopping time. Wald's identity (Feller, 1971) then implies the expected length of an apparent greedy island,

$$\begin{aligned} \mathbb{E}\left(\sum_{j=1}^{J} X_j\right) &= \mathbb{E}(J)\mathbb{E}(X_1) \\ &= \frac{e^{3c\sigma l} - 1}{c} + \theta l + (1 - l - 2\sigma l)e^{2c\sigma l}. \end{aligned} \quad [11]$$

(v) The coverage of the genome by greedy islands is the same as that for traditional recombinant libraries, included here for completeness. ■

Recall that a contig is an apparent greedy island with

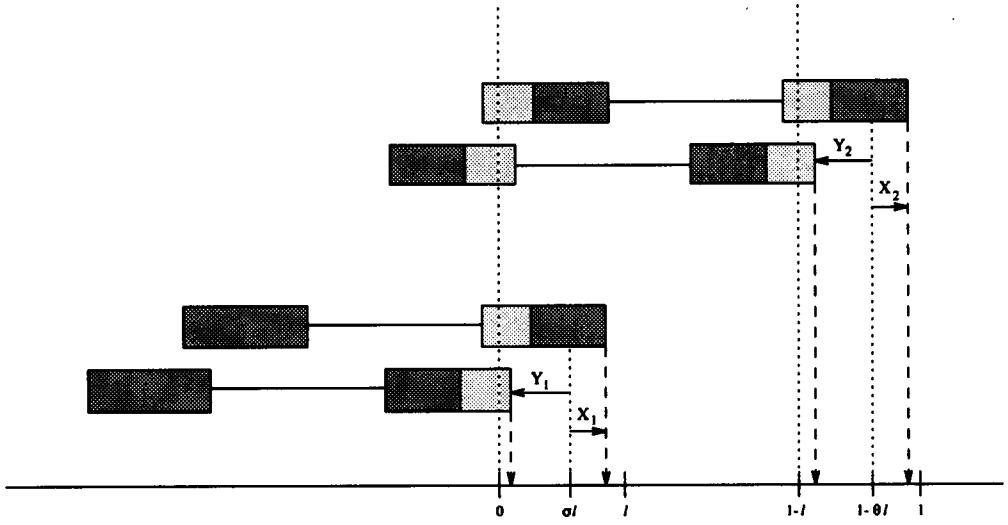

FIG. 6. Two block islands covering 0.

$j > 1$ clones. The next theorem gives the corresponding results for contigs. We were unable to derive a formula for coverage by contigs.

THEOREM 3 (apparent greedy contigs). *Under the same assumptions and notation as Theorem 2 and* $p = e^{-3c\sigma l}$.

(i) The expected number of contigs is $Np(1-p)$.

(ii) The expected number of clones in a contig is $1 + 1/p$.

(iii) The expected length of a contig is $\lambda' L$, *where*

$$\lambda' = \frac{1}{1 - e^{-3c\sigma l}} \times \left(\frac{e^{3c\sigma l} - 1}{c} + \theta l + (1 - l - 2\sigma l)e^{2c\sigma l} - e^{-3c\sigma l} \right).$$

Proof. (i) The expected number of contigs equals the expected number of islands minus the expected number of singletons. The expected number of singletons is Np^2.

(ii) From the proof of Theorem 2 (iii) we see that $\mathbb{E}J = 1/p$, and

$$\mathbb{P}(J = k \mid J \geq 2) = (1-p)^{k-2}p, \quad k = 2, 3, \ldots.$$

Thus, $\mathbb{E}[J \mid J \geq 2] = (1+p)/p$. This proves (ii).

(iii)

$$\begin{aligned}
\mathbb{E}\Big(\sum_{i=1}^{J} X_i \mid J \geq 2\Big) &= \frac{1}{\mathbb{E}(J \geq 2)}\, \mathbb{E}\Big(\sum_{i=1}^{J} X_i \mathbb{I}_{\{J \geq 2\}}\Big) \\
&= \frac{1}{1 - e^{-3c\sigma l}}\, \mathbb{E}\Big(\sum_{i=1}^{J} X_i - \sum_{i=1}^{J} X_i \mathbb{I}_{\{J=1\}}\Big) \\
&= \frac{1}{1 - e^{-3c\sigma l}}\, \mathbb{E}\Big(\sum_{i=1}^{J} X_i - 1 \cdot \mathbb{P}(J = 1)\Big) \\
&= \frac{1}{1 - e^{-3c\sigma l}}\, \mathbb{E}\Big(\sum_{i=1}^{J} X_i - 1 \cdot e^{-3c\sigma l}\Big).
\end{aligned}$$

The result follows from Theorem 2(iv). ■

3.2. Coverage by Block Islands

The $2N$ characterized blocks will themselves form islands and contigs based on their overlap. See Fig. 3. The block islands are dependent due to the coupling of the two blocks of the same clone. We emphasize the fact that the process of the $2N$ right block ends is *not* a Poisson process. Nevertheless, the results below are identical to those of Theorem 1 with $2N$ clones of length l (or lL, if clone length is L). The variance of these quantities is increased, however.

THEOREM 4 (apparent block islands). *With the same notation as above,*

(i) The expected number of apparent block islands is $2Ne^{-2c\sigma l}$.

(ii) The expected number of blocks in an apparent block island is $e^{2c\sigma l}$.

(iii) If $\theta < \frac{1}{2}$, *the expected length of an apparent block island is* $\lambda l L$ *where*

$$\lambda = \theta + \frac{e^{-2c\sigma l} - 1}{2cl}.$$

(iv) The proportion of the genome covered by apparent block islands is $1 - e^{-2cl}$.

Proof. (i) We label the right ends of apparent block islands by

$$\cdots E_{-2} < E_{-1} < E_0 \leq 0 < E_1 < \cdots < E_K < g \leq E_{K+1} < \cdots$$

so that K is the number of apparent islands that have their right ends in the genome $(0, g)$.

Let $p(t) = \mathbb{P}$ (a block ending at t is the right end of an apparent island | a block ends at t). Stationarity implies that $p(t)$ is independent of t. Let p_1 denote the common value of $p(t)$. We next calculate the value of p_1. Without loss of generality we set $t = 0$. It is the

right end of an apparent block island if and only if there are no clones ending in $(1 - l, 1 - \theta l)$ or $(0, \sigma l)$, and this occurs with probability $e^{-2c\sigma l}$. Then

$$\mathbb{E}(\text{number of apparent block islands in } (0, g)) = \int_0^g 2cp(t)dt = 2cgp_1.$$

The factor of 2 comes from the fact that there are two blocks in each clone. This implies Theorem 4(i).

(ii) Let M_j be the number of blocks in the jth apparent island whose right end is in the genome $(0, g)$. The average number of blocks per island is

$$\bar{M}_g = \frac{1}{K}\sum_{j=1}^{K} M_j.$$

Just as in Arratia *et al.* (1991), we can ignore boundary effects. Then $\Sigma_{j=1}^{K} M_j$ is equal to the number of blocks in $(0, g)$. Thus,

$$\lim_{g\to\infty} \frac{1}{g}\sum_{j=1}^{K} M_j = \lim_{g\to\infty} \frac{1}{g}(2N) = 2c,$$

and by the ergodic theorem,

$$\lim_{g\to\infty} \frac{K}{g} = 2ce^{-2c\sigma l}.$$

Hence

$$\lim_{g\to\infty} \bar{M}_g = \lim_{g\to\infty} \frac{g}{K}\frac{1}{g}\sum_{j=1}^{K} M_j = e^{2c\sigma l}.$$

This implies Theorem 4(ii).

(iii) Suppose that the jth apparent block island has length S_j. Let

$$\bar{S}_g = \frac{1}{K}\sum_{j=1}^{K} S_j.$$

Since for $\theta < \frac{1}{2}$ a point can fall in at most two apparent block islands, it follows that $r_0 + r_1 + r_2 = 1$, where r_i is the probability that a point is covered by precisely i apparent block islands. Then

$$\lim_{g\to\infty} \frac{1}{g}\sum_{j=1}^{K} S_j = r_1 + 2r_2 = 1 - r_0 + r_2.$$

Fix a point t in the genome that as usual we suppose has coordinate 0. 0 does not belong to any apparent block island if and only if there are no clones ending in $(0, l)$ or $(1 - l, 1)$, and this event occurs with probability $e^{-2cl} \equiv r_0$.

The calculation of r_2 is similar to the calculation made in the proof of Theorem 1(v), but is more involved. We begin by considering all the ways that 0 can be covered by two apparent block islands. Figure 6 demonstrates that 0 is covered by a block if and only if there is a right block end in $(0, l)$ or in $(1 - l, 1)$. Note that the latter case corresponds to a left block of a clone covering 0. Recall that the set of right clone ends $\{A_i : i \in \mathbb{Z}\}$ is a Poisson process with rate c. Let

$$\mathcal{A} = \{A_i : A_i \in (0, l)\},$$
$$\mathcal{B} = \{A_i : A_i \in (1 - l, 1)\}$$

be the subsets of clone ends corresponding to all blocks covering 0. Since $(2\sigma + 1)l < 1$, $\mathcal{A}$ *and* $\mathcal{B}$ *are disjoint. To compute the probability* r_2, we define the following four random variables analogous to those in the proof of Theorem 1(v). Let

$X_1 = \{$distance from σl to the first A_i in $\mathcal{A}$ after $\sigma l\}$,

$Y_1 = \{$distance from σl to the last A_i in $\mathcal{A}$ before $\sigma l\}$,

$X_2 = \{$distance from $1 - \theta l$ to the first A_i in $\mathcal{B}$ after $1 - \theta l\}$,

$Y_2 = \{$distance from $1 - \theta l$ to the last A_i in $\mathcal{B}$ before $1 - \theta l\}$.

To see how the block islands overlap in terms of these variables, start at σl and move toward 0. The first right block end encountered is at a distance $Y = \min\{Y_1, Y_2\}$ from σl. Moving in the other direction we encounter the first right block end at a distance of $X = \min\{X_1, X_2\}$ from σl. We next compute the joint distribution of X and Y. Fix x and y such that $0 \leq x, y < l$ and let $I_1 = (\sigma l - y, \sigma l + x)$, $I_2 = (1 - \theta l - y, 1 - \theta l + x)$. Then $(2\sigma + 1)l < 1$ implies $|I_1 \cup I_2| = 2(x + y)$. Hence

$$\begin{aligned}\mathbb{P}(X > x, Y > y) &= \mathbb{P}(X_1 > x, Y_1 > y, X_2 > x, Y_2 > y)\\ &= \mathbb{P}(\text{no right clone ends in } I_1 \cup I_2)\\ &= e^{-2c(x+y)},\end{aligned}$$

so for $0 \leq x, y \leq l$, X and Y have joint density

$$f(x, y) = 4c^{-2}e^{-2c(x+y)}. \qquad [12]$$

As in the proof of Theorem 1(v), 0 is covered by two apparent block islands if and only if $X \in (0, \sigma l)$, given $X = x$, $Y \in (\sigma l - x, \sigma l)$. Hence

$$\begin{aligned}r_2 &= \int_0^{\sigma l}\int_{\sigma l - x}^{\sigma l} f(x, y)dydx\\ &= (2c\theta l - 1)e^{-2c\sigma l} + e^{-2cl}. \qquad [13]\end{aligned}$$

Thus

$$\lim_{g\to\infty}\frac{1}{g}\sum_{j=1}^{K} S_j = 1 - r_0 + r_2$$
$$= 1 + 2c\theta l e^{-2c\sigma l} - e^{-2c\sigma l}, \qquad [14]$$

and

$$\lim_{g\to\infty}\frac{1}{K}\sum_{j=1}^{K} S_j = \lim_{g\to\infty}\frac{g}{K}\frac{1}{g}\sum_{j=1}^{K} S_j$$
$$= \frac{1 + 2c\theta l e^{-2c\sigma l} - e^{-2c\sigma l}}{2ce^{-2c\sigma l}}$$
$$= \theta l + \frac{e^{2c\sigma l} - 1}{2c}. \quad \blacksquare$$

THEOREM 5 (apparent block contigs). *(i) The expected number of apparent contigs is* $2Ne^{-2c\sigma l}(1 - e^{-2c\sigma l})$.

(ii) The expected number of blocks in a contig is $1 + e^{2c\sigma l}$.

(iii) If $\theta < \frac{1}{2}$, *the expected length of an apparent contig is* $\lambda' lL$ *where*

$$\lambda' = (1 - e^{-2c\sigma l})^{-1}\left(\theta + \frac{e^{2c\sigma l} - 1}{2cl} - e^{-2c\sigma l}\right).$$

(iv) If $\theta < \frac{1}{2}$, $l < (4 - 2\theta)^{-1}$, *then the proportion of the genome covered by contigs is*

$$1 + e^{-4c\sigma l}[(2c\theta l - 1)(2 - e^{-2c\sigma l}) - 2cl] - e^{-2cl}(1 - e^{-2c\sigma l})^2.$$

Proof. (i) As in the proof of Theorem 4(i), label the right ends of contigs of blocks by

$$\cdots E'_{-2} < E'_{-1} < E'_0 \leqslant 0 < E'_1 < \cdots < E'_{K'} < g \leqslant E'_{K'+1} < \cdots$$

so that K' is the number of contigs that have their right ends in the genome $(0, g)$. Recall that characterized block ends appear at rate $2c$ in $(0, g)$, so the expected number of contigs is $2cgp'_1$, where

$p'_1 = \mathbb{P}$ (a block ending at t is the right end of a contig | a block ends at t).

To calculate p'_1 fix a point $t \in (0, g)$ and label it with coordinate 0. The event for p'_1 occurs if and only if no clones end in $(0, \sigma l)$ or in $(1 - l, 1 - l + \sigma l)$, and there exists a clone ending in $(-\sigma l, 0)$ or $(1 - l - \sigma l, 1 - l)$. Since clone ends occur at rate c,

$$p'_1 = e^{-2c\sigma l}(1 - e^{-2c\sigma l}).$$

This implies Theorem 5(i).

(ii) Let M'_j be the number of blocks in the jth contig with right ends in the genome $(0, g)$, and let

$$\bar{M}'_g \equiv \frac{1}{K'}\sum_{j=1}^{K'} M'_j = \frac{g}{K'}\frac{1}{g}\sum_{j=1}^{K'} M'_j.$$

As before we have

$$\lim_{g\to\infty}\frac{K'}{g} = 2cp'_1 = 2ce^{-2c\sigma l}(1 - e^{-2c\sigma l}).$$

A block ending at $t = 0$ is a singleton if and only if there are no clones ending in either $(-\sigma l, +\sigma l)$ or $((1 - l) - \sigma l, (1 - l) + \sigma l)$. The probability of this event is $e^{-4c\sigma l}$. Then

$$\lim_{g\to\infty}\frac{1}{g}\left(\sum_{j=1}^{K'} M'_j\right) = \lim_{g\to\infty}\frac{1}{g}(2N - (\text{No. of singleton block islands}))$$
$$= 2c - 2ce^{-4c\sigma l} = 2c(1 - e^{-4c\sigma l}).$$

Thus,

$$\lim_{g\to\infty}\bar{M}_g \equiv \lim_{g\to\infty}\frac{g}{K'}\frac{1}{g}\sum_{j=1}^{K'} M'_j$$
$$= 1 + e^{2c\sigma l}.$$

(iii) As in the proof of Theorem 4(iii), let S'_j be the length of the jth contig. Let

$$\bar{S}'_g \equiv \frac{1}{K'}\sum_{j=1}^{K'} S'_j = \frac{g}{K'}\frac{1}{g}\sum_{j=1}^{K'} S'_j.$$

Then using Eq. [14]

$$\lim_{g\to\infty}\frac{1}{g}\sum_{j=1}^{K'} S'_j = \lim_{g\to\infty}\frac{1}{g}\left[\sum_{j=1}^{K} S_j - l(\text{No. of singletons})\right]$$
$$= 1 + 2c\theta l e^{-2c\sigma l} - e^{-2c\sigma l} - 2lce^{-4c\sigma l}. \qquad [15]$$

Thus

$$\lim_{g\to\infty}\bar{S}'_g = \frac{1 + 2c\theta l e^{-2c\sigma l} - e^{-2c\sigma l} - l2ce^{-4c\sigma l}}{2ce^{-2c\sigma l}(1 - e^{-2c\sigma l})}$$
$$= \frac{1}{1 - e^{-2c\sigma l}}\left(\theta l + \frac{e^{2c\sigma l} - 1}{2c} - \frac{l}{e^{2c\sigma l}}\right). \qquad [16]$$

(iv) We assume that $l < (4 - 2\theta)^{-1}$ and $\theta < \frac{1}{2}$. We define the approximate proportion of the genome covered by contigs to be

$$\bar{S}'_g = \frac{1}{g}\left(\sum_{j=1}^{K'} S'_j - R'\right), \qquad [17]$$

where R' is the length of the genome covered by two contigs. The first term is given by Eq. [15]. For the second term,

$$\lim_{g\to\infty} \frac{R'}{g} = r'_2,$$

where r'_2 is the probability that a point is covered by two contigs.

Let A be the event that 0 is covered by two block islands, and B be the event that 0 is covered by two contigs. Equation [13] in the proof of Theorem 4(iii) tells us that $\mathbb{P}(A) = r_2$. Referring to Fig. 6 and the proof of Theorem 4(iii), let X be the distance to the first block with a right end after σl, and call this block D_1. Let Y be the distance to the last block with a right end before σl, and call this block D_2. Then A is equivalent to the event that D_1 and D_2 both cover 0 and are in different apparent islands, and B is equivalent to the event that A occurs and neither D_1 nor D_2 are singletons.

Recall that a point can be covered by at most two contigs, since $\theta < \frac{1}{2}$. To calculate r'_2, fix a point $t \in (0, g)$ and label it with 0. Then as in the proof of Theorem 1(v'), B given A occurs if and only if D_2 is the rightmost block of a contig and D_1 is the left most block of a contig. These intervals are disjoint when $l \leq (4 - 2\theta)^{-1}$. Thus,

$$\mathbb{P}(B|A) = (1 - e^{-2c\sigma l})^2,$$

and since $B \subseteq A$,

$$r'_2 = \mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) = (1 - e^{-2c\sigma l})^2 r_2. \qquad [18]$$

So by Eqs. [13], [15], [17], and [18], we get

$$\lim_{g\to\infty} \bar{S}'_g = 1 + e^{-4c\sigma l}[(2c\theta l - 1)(2 - e^{-2c\sigma l}) - 2cl] - e^{-2cl}(1 - e^{-2c\sigma l})^2. \qquad [19]$$

■

4. DISCUSSION OF THE RESULTS

4.1. Lander–Waterman

In Section 2 we gave an addition to the Lander–Waterman results in Theorem 1(vi), the expected coverage by contigs. Of course, expected coverage by clones is $1 - e^{-c}$, the Carbon–Clark formula, which by itself is not a very revealing indication of progress. The quantity $1 - e^{-c}$ is an upper bound for coverage in Fig. 7, which in addition to $1 - e^{-c}$ shows contig coverage for $\theta = 0, 0.1, 0.25, 0.5$. These quantities are of interest in comparing genome coverage by clones ($1 - e^{-c}$) with that by contigs.

We now investigate the effects of an inhomogeneous clone rate $c(t)$ on a physical mapping project. To relate this discussion to what follows, we suppose that clones correspond to 1000 bp of sequenced DNA. We use the parameters $G = 200$ kb, $L = 1000$ bp, $N = 1000$, $\theta = 25/1000$. θ is derived from the assumption that 25-bp overlaps can be detected. We consider clone ends occurring with rate

$$\text{(ii)} \qquad c(t) = \begin{cases} 2.5, & 0 \leq t < G/2 \\ 7.5, & G/2 \leq t \leq G \end{cases}$$

or with rate

$$\text{(iii)} \qquad c(t) = \begin{cases} 0.5, & 0 \leq t < G/2 \\ 9.5, & G/2 \leq t \leq G. \end{cases}$$

In both of these cases the expected number of clones $N = (1/L)\int_0^G c(t)dt = 1000$ and average clone rate $\bar{c} = (1/G)\int_0^G c(t)dt = 5$. In Table 1 we compare results for physical mapping with clone end rates described by (ii) and (iii), and Lander–Waterman mapping with constant rate $c \equiv 5$ (i).

Table 1 contains several features of interest. We are able to tabulate quantities not given in Theorem 1' since our model (ii), for example, essentially divides into two Theorem 1 regions with $N_1 = 250$ and $N_2 = 750$. We have separated the calculations into two components, those resulting from [0, $G/2$] and [$G/2$, G], to reveal the difference between those two intervals. For example, while in model (ii) genome coverage by contigs is 94.9%, that for [0, $G/2$] is 89.9% and for [$G/2$, G] is 99.9%. In model (iii) where the ratio of coverage is 9.5/0.5 = 19, the genome coverage by contigs is 60.2% while only 20.5% in [0, $G/2$]. The second interval is saturated by clones, while obviously the first interval is not. We remark that the estimate of the number of islands is not accurate for extremely large N as $Ne^{-c\sigma} = Ne^{-NL\sigma/G} \to 0$ as $N \to \infty$. This is because islands are counted by right-hand ends. If an island covers G, that island is not counted. Therefore, in the limit, there is one more island than the expression $Ne^{-c\sigma}$ counts.

4.2. Gapped Clones

Despite the increased complexity of the mathematical models for gapped clones, the results in Section 3 about number of islands and contigs are closely related to the Lander–Waterman formula. Thus, for the number of islands, the relevant graphs differ just by scale changes. See Figs. 8 and 9. We give some intuition for these results as follows. In the Lander–Waterman clone overlap model, there are N clones each with end of island or exit probability of $e^{-c\sigma}$. For greedy islands

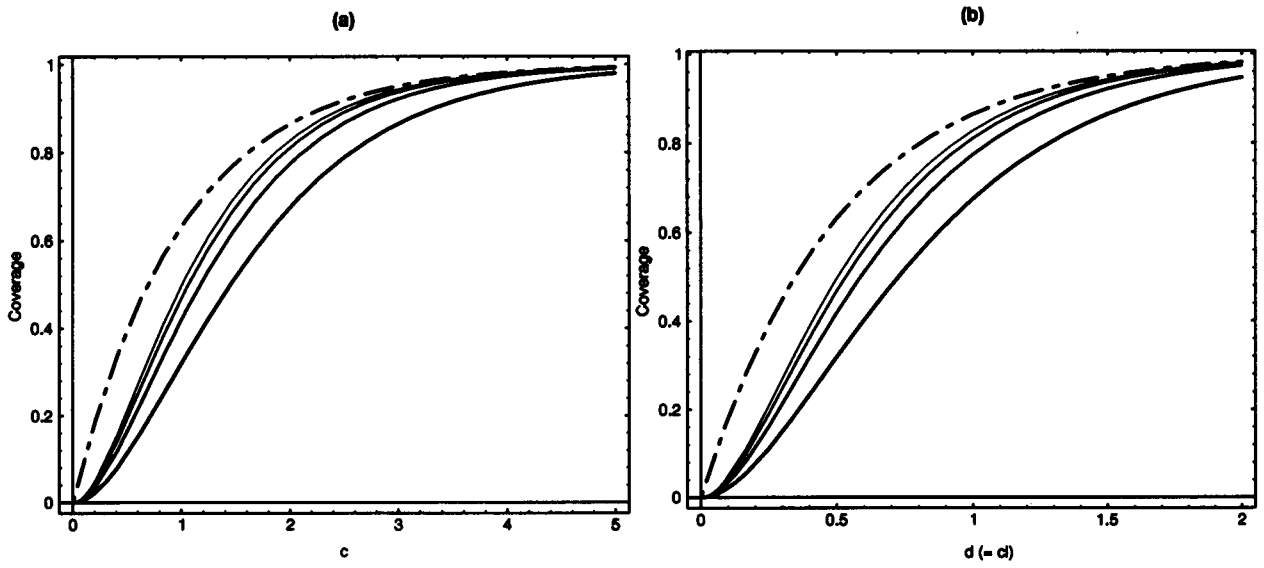

FIG. 7. Expected genome coverage by clones (dashed curves) and by contigs ($\theta = 0, 0.1, 0.25, 0.5$). The remaining curves corresponding to θ can be distinguished by line thickness, which varies with θ. The thinnest line is $\theta = 0$, e.g., (**a**) Lander–Waterman and (**b**) coverage by block contigs. Note that the horizontal axis is in units of c in **a** and $d = cl$ in **b.** In **b** we require $l < (4 - 2\theta)^{-1}$, so $c > (4 - 2\theta)d$.

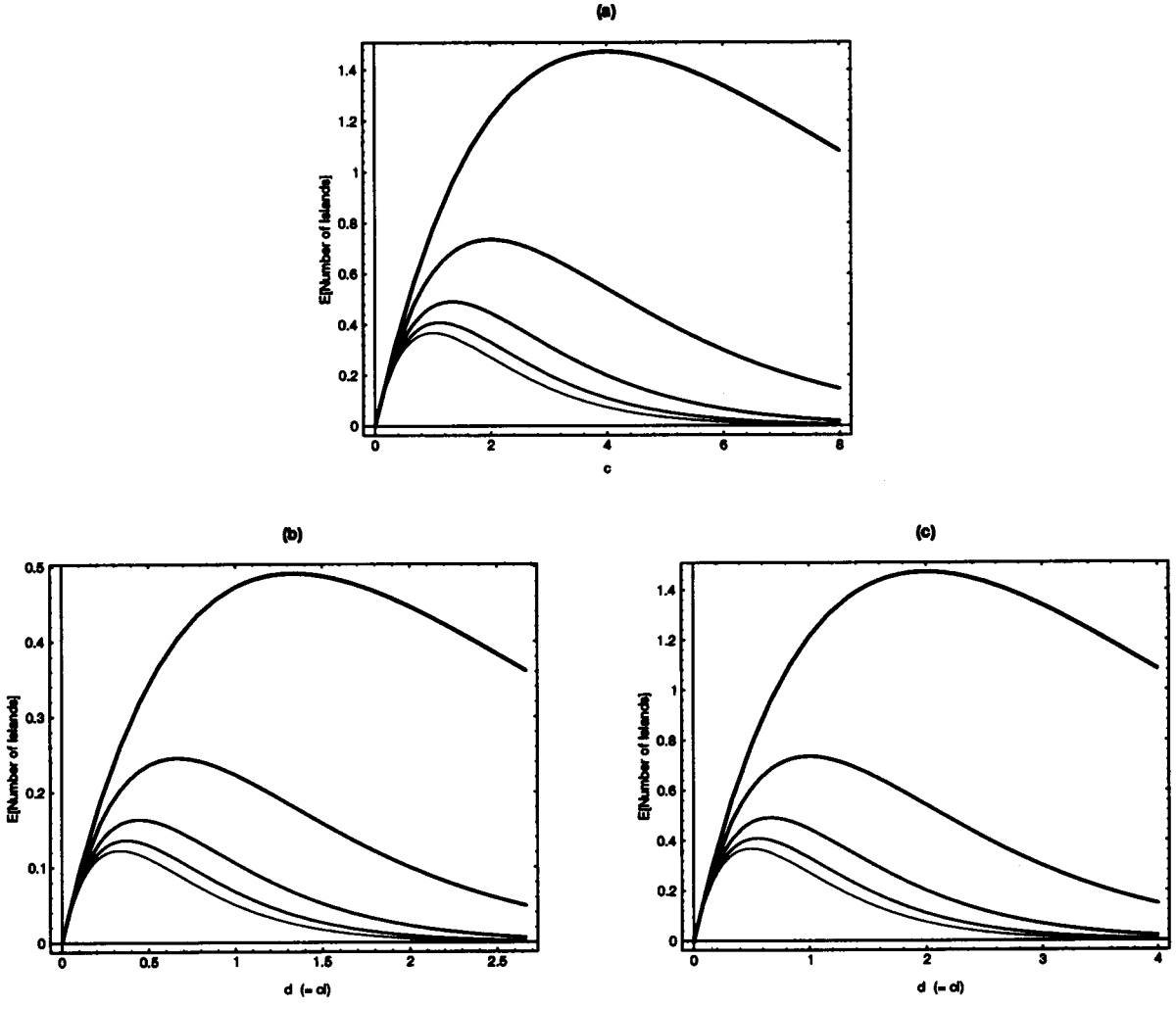

FIG. 8. Expected number of islands $\theta = (0, 0.1, 0.25, 0.5, 0.75)$. (**a**) Lander–Waterman in units of G/L, (**b**) greedy islands in units of G/lL, and (**c**) block islands in units of G/lL. The horizontal axis is in units of c in **a** and $d = cl$ in **b** and **c.**

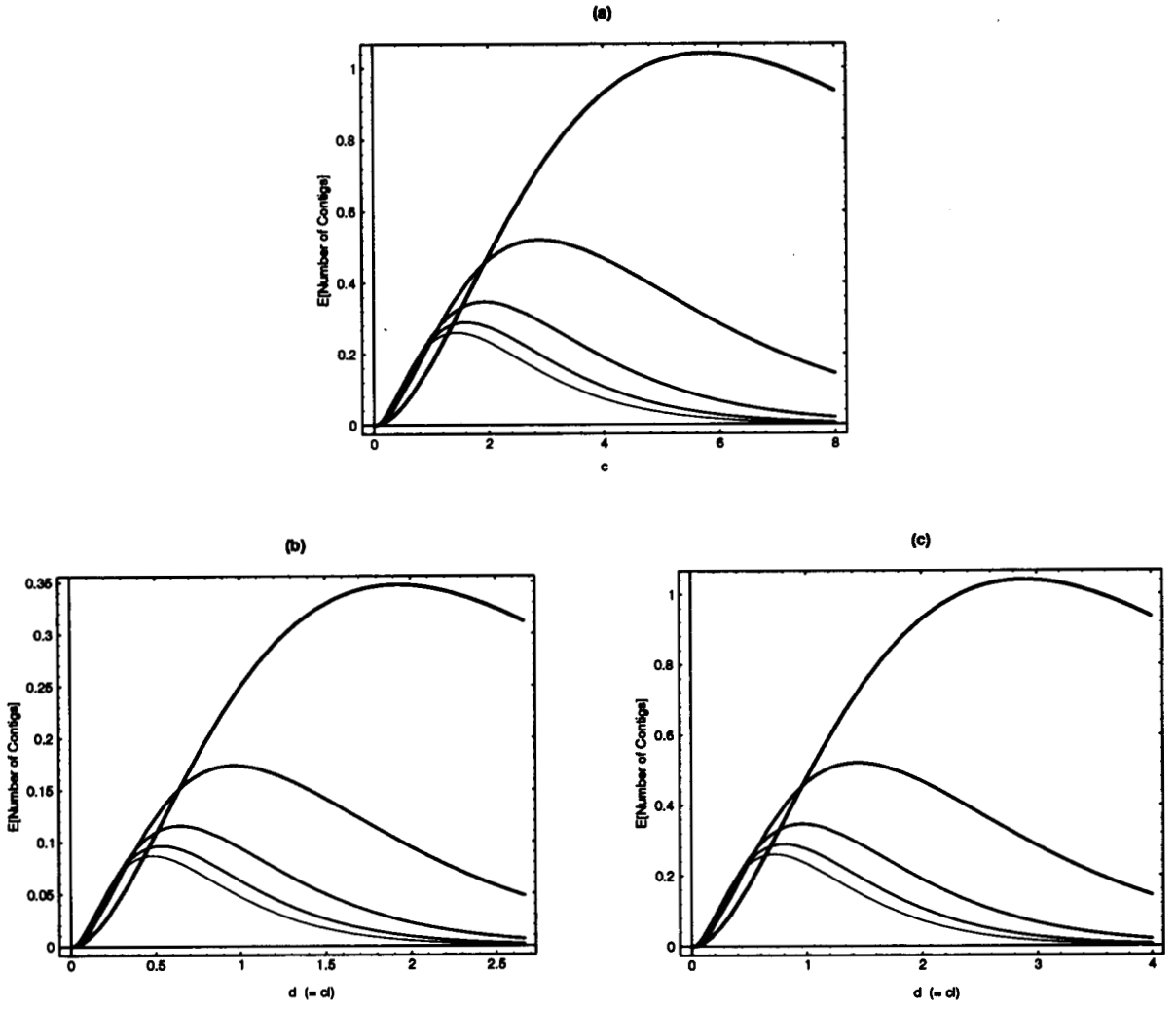

FIG. 9. Number of contigs θ = (0, 0.1, 0.25, 0.5, 0.75). (**a**) Lander–Waterman in units of G/L, (**b**) greedy islands in units of G/lL, and (**c**) block islands in units of G/lL. The horizontal axis is in units of c in **a** and $d = cl$ in **b** and **c**.

the interplay between the two clone ends gives an exit probability of $e^{-3c\sigma l}$. (In the gapped clones, σl plays the role of σ). For block islands, there are $2N$ blocks that occur in pairs at rate c, hence naively with an exit probability of $e^{-2c\sigma l}$. The expected number of contigs follow $Np(1 - p)$, where p is the exit probability.

The formulas for expected island length are not so easily related to one another. In most of the formulas there is a natural parameter $d = cl$, but in the case of expected length for greedy islands a separate graph must be drawn for each value of l. In Fig. 10 we graph expected island length and in Fig. 11 expected island length divided by expected contig length.

Edwards and Caskey (1991) introduce the ideas of mapped gap sequencing. The ends of clones are sequenced, and in our tables we assume that the sequence lengths (block lengths) are $T = lL$ = 250, 500, and 1000 bp. We assume that 25 bp is a minimum overlap required. Therefore, θ = 0.1, 0.05, and 0.025, respectively. Table 2 considers a G = 60 kb project with plasmid inserts of L = 1500 bp and depth c = 5. The coverage by greedy islands is high, 99%, with 21 islands for θ = 0.1 and 2 islands for θ = 0.05. The block islands of sequence show a different picture. For θ = 0.1, there are 89 sequence islands and 81% sequence island coverage. For θ = 0.05, there are only 17 islands and 96% coverage.

In Table 3, we consider a larger project. Here, G = 900 kb, and we have cosmids with L = 40,000. Because the sequenced ends are such a small fraction of the cosmid inserts, we take c = 20. Clearly the entire 900 kb will be covered by greedy islands. The block islands show interesting features. The block island coverage is 22% for θ = 0.1, 40% for θ = 0.05, and 63% for θ = 0.025. The number of islands also varies at 719, 560, and 540, respectively.

In Chen *et al.* (1993), ordered shotgun sequencing is proposed in which a YAC is subcloned into plasmids, plasmid ends are sequenced, and the end sequences (our blocks) are overlapped to create a plasmid map. The remainder of the strategy includes complete sequencing beginning with the plasmids that have been overlapped. The idea is to sequence the YAC with minimal redundant sequencing. The initial steps of the or-

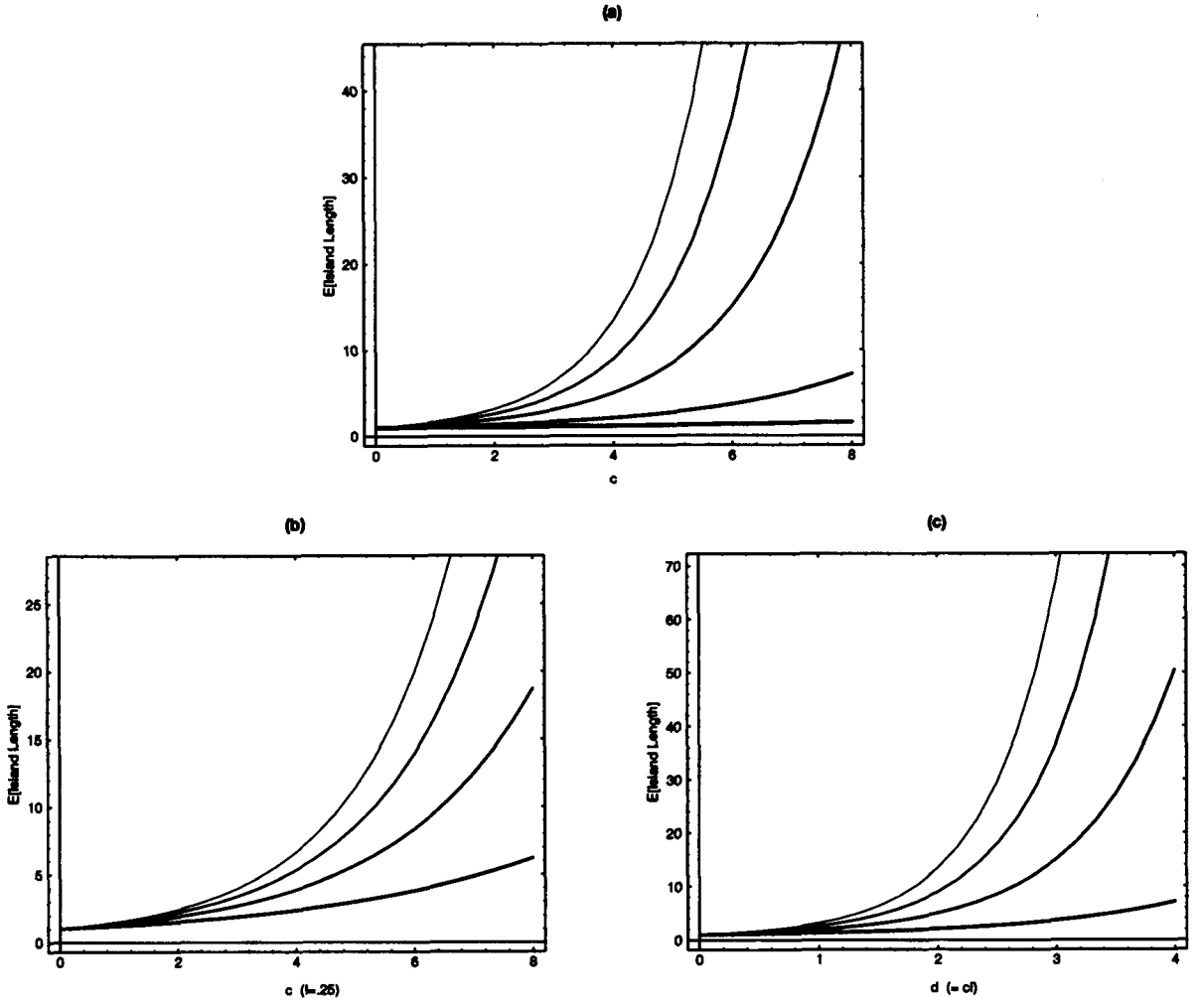

FIG. 10. Expected island length (**a**) Lander–Waterman θ = (0, 0.1, 0.25, 0.5, 0.75) in units of L, (**b**) greedy islands θ = (0, 0.1, 0.25, 0.5) with l = 0.25 in units of L, and (**c**) block islands in units of $lL\theta$ = (0, 0.1, 0.25, 0.5). The horizontal axis is in units of c in **a, b,** and units of $d = cl$ in **c.**

dered shotgun strategy clearly coincide with the models that we analyze in Section 3. Next, we show some associated numerical results. Let the YAC have G = 100 kb. The parameters are coverage c = 5, plasmid length L = 5 kb, $T = lL$ = 250 or 500 with θ = 0.1 or 0.05, respectively. Here N = 100, so there are 200 end sequences from plasmids. Theoretical results for both gapped plasmid maps and the end sequence maps appear in Tables 4a and 4b. Chen *et al.* (1993) carried out simulations using four published sequences and obtained values very close to those given here. Exact comparisons are unavailable, as their results are given by a graph.

4.3. The Models

Experimental results seldom fit a model's prediction perfectly. In the experiments modeled in this paper, there are many possible sources for differences. Bias in cloning efficiency has already been mentioned. The perfect θ-overlap detection is not achieved in practice. Clone lengths are not constant. Mathematical models that realistically account for these features have not yet been studied in detail and except for simple cases like Theorem 1′ are likely to be very difficult to study analytically. See Port (1994) for generalizations of Theorem 1′ corresponding to some of our results on gapped clones.

Recall that in Section 3.1 we were able to establish results only for greedy islands. For any $\theta \geq 0$, it appears to be very difficult to obtain the corresponding results for gapped islands. We conjectured the expected number of gapped islands to be of the form $Ne^{-\lambda c\alpha l}$, where $\lambda > 3$.

One realistic generalization for the gapped clone

Number of islands	Number of contigs	Model
$Ne^{-c\alpha}$	$Ne^{-c\alpha}(1 - e^{-c\alpha})$	Lander–Waterman
$Ne^{-3c\alpha l}$	$Ne^{-3c\alpha l}(1 - e^{-3c\alpha l})$	Greedy islands
$2Ne^{-2c\alpha l}$	$2Ne^{-2c\alpha l}(1 - e^{-2c\alpha l})$	Block islands

TABLE 1

Map Characteristics for Three Models of Cloning Efficiency

	(i) $c = 5$	(ii) $c(t) = \begin{cases} 2.5, & 0 \leq t < G/2 \\ 7.5, & G/2 \leq t \leq G \end{cases}$	(iii) $c(t) = \begin{cases} 0.5, & 0 \leq t\ G/2 \\ 9.5, & G/2 \leq t \leq G \end{cases}$
Number of clones	1000	1000	1000
Number of islands	7.63	21.84 + 0.50 = 22.35	30.7 + 0.09 = 30.80
Island length	26 kb	$\frac{(4.2)(21.84) + (199.7)(0.50)}{22.35} = 8.57$ kb	$\frac{(1.3)(30.7) + (1108.9)(0.09)}{30.80} = 4.54$ kb
Coverage by islands	0.993	$\frac{0.9179 + 0.9995}{2} = 0.9586$	$\frac{0.3932 + 0.9999}{2} = 0.6967$
Number of contigs	7.58	19.94 + 0.50 = 20.43	11.85 + 0.09 = 11.93
Contig length	26 kb	$\frac{89.73 + 99.95}{20.43} = 9.28$ kb	$\frac{20.145 + 99.80}{11.93} = 10.05$ kb
Number of clones/island	133	$\frac{249.95 + 749.45}{22.35} = 44.71$	$\frac{50.04 + 948.186}{30.80} = 32.41$
Coverage by contigs	0.993	$\frac{0.8988 + 0.9995}{2} = 0.949$	$\frac{0.2049 + 0.9999}{2} = 0.602$

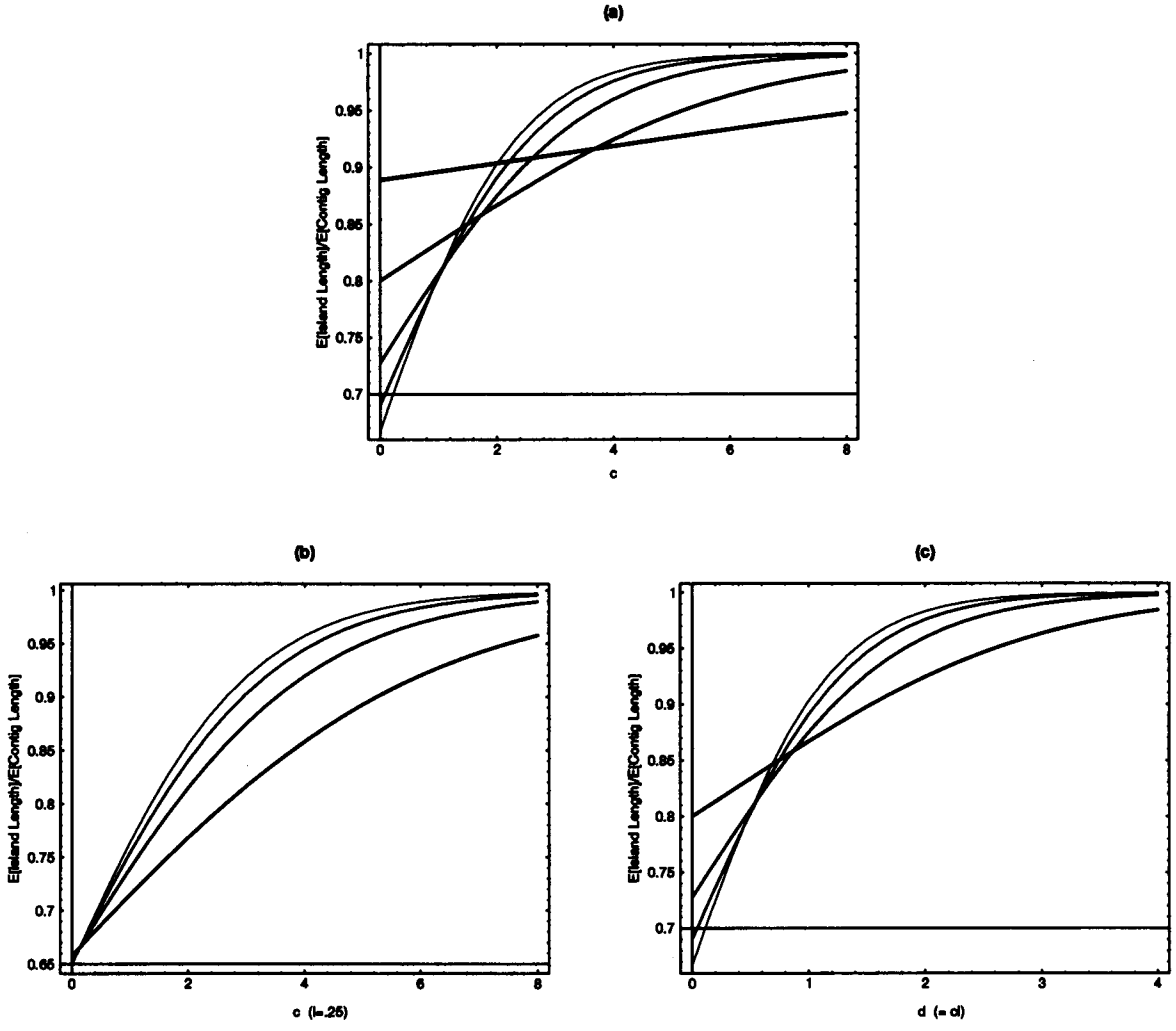

FIG. 11. Expected island length divided by expected contig length. (**a**) Lander–Waterman $\theta = (0, 0.1, 0.25, 0.5, 0.75)$, (**b**) greedy islands $\theta = (0, 0.1, 0.25, 0.5)$ with $l = 0.25$, and (**c**) block islands $\theta = (0, 0.1, 0.25, 0.5)$. The horizontal axis is in units of c in **a** and **b** and in units of $d = cl$ in **c**.

TABLE 2

Results for G = 60 kb, N = 200, c = 5, L = 1500

lL	250	500
θ	0.1	0.05
(a) Greedy islands		
Total number of plasmids	200	200
Number islands	21	1.73
Island length	6157 bp	35,586 bp
Coverage by islands	0.993	0.993
Number contigs	19	1.715
Contig length	6705 bp	35,884 bp
Number of clones/island	9.5	116
(b) Block islands		
Total number of plasmids	200	200
Number islands	89	17
Island length	547 bp	3434 bp
Coverage by islands	0.81	0.96
Number contigs	69	16
Contig length	633 bp	3563 bp
Number of clones/island	4.48	23.73
Coverage by contigs	0.79	[a]

[a] The formula fails because $l = 500/1500 = \frac{1}{3} > (4 - 2\theta)^{-1}$.

model is to keep the blocks with fixed length l but to let clone length L be a random variable. End sequenced cosmids certainly have this property. This creates mathematical difficulties that invalidate all of our proofs. It is easy to believe that the expected number of block islands remains $2Ne^{-2c\sigma l}$. If L is variable enough relative to l, then the expected number of greedy islands should be about $Ne^{-4c\sigma l}$. We are not sure of the technical conditions required to make this true or what

TABLE 3

Results for G = 900 kb, N = 400, c = 20, L = 40,000

lL	250	500	1000
θ	0.1	0.05	0.025
(a) Greedy islands			
Total number of cosmids	400	400	400
Number islands	321	221	104
Island length	50 kb	64 kb	105 kb
Coverage by islands	1.0	1.0	1.0
Number contigs	92	112	80
Contig length	75 kb	87 kb	124 kb
Number of clones/island	1.40	2.04	4.32
(b) Block islands			
Total number of cosmids	400	400	400
Number islands	719	560	340
Island length	2.8 kb	6.3 kb	16.8 kb
Coverage by islands	0.22	0.40	0.63
Number contigs	145	212	211
Contig length	3.6 kb	8.5 kb	20.9 kb
Number of blocks/island	1.25	1.61	2.65
Coverage by contigs	0.085	0.26	0.58

TABLE 4

Results for G = 100 kb, N = 100, c = 5, L = 1500

lL	250	500
θ	0.1	0.05
(a) Greedy islands		
Total number of plasmids	100	100
Number of islands	51	24
Island length	7,733 bp	12,362 bp
Coverage by clones	0.99	0.99
Number of contigs	25	18
Contig length	10,567 bp	14,693 bp
Number of clones per island	1.96	4.16
(b) Block islands		
Total number of plasmids	100	100
Number of islands	128	77
Island length	309 bp	818 bp
Coverage by blocks	0.39	0.63
Number of contigs	46	47
Contig length	413 bp	1018 bp
Number of blocks per island	1.57	2.59
Coverage by contigs	0.19	0.48

the general result is. We suggest variable clone length L as another area for future research.

The results in this and other papers are expected or mean values: expected number of islands, expected coverage, etc. The distribution of the random variables would be of interest instead of just the expected values. Hall's (1988) book has some results related to these difficult questions. We hope that future research addresses them, as they could provide an answer to the variation expected from the models.

ACKNOWLEDGMENTS

Michael Waterman learned of this set of mapping strategies from Tom Caskey, who inspired this work. We are grateful to Ellson Chen, David Cox, Simon Tavaré, and the two referees for their comments and corrections. We thank David Torney, who showed us a formula and derivation for expected coverage by L-W Contigs for $\theta \geq \frac{1}{2}$. We are grateful to Jared Roach, who pointed out to us that greedy islands and gapped islands are not equivalent. This work was supported by grants (to M.S.W.) from the National Science Foundation (DMS-90-05833) and the National Institutes of Health (GM-36230). Some of this work was performed while three of us were visiting DIMACS at Rutgers University (F.S., D.M., and M.S.W.) and was supported by the National Science Foundation (STC-91-19999 and BIR-9412594).

REFERENCES

Arratia, R., Lander, E., Tavaré, S., and Waterman, M. (1991). Genomic mapping by anchored random clones: A mathematical analysis. *Genomics* **11:** 806–827.

Barillot, E., Dausset, J., and Cohen, D. (1991). Theoretical analysis of a physical mapping strategy using random single-copy landmarks. *Proc. Natl. Acad. Sci. USA* **88:** 3917–3921.

Chen, E., Schlessinger, D., and Kere, J. (1993). Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones. *Genomics* **17:** 651–656.

Coulson, A., Sulston, J., Brenner, S., and Karn, J. (1986). Toward a

physical map of the genome of the nematode, *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* **83:** 7821–7825.

Daley, D. J., and Vere-Jones, D. (1988). "An Introduction to the Theory of Point Processes," Springer-Verlag, New York.

Edwards, A., and Caskey, C. T. (1991). Closure strategies for random DNA sequencing. *Methods* **3:** 41–47.

Ewens, W. J., Bell, C. J., Donnelly, P. J., Duinn, P., Matallana, E., and Ecker, J. R. (1991). Genome mapping with anchored clones: Theoretical aspects. *Genomics* **11:** 799–805.

Feller, W. (1991). "An Introduction to Probability Theory and Its Applications," Vol. I, Wiley, Canada.

Goldstein, L., and Waterman, M. (1987). Mapping DNA by stochastic relaxation. *Adv. Appl. Math.* **8:** 194–207.

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

Hoel, P., Stone, C., and Port, S. (1971). "Introduction to Probability Theory," Vol. 3, Houghton Mifflin, Boston.

Karlin, S., and Macken, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Am. Stat. Assoc.* **86**(413)**:** 27–35.

Kingman, J. (1993). "Poisson Processes," Academic Press, San Diego, CA.

Kohara, Y., Akiyama, A., and Isono, K. (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.

Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.

Lee, W. (1992). "A Mathematical Analysis of Genome Physical Mapping," Masters Thesis, University of Southern California.

Marr, T. G., Yan, X., and Yu, Q. (1992). Genomic mapping by single copy landmark detection: A predictive model with a discrete mathematical approach. *Mamm. Genome* **3:** 644–649.

Nelson, D. O., and Speed, T. P. (1994). Predicting progress in direct mapping projects. Manuscript in preparation.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., and Frand, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* **83:** 7826–7830.

Port, E. (1994). "Statistical Analysis of Physical Mapping Strategies," Ph.D. dissertation, University of Southern California.

Richards, S., Muzny, D. M., Civitello, A. B., Lu, F., and Gibbs, R. A. (1994). Sequence map gaps and direct reverse sequencing for the completion of large sequencing projects. *In* "Automated DNA Sequencing and Analysis Techniques" J. C. Venter, Ed., Chap. 28, pp. 191–198, Academic Press, San Diego, CA.

Smith, M. W., Holmsen, A. L., Wei, Y. H., Peterson, M., and Evans, G. A. (1994). Genomic sequence sampling: A strategy for high resolution sequence-based physical mapping of complex genomes. *Nature Genet.* **7:** 40–47.

Torney, D. C. (1991). Mapping using unique sequences. *J. Mol. Biol.* **217:** 259–264.