Optimizing Restriction Fragment Fingerprinting Methods for Ordering Large Genomic Libraries[1]

E. BRANSCOMB,* T. SLEZAK,* R. PAE,* D. GALAS,† A. V. CARRANO,* AND M. WATERMAN†'‡

*Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, California 94550; and Departments of †Molecular Biology and ‡Mathematics, University of Southern California, Los Angeles, California 90089-1113

Received September 18, 1989; revised May 5, 1990

We present a statistical analysis of the problem of ordering large genomic cloned libraries through overlap detection based on restriction fingerprinting. Such ordering projects involve a large investment of effort involving many repetitious experiments. Our primary purpose here is to provide methods of maximizing the efficiency of such efforts. To this end, we adopt a statistical approach that uses the likelihood ratio as a statistic to detect overlap. The main advantages of this approach are that (1) it allows the relatively straightforward incorporation of the observed statistical properties of the data; (2) it permits the efficiency of a particular experimental method for detecting overlap to be quantitatively defined so that alternative experimental designs may be compared and optimized; and (3) it yields a direct estimate of the probability that any two library members overlap. This estimate is a critical tool for the accurate, automatic assembly of overlapping sets of fragments into islands called "contigs." These contigs must subsequently be connected by other methods to provide an ordered set of overlapping fragments covering the entire genome. © 1990 Academic Press, Inc.

1. INTRODUCTION

We consider the problem of constructing ordered covering libraries for relatively large genomic regions such as chromosomes. Such libraries are made by fragmenting the DNA and cloning the fragments in a microbial host. The task is then to extract a subset of these clones that together contain all of the source DNA and to reconstruct the native order the fragments had in the genome.

Our analysis is confined essentially to approaches based on libraries formed by "random" overlapping fragmentation of the DNA. These approaches seek to produce fragments having sufficient overlaps that a continuous path can be constructed by passing through a series of overlapping nearest neighbors. Discovering the original order of the cloned fragments is then dependent on detecting the necessary overlaps. Our analysis is further restricted largely to approaches in which overlap detection is based on "restriction fingerprinting." This involves digesting each cloned fragment to completion with restriction enzymes and measuring the lengths of the fragments produced using electrophoresis.

Our study of this problem was undertaken primarily to support an effort under way at Livermore to construct an ordered cosmid library of human chromosome 19 (Carrano *et al.*, 1989). Previous ordering efforts of this type have been undertaken by others, principally that of Coulson and co-workers (1986) in the nematode *Caenorhabditis elegans* and that of Olson and co-workers (1986) in the yeast *Saccharomyces cerevisiae.*

2. THE MAGNITUDE OF THE PROBLEM

Considerations of strategy for these problems are dominated by the effective size S of the genomic region defined as the ratio of the actual size of the genomic domain being analyzed G to the average size of a cloned fragment L: $S \equiv G/L$. For reasonable coverage efficiency, approaches based on randomly selected clones require that a minimum of $5 \times S$ cloned elements be analyzed, while some strategies require many times this number (Michiels *et al.*, 1987). Thus, even a small human chromosome (ca. 50 Mb), cloned in a cosmid vector (average insert sizes of 40 kb), involves the characterization of at least 7×10^3 clones, and each of the approximately 25×10^6 distinct pairwise combinations of these must be analyzed for possible overlap. Only about 25×10^3 true overlaps (at a five-fold covering) would be expected in this collection, so that a false positive error rate of one in a thousand would produce nearly as many false as true overlaps. A very stringent rejection of false pos-

[1] The U.S. Government's right to retain a nonexclusive royalty-free license in and to the copyright covering this paper, for governmental purposes, is acknowledged.

0888-7543/90 $3.00

itives must therefore be achieved at the inevitable cost of an increased false negative rate.

These limitations become more severe as the effective size of the genome region is increased. The ratio of the number of false positive overlap to the number of true overlaps increases linearly with effective genome size ($\propto S$). Further, even when about $5 \times S$ clones are analyzed, and perfectly unbiased cloning is assumed, a large number of gaps (several hundred in our example) will remain for statistical reasons alone. Inefficient overlap detection can increase the number of elements required for even this relatively limited degree of reconstruction by an order of magnitude or more. Moreover, in most schemes, all elements must be characterized by the same procedure so that they can all be compared in pairwise combinations for indications of overlap.

3. PREDICTING THE RATE OF PROGRESS

The number of clones one must analyze to achieve a given degree of map completion is critically dependent on the sensitivity of the method used to detect overlap (that is, on the average amount by which clones must overlap for the overlap to be detected). This was first shown by Lander and Waterman (1988), who analytically analyzed a probabilistic model of ordering random clonal libraries. Their results relate quantities that characterize the progress achieved to the number of cloned elements analyzed and other parameters. We have extended their analysis somewhat by performing Monte Carlo simulations of the same problem, and begin our discussion by summarizing the main points that result from these two complementary approaches.

The relevant result of Lander and Waterman expresses the expected number of contigs in terms of the number of library elements analyzed and the overlap fraction needed for detection. (A "contig" (Staden, 1980) is a collection of two or more cloned segments each of which is connected to all others by at least one path of pairwise overlapping elements.) In a form that is independent of the size of the domain being mapped, this expression can be written

$$N = Ce^{-C(1-\theta)}(1 - e^{-C(1-\theta)}), \qquad [1]$$

where N is the ratio of the number of contigs found to S, the mapping size of the genome defined above; C is the redundancy of coverage, i.e., the ratio of number of cloned elements analyzed n to S ($C \equiv n/S = nL/G$); and θ is the overlap fraction required for detection. The major assumptions involved in deriving this formula are: (a) all inserts are of the same length; (b) each cloned insert is considered to result from a fresh random sampling from all possible positions in the genome being mapped to select the location of one end of the insert (where all possible locations are equally probable); and (c) the overlap of two inserts is detected if and only if the overlap is larger than a specified fraction of their length. That is, there are no false positives and all false negatives are those whose overlap is less than a specified fraction of the length of the inserts (Lander and Waterman also analyzed a more complex case in which both the insert length and the required overlap fraction were random variables).

The above assumptions (particularly "b" and "c") simplify the real experimental situation in a manner that introduces an optimistic bias. We have attempted to estimate this bias and to investigate other aspects of the experimental situation by performing Monte Carlo simulations of the same problem.

In the present simulations, we begin with a complete "restriction site" model of the chromosome. Assuming, as in the Lander–Waterman model, that such sites are uniformly distributed, we assign both the location of the "cloning" restriction sites (i.e., those involved in producing the cloned segments by partial digestion) and the location of the fingerprinting sites (i.e., those that, by complete digestion, form the fragments whose lengths constitute the restriction fingerprint used to characterize the cloned insert). A restriction fingerprint in this model is a list of lengths that indicate that at least one fragment of that length both was produced by the digest of the cloned segment being fingerprinted and was of a length that fell within the observational range of the electrophoresis method used to measure them. The details of the Monte Carlo simulation calculations are described in the first section of the Appendix.

Perhaps the most important simplifications common to both sets of results are the assumptions that restriction sites are uniformly distributed and that all fragments in the allowed size range are equally clonable. In the simulations presented here, the process of contig construction is also simplified to one of simple assembly based on pairwise determination of overlap. In the simulations, however, overlap determination is based on a statistical assessment of the fingerprints involved not merely on the actual amount of overlap present. As a result, both false positive and false negative overlaps occur in the simulations. More importantly, because the simulations employ fixed restriction sites, the effects of the statistical clustering of these sites persist independent of the level of sampling. Certain regions can therefore have little or no probability either of being represented in the library or of being spanned by members having detectable overlap.

In Fig. 1, "progress curves" that give the number of contigs found as a function of the number of cloned inserts analyzed are presented. Such curves initially rise when the process is dominated by finding overlaps between previously unplaced elements, and then fall

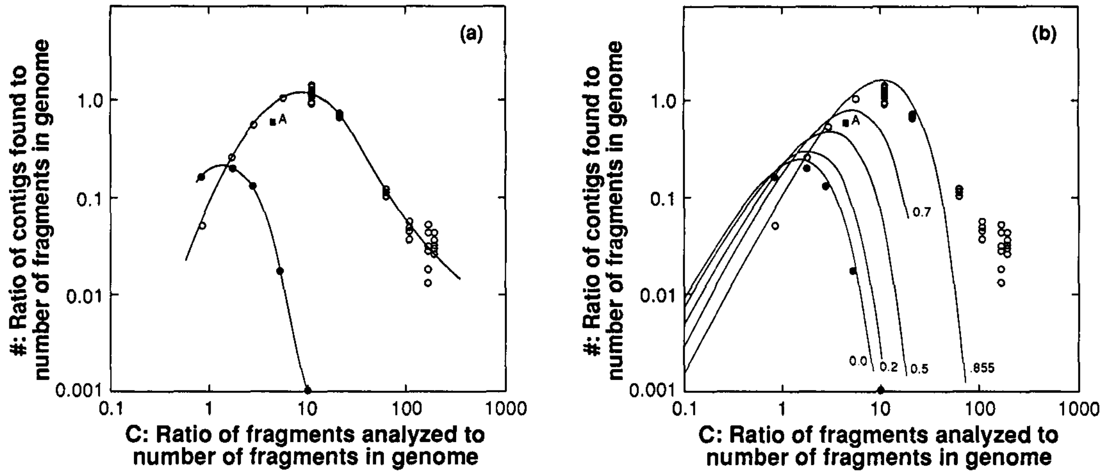

FIG. 1. Ordering progress curves: progress in number of contigs found versus effort in number of cloned fragments analyzed. Both axes are normalized to the size of the genome; ordinate, $\log_{10}$ ((number of contigs found)/S); abscissa, $\log_{10}$ ((number of cloned inserts analyzed)/S); S = (size of the genomic domain in bases)/(average size of cloned inserts in bases). (**a**) The results of our simulations of the yeast experiments of Olson's group (20) (upper curve, open circles), and the same simulations except assuming that all overlaps are detected (lower curve, filled circles). Multiple points on the upper curve with the same abscissa show the results of independent simulations starting with different random number seeds (the points at $C = 10$ represent 11 independent calculations). The single point indicated with the filled square is the state of progress reported by Olson (20). (**b**) The curves obtained using the analytic formula of Lander and Waterman (16) (Eq. [1] in text), with the simulation points from (**a**) superimposed for ease of comparison. Each of the curves corresponds to a different value of the parameter θ, the fraction of overlap assumed to ensure overlap detection; curves for five values are shown: θ = 0.0, 0.2, 0.5, 0.7, and 0.855.

when the process is dominated by finding overlap connections between previously formed contigs.

In Fig. 1a, the results of two simulated progress plots are shown. Both curves are based on a simulated library that was constructed to reflect the properties of the one used by Maynard Olson and his colleagues in their ordering of the yeast genome (Olson *et al.*, 1986; see Fig. 1 legend for details). The curve connecting the solid circles shows the progress obtained when it is assumed that all true overlaps in the library are detected. The curve through the open circles shows the same library analyzed by methods simulating those reported by Olson *et al.* (1986), which include (1) the size distribution of the fingerprint fragments and the accuracy and size limitations of the electrophoretic methods and (2) the statistical criteria used for detecting overlap (except, however, for the imposition of what these authors termed "topological constraints"). In this latter curve, multiple points with the same abscissa correspond to the different outcomes obtained from repeat calculations using different initial settings of the random number generator. They give an indication of the statistical fluctuations of the points in this curve.

In this situation, about 10 times more inserts must be analyzed to achieve the same level of closure as the ideal case and complete closure is not obtained even after over 100 genome equivalents have been analyzed. The dramatic difference between the two curves shown in this graph means that a large fraction of the true overlaps are not being detected by the simulated procedure used here. Only about 15% of the overlaps present in the simulated data were detected; this corresponds roughly to those pairs that overlap by about 85% or more of their length. We see, therefore, that not only are most overlaps being missed, but those missed have moderate or small overlaps, exactly those most valuable in generating order. The single point marked with the solid square in Fig. 1a indicates the progress reported by Olson *et al.* (1986). We emphasize that only-"statistical" limitations are represented in these results, since we have assumed that no cloning bias exists.

These results suggest that it is worthwhile to know how the efficiency of contig construction varies with overlap detection efficiency and whether economical means exist for achieving high overlap detection efficiency.

We address the first of these in Fig. 1b, which shows the results obtained using Eq. [1] (smooth curves) superimposed on the simulation results presented in Fig. 1a (the solid and open circles) to aid comparison. The analytic curves are parameterized by the quantity θ, the fractional amount by which two inserts must overlap for the overlap to be detectable. Curves for five values of θ, 0, 0.2, 0.5, 0.7 and 0.855, are shown. As was emphasized by Lander and Waterman, the curves show clearly that the efficiency of contig assembly exponentially worsens as the amount of overlap required for detection increases.

The general agreement between the analytic and the simulation results, particularly in the limit of perfect overlap detection, is evident. Also evident is the significantly poorer progress achieved in the simulations particularly as the number of cloned elements analyzed increases and as larger amounts of overlap are required for detection. This discrepancy arises primarily because of two consequences of the simulations' use of fixed (albeit randomly distributed) locations of restriction sites. First, because of fluctuations in the density of cloning restriction sites, some locations in the genome will be too sparsely populated with such sites to be spanned, with reasonable probability, by a clonable segment. Most of the discrepancy in this simulation, however, is due to the analogous clustering of the restriction sites used for fingerprinting. Some small regions of the genome are too sparsely populated with these sites for an overlap within them to be detected with sufficient statistical confidence.

The above results emphasize the value in being able to detect relatively small overlaps. A scheme capable of detecting 20% overlap requires analysis of roughly 10-fold fewer elements than one requiring 80% overlap. Further, there is relatively little gain in being able to detect overlaps of less than 20%. The efficiency of contig construction at that level is nearly as high as the efficiency at one in which all overlaps were detected.

The results also allow the value of a change in procedure to be assessed. We should adopt a more "costly" alternative procedure that reduces θ only if the consequent reduction in the total number of clones it was necessary to analyze was worth the increase in cost per clone. The results also make it clear that ordering methods based on random clone selection become exponentially less efficient after covering depths much over fivefold are reached.

The remainder of this paper is devoted to an analysis of the problem of designing an efficient contig assembly strategy based on restriction fingerprinting. Our analysis is based on using the likelihood ratio as the statistic for detecting overlap together with an "information–theoretic" extension of the likelihood ratio formalism to obtain a means for quantitatively assessing and comparing alternative experimental strategies (Kullback, 1968; Good, 1983).

4. INFORMATION PRESENT IN THE FINGERPRINTS FOR DETECTING OVERLAP

4.1. Detecting Overlap from Statistical Data

Our problem is to assess the weight of the evidence presented in a pair of restriction fingerprints for or against the hypothesis that they overlap. This could be treated as a conventional problem of statistical hypothesis testing in which the goal is to develop criteria for accepting or rejecting the hypothesis of overlap.

For example, in the standard Neyman–Pearson approach to this problem, the likelihood ratio is used as a test statistic for an acceptance/rejection criterion that achieves minimum false negative errors for fixed false positive errors (see, for example, Hoel *et al.* 1971; Cox and Hinkley, 1986; Kiefer, 1987). We have adopted a different approach in which the likelihood ratio statistic is used to capture the strength of the evidence for or against overlap (Good, 1983; Kullback, 1968). Having the overlap likelihood ratio for all pairs of fingerprints proves to be useful both for contig assembly and for assessment of the reliability of the results.

One advantage to this choice of statistic for this problem is the relative ease with which it permits complex and nonideal properties of the data to be properly taken into account. Experimental errors and the observed statistical characteristics of the fingerprint-generating process are examples. This use of the likelihood ratio also permits the effects of additional independent experimental evidence bearing on the same question to be easily calculated.

Our likelihood ratio approach at this point is common statistical practice (Cox and Hinkley, 1986). A biological example is the now common use of the lod score in genetic linkage analysis (see, for example, Lander and Botstein, 1986; Conneally and Rivas, 1980; Ott, 1974). The likelihood ratio further leads to an information–theoretic formulation with which the efficiency of different experimental approaches can be measured and compared.

4.2. The Likelihood Ratio Statistic

We must decide between two alternative explanations (that two cloned DNA fragments overlap or not) for an experimental outcome (their restriction fingerprints) where the data are statistical in character and where both alternatives could produce the result. The intuitive notion that explanations should be favored in the proportion that they predict the outcome is given formal expression through the quantity called the likelihood ratio, defined next (see Cox and Hinkley, 1986; Edwards, 1987, presents an extended discussion).

We discuss below (Section 5.1) how a pair of restriction fingerprints characterizing two clones will be abstracted into a "datum" for the purpose of detecting overlap. Here we merely symbolize that datum by the notation $x_{i,j}$, where i and j label the two segments being fingerprinted. The likelihood ratio $L(x_{i,j})$ "in favor of overlap" is the ratio

$$L(x_{i,j}) \equiv \frac{p(x_{i,j}|O)}{p(x_{i,j}|N)}, \qquad [2]$$

where $p(x_{i,j}|O)$ and $p(x_{i,j}|N)$ are, respectively, the probability that the restriction fingerprint pair $x_{i,j}$ would occur given that the two cloned segments either did, O, or did not, N, overlap. (Michiels *et al.*, 1987, used a similar approach in the analysis of genomic ordering strategies based on probing with random sequence oligonucleotides.) A useful experimental procedure will generate data such that $L(x_{i,j})$ is $\gg 1$ for virtually all overlapping pairs and <1 for most nonoverlapping pairs. Our ability to classify overlapping and nonoverlapping pairs will depend on how the distribution of this statistic for nonoverlapping pairs intersects with that for overlapping pairs. We address this question in Section 7 below.

To employ this approach, however, we must be able to calculate the probability distribution function not only for what we might call the "null" hypothesis, i.e., $p(x_{i,j}|N)$, given any pair of fingerprints, but also for that of the non-null hypothesis, $p(x_{i,j}|O)$.

We note that the likelihood ratio is a "sufficient statistic" for distinguishing between the two hypotheses appearing in its definition (Kullback, 1968, pp. 43–45; Cox and Hinkley, 1986, pp. 18–25). Using the likelihood ratio also leads to a way of measuring the hypothesis-testing efficiency of alternative experimental approaches as outlined in the next section.

4.3. *Weighing the Efficiency of Different Experimental Designs*

The log of the likelihood ratio $I(x_{i,j}) \equiv \log(L(x_{i,j}))$ can be interpreted as the information in the observation $x_{i,j}$ in favor of the hypothesis that the two fragments overlap versus the alternative hypothesis that they do not (Kullback, 1968. See also: Basu, 1988; Good, 1983). If, for example, the outcome $x_{i,j}$ is as likely to occur with nonoverlapping clones as with overlapping ones, then $L(x_{i,j}) = 1$, and $I(x_{i,j}) = 0$; i.e., the observation yields no information on the question of overlap. The information measure defined in this way has the expected additive property in that if two statistically independent fingerprinting experiments are performed on the same pair of elements, the information provided by the two experiments considered together is the sum of the information provided by the two individual experiments. (For a discussion of a generalized class of such measures, see Särndal, 1970.)

Further, the expectation value $I(O{:}N;\mathbf{x})$ of the information of the set of observations $\mathbf{x} = \{x_{i,j}\}$ distributed according to the non-null hypothesis is given by

$$I(O{:}N;\mathbf{x}) \equiv \sum_{x_{i,j}} p(x_{i,j}|O)I(x_{i,j}). \qquad [3]$$

This quantity was introduced by Kullback and Leibler (1951) as the average information per observation from the distribution of the hypothesis O for discriminating in favor of the hypothesis O against the hypothesis N. In the expression $I(O{:}N;\mathbf{x})$, $\mathbf{x}$ makes explicit the dependence of average information per observation on the particular experimental method employed. This same expression has been used in other contexts under many different names, principally that of "relative entropy" (Arratia and Gordon, 1989). We refer to it as the "Kullback–Leibler information," or K-L information, produced by the experiment and use it, through its dependence on the experiment used, as a way of measuring the hypothesis-testing efficiency of an experimental procedure (in the Appendix we present a brief discussion of the reasons for giving it this interpretation).

To illustrate one application of this concept, we make use of the additive property of information mentioned above and consider that two statistically independent experimental procedures $\mathbf{x}_1$ and $\mathbf{x}_2$ are performed to detect the overlap of two cloned fragments. The final outcome can be interpreted as resulting from two sequential applications of Eq. [3] (see Eq. [5] below and the discussion in Section 3 in the Appendix), with the result that the likelihood ratio of the total result is the product of the likelihood ratios of the two component experiments:

$$L(\mathbf{x}_1\mathbf{x}_2) = L(\mathbf{x}_1) \times L(\mathbf{x}_2). \qquad [4a]$$

It then follows directly that if n statistically independent repeats of the same procedure are performed, we have

$$I(O{:}N{:}n \times \mathbf{x}) = n \times I(O{:}N;\mathbf{x}). \qquad [4b]$$

From this we can conclude that if an "improved" experimental approach $\mathbf{y}$ is found which yields approximately twice the K-L information in favor of overlap versus nonoverlap as does the original $\mathbf{x}$,

$$I(O{:}N;\mathbf{y}) \approx 2 \times I(O{:}N;\mathbf{x}), \qquad [4c]$$

the same amount of information could alternatively be produced by repeating the first procedure twice (using different restriction fingerprinting enzymes to achieve statistical independence). Which alternative we choose would presumably depend on whether we find it more desirable experimentally to perform the new procedure once per clone or the first twice per clone.

4.4. *Computing Overlap Probabilities*

Given the likelihood ratio for any pair of cloned fragments, the "posterior" probability that they overlap can be calculated using Bayes' theorem (Box and Tiao, 1973; Good, 1983; see also Appendix Section 3). Let $p(O)_{i,j}$ represent the "prior" probability of overlap (the probability that two fragments i, j drawn at random from the collection of clones overlap). As explained in Section 3 in the Appendix, this probability is the same for all pairs and is approximately $2/S$, where S, defined in Section 2 above, is the ratio of the size of the DNA domain being mapped to the average size of the cloned fragments. Given fingerprint data for this pair, i.e., $x_{i,j}$, Bayes' theorem says that the posterior probability that the pair overlap $(p(O|\,x_{i,j}))$ is given by

$$OD(O|\,x_{i,j}) = OD(x_{i,j}) \times L(x_{i,j}), \qquad [5]$$

where $OD(x) \equiv p(x)/(1 - p(x))$ is called the "odds" of the event x (see Section 3 of the Appendix).

In the next section, we show how restriction fingerprint data can be used to compute values for the likelihood ratio statistic.

5. CALCULATING THE LIKELIHOOD RATIO

The restriction fingerprints that are the raw data for this type of problem typically consist of a "signal" curve containing a series of peaks that reflect the presence of DNA fragments separated by length. Such data are produced from autoradiographs by gel scanners (Sulston *et al.*, 1988) or directly in automated fluorescence-based electrophoresis systems (Carrano *et al.*, 1989). We ignore for the present the important problems of feature abstraction and fragment length determination that such data present. We are still left with a range of possibilities about how the data should be characterized for the purpose of computing likelihood ratios.

5.1. *The Data Abstraction*

At one extreme, the attempt could be made to frame the competing hypotheses of overlap and nonoverlap so that fairly detailed features of such data, including peak area and peak shape, for example, were predicted. However, such features generally show poor reproducibility, and taking them into account entails very considerable computational costs. At the other extreme, we might specify simply the number of bands that appear to be in common between the two fingerprints (as done by Sulston *et al.*, 1988), or count the number of apparently common bands together with the total number of bands as was done by Olson *et al.* (1986).

There is also the question of whether the hypothesis of overlap is framed in terms of a measure of the extent of agreement and disagreement between two fingerprints or as a specific partitioning of the fragments in the two fingerprints into those that are presumed to be shared between the two DNA segments being fingerprinted and those not shared.

In the approach described here we adopt the former choice, and we have taken a middle course as to how detailed a description of the pattern of agreement and disagreement between fingerprints is made. Our reasons for this are as follows.

(1) The dependence of the probability of generating a restriction fragment on the fragment's length can lead to an order of magnitude difference in the probability of finding two bands in common between the shortest and the longest fragments in a fingerprint.

(2) The probability of obtaining a fragment of some length depends on the length of the piece of DNA being digested; even with cosmid cloning, the difference in insert lengths can be as large as 30%.

(3) In a comparison of two restriction fingerprint patterns, at least three types of events occur that are informative about overlap: positions at which there are common bands, positions at which neither gel has a band, and positions at which one of the gels has a band but the other does not. Only the latter events provide information disfavoring the overlap hypothesis. Both "band agreements" and "blank agreements" contribute information favoring the overlap hypothesis.

(4) The underlying probabilities on which all of the relevant statistical inferences depend are those for finding a band corresponding to any specific fragment length in an individual fingerprint. While various a priori assumptions (e.g., randomly distributed cutting sites) about these probabilities and their dependence on fragment length can be made, experience indicates that the frequency distributions in real biological samples deviate significantly from such expectations. Moreover, certain specific fragment lengths may occur at highly anomalous frequencies (i.e., those produced by repetitive elements) and may have correlated frequencies of occurrence. For these reasons, it is worthwhile to be able to incorporate the actual measured frequencies for bands at all different fragment lengths from the total data set, along with whatever significant interband correlation frequencies are found.

(5) As the number of bands in a gel pattern increases, so does the probability that two fragments are present in a single band. This effect reduces the significance of observing common bands between fingerprints. On the other hand, reducing the band density to the point where this effect can be neglected makes inefficient use of the information-producing capacity of the system (see Section 6.3). An approach that can properly take the probability of coincident fragments into account is therefore needed.

(6) Finally, more detailed experimental information characterizing the digests could be usefully taken into account in computing the likelihood ratio for any given pair of fingerprints. Most important examples are the experimentally determined performance parameters characterizing the quality of the data: the repeatability of band position estimates and false positive and false negative frequencies in band detection.

In our experimental work we have adopted a data abstraction and a related method of computing the likelihood ratio that takes into account, at least in approximation, all of the above issues. For our present purposes, however, we employ a relatively simple data abstraction, called here the "trinomial model," which addresses only a few of the issues mentioned above.

5.2. *The Trinomial Model*

Suppose, first, that fragment lengths can be reproducibly resolved to one base (a fairly good approximation for polyacrylamide gels in the range below 300 bp), second, that the probability that a fingerprint digest will produce a fragment of a given length is the same for all lengths, and, third, that all cloned inserts have the same length. Thus, in this model, each pair of fingerprints can be characterized in terms of just three numbers: n, the number of places on the gel where both fingerprints have bands, m, the number of places on the gel where separate bands could be resolved and detected but at which neither fingerprint has a band, and l, the number of places on the gel where one fingerprint has a band but the other does not; $l = M - m - n$, where M is the total number of bands that could be resolved in the electrophoretic analysis. Put another way, at each band position only one of three possible events can occur: a band either is present in both fingerprints or is present in neither, or is present in only one (this is diagrammed in Fig. 2). Thus, the outcome at all gel positions corresponds to tossing a three-sided coin, with the same probabilities at all positions (fragment lengths). Therefore (see Section 2 in the Appendix), entire coincidence patterns correspond to M such samplings and their probabilities are distributed according to the trinomial distribution.

In subsequent calculations, it is usually assumed that the electrophoresis system is capable of reproducibly resolving about 400 separate fragment lengths, i.e., $M = 400$ (this seems approximately what fluorescence-based polyacrylamide systems can deliver: Carrano *et al.,* 1989). In this case it is feasible to compute results for all possible outcomes. An outcome is a choice for the three numbers n, m, and l consistent with $n + m + l = M$ (see above) of which there are, with $M = 400$, only about 80,000 ($M(M - 1)/2$).

Computing the probability of a coincidence pattern under the assumption of overlap requires an indirect step because the probabilities for the band-specific coincidence events (a band in common, for instance) depend strongly on the amount by which two fragments overlap. The desired probability can, however, be directly expressed if an arbitrary but specific amount of overlap is assumed. The probability of the evidence given that there is any overlap at all can then be written as a weighted sum over the overlap-specific probabilities. The details of the computation of the likelihood ratio in the trinomial model are presented in the Appendix.

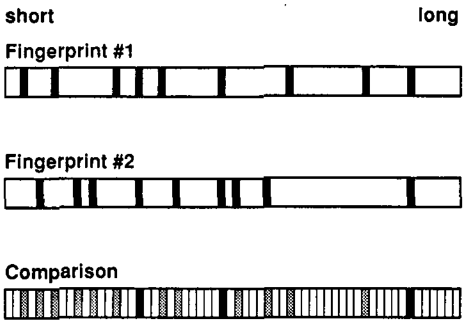

FIG. 2. Abstraction of fingerprint data and coincidence patterns in the trinomial model. The fingerprint data are first size-calibrated to assign an equivalent fragment size in terms of base pairs to each position in the fingerprint pattern. The presence or absence of a band at each resolvable position (assumed to be one such for each different number of base pairs between the smallest and largest fragments detected) is then noted in each fingerprint. In the trinomial model approximation, we then compare two fingerprints by determining three numbers: n the number of positions at which they both have a band (band agreements, ■), m the number of positions at which neither has a band (blank agreements, □), and l the number of positions at which one has a band and the other does not (disagreements, ■).

6. RESULTS USING THE TRINOMIAL MODEL

We next describe the results of calculations obtained using the trinomial model. The first point addressed is how overlapping and nonoverlapping events are distinguished by the statistical distribution of their likelihood ratio statistic.

6.1. *Distinguishing Overlapping from Nonoverlapping Pairs of Clones*

The problem of contig assembly is strongly affected by the statistical properties of overlap detection. Insight into this connecting is gained by determining how the probability of obtaining a specific value of the test statistic L depends on whether overlapping or nonoverlapping clones are considered. We expect, of course, that L values very much larger than 1 should result much more frequently from overlapping pairs than from nonoverlapping ones, while the opposite should be true for L values less than 1. Formally, the transformation from a datum x to the statistic $L(x)$ for that datum, $x \twoheadrightarrow L(x)$, defines a new random variable, L, whose distribution under the two contending hypotheses we now examine. We recall that in the trinomial model a datum is a triplet $x = (n, m, l)$, where $n + m + l = M$.

A representative example is shown in Figs. 3a and 3b. In Fig. 3a, values of $p(L(x)|O)$ and $p(L(x)|N)$ are plotted against their corresponding log L values; i.e., two points are plotted $(p(L(x)|O), \log L(x))$ and $(p(L(x)|N), \log L(x))$ for each of the possible coincidence patterns $x = (n, m, l)$. These plots have a seemingly multivalued, space-filling character because events with very close L values can have quite different probabilities, ranging up to a maximum that varies smoothly with log L.

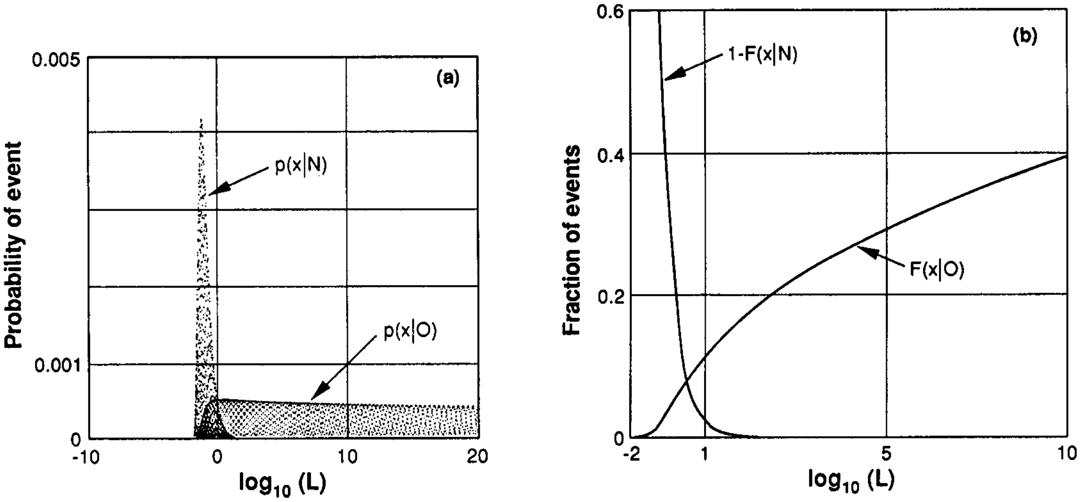

FIG. 3. Distribution of overlapping and nonoverlapping events for the likelihood ratio statistic. (**a**) The probability of getting an outcome x (a specific (n, m, l) triplet) assuming either overlap or no overlap ($p(x|O)$ and $p(x|N)$ are the probabilities of the outcome $x = (n, m, l)$, given, respectively, overlap and nonoverlap) is plotted as a function of the log of the likelihood ratio L associated with that outcome. The nonoverlapping events form the sharply peaked distribution on the left and the overlapping events form the nearly uniform distribution extending to the right. These plots have a space-filling character because different (n, m, l) outcomes with nearly equal L values can have sharply differing probabilities, ranging from zero to a maximum value that defines the smoothly varying envelope of these plots. Under the assumptions made, the distribution for overlapping events continues as depicted up to nearly 10^{100}. (**b**) The same data are represented using cumulative distributions. For overlapping events the cumulative distribution itself is plotted ($F(x|O) \equiv p(\log L < x|O)$, which is the probability that the log likelihood is less than the value of the abscissa; for nonoverlapping events one minus the cumulative corresponding distribution, $1 - F(x|N)$, is plotted. The calculations are based on the trinomial model (see text and the Appendix), assuming 400 resolvable events in each fingerprint and a probability of 0.77 that any given position in a single restriction fingerprint will be empty.

We see in Fig. 3a that the probability of obtaining a particular likelihood ratio from nonoverlapping events peaks very sharply between $\log L = -2$ and $\log L = 0$ and drops off quickly for values above 0. In sharp contrast, the overlapping events have a nearly flat distribution in $\log L$ from roughly -1 to above $+20$ and only a very small fraction of this distribution overlaps visibly with the nonoverlapping events.

In Fig. 3b cumulative plots for these same data are shown to give a clearer perception of the magnitude of the overlap between the two distributions. We show the cumulative distribution, $F(x|O) = p(\log L < x|O)$, for overlapping events, and one minus the cumulative distribution, $1 - F(x|N) = p(\log L > x|N)$, for nonoverlapping ones. We see that almost 90% of the overlapping events have $\log L$ values above 1, while the same is true of only about 3% of the nonoverlapping events. However, if false positive (classifying nonoverlapping events as overlapping) errors must be held below 1 in 10^6 pairwise tests, then, in these model calculations, events with $\log L$ values below 4 must be rejected. Using this critical value would nevertheless correctly identify over 74% of the overlapping events for a false negative rate of less than 27%. One in a million false positive errors would imply about 30 false overlap determinations in ordering 7.5K elements, although we argue in Section 8 that a contig assembly strategy exists that can detect and exclude the majority (about 80% in a 5X library) of these errors.

The favorable tradeoff between false positives and false negatives degrades rapidly as lower L values are considered. To reduce the false negative rate by only 3.5%, we would have to accept about 13-fold more false positives. On the other hand, almost all overlapping pairs have overwhelming likelihood ratios; over 60% of such events have L values above 10^{10} and, in a problem where $S = 1500$, yield posterior odds in favor of overlap above 10^7 to 1.

6.2. *Detection Probability as a Function of Overlap*

Some idea of what this means in terms of overlap detection efficiency (expressed in the Lander–Waterman terms of the minimum overlap required for detection) is obtained by calculating the L values for the most probable coincidence state (n, m, l) as a function of the amount of overlap. This plot is shown in Fig. 4. It reveals an extremely steep rise in L with overlap; the most probable L value for events with an overlap fraction of about 27.5% is 10^5. We might then expect that this model system would function as a good overlap detector for overlaps at or above about 25%.

6.3. *The Optimal Density of Bands*

To illustrate the use of the K-L information measure in optimizing experimental design, we consider the

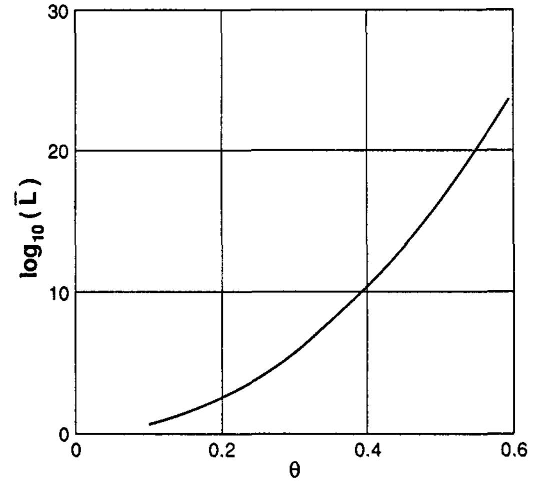

FIG. 4. $\bar{L}$ as function of the amount of overlap. The most probable value of the likelihood ratio defined for a specific fractional overlap (i.e., the likelihood ratio defined with respect to overlap by a specific fractional amount θ is evaluated for the most probable event (n, m, l) given that overlap; see Section 2 of the Appendix) is plotted for overlap as a function of θ.

question of the optimal number of fragments in the fingerprints. We seek the optimum choice between the extremes of too many fragments (with degraded significance of fragment agreements between fingerprints) and too few (with increased risk of having no informative fragment in the region of overlap). This can be found by evaluating the expression given in Eq. [5] for the K-L information for different choices of the probability that the restriction digest, labeling, and electrophoretic procedures produce a recognizable fragment of any specific length. The results of such a calculation, within the assumptions of the trinomial model, are shown in Fig. 5. The conclusions can be summarized as follows. The optimal density of cuts is that which yields, on average, a band in about 30–35% of the resolvable segment of the electrophoresis run. This corresponds to about 120 to 140 bands in a system that can resolve 400 different fragment lengths. However, informativeness turns out to depend rather weakly on the density of fragments so that $I(X)$ has dropped by less than 20% if as few as 10% of the resolvable slots are filled. And although the rate of loss from that point on becomes increasingly steep, one must drop to as few as 2.5% of the resolvable slots filled before $I(X)$ has decreased to $0.5I(X)_{\max}$.

More rigorous calculations (not presented) show that most of the complicating aspects of real fingerprint data, such as the exponential dependence on fragment length of the probability of getting a fragment, modify these conclusions only slightly, placing the optimal density somewhat lower. However, significantly lower fragment densities (reduced by at least 40 to 50%) are indicated if realistic estimates of the errors in fragment ascertainment and fragment length determination are taken into account. Because the magnitudes of such errors depend sensitively on the experimental methods used, precise general conclusions cannot be given.

Another perspective on the significance of band density is shown by comparing the cumulative probability distributions for overlapping events assuming two well-separated choices (23 and 5% filled) for the density of bands; these results are shown respectively in Fig. 6, curves A and C. Whereas a cutoff of $L \geqslant 10^5$ would lead to a false negative rate of 30% for the higher band density, it would produce a false negative rate of over 40% for the lower density. Because all extents of overlap are equally probable, this implies that the more dense design would correspond, in the Lander–Waterman model, to a 30% overlap detector, while the less dense design would correspond to a 40% overlap detector.

6.4. *The Contribution of Blank Agreements*

In characterizing the degree of similarity between fingerprints we have counted "blank agreements," i.e., the number of locations at which neither fingerprint had a band, as well as the number of band agreements and band disagreements. How much additional information is brought in by considering the blank agreements? In Section 2 of the Appendix, we outline how the formulas presented there can be used to estimate this contribution. We find that at "optimal" band frequencies, the blank agreements contribute about as

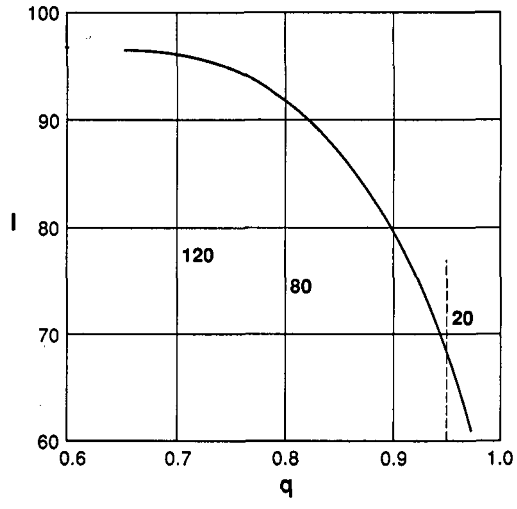

FIG. 5. Informativeness versus band density. The informativeness I of the procedure by the Kullback–Leibler measure (the expectation value of the log likelihood ratio in the overlapping distribution; see Eq. [3] in text) is plotted as a function of the average fraction of empty positions in the fingerprints q. Calculations are made using the trinomial model assuming 400 resolvable events in each fingerprint.

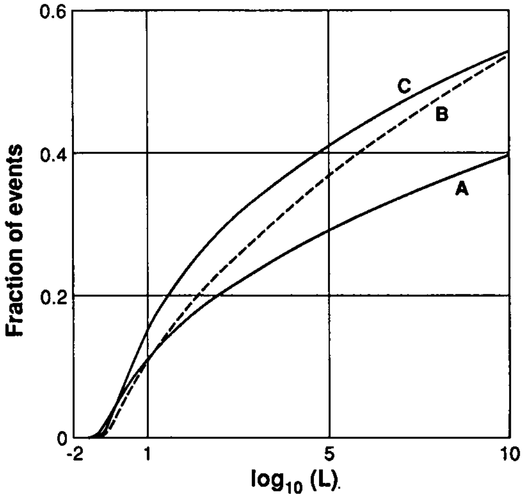

FIG. 6. The effects of reduced resolution and suboptimal numbers of bands. Cumulative probability distributions for overlapping events, as defined in the legend to Fig. 3b, are plotted under the assumption of (**A**) 400 resolution bins, 23% of which are filled on the average; (**B**) 200 resolution bins, 36% of which are filled (assumes that the same number of fragments as in A is distributed among half the number of bins); and (**C**) 400 resolution bins, 5% of which are filled.

much of the favoring overlap information (40 to 60%) for discriminating overlaps (those overlapping by 20 to 30%) as do the band agreements.

6.5. *The Effect of Electrophoresis Resolution*

Perhaps the single most important parameter in the design of a system for overlap detection is the resolving power of the electrophoresis system used. This issue reduces to two questions: (1) how reproducible is fragment length determination, and how closely can it approximate the ideal of single-base definition; (2) how many different fragments can be resolved in a single electrophoresis run? These two issues can trade off against each other to some extent and one advantage of the likelihood ratio statistic is that it allows the direct incorporation of the actual performance of the experimental system in both these respects. A detailed analysis of this issue is rather complex and will not be undertaken here. However, a rough upper limit on the consequences of reduced resolution can be obtained within the confines of the trinomial model by approximating the effects as simply reducing the number or resolvable bins in a fingerprint. The informativeness, expressed in terms of the K-L information measure ($I(O{:}N;\mathrm{X})$), can be calculated as a function of the number of resolution bins M using, as before, Eq. [6]. Such calculations show, as expected, that the informativeness of the procedure, per electrophoresis lane, is roughly proportional to the number of resolvable elements that can be distinguished in that lane. Comparing curves A and B in Fig. 6 shows the effects of resolution loss modeled in this way on the cumulative probability distributions for overlapping events.

In practice, however, the consequences of deviations from perfect fragment length determination are significantly more troublesome. This is largely the result of fragment length imprecision compounded by the imperfect correlation between fragment length and electrophoretic migration velocity. Single-stranded fragments produced in high-resolution denaturing gels migrate in a sequence-dependent manner so that the measured lengths of two fragments with the same length often differ significantly, occasionally by more than 1%. This has the consequence that fragments of different length can appear to have the same length no matter how accurate the measurement. It also implies that the best estimate of a fragment's length may involve fractions of a nucleotide, e.g., 100.6 ± 0.25 nt. As a result, two fragments whose real lengths differ by, for example, 2 nt will often have apparent lengths that differ by less than 1 nt and confidence intervals with substantial overlap. Further, since band agreement or disagreement must be judged by how far separated two nearby bands appear to be, there is an unavoidable tradeoff between missing true band agreements and including false agreements. As a result, the information contribution is reduced for all classes of comparison events (band agreements, blank agreement, and band disagreement). In our experience, these and other error effects make a large contribution to the "true" value of the likelihood ratio and it is correspondingly important that they be taken into account in the calculation of the overlap statistic. A treatment of this issue, however, is beyond the scope of the present paper.

7. ALTERNATIVE FINGERPRINTING SCHEMES

Quite a few of the alternative methods of overlap detection that have been proposed, or are being used, are based on direct hybridization (Evans and Lewis, 1989), recombination, oligo probing (Lehrach, 1987), or other methods of restriction fingerprinting such as partial digestion restriction mapping (Kohara, 1987).

We briefly consider partial digestion strategies. These are attractive because they reveal the order of the digest fragments produced and this greatly increases the information on the question of overlap provided by a given number of fragments. However, this gain is set off against a number of disadvantages. One is the difficulty in doing partial digestions reliably on many different DNA samples without laborious titration and testing trials. In addition, if one attempts to detect most of the true overlaps, partial maps that extend from the end of each insert well toward the middle would be required. Otherwise, very large overlaps would go undetected. Fingerprinting such digests would have to be done on agarose with the consequence that the fragments generated be both few and large. This has several negative consequences. First, the probability would be increased that the most important overlaps, i.e., small ones, would be missed because they would not contain enough partial digest sites in their region of overlap. Also, the accuracy of fragment length determination is significantly degraded (1) because length determination is relatively poor in agarose and (2) the lengths that must be compared between digests are the small differences between large, poorly determined numbers. Our current estimate of the magnitude of these factors and their consequence for the amount of ordering information achievable by partial digest methods, while preliminary, have led us in our own efforts to prefer the limit digest approach, at least for a first-pass processing of a library of cosmid clones.

Finally, suppose we find an alternative experimental approach that could deliver near perfect overlap detection. Under what conditions should we adopt it? On the basis of the progress curves in Fig. 1, we could achieve the same degree of contig closure with this alternative by analyzing only about half the number of clones as the 30% overlap detecting method. If, however, the "better" method was more than twice as costly (in whatever sense matters to us) to perform as the original 30% method for each clone analyzed, we would not be advised to adopt it. Such consideration argues against some alternative fingerprinting methods although they offer much greater intrinsic detection efficiency.

8. CONTIG ASSEMBLY

For genomic domains thousands of times longer than the size of the average cloned fragment, the problem of contig assembly is, in principle at least, daunting. In the "small chromosome" example considered above, one would be faced with "ordering" at least 7000 to 8000 elements, which amounts to seeking the best solution out of, for example, $7000! \approx 10^{27K}$ alternatives. Even with a good way to define the quality of a candidate solution, searching all alternatives is clearly impossible computationally. How much easier the real problem is than this worst case depends mainly on the frequency with which contending alternative placements must be tested and decided between. Of course, smaller overlaps are most likely to be ambiguously indicated in the data, while being, at the same time, the most valuable to detect.

While this issue is not discussed in detail here, we want to emphasize two points. First, having the ability to rank all possible overlaps as to their probability of overlap is very useful for automatic contig assembly. As will be discussed in detail in a separate publication, building contigs by assembling the pieces in decreasing order of confidence of overlap aids substantially in avoiding errors, reducing the number of gaps, and permitting order within contigs to be defined automatically. This is because this method permits information gained from the placement of the higher confidence overlaps to be used in avoiding misplacements of the lower confidence ones. Simulation studies indicate that highly reliable contig assembly can be achieved with this approach, and within these contigs, near-minimal "spanning subsets" that allow the end elements of the contig to be identified with high confidence can be accurately defined. The latter fact is of practical importance because it allows subsequent computational and experimental investigation to focus on the end elements and their potential overlaps. Moreover, the confidence ranking of all possible overlaps between these end clones, or involving end elements and isolated clones, ranks these clones in the order in which any further experimental characterization would be most fruitfully accomplished. We note that contig assembly can be viewed as a problem in interval graphs for which it has been shown that a solution can be found in linear time when overlap information is perfect (Waterman and Griggs, 1986).

Second, the very sharply peaked nature of the distribution of L values among nonoverlapping pairs of clones (Fig. 3) has the consequence that only a small fraction of all candidate overlaps, those whose L values fall in a relatively narrow range, present the possibility of ambiguous but resolvable placements, especially when placement is carried out in order of decreasing confidence. Thus, further computational and experimental effort can be focused on these few cases.

Finally, we note two significant issues neglected in the treatment of overlap detection and its relation to contig assembly presented here. The first concerns the desirability of having a statistical measure of the confidence to be attached to a complete contig (rather than just that of its constituent pairwise overlaps). A derivative point is that of giving the proper statistical weight to the evidence concerning a particular overlap contributed by all other members of the contig that cover the overlap segment. These issues also relate to the second issue neglected in our treatment. A hypothesized contig structure implies a partitioning of the fragments found in its members and it is desirable to make proper statistical use of the consistency requirements that result from this partitioning. That is, a particular contig hypothesis may be viewed as a sequence of contiguous regions each of which is covered by a different subset of the contig membership. All of the fragments found in the contig's members are then assigned by the hypothesis to exactly one of these regions (see Olson *et al.*, 1986). This assignment supplies additional information that bears on the probability that a given contig structure is correct.

APPENDIX

1. *Methods Used in the Monte Carlo Simulations*

The stochastic simulation calculations were carried out as follows. Assuming a specific genome size, a library of potentially overlapping "inserts" of this genome was produced (a list of pairs of numbers giving the order in the genome of the first and last base in the insert). Potential digest sites were assigned randomly along the genome at a specified probability per base pair. Inserts based on these sites were generated, simulating a partial digest, by sequentially choosing (with a predetermined probability) which of these sites would be "cut"; two consecutive cuts define the ends of a digest fragment. This was repeated in serial passages over the genome until an insert collection of a specified number was obtained (in the simulations presented here about 7.5K elements were selected, corresponding to a multiplicity or coverage of roughly 5X).

A collection of randomly terminated, potentially overlapping segments with a specified average length is thus obtained. The cloning step was then modeled by picking randomly, from the partial digest collection, members whose size fell within a predetermined range of clonable sizes until a specified number was selected. This method allows for repeat selections of the same element, and was adjusted, in the present simulations, to reflect the frequency of duplicate selections (approximately 20% at a 5X coverage) reported by Olson *et al.* (1986), i.e., approximately one chance in five for a coverage of 5X.

Similarly, cut points for the restriction enzyme sites corresponding in frequency to the restriction enzyme(s) used for the fingerprinting step were randomly assigned along the genome. The "limit digest" (i.e., complete not partial) fingerprint fragment lengths were then determined, as the segments between consecutive restriction sites, for each member of the cloned library. Those whose size fell within the range assumed to be observable by the electrophoretic method being used are recorded as the restriction fingerprint for that cloned segment.

Statistical criteria designed to detect overlap were subsequently applied to pairs of such fingerprint patterns, and the true existence and extent of overlap, if any, recorded for each pair. Finally, in the present simulations, contigs were assembled simply by placing together all elements connected by "statistically certain" overlap (i.e., whose overlap likelihood ratio was greater than 10^5).

2. *Likelihood Ratio Calculations based on the Trinomial Model*

In the trinomial model we assume that the probability of a band occurring is independent of the position on the gel, and that both segments being analyzed have the same number of fragments ν (the latter assumption is made for convenience in the derivations and calculations). The model is based on the assumption that individual fragments lengths can be reproducibly resolved to the base pair and that two such fingerprints can be aligned and compared to determine how many of the locations have each of three possible outcomes:

(a) both have bands, $\equiv a$, the number of which is called n,

(b) neither has bands, $\equiv b$, the number of which is called m,

(c) only one has a band, $\equiv c$, the number of which is called l,

where the sum of these numbers is equal to the number of resolvable locations in the fingerprint, $M = n + m + l$, and a possible experimental outcome x is specified by a triplet of integers $x = (n, m, l)$. The trinomial distribution gives the probability of obtaining such an outcome under any assumption concerning overlap Y as

$$p(x/Y) = p(n, m, l|Y) \equiv \binom{M}{n, m, l} p(a|Y)^n p(b|Y)^m p(c|Y)^l,$$

where

$$\binom{M}{n, m, l} \equiv \frac{M!}{n!m!l!}, \quad \text{and} \quad M = n + m + l.$$

The quantities $p(e|Y)$ for $e = a$, b, or c are the individual-band comparison event probabilities to be calculated below, and the assumption concerning overlap can be overlap by any amount ($Y = O$), overlap by a specific amount θ ($Y = O_\theta$), or no overlap ($Y = N$).

The likelihood ratio in this notation is

$$L(x) = L(n, m, l) = \frac{p(n, m, l|O)}{p(n, m, l|N)}.$$

And, as noted in the text, the probability $p(n, m, l|O)$ can be expanded into the sum over all possible extents of overlap θ (since they form an exhaustive and mutually exclusive set) and written as

$$p(n, m, l|O) = \sum_\theta p(n, m, l|O_\theta)p(O_\theta|O).$$

We make the approximation that θ is measured in number of fragments and not in bases; i.e., $\theta = 1/\nu, 2/\nu, \cdots, \nu/\nu = 1$, where ν is the expected number of fragments in the digest.

We further make the approximation that cloned inserts of the same length have the same number of fragments. In this case, the probability that the inserts overlap by a specific amount, given that they overlap at all, is independent of the amount of overlap so that

$$p(O_\theta | O) = 1/\nu,$$

and we obtain

$$p(n, m, l | O) = \frac{1}{\nu} \sum_\theta p(n, m, l | O_\theta).$$

From this equation it follows directly that the L for arbitrary overlap, $L(x, O)$, is given by the corresponding sum of the L values for the specific amounts of overlap $L(x, O_\theta)$,

$$L(x, O) = 1/\nu \sum_\theta L(x, O_\theta).$$

The task is now to define the individual comparison event probabilities $p(e | X)$, where e stands for any one of the three possible band comparison events (i.e., a, b, or c), and X stands for one of the hypotheses (e.g., either O_θ or N). Define

$q \equiv$ probability that a gel slot is blank.

It follows easily that, for the case of no overlap

$$p(a|N) = (1 - q)^2 = 1 - 2q + q^2$$
$$p(b|N) = q^2$$
$$p(c|N) = 2q\,(1 - q).$$

To compute the probabilities assuming overlap, q, the probability of not having a fragment of a given length in the digest, must be related to the length of DNA being digested. Since we have assumed that the lengths of the fragments produced in a given digest are uncorrelated, digesting a stretch of DNA into ν fragments can be regarded as approximately equivalent to sampling ν times from the fragment length distribution. Because of the finite length of the DNA being digested, one fragment in the collection will be shorter than it would be under this assumption. If g is the probability of not getting a fragment of a given length in a single sampling, then one can write

$$q(\nu) = g^\nu.$$

We then assume that the two cloned inserts being compared, both of which are ν fragments long, overlap by some integral number of fragments ν_o. The comparison events whose probabilities are desired can then arise either from sampling from the two independent, nonoverlapping domains (each being $\nu - \nu_o$ fragments long) or from sampling from the single overlapping domain ν_o fragments long. Thus, the probability of getting a blank agreement assuming overlap by ν_o fragments can be written as the probability of not getting the fragment from either of the two nonoverlapping regions times the probability of not getting it from the single overlapping region,

$$p(b | O_\theta) = q(\nu - \nu_o)^2 \times q(\nu_o) = q^2 q^{-\theta}$$

(where $q \equiv q(\nu)$, and $\theta = \nu_o/\nu$). Similarly, in the case of mismatch, the probability of getting a fragment from the nonoverlapping part of either of the two inserts while having no matching fragment from the entire length of the other insert is

$$p(c | O_\theta) = 2q(\nu) \times (1 - q(\nu - \nu_o)) = 2\,(1 - qq^{-\theta})q.$$

Finally, the probability of getting a band agreement assuming overlap by θ can be obtained by invoking the conservation equation

$$p(a | O_\theta) + p(b | O_\theta) + p(c | O_\theta) = 1.$$

The probability of obtaining the outcome (n, m, l) for a specific amount of overlap θ can therefore be expressed in terms of these quantities as

$$p(n, l, m | O_\theta) = \binom{M}{n, m, l} p(a | O_\theta)^n p(b | O_\theta)^m p(c | O_\theta)^l.$$

The likelihood ratio with respect to the hypothesis of overlap by θ (versus no overlap) is therefore

$$L(n, m, l; O_\theta) \equiv \frac{p(n, m, l; O_\theta)}{p(n, m, l; N)} = L(\theta)_a^n L(\theta)_b^m L(\theta)_c^l,$$

where

$$L(\theta)_a \equiv \frac{p(a | O_\theta)}{p(a|N)},$$

$$\mathrm{L}(\theta)_b \equiv \frac{p(b | O_\theta)}{p(b|N)}, \quad L(\theta)_c \equiv \frac{p(c | O_\theta)}{p(c|N)}.$$

It is useful to calculate the most probable value of this likelihood ratio among the events that overlap by a specific amount Θ. This quantity, which we denote $\bar{L}(\Theta)$, is by definition $L(\bar{n}(\Theta), \bar{m}(\Theta), \bar{l}(\Theta) | O_\theta)$, the value of the likelihood ratio for the most probable outcome $(\bar{n}(\Theta), \bar{m}(\Theta), \bar{l}(\Theta))$, given a specific amount of overlap Θ (where Θ need not be the same as the θ referred to in the definition of the likelihood ratio). We can approximate this quantity by noting that the function $p(n, m, l | O_\theta)$, the probability of the outcome (n, m, l) given θ, is approximately maximal at the values of n, m, and l given by $\bar{n}(\theta) = Mp(a|O_\theta)$, $\bar{m}(\theta) = Mp(b|O_\theta)$, and $\bar{l}(\theta) = Mp(c|O_\theta)$ (the approximation involves assuming the crude form of the Stirling approximation for the logs of the factorials of n, m, and l, i.e., $\log(n!) = n\log(n)$, etc.). It follows that

$$\bar{L}(\Theta) \approx (L(\Theta)_s)^M,$$

where

$$L(\Theta)_s \equiv L(\theta)_a^{p(a|O_\Theta)}\, L(\theta)_b^{p(b|O_\Theta)}\, L(\theta)_c^{p(c|O_\Theta)}.$$

We can, for example, use this expression to estimate the relative amounts of information in favor of overlap contributed by band agreements and blank agreements, respectively. By taking the log of the equation for $\bar{L}(\Theta)$, we see that $Mp(a|O_\Theta)\log(L(\theta)_a)$ is the information in favor of overlap brought in by the band agreements in the "most probable" event involving overlap by Θ, and $Mp(b|O_\Theta)\log(L(\theta)_b)$ is the same for blank agreements. Thus, for example, assuming $q = 0.8$, and $\Theta = 0.2$, we can determine that in such events, almost half of the total overlap favoring information in these fingerprints is due to the blank agreements.

3. *Posterior Overlap Probabilities and the Likelihood Ratio Statistic*

The connection between the posterior probabilities of overlap and the likelihood ratio is outlined below.

Let $p(O)_{i,j}$ be the prior probability that two cloned elements (i, j) overlap, and $p(O|x_{i,j})$ be the posterior probability given the data $x_{i,j}$. In our case $x_{i,j}$ stands for the pair of restriction fingerprints i and j.

The relationship between the posterior and prior probabilities involves the likelihood ratio and is called the Bayes rule (Box and Tiao, 1973), which is a consequence of the definition of conditional probabilities,

$$\frac{p(O|x_{i,j})}{p(N|x_{i,j})} = \frac{p(O \,\&\, x_{i,j})}{p(N \,\&\, x_{i,j})} = \frac{p(x_{i,j}|O)p(O)}{p(x_{i,j}|N)p(N)},$$

which we can rewrite as

$$OD(O|x_{i,j}) = OD(O) \times L(x_{i,j}),$$

where

$$L(x_{i,j}) \equiv \frac{p(x_{i,j}|O)}{p(x_{i,j}|N)}$$

is the likelihood ratio and where $OD(A) \equiv p(A)/(1 - p(A))$ defines the "odds" of an event A whose probability is $p(A)$ (Section 4.4). The above equation is the form of the Bayes rule applicable to our problem.

To a good approximation, the prior overlap probabilities for two otherwise uncharacterized, randomly chosen elements are all the same and roughly equal to $2/S$, where S is the "effective" genome size defined above (Section 2): $p(O) \approx 2/S$ so that $OD(O) \approx 2/S$. That is, with a chosen segment of L bases in a genome of size G bases, the chance that a second such segment will overlap is the chance that its left end falls either within the L bases of the first segment or within the nearest $L - 1$ bases to the left of the first segment. The probability is thus $=(2L - 1)/G \approx 2/S$. The probability of nonoverlap is $1 - 2/S \approx 1$.

Therefore if we want the posterior odds for some overlapping pair to be >0.9 (an odds of 9), then we need a likelihood ratio of approximately $9 \times (S/2) \approx 5400$. But, as argued above, the real issue for the present problem is the avoidance of false positives, and therefore the distribution of $L(x)$ values for nonoverlapping pairs. As we will see later, this, in general, requires that we generate substantially larger L values for the overlapping pairs we wish to detect.

4. *The Kullback-Leibler Distance and Experimental Efficiency*

The K-L information (see Eq. [3]) can be viewed as a measure of the distance between two probability distributions and as a measure of the difficulty distinguishing between the distributions, or equivalently of the two hypotheses they represent, using the specified procedure. It has also been characterized, in the terminology of our problem, as the information per observation in the experiment for distinguishing in favor of overlap as against nonoverlap, and as the "relative entropy" in the non-null distribution with respect to the null distribution (Kullback, 1968).

In the text, we have used the K-L information measure, through its dependence on the experimental procedure used, as a linear measure of the hypothesis-discriminating efficiency of alternative experimental methods. Our arguments supposed, for example, that two different experimental methods for answering the same question are of equal efficiency if the K-L distances between the two hypotheses at test are the same under the two methods. This use of the K-L information measure was adopted in the context of a hypothesis-testing problem, which we argued was not simply one of classification but rather one in which the accurate assignment of posterior probabilities of overlap for each pair of clones was also important.

Two types of arguments support this use of the K-L distance. First, we can take the log of the odds ratio given in the previous section, and then average these terms over the distribution of overlapping events to obtain the result

$$I(1{:}2;X) \equiv \sum_x p(s|O)L(x;X)$$

$$= \sum_x p(x|O)\ln\frac{p(O|x)}{p(N|x)} - \ln\frac{p(O)}{p(N)},$$

which we can rewrite as

$$\langle \ln OD(O|X)\rangle_o = I(O{:}N;X) + \ln OD(O).$$

In these expressions, X stands for the particular experimental procedure used to produce the randomly distributed experimental outcomes x. The above equation says that the K-L distance is equal to the difference between the expected value (with respect to the distribution of overlapping events) of the log of the posterior odds in favor of overlap and the log of the prior odds in favor of overlap. Thus the K-L information, or "distance," is equal to the amount by which the experiment has increased the average log-weighted odds in favor of overlap for overlapping events. This quantity is a measure of the average confidence with which overlapping events are correctly identified as weighted to value easy calls (exponentially) less than hard ones. Notice also that this quantity increases linearly with the number of independent repeats of a given method:

$$\langle \ln OD(O|n \times X)\rangle_o = OD(O) + n \times I(O{:}N;X),$$

where $n \times X$ denotes n independent repeats of the procedure X. The K-L information measure can be used as a means of comparing one alternative experimental strategy against another, since, given two alternative experimental strategies X and Y, we can imagine replicating the X experiment n times and the Y experiment m times until $n \times I(X)$ is as close to $m \times I(Y)$ as we care, and then ask ourselves, on the basis of effort or other criteria, which of the two alternative methods of obtaining the same K-L information we prefer.

The second line of argument comes from adopting the approximation that our problem is essentially one of classification and thus its performance is adequately characterized by the type 1 and type 2 misclassification rates that result. While it is not true that two procedures that yield the same K-L information will necessary yield, for example, the same type 1 errors for fixed type 2, such a connection does exist as a form of asymptotic relation for large samples.

For example, if $I(O{:}N)$ is the K-L information from a single repeat of an experiment to distinguish O from N, then

$$\lim_{n\to\infty}\left(\frac{1}{n}\ln(1/\alpha_n^*)\right)_{\beta_0} = I(O{:}N),$$

where α_n^* is a lower bound on the type 1 errors for fixed type 2, say $\beta = \beta_0$ (Kullback, 1968, pp. 74–77). Therefore, if for two alternative experimental methods, X and Y, we have that, respectively, n_x and n_y are "asymptotic" in the above sense, and if

$$n_x I(1{:}2;X) \approx n_y I(1{:}2;Y),$$

then n_x repeats of X yields the same performance in terms of type 1 error rates for fixed type 2 errors as does n_y repeats of Y.

Finally, consider the "ergodic" assumption that, averaged over pairs of clones from the same distribution, the K-L information per observation produced by n independent repeats of the "same" fingerprinting procedure on the same pair of clones is equal, for n large enough, to the average information produced by using the procedure on n different pairs of clones randomly selected from the same population. To the extent that this is a valid approximation, we may also use the above result to estimate that "on average" a single execution of the X experiment is worth

$$\approx \frac{I(O{:}N;X)}{I(O{:}N;Y)}$$

repeats of the Y experiment.

ACKNOWLEDGMENTS

The authors thank Louis Gordon, Eric Lander, Richard Langlois, and P. J. E. Compton for helpful discussions about the mathematical aspects of this problem and Maynard Olson for generous discussions of the yeast ordering effort. The efforts of the LLNL authors on this work were made under the auspices of the U. S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-ENG-48. D.G. was supported by a grant from NIH, and M.W. was supported by grants from NIH and SDF.

REFERENCES

1. Arratia, R., and Gordon, L. (1989). Tutorial on large deviations for the binomial distribution. *Bull. Math. Biol.* **51:** 125–132.

2. BASU, D. (1988). Statistical information and likelihood, a collection of critical essays. *In* "Lecture Notes in Statistics," (J. Berger, S. Fienberg, J. Gani, and K. Krickerberg, Eds.), Vol. 45. Springer-Verlag, New York.
3. BOX, G. E. P., AND TIAO, G. C. (1973). "Bayesian Inference in Statistical Analysis," Addison-Wesley, Reading, MA.
4. CARRANO, A. V., LAMERDIN, J., ASHWORTH, L. K., WATKINS, B., BRANSCOMB, E., SLEZAK, T., RAFF, M., DE JONG, P. J., KEITH, D., MCBRIDE, L., MEISTER, S., AND KRONICK, M. (1989). A high-resolution, fluorescence-based, semiautomated method for DNA fingerprinting. *Genomics* **4:** 129–136.
5. CONNEALLY, P. M., AND RIVAS L. R. (1980). Linkage analysis in man. *In* "Advances in Human Genetics" (H. Harris and Kurt Hirschhorn, Eds.), Vol. 10, Plenum, New York.
6. COULSON, A., SULSTON, J., BRENNER, S., AND KARN, J. (1986). Towards a physical map of the genome of the nematode *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* **83:** 7821–7825.
7. COX, D. R., AND HINKLEY, D. V. (1986). "Theoretical Statistics," Chapman and Hall, New York.
8. EDWARDS, A. W. F. (1986) "Likelihood," Cambridge Univ. Press, Cambridge, UK.
9. EVANS, G. A., AND LEWIS, K. A. (1989). Physical mapping of complex genomes by cosmid multiplex analysis. *Proc. Natl. Acad. Sci. USA* **86:** 5030–5034.
10. GOOD, I. J. (1982). "Good Thinking: The Foundations of Probability and Its Applications," Univ. Of Minnesota Press, Minneapolis.
11. HOEL, P. G., PORT, S. C., AND STONE, C. J. (1971). "Introduction to Statistical Theory," Houghton Mifflin, Boston.
12. KIEFER, J. C. (1987). "Introduction to Statistical Inference," Springer-Verlag, New York.
13. KOHARA, Y., AKIYAMA, K., 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.
14. KULLBACK, S. (1968). "Information Theory and Statistics," Dover Public [Reprinted by Peter Smith Press, Gloucester, MA, 1978].
15. KULLBACK, S., AND LEIBLER (1951).
16. LANDER, E. S., AND WATERMAN, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.
17. LANDER, E. S., AND BOTSTEIN, D. (1986). Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. *Proc. Natl. Acad. Sci. USA* **83:** 7353–7357.
18. MELSA, J. L., AND COHN, D. L. (1978). "Decision and Estimation Theory," McGraw-Hill, New York.
19. MICHIELS, F., CRAIG, A. G., ZAHETNER, G., SMITH, G. P., AND LEHRACH, H. (1987). Molecular approaches to genome analysis: A strategy for the construction of ordered overlapping clone libraries. *Cabios* **3:** 203–210.
20. OLSON, M. V., DUTCHIK, J. E., GRAHAM, M. Y., BRODEUR, G. M., HELMS, C., FRANK, M., MACCOLLIN, M., SCHEINMAN, R., AND FRANK, T. (1986). A random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* **83:** 7826–7830.
21. OTT, J. (1974). Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage analysis. *Amer. J. Hum. Genet.* **26:** 588–597.
22. SÄRNDAL, C. E. (1970). A class of explicata for "information" and "weight of evidence," *Rev. Int. Stat. Inst.* **38:** 223–235.
23. STADEN, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* **8:** 3673–3694.
24. SULSTON, J., MALLETT, F., STADEN, R., DURBIN, R., HORSNELL, T., AND COULSON, A. (1988). Software for genome mapping by fingerprinting techniques. *Cabios* **4:** 125–132.
25. WATERMAN, M. S., AND GRIGGS, J. R. (1986). Interval-graphs and maps of DNA. *Bull. Math. Biol.* **48:** 189–195.