GENOMICS 2, 231–239 (1988)

Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis

ERIC S. LANDER*,† AND MICHAEL S. WATERMAN‡

*Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142; †Harvard University, Cambridge, Massachusetts 02138; and ‡Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089

Received January 13, 1988; revised March 31, 1988

Results from physical mapping projects have recently been reported for the genomes of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, and similar projects are currently being planned for other organisms. In such projects, the physical map is assembled by first "fingerprinting" a large number of clones chosen at random from a recombinant library and then inferring overlaps between clones with sufficiently similar fingerprints. Although the basic approach is the same, there are many possible choices for the fingerprint used to characterize the clones and the rules for declaring overlap. In this paper, we derive simple formulas showing how the progress of a physical mapping project is affected by the nature of the fingerprinting scheme. Using these formulas, we discuss the analytic considerations involved in selecting an appropriate fingerprinting scheme for a particular project.

1. INTRODUCTION

Given a gene of interest in a higher eukaryote, an increasing array of techniques are becoming available for isolating a DNA clone located within 1–2 million base pairs of the gene locus. Cloning the gene then requires traversing the remaining gap. The traditional approach of chromosomal walking, however, is ill-suited to such long distances.

Accordingly, attention has focused recently on the potential advantages of constructing a complete physical map of the DNA of an organism, consisting of overlapping clones spanning the entire genome. In principle, the overlapping clones could consist of phage with 15-kb inserts, cosmids with 40-kb inserts, or artificial chromosomes propagated in yeast with 100- to 1000-kb inserts (Burke *et al.*, 1987). By eliminating the need for tedious walking, a physical map would allow molecular biologists to focus on the more challenging issue of pinpointing the gene within an available region of up to several megabases and of studying its properties. In addition, the overlapping clones comprising the physical map would constitute the logical substrate for efforts to sequence an organism's genome.

Recently, three pioneering efforts have investigated the feasibility of assembling physical maps by means of "fingerprinting" randomly chosen clones. The fingerprints consisted of information about restriction fragment lengths. Overlaps between clones were inferred when the fingerprints of two clones were sufficiently similar. The three groups used different specific criteria for declaring overlap, but, broadly speaking, the criteria amount to the requirement that the clones overlap in a sufficiently long region. In particular:

(i) Olson *et al.* (1986) fingerprinted 5000 λ clones containing approximately 15-kb inserts of genomic DNA from *Saccharomyces cerevisiae*, by measuring the restriction fragment lengths obtained upon double digestion with *Eco*R1 and *Hin*dIII.

(ii) Coulson *et al.* (1986) adopted a somewhat different protocol to fingerprint nearly 8000 cosmids containing approximately 34-kb inserts of genomic DNA from *Caenorhabditis elegans*. They digested cosmid DNAs with the six-cutter *Hin*dIII, filled the 5′-overhang with radioactive nucleotides, digested with the four-cutter *Sau*3A, and then determined the size of the labeled fragments by electrophoresis in a sequencing gel followed by autoradiography.

(iii) Kohara *et al.* (1987) used a more elaborate fingerprint to analyze 1025 phage clones containing 15.5-kb inserts of genomic DNA from *E. coli*: for each clone, they constructed a complete restriction map involving sites for eight restriction enzymes. See Daniels and Blattner (1987) for another early *E. coli* mapping project.

As expected by each group of investigators, the random fingerprinting procedure did not yield complete physical maps. Instead, it produced many "islands," each a fragment of the desired map consisting

0888-7543/88 $3.00

of one or more overlapping clones. Completing the map requires finding clones that join the islands. Kohara *et al.* have already essentially completed this task for *E. coli* by closing 63 of 70 gaps via DNA hybridization, but this approach would be more formidable for higher organisms in which thousands of islands would occur.

Planning a physical mapping project involves estimating how the expected number of islands varies with the number of clones fingerprinted and the type of fingerprinting scheme used. Computer simulations have previously been used for this purpose (Coulson *et al.*, 1986).

In this paper, we derive some simple mathematical formulas that characterize the properties of islands in a physical mapping project. The formulas make explicit how the fingerprinting scheme itself—in particular, the extent of overlap that can be reliably detected between two clones—determines the progress of the project. In light of these results, we then present a simple analytical approach for comparing different fingerprinting schemes which might be considered in a mapping project.

The paper is organized as follows. Section 2 presents the basic formulas, Section 3 considers their agreement with experimental results, and Section 4 discusses some basic considerations in designing a fingerprinting scheme. All mathematical proofs are deferred until Section 5, so as not to interfere with the main points of the text.

In a sense, the results below may be considered as analogs of the well-known formulas given by Clarke and Carbon (1976) for the completeness of a recombinant library. Just as those formulas provide the approximate probability that any given sequence will be present in the library, the results here provide the approximate probability that contiguous stretches of any given size will be found when clones from such a library are fingerprinted.

In addition, we should note that the results below also apply to the progress of "shotgun" DNA sequencing, since the DNA sequence of the individual fragments can be thought of as the most detailed possible fingerprint.

2. THE DISTRIBUTION OF ISLANDS

A fingerprinting scheme consists of two parts: (i) a method for fingerprinting a clone to obtain certain partial information about it, and (ii) a rule for declaring overlaps between clones sufficiently strict that false positives are rare.

For the purpose of analysis, we will make certain simplifying assumptions, which we later relax. First, we will abstract away the details of the particular fingerprinting scheme by simply considering an idealized scheme capable of detecting overlap between two clones whenever they share at least a fraction θ of their length. (In reality, the minimum detectable overlap for most fingerprinting schemes will vary somewhat from clone to clone, depending on the number of restriction fragments in the clone. Nevertheless, we may think of θ as the *expected* minimum fractional overlap required between two clones.) In addition, suppose that the criteria for overlap are sufficiently stringent that false positives are rare.

Suppose that we have a perfectly representative genomic library, with all inserts of equal size. Define the following symbols:

G = haploid genome length in bp;
L = length of clone insert in bp;
N = number of clones fingerprinted;
$\alpha = N/G$ = probability per base of starting a new clone;
T = amount of overlap in base pairs needed to detect overlap;
$\theta = T/L$;
c = redundancy of coverage = LN/G.

Clones fall into "apparent" islands consisting of one or more members, based on their fingerprints. The islands are only "apparent" because some actual overlaps will go undetected. Islands with two or more members will be called "contigs," a coinage due to Staden (1980), and the gaps between islands will be called "oceans."

The following results describe some expected properties of islands and oceans, both apparent and actual, as the mapping project proceeds.

Proposition 1. Let θ be the fraction of length which two clones must share in order that the overlap be detectable given the fingerprinting scheme, let N be the number of clones fingerprinted, and let c be the redundancy of coverage. Also, let $\sigma = 1 - \theta$.

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

(ii) The expected number of apparent islands consisting of j clones ($j \geq 1$) is

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

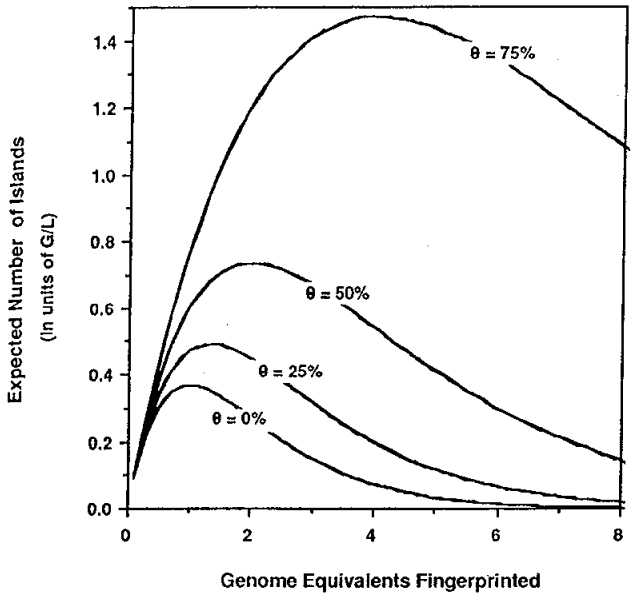
(ii′) The expected number of apparent islands consisting of at least two clones (i.e., contigs) is

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

(iii) The expected number of clones in an apparent island is $e^{c\sigma}$.

(iv) The expected length in base pairs of an apparent island is

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



	Approximate value of G/L		
	Phage (15kb)	Cosmid (40kb)	Yeast (1Mb)
E. coli	267	100	4
S. cerevisiae	1333	500	20
C. elegans	5,667	2,125	85
Human	200,000	75,000	3,000

FIG. 1. The graph shows the expected number of islands as a function of the number of genome equivalents fingerprinted, for four values of the minimum detectable overlap, θ. In order to make the graph independent of the genome size, the number of islands is expressed as a multiple of G/L, the size of the genome divided by the size of a cloned insert. The table lists the values of G/L for certain representative organisms and cloning vectors.

(v) The corresponding results for the actual islands that would result if all overlaps could be detected are obtained by setting $\sigma = 1$. For example, the expected number of actual islands is Ne^{-c}.

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

The minimum detectable overlap θ clearly has a major effect on the progress of a mapping project. Figure 1 shows the expected number of islands as a function of the number c of genome equivalents of DNA fingerprinted, and Fig. 2 shows expected average island length. At the beginning of the project, the number of islands increases because new clones are unlikely to overlap others. The maximum number of islands occurs at $c = (1 - \theta)^{-1}$ and is equal to $(G/L)e^{-1}(1 - \theta)^{-1}$. After this point, the number of islands declines as gaps are closed. After some point, a directed strategy for bridging gaps must be employed, since it would require a huge amount of work to close all the gaps by fingerprinting random clones.

Notice how decreasing the minimum detectable overlap from 50 to 25% greatly speeds the progress of the project. By contrast, the decrease from 25% minimum detectable overlap to the theoretical limit at 0% has relatively less effect. These results suggest that a fingerprinting scheme with $\theta = 0.15$–0.20 may be a sensible goal, with further decrease being of limited value. Of course, the advantage of smaller θ must be balanced with the increased effort to obtain a more sensitive fingerprint.

How do the results change if we relax our simplifying assumptions about θ and L? We briefly summarize the effect here; a more precise statement of the results is given in Section 5.

Proposition 2. Suppose that θ and L are allowed to vary from clone to clone. Compared to the situation in

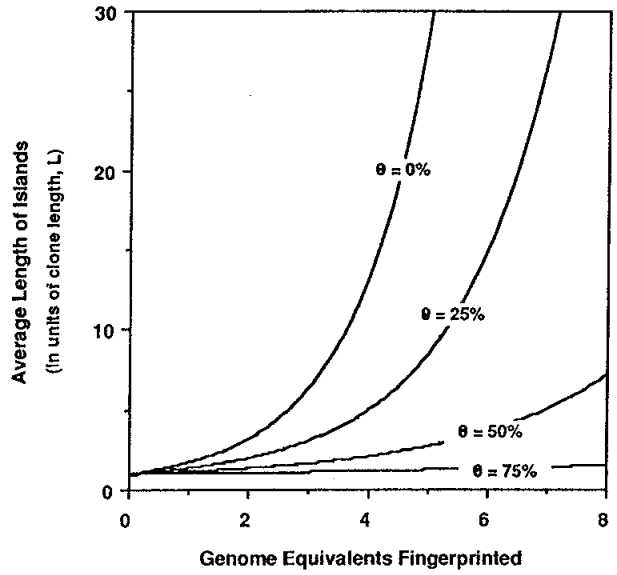


FIG. 2. The graph shows the expected length of an island as a function of the number of genome equivalents fingerprinted, for four values of the minimum detectable overlap, θ.

which θ and L were fixed at their average value, the following results hold:

(i) at any stage of the project, the expected number of islands is somewhat greater;

(ii) at any stage of the project, the expected number of clones per island is somewhat smaller;

(iii) the expected number of contigs is somewhat smaller at early stages of the project and somewhat greater at later stages of the project.

Finally, we ask: Is there an advantage to constructing a physical map chromosome by chromosome, rather than all at once? At first glance, there appears to be no advantage: the formulas of our model are linear in N, the number of clones fingerprinted. In an organism with two chromosomes of equal length, one would expect the same number of islands if N clones were fingerprinted from a whole genomic library or if $N/2$ clones were fingerprinted from each of two chromosome-specific libraries—provided that, in each case, the fingerprinting scheme could detect matches between clones overlapping in a proportion θ of their length.

In fact, there are some second-order considerations that favor subdividing the project:

(i) If the rule for declaring overlaps were kept constant, the rate of false positives would be greater for genomes of larger size. In order to maintain the same rate of false positives, a greater proportion of overlap θ must be required for declaring overlap in a larger genome. However, the effect is not large (cf. Section 4 below).

(ii) If the genome were divided into two parts, an investigator might decide not to fingerprint an equal number of each half: as the project progressed, the investigator could fingerprint more clones from whichever half was progressing more slowly. However, the law of large numbers assures roughly comparable progress in each half (unless systematic cloning bias caused one half to be significantly less representative.) Only a slight increase in efficiency would result (unless the genome were decomposed into so many parts that the expected number of clones required to cover each part was small).

In addition to these mathematical considerations, there are various practical issues that might favor subdividing the project as well. For example, subdividing the project permits the adoption of improved fingerprinting strategies for later parts, as they become available.

3. AGREEMENT WITH EXPERIMENTAL DATA

The formulas in the previous section can be used to predict the expected progress of the recent physical mapping projects in bacteria, yeast, and nematodes. The extent of agreement will depend, of course, on how closely the assumptions above are satisfied.

One key consideration is whether the minimum detectable overlap may be taken to be a constant θ, independent of the clone under consideration. Proposition 2 gives some implications of nonconstant θ. This assumption will be satisfied by fingerprinting methods involving information derived from a large number of fragments (such as that used in *E. coli*): in this case, clones will all have about the same density of fingerprint information and therefore will require roughly the same minimum overlap needed to recognize a match. By contrast, our assumption may not fit well for more rudimentary fingerprinting schemes: some clones might contain too little fingerprint information to make it possible ever to declare overlap. For example, since the yeast project required that clones share five restriction fragment lengths in order to declare pairwise overlap, those clones containing fewer than six fragments could never be joined into islands. Thus, our formulas would be expected to agree closely with the data for the *E. coli* project, but significantly to underestimate the number of singleton clones for the yeast project. This is indeed the case, as we shall see presently.

A second important assumption is that the libraries represent unbiased samples from the genome. Cloning bias will obviously slow progress, to a degree that depends rather sensitively on the nature and extent of the bias. For example, some bias was reported in the *C. elegans* project, as noted below.

(i) E. coli

On the basis of pairwise fingerprint comparison, Kohara *et al.* (1987) arranged 1025 clones into 70 islands, of which 7 were isolated singletons. A hybridization method was then used to find clones spanning the gaps, resulting in all but 6 gaps being closed.

The authors reported that the cloned inserts were 15.5 kb on average and that an overlap of about 3 kb could be detected. Taking the minimum detectable overlap to be $\theta = 3/15.5 \approx 0.19$, the genome size to be 4704 kb, and the genome coverage $c = (1025)\ (15.5)/4704 \approx 3.38$, the formulas predict 67.16 islands of which 4.39 should be singletons. The agreement with the observed 70 and 7 is quite close. The small disagreement is even in accord with the results of Proposition 2 on variable θ and L.

Moreover, the formulas predict that most gaps will be small enough to be closed in one step via hybridization, but that a few should remain. Given an infinite library to screen via hybridization, Proposition 1(6) predicts that about 2 of 70 gaps could not be closed because they would be longer than 15.5 kb. If a 1-kb overlap on each end were required to obtain a positive hybridization signal, about 4 of 70 gaps would be expected to remain because their lengths exceed 13.5 kb. Since only a finite library containing only 2344 clones

was screened by hybridization (including the clones that had been fingerprinted), a few additional gaps might remain simply because an optimally situated clone did not occur in the finite library. Thus, the six remaining gaps are well within expectation.

(ii) S. cerevisiae

Using a much simpler but less sensitive fingerprint, Olson *et al.* (1986) arranged 4946 clones into 1422 islands, consisting of 680 contigs and 742 singletons. The average number of fragments per clone was reported to be 8.36 and the criteria for declaring overlap included the requirement that a pair of clones share at least 5 fragments. The authors noted that most of the 742 singletons contained 5 or fewer fragments and thus even large overlaps involving such clones would be ignored in pairwise comparisons. Accounting for the problem of clones with 5 or fewer restriction fragments, what do our formulas predict for such a project?

A simple calculation shows that somewhat less than 14% of the clones would be expected to have 5 or fewer restriction fragments. (Suppose that restriction sites are distributed according to a Poisson process with a mean of 8.36 fragments, or 7.36 restriction sites, per clone. Letting K denote the actual number of restriction sites in a clone, then, for an any integer K, $\text{Prob}(K = k) = \lambda^k e^{-\lambda}/k!$ with $\lambda = 7.36$. Accordingly, about 14% of all clones will have 4 or fewer sites. This estimate must be reduced slightly, since fragments smaller than 400 bp were not scored in deriving the estimate $\lambda = 7.36$.) Taking this proportion somewhat arbitrarily at 13%, about 4300 clones would be expected to have 6 or more fragments and about 650 would have 5 or fewer.

With 5 out of an average of 8.36 common fragments being required to declare overlap, we roughly estimate θ as $5/8.36 \approx 0.60$. Since a few other technical conditions were required, the estimate should be increased slightly. Somewhat arbitrarily, we take $\theta = 0.63$.

With $N = 4303$, $\theta = 0.63$, and $c = 4.5$ (as reported by the authors), we expect 660 contigs and 154 singletons. Adding the 643 singletons expected from clones with 5 or fewer fragments yields a total of 660 contigs and 797 singletons. This agrees fairly well with the observed 680 contigs and 742 singletons, given the rough approximations involved.

Since the yeast project was a first-generation mapping project, it is interesting to compare the results to the expected progress if one instead used the fingerprint subsequently developed for *E. coli*. With the same number of clones ($N = 4946$, $\theta = 0.20$, $c = 4.5$), one would expect only about 131 contigs and 4 singletons. Thus, as noted above, the use of a more informative fingerprint can greatly speed progress toward the completion of a physical map.

(iii) C. elegans

Coulson *et al.* (1986) used Monte Carlo computer simulations to determine expected progress curves, assuming unbiased cosmid libraries. Our formulas above reproduce these curves.

Unfortunately, their experimental progress was significantly slower than expectation. Some of the explanation may lie in the bias in cloning efficiency that the authors reported: certain sequences occurred much more often than expected in their clone bank (e.g., 100-fold overrepresentation of ribosomal DNA). Without knowing the precise nature of the cloning bias, it is difficult to estimate its effect on the expected progress.

Clearly, it is important to minimize bias in cloning efficiency, although this is easier said than done. The authors speculate that *Eco*K restriction endonuclease activity in packaging extracts might contribute to the bias. It is also possible that cosmid libraries are subject to greater cloning bias than phage libraries, such as were used in the *E. coli* and *S. cerevisiae* mapping projects.

4. EVALUATING A FINGERPRINTING SCHEME

A good fingerprinting scheme should be able to detect relatively small overlaps between clones, while allowing an acceptably low rate of false positives given the size of the genome to be mapped. Here, we discuss how the choice of the fingerprint itself and the rule for declaring overlaps determine the minimum detectable overlap and the false positive rate, with the aim of providing general guidance for those designing a mapping project.

We consider two basic types of fingerprinting schemes:

Type (a)

The fingerprint consists of the lengths of the restriction fragment lengths produced following digestion by a single enzyme (or a combination of enzymes used consecutively, as by Olsen et al. and Coulson *et al.*) that produces an average of n fragments. The matching rule declares an overlap when two fingerprints share at least k fragment lengths. Since measurement error is roughly proportional to fragment length, two restriction fragments will be assumed to match if their lengths differ by at most $100\beta_1\%$. Typically, $0.01 \leq \beta_1 \leq 0.05$, depending on the gel system.

Type (b)

The fingerprint consists of a restriction map for a single enzyme (or enzyme combination) that produces an average of n fragments. The matching rule declares an overlap when the lengths of k terminal fragments in the two maps agree. Fragments will be assumed to match if their lengths differ by at most

$100\beta_2\%$, where β_2 may be larger than, smaller than, or equal to β_1, depending on how the restriction maps are made (partial digestion or double digestions). In our examples below, we will take $\beta_1 = \beta_2 = 0.03$.

These two examples are meant to be illustrative, not exhaustive. Other fingerprints, as well as more elaborate rules for declaring matches, clearly deserve further exploration.

Intuitively, it is clear that fingerprints of type (b) contain more information than fingerprints of type (a), provided that $\beta_1 = \beta_2$. The restriction map drastically restricts which fragment length matches are to be considered meaningful—namely, those corresponding to overlaps between terminal segments of the two restriction maps.

The following result estimates the minimum detectable overlap and the chance of false positives for the two fingerprinting schemes specified above.

Proposition 3. With the assumptions above, the minimum detectable overlap θ for both fingerprinting methods is approximately k/n.

(i) For a fingerprint of type (a), the expected number of fragments shared by fingerprints of two nonoverlapping clones will be about $\mu_1 = \frac{1}{2}\beta_1 n^2$. Provided that μ_1 is relatively small compared to n, the distribution of fragments shared by two nonoverlapping clones will be approximately Poisson; that is, the probability that a declared overlap is a false positive is approximately

$$\pi_1(n, \beta_1, k) \approx \text{Prob}(X \geq k),$$

where X is a Poisson random variable with mean μ_1.

(ii) For a fingerprint of type (b), the probability that a declared overlap is a false positive is approximately

$$\pi_2(n, \alpha, k) = 4\,(\tfrac{1}{2}\beta_2)^k(1 + \tfrac{1}{2}\beta_2).$$

The two fingerprinting schemes produce roughly the same minimum detectable overlap $\theta \approx k/n$, but the rate of false positives is considerably higher for type (a) fingerprints than for type (b) fingerprints. (Indeed, the rate of false positives increases with n for type (a) fingerprints, but is essentially independent of n for type (b).) In order to achieve a comparable rate of false positives, a larger value of k must be used for type (a) fingerprints, which would increase the overlap θ required for detection.

For example, suppose that we use an enzyme (or enzyme combination) that yields an average of 10 fragments per insert and gels that can resolve fragment lengths to within $\beta_1 = \beta_2 = 0.03$. If we simply measure fragments lengths and require that two such type (a) fingerprints match at 7 fragments, then the chance of a false positive will be about 0.0009 according to Proposition 3(i) and the minimum detectable overlap θ will be approximately $7/10 = 0.70$.

If we instead made a restriction map of these fragments, then the same false positive rate according to Proposition 3(ii) could be obtained by requiring that fingerprints share only 2 fragments in common. Thus, the minimum detectable overlap yielding roughly the same false positive rate would be $\theta \approx 2/10 = 0.20$—a substantial improvement in view of the results of Section 2. In general, the false positive rate when an overlap of 2 fragments is required is $\pi_2(n, \beta_2, k = 2) \approx \beta_2^2$ which may be adequate for most purposes and yields $\theta = 2/n$.

One can clearly construct more detailed fingerprints by combining multiple fingerprints of type (a) or type (b) (as Kohara *et al.* did using eight different restriction maps), determined using separate gel lanes. If we require that each of the component fingerprints match in at least k fragments in order to declare an overlap, then the chance of a false positive is roughly the product of the false positive rates for each of the component fingerprints.

For example, if we wished to avoid constructing a restriction map in the situation described above, we could nevertheless achieve roughly the same false positive rate and the same minimum detectable overlap (of $\theta = 0.20$) by using multiple type (a) fingerprints involving a number of independent enzymes, each yielding about 10 fragments. Since each fingerprint would have a false positive rate of about 0.44 if overlaps were declared whenever fingerprints shared two bands, straightforward calculation shows that about nine independent enzymes are required. The choice of whether to construct a single restriction map (say, via partial digestion as in Kohara *et al.*) or to determine restriction fragment lengths for nine different enzymes would be governed in practice by an investigator's estimate of the work involved in each approach and of the acceptable rate of false positives for the project. Also, note that the analysis above depends on the resolving power (β_1 and β_2) of the gels used.

By following the basic approach described here, one can compare the features of various fingerprinting schemes under consideration for a project. Since the discussion in this paper makes a few simplifying assumptions (which may not be satisified, for example, by fingerprints containing few fragments, as mentioned in Section 3), we would recommend that computer simulations be performed as a final step, taking into account any significant deviations from the assumptions.

A final remark concerning the comparison of type (a) and type (b) fingerprints above is in order. We have analyzed the rate of false positives under the assumption that only pairwise comparisons are made between fingerprints. Although this is appropriate to the early stages of a mapping project, it should be noted that stronger matching rules can be invoked if many genome equivalents are eventually finger-

printed: multiple overlap clones in a region impose additional "topological" constraints which, in the limit, allow the entire restriction map to be reconstructed (Olson *et al.*, 1986). Thus, our estimate of the false positive rate is too pessimistic in the event that many genome equivalents are to be fingerprinted. Further attention should be given to quantifying the nature and extent of this additional information.

5. MATHEMATICAL PROOFS

This section contains the mathematical proofs for Propositions 1, 2, and 3. Readers interested primarily in applications may wish to omit this section on first reading.

Proof of Proposition 1. Imagine that we move from base pair to base pair through the genome, starting at one end. The probability that we encounter the beginning of a cloned insert at any base pair is α. An island begins when we encounter a cloned insert and continues while we detect overlapping clones. The probability that we begin a cloned insert and fail to detect an overlapping clone is $\alpha(1-\alpha)^{L\sigma} = \alpha(1 - N/G)^{(G/N)c\sigma} \approx \alpha e^{-c\sigma}$. Since the number of islands is equal to the number of times we exit a clone without detecting overlap, the expected number of islands is $G\alpha e^{-c\sigma} = Ne^{-c\sigma}$ and we have shown (i).

The above reasoning shows that the number of clones in an island follows a geometric distribution with stopping probability $e^{-c\sigma}$. Thus, the probability that an island contains exactly j clones is $(1-e^{-c\sigma})^{j-1}e^{-c\sigma}$. Multiplying this probability by the expected number of islands gives (ii), while the mean of the distribution required by (iii) is $e^{c\sigma}$.

To prove (iv), consider an island consisting of J clones, where J has the geometric distribution described in the last paragraph. Let X_i denote the length (in base pairs) of the coverage of the ith clone ($1 \leqslant i \leqslant J-1$). Since the ith clone either ends with the beginning of the next clone (if a new clone is encountered within the first $L\sigma$ bases) or is the last clone and contributes length L (if no new clone is encountered), then the random variable X_i has the distribution

$$P(X_i = m) = \alpha(1-\alpha)^{m-1}, \quad \text{for} \quad 1 \leqslant m \leqslant L\sigma$$

$$P(X_i = m) = 0, \quad \text{for} \quad L\sigma < m < L$$

and

$$P(X_i = L) = (1-\alpha)^{L\sigma}.$$

The expected length of an apparent island is then $E(\Sigma_{1\leqslant i\leqslant J} X_i)$. Evaluating this expectation requires a bit of special theory. The random variable J is a stopping time for $X_1, X_2, \cdots$ (that is, the event $\{J=j\}$ is independent of $X_{j+1}, X_{j+2}, \cdots$). The expectation can then be evaluated by Wald's equation (Ross, 1970, p. 38): $E(\Sigma_{1\leqslant i\leqslant J} X_i) = E(X)E(J)$. Straightforward calculation shows that

$$E(X) = L[((1-e^{-c\sigma})/c) - (1-\sigma)e^{-c\sigma}],$$

and, since J has a geometric distribution, $E(J) = e^{c\sigma}$. Thus, (iv) follows directly. Statement (v) follows from the definition of σ. For (vi), we require the probability that an ocean of length at least kL occurs at the end of an apparent island. This means that no new clones begin within $\theta L + kL$ bases, and the probability of this event is $(1-\alpha)^{\theta L + kL} \approx e^{-c(\theta+k)}$. ■

Next, we develop a more precise version of Proposition 2, which relaxes the condition that L and θ are constant. Let the size of the cloned insert be chosen according to some probability distribution with mean $E(L)$. The overlap between two clones necessary to detect overlap will be $\theta E(L)$ base pairs, where θ is chosen according to some probability distribution with mean $E(\theta)$. The distribution of θ is meant to model the fact that the fingerprints of certain clones make it easier to detect overlaps (i.e., by having more than the expected number of restriction fragments).

Let G, N, and α be defined as before. Define the redundancy by $c = E(L)N/G$. It will become evident that $\sigma = L/E(L) - \theta$ is the correct formula for σ.

The probability that overlap is not detected in a clone of length L is $(1-\alpha)^{L-E(L)\theta} \approx e^{-c(L/E(L)-\theta)} = e^{-c\sigma}$. Let $f(\sigma)$ denote the probability density function of σ. The average stopping probability is then $\int e^{-c\sigma} f(\sigma)d\sigma$. Replacing $e^{-c\sigma}$ by this integrated form gives the following proposition, which generalizes Proposition 1(i–iii) and provides a precise version of Proposition 2.

Proposition 2'. With assumptions and notation as above,

(i) The expected number of apparent islands is

$$N\int e^{-c\sigma} f(\sigma)d\sigma.$$

(ii) The expected number of apparent islands consisting of j clones ($j \geqslant 1$) is

$$N\left\{\int e^{-c\sigma} f(\sigma)d\sigma\right\}^2\left\{1 - \int e^{-c\sigma} f(\sigma)d\sigma\right\}^{j-1}.$$

(iii) The expected number of clones in an apparent island is

$$\left\{\int e^{-c\sigma} f(\sigma)d\sigma\right\}^{-1}.$$

To justify the generalization of Proposition 1(iv), we need to assume that the X_i, the coverage of the ith clone in base pairs, define statistically independent random variables. This was easily the case when L and θ were constant and is a reasonable assumption here. Under this assumption, we have:

(iv) The expected length in base pairs of an apparent island is

$$\left\{\int e^{-c\sigma} f(\sigma)d\sigma\right\}^{-1}[\alpha^{-1} - \int\int L\sigma e^{-c\sigma}/(1 - e^{-c\sigma})f(L, \sigma)\, dLd\sigma] + \int\int L(1-\sigma)f(L, \sigma)dLd\sigma,$$

where $f(L, \sigma)$ is used to indicate the random nature of both L and σ.

Although the precise results depend upon the distributions of L and θ, some general conclusions can be drawn about the effect of letting L and θ vary compared to fixing them at their average values. For example, Jensen's inequality states that a convex function of the average of a random variable is less than or equal to the average of that convex function evaluated at the random variable. Since $e^{-c\sigma}$ is a convex function, (i) implies that

$$\text{Avg number of islands} \geqslant Ne^{-cE(\sigma)} = Ne^{-c(1-E(\theta))},$$

and (iii) implies that

$$\text{Avg number of clones per island} \leqslant e^{cE(\sigma)} = e^{c(1-E(\theta))}.$$

In other words, using the average values of L and θ underestimates the number of islands and overestimates the number of clones in an island.

Unfortunately, the effect on the number of contigs (islands consisting of two or more clones) cannot be determined via Jensen's inequality, since the number of contigs is the difference between two convex functions, but not itself a convex function.

An alternative approach to approximating (i), (ii), and (iii) is to expand the exponential functions into a Taylor series. Let $\text{Var}(\sigma)$ denote the variance of σ. Expanding out to second-order terms yields

(i′) Avg number of islands
$$\approx Ne^{-cE(\sigma)}\{1 + \tfrac{1}{2}c^2\text{Var}(\sigma)\};$$

(ii′) Avg number of contigs
$$\approx (Ne^{-cE(\sigma)} - Ne^{-2cE(\sigma)}) + Nc^2e^{-cE(\sigma)}\{\tfrac{1}{2}\text{Var}(\sigma)\}\{1 - 2Ne^{-cE(\sigma)} - \tfrac{1}{2}Nc^2e^{-cE(\sigma)}\text{Var}(\sigma)\};$$

(iii′) Avg number clones per island
$$\approx \{e^{-cE(\sigma)}\{1 + \tfrac{1}{2}c^2\text{Var}(\sigma)\}\}^{-1}.$$

It is clear from the Taylor expansion that ignoring the variability of L and θ underestimates the number of islands and overestimates the number of clones in an island, as we noted from Jensen's inequality. Inspecting the second term in (ii′), we see that the number of contigs decreases under variation for small c and increases for large c.

How large are these changes? Each effect involves Var (σ). From the definition of σ, we find that

$$\text{Var}(\sigma) = \text{Var}(L)/E(L)^2 + \text{Var}(\theta),$$

which is not likely to be very large. The first term is unlikely to contribute significantly, while $\text{Var}(\theta)$ might be as large as 0.01. Still, multiplying by Nc^2 in (ii′) increases the proportional effect.

Finally, we turn to the properties of fingerprinting schemes asserted in Proposition 3.

Proof of Proposition 3. If we require that the fingerprints of two clones share at least k out of an expected n fragments, then the expected required overlap θ will clearly be about k/n. In fact, θ is slightly less than k/n, because, with positive probability, the nonoverlapping portions may contribute matching fragments. One may adjust the value of θ by simply accounting for the expected number of accidental matches among the fragments in the $(1 - \theta)$ of the clones which do not overlap. However, the effect is small and the approximation $\theta = k/n$ will suffice for most purposes.

To compute the probability of a false match being declared between two nonoverlapping clones, we first calculate the chance that two randomly chosen restriction fragments have matching lengths. If the restriction enzyme yields fragments with mean length λ^{-1}, then the restriction fragment lengths are well approximated by a continuous exponential distribution with density $f(x) = \lambda e^{-\lambda x}$, for $x > 0$. Suppose that we pick two fragments at random from such a distribution, that the first fragment has length x chosen from the distribution $f(x)$, and that the second fragment will match to within $100\beta\%$ provided that its length is between $x(1 - \beta)$ and $x(1 + \beta)$. Thus, the chance that two random fragments match is

$$\int_0^\infty \left[\int_{x(1-\beta)}^{x(1+\beta)} (\lambda e^{-\lambda y}dy)\right]\lambda e^{-\lambda x}dx = 2\alpha/(4 - \beta^2) \approx \tfrac{1}{2}\beta.$$

To be more precise, we could use the actual upper and lower limits for fragment sizes resolvable on the gel as the limits for the first integral. However, the simple approximation $\frac{1}{2}\beta$ is usually precise enough for use.

In comparing two fingerprints of type (a), there are n^2 ways to pick one fragment from each fingerprint, when each fingerprint consists of n fragments. Thus, the expected number of matching pairs will be $\frac{1}{2}\beta_1 n^2$. Since matches between any given pair of fragments are rare, it is plausible that the distribution of fragments will be roughly Poisson with mean $\frac{1}{2}\beta_1 n^2$. In fact, a proof can easily be obtained by using Proposition 1 in Arratia *et al.* (1988), from which one can even derive a bound on the variation distance from the Poisson distribution.

In comparing two fingerprints of type (b), the situation is much more limited. The chance that two restriction maps will match in exactly k fragments is roughly 4 $(\frac{1}{2}\beta_2)^k$, there being four orientations for the two maps and, once the orientations are fixed, each terminal k fragments in the two maps must match

exactly. The chance that the maps will match in at least k fragments is then

$$4(\tfrac{1}{2}\beta_2)^k + 4(\tfrac{1}{2}\beta_2)^{k+1} + 4(\tfrac{1}{2}\beta_2)^{k+2} + \cdots = 4(\tfrac{1}{2}\beta_2)^k(1 - \tfrac{1}{2}\beta_2)^{-1} \approx 4(\tfrac{1}{2}\beta_2)^k(1 + \tfrac{1}{2}\beta_2).$$

This completes the proof of Proposition 3. ■

6. CONCLUSION

In the next few years, physical mapping projects of a number of organisms, including bacteria, fungi, fruit fly, mouse, and human, will likely be started. Many factors bear upon the design of such projects, only some of which are mathematical. Practical matters will be important, including the ability to streamline various laboratory procedures or the availability of chromosome-specific libraries. Biological considerations may also be significant, including the extent of highly repetitive DNA in a genome which may prove difficult to fingerprint uniquely or extreme differences in the cloning efficiencies of various fragments.

While recognizing these factors, it is clear that significant improvements in the efficiency of physical mapping projects can be made through the careful design of a fingerprinting scheme. The analysis and formulas here should be of some value in designing and monitoring the progress of physical mapping projects.

ACKNOWLEDGMENTS

The authors are indebted to Lou Gordon and Elbert Branscomb for many discussions about mathematical aspects of genomic mapping. We are grateful to Fotis Kafatos and Babis Savakis for helpful conversations concerning the practical planning of physical mapping projects. We thank the Foundation les Treilles and Mme. Gruner-Schlumberger for providing an opportunity to develop the ideas discussed here. This work was partially supported by grants from the System Development Foundation (to E.S.L. and to M.S.W.), the National Science Foundation (DCB-8611317 to E.S.L.), and the National Institutes of Health (GM-36230 to M.S.W.).

REFERENCES

1. ARRATIA, R., GOLDSTEIN, L., AND GORDON, L. (1988) Two moments suffice for Posisson approximations: The Chen–Stein method. *Ann. Prob.*, in press.
2. BURKE, D., CARLE, G. F., AND OLSON, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806–812.
3. CLARKE, L., AND CARBON, J. (1976). A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire *E. coli* genome. *Cell* **9:** 91–101.
4. 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.
5. DANIELS, D. L., AND BLATTNER, F. R. (1987). Mapping using gene encyclopedias. *Nature (London)* **325:** 831–832.
6. 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.
7. 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.
8. ROSS, S. M. (1970). "Applied Probability Models with Optimization Applications," Holden-Day, San Francisco.
9. STADEN, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* **8:** 3673–3694.