Statistical Science
1994, Vol. 9, No, 3, 334–354

Statistical Issues in Constructing High Resolution Physical Maps

David O. Nelson and Terence P. Speed

Abstract. One of the great success stories of modern molecular genetics has been the ability of biologists to isolate and characterize the genes responsible for serious inherited diseases like Huntington's disease, cystic fibrosis and myotonic dystrophy. Instrumental in these efforts has been the construction of so-called physical maps of regions of human chromosomes.

A major goal of the Human Genome Project is to construct physical maps of the entire human genome. Such maps will reduce the time and expense required to isolate and study interesting chromosomal regions by many orders of magnitude. This article describes what physical maps are and how they have been used, and it outlines some of the statistical issues involved in making them.

Key words and phrases: Physical mapping, recombinant DNA techniques, combinatorial optimization.

1. INTRODUCTION

One of the great success stories of modern molecular genetics has been the ability of biologists to isolate and characterize the genes responsible for serious inherited diseases like Huntington's disease, cystic fibrosis and myotonic dystrophy. These efforts have been massive in scope and expense, involving many years of labor by multiple laboratories. Much of this effort has involved determining with some precision where genes of interest reside on a chromosome, and then constructing a so-called physical map of the region to guide subsequent biochemical analyses.

One major goal of the Human Genome Project (Olson, 1993) is to reduce the time and expense required to isolate and study regions of biological interest by constructing physical maps of the entire human genome. Such maps could then be used by other molecular biologists. Not only would gene-finding be assisted by such maps. Isolating and cloning an interesting chromosomal region is a necessary first step in nearly any research project involving a molecular analysis of chromosomes. As Walter Gilbert noted at Human Genome II, when the Human Genome Project is complete, isolating a region or gene will become a semester project instead of a decade's work. Biologists will be able to concentrate on the more interesting and difficult task of understanding how the approximately 100,000 genes buried in our chromosomes conspire to make us human beings.

In this article we will concentrate on issues involved in constructing high-resolution physical maps. In the following sections we will:

- explain in more detail what physical maps are and why they are necessary;
- describe current methods for constructing physical maps;
- describe in some detail how the group at Lawrence Livermore National Laboratory's Human Genome Center (LLNL) are going about constructing a high resolution physical map of human chromosome 19.

For another, complementary discussion of statistical issues in physical mapping, see Balding (1994).

David O. Nelson is Computer Scientist, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, L-452, Livermore, California 94551, and Ph.D. candidate, Statistics Department, University of California, Berkeley, California 94720. Terence P. Speed is Professor, Statistics Department, University of California, Berkeley, California 94720.

2. DNA, GENES AND MAPS

The genetic complement of a human being (a "human genome") consists of 23 pairs of chromosomes

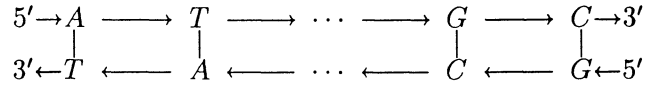

FIG. 1. *Typical diagram of a part of a double strand of DNA.*

(Watson et al., 1987). Each chromosome is composed of a single large molecule of DNA, as well as other supporting structures. Each molecule of DNA is itself composed of two complementary polymers called *strands*, bound together by hydrogen bonds between the monomers in the two strands. Each strand is composed of a long sequence of four different monomers containing *bases* denoted by A, T, G and C. The instructions and data necessary to transform a fertilized egg in utero into a developed human being are encoded in the sequence of bases making up the two complementary strands of the DNA in each of the zygote's 23 pairs of chromosomes.

The strands are called complementary because the configuration of hydrogen bonds between the bases ensures that, under normal conditions, base A in one strand is only paired with base T in the other strand. Likewise, base G in one strand is only paired with base C in the other strand. Thus, a portion of a DNA molecule is often diagramed as in Figure 1. The short vertical lines represent the hydrogen bonds between the strands, and the horizontal arrows represent bonds between the monomers within a strand. The bonds between the monomers are drawn as arrows because each strand has a *direction*, and the messages encoded in the bases are to be read from the 5′ end to the 3′ end of a strand. (This notation for the ends of a strand of DNA is standard and is based upon molecular labeling conventions.) Most messages are coded in the form of *genes*, which are sections of DNA containing instructions for the synthesis of proteins.

Each of the 23 human chromosomes differs in size, and the total number of pairs of bases in one set of 23 chromosomes is approximately 3×10^9. The Human Genome Project is an ambitious (some would say audacious) "...international effort to develop genetic and physical maps and determine the DNA sequence of the human genome and the genomes of several model organisms" (Collins and Galas, 1993).

What are physical maps? The answer is not as precise as one would like. To understand this, we must first understand something about recombinant DNA techniques as well as current limitations in how regions of DNA can be analyzed by molecular geneticists. [See Brown (1990) for a readable introduction to recombinant DNA techniques and genetic analysis.] The twin overriding facts of life are the following:

- Current methods of chemically analyzing substantial stretches of DNA require a sample containing a large number of identical molecules, typically produced by recombinant DNA amplification.
- However, the maximum size of a region that can be amplified by current techniques is orders of magnitude smaller than even the smallest human chromosome.

For example, the size of the longest contiguous fragment of DNA that can be reliably amplified by a recombinant DNA process called cloning (see below) ranges from around 4×10^4 to 1×10^6, depending on the vector and host. Similarly, the longest stretch of DNA that can be reliably amplified by a purely chemical technique known as polymerase chain reaction (PCR) is approximately 1×10^3 bases (Navidi and Arnheim, 1994). In contrast, the 22 human autosomes range in size from around 3×10^8 bases for chromosome 1 down to about 5×10^7 bases for chromosome 21. Because of this mismatch in sizes, producing enough DNA to permit biochemical analyses currently requires a process called *cloning*, in which the following steps are performed:

- A large number of identical chromosomes are broken randomly into fragments by one or more of a class of enzymes known historically as *restriction* enzymes.
- Individual fragments of appropriate size are incorporated by biological or chemical mechanisms into the DNA of host organisms such as *E. coli* or yeast.
- The individual hosts are separated from each other and allowed to grow into colonies, with the fragment in each host being replicated along with the DNA of the host during cell division (*mitosis*).

In this way, the natural DNA replication machinery of the host organism is exploited to replicate the fragment along with the host's chromosomes. After enough mitoses, each host colony can be harvested. The result of this process is a *library* of cloned chromosome fragments, where each fragment is present in large enough quantities to permit isolation and purification of the fragment and subsequent biochemical analyses. Unfortunately, the library contains no information about the relative positions of the fragments along the chromosome. *Physical maps* are data structures which provide the necessary information to enable the order and distance among fragments to be deduced. Hence, they are essential if a collection of overlapping cloned chromosome fragments (a *contig*) is to be treated as though it were a contiguous region of DNA.

Sometimes the ordering information about a contig is exactly that: the relative order of the cloned fragments themselves. In this case, the physical map is intimately related to the clone library from which it arose. However, this need not be the case. Olson et al. (1989) propose a "common language" for physical mapping based on the notion of a *sequence-tagged site*, or STS. An STS is a small sequence of DNA that occurs precisely once in the human genome and can be reliably assayed by PCR. In their proposal, a physical map would consist of a linear sequence of STS's, along with the instructions necessary to construct reliable assays for each STS. In this case, the map would not be tied to a particular set of clones, but could be used to order *any* subsequently generated library. The needed resolution, or distance between STS's, would depend on the average size of cloned fragments to be ordered: "The main practical requirement is that the resolution should be high enough to make regeneration of cloned coverage of any region straightforward" (Olsen et al., 1989). Indeed, even if a physical map is constructed for a particular library, one can then create a map of STS's from the ordered library to produce an STS map for subsequent general use.

As the above paragraph indicates, the critical requirement for an STS map to be useful is that the distance between neighboring STS's be suitably matched to the average size of a cloned fragment: the larger the cloned fragments, the farther apart the STS's can be. Not surprisingly, then, most of the maps published to date which use STS techniques are for libraries containing very large fragments (see Section 2.1).

In summary, the main purpose of a physical map is to enable a set of DNA fragments to be treated as though it were a contiguous stretch of a chromosome. Currently published physical maps provide order information for a particular set of clones. These maps typically contain STS's, which can, in principle, be used to generate contigs from other libraries. However, the density of STS's in these maps is such that they require large cloned inserts. Despite that, an ultimate goal is to produce a detailed set of abstract landmarks which can be maintained in a database and not be associated with any particular library at all.

Finally, knowing the order among clones also allows one to construct another kind of map, known as a *restriction map*, which can in turn help verify the ordering information. (See Figure 9 for an example of a restriction map.) Restriction maps document the order and distance along the genome between sequences of DNA known as restriction sites. A restriction site is the location of a sequence, typically four to six bases long, where a particular restriction enzyme will cut the DNA. Different enzymes recognize and cut different sequences. The distances between restriction sites in a cloned segment of DNA can be determined by digesting the cloned segment with the enzyme and observing the sizes of the resulting pieces by a process known as *gel electrophoresis*. Since overlapping cloned segments must share restriction sites, ordering information can be used to generate restriction maps, and conversely, restriction maps can be used to validate orderings. Restriction maps can also be vital in identifying features known as polymorphisms, as we shall see below. (Polymorphisms are regions of DNA which tend to be different in different individuals.)

2.1 Constructing Physical Maps

In this section, we describe in more detail some of the issues involved in constructing a large-scale physical map. Some early efforts to explore these issues include those by Coulson et al. (1986), Olson et al. (1986) and Kohara, Akiyama and Isono (1987). Constructing a map of a chromosome, or of a large region on a chromosome, can be broken down into several basic steps:

1. create a library of cloned fragments as described above;
2. produce a data "fingerprint" for each cloned fragment;
3. use the information in each fragment's fingerprint to assemble a physical map of the region of interest.

By *fingerprinting* a clone, we mean performing one or more experiments on that clone, the results of which depend in some way on the underlying DNA sequence. Hence, the results of these experiments can help identify or characterize the clone. Cloned fragments which overlap, that is, share a portion of the genome, *may* produce fingerprints more similar to one another than clones which do not overlap. We then use the similarity between fingerprints as a measure of similarity between clones.

Differences between maps and map-making methods boil down to making different choices in these steps. For instance, in STS-content maps (Green and Green, 1991), a "fingerprint" consists of an enumeration of the STS's contained in a cloned fragment. Hence two clones overlap whenever they share an STS, but may very well overlap without sharing any STS's, if the resolution is too coarse. On the other hand, in the high resolution map of chromosome 19 being produced by LLNL, a fingerprint is a list of observed fragment sizes resulting from a complicated digestion of the cloned fragment. In this situation, statistical methods are used to compute a posterior

probability of overlap, given the data from the fingerprint. (We will describe LLNL's approach in much more detail in Section 3.) Let us now look at each of the steps enumerated above in a little more detail.

2.1.1 *Producing a library of cloned fragments*

From the point of view of subsequent analyses, a library is just a "random" sample of overlapping DNA intervals. However, the choices made at this step are critical in determining the subsequent uses to which the map can be put.

Molecular cloning requires that the DNA to be cloned (the *insert*) be joined to another DNA molecule (the *vector*) that can replicate in *host* cells. This joining is carried out in vitro and the resulting recombinant DNA molecules are then introduced into the host cells. Common hosts include the bacterium *E. coli* and baker's yeast *S. cerevisiae*. Vectors which are used with *E. coli* include naturally occurring genetic elements known as plasmids, of which pBR322 is perhaps the most well known, and bacteriophages (bacterial viruses) such as the phage λ. The LLNL library we discuss below consists of *cosmids*, hybrid vectors which replicate like plasmids but can be packaged in vitro into λ coats. There are also naturally occurring yeast plasmids, but more important are *yeast artificial chromosomes*, or YAC's, which consist of the insert joined to a specially designed piece of DNA which functions as a synthetic yeast chromosome.

The importance of all of this is that different hosts and methods of incorporation can vary widely in the average size, the variability in size and the subsequent stability of the incorporated fragment, as in the following examples:

- YAC's in yeast may have DNA fragments ranging from about $0.1–1 \times 10^6$ bases in length.
- Cosmids in *E. coli* may have fragments ranging from about $3.5–4.5 \times 10^4$ bases in length.
- YAC's are much more prone than cosmids to *chimerism*, where two or more fragments are combined in the same clone.
- Some host strains can exhibit *cloning bias*, preferentially amplifying fragments from certain regions of the genome over others.

How important each of these characteristics is depends on how the map will be used. One can see that, as far as efficiency of coverage is concerned, the order-of-magnitude difference between the size of an insert in YAC's and other methods make YAC-based libraries appealing. However, YAC-based maps have their weaknesses, too. For instance, YAC-based maps are, by themselves, too coarse-grained for many purposes, such as sequencing. Consequently, if the map is to be used to guide a large-scale sequencing effort, individual YAC's must themselves be subcloned into smaller pieces of a size suitable for sequencing. In addition, the high rate of chimerism in YAC libraries can create considerable complications in map assembly. On the other hand, YAC-based maps have proved immensely useful in guiding the construction of higher resolution maps. For instance, Baxendale et al. (1993) and Hoheisel et al. (1993) have produced cosmid-based maps of the Huntington region (2×10^6 bases) and the *S. pombe* genome (1.4×10^7 bases), respectively. In both of these cases, they had a preexisting YAC map to guide the construction of a higher level map. One can imagine that the other "first-generation" YAC-based maps, such as those described by Bellarme-Chantelot et al. (1992), Chumakov et al. (1992), Foote et al. (1992) and Cohen, Chumakov and Weissenbach (1993), will prove similarly valuable in guiding and focusing efforts toward higher resolution maps of regions of interest, including whole human chromosomes.

2.1.2 *Producing "fingerprint" data for each clone*

Recall that, given a library of clones, our task is to determine the ordering relationships between them so they can be treated as though they were a single, contiguous piece of DNA. In most cases, one can imagine this task proceeding in two logical steps:

- analyze each clone to produce a vector of data which depend on its DNA content;
- then use this information to assemble contigs.

We now examine some of the kinds of data produced in mapping projects. Most types of data record the result of probing clones for the presence of particular sequences of DNA. Such probing is usually carried out by PCR, restriction digestion or a process called *hybridization*, a kind of chemical analogue of the process of aligning strings of letters so they match (Hames and Higgins, 1985). In a hybridization experiment, the similarity between two single strands of DNA (a *probe* and a *target*) is measured by observing the extent to which the two single strands can form a duplex of complementary base-pair sequences at a particular temperature, pH and so forth. Generally speaking, strands which are perfectly complementary will remain bound into a duplex at measurably higher temperatures than strands containing base-pair mismatches. By measuring the amount of probe that sticks to a target, one can (imperfectly) estimate whether or not a probe and a target share DNA.

Some data represent events that should happen at most once in the genome. For instance, one can assay clones for the presence of an STS. Because of the uniqueness of the sequence represented by the

STS, if two clones both test positive for the STS, then we can conclude (barring experimental error) that the two clones overlap. Let us call this kind of data *unique sequence* data. Other data represent events that can happen more than once in the genome. For instance, we can assay clones for the presence of a particular small sequence like GGAATTCC. For any given clone, barring experimental error, the sequence either will or will not appear. However, as the sequence is likely to occur by chance in many different places in the genome, we cannot conclude that two clones overlap simply because they both test positive for this sequence. Such an event *does* provide some evidence for overlap. Let us call this kind of data *repetitive sequence* data. Finally, we can digest the clone using one or more restriction enzymes and observe the sizes of the resulting fragments. Again, the two clones will both contain a fragment of size k whenever they both have two consecutive restriction sites that are k bases apart. In this case, some experimental error is inevitable, as fragment sizes can only be measured with a relative error of several percent at best. As with the repetitive sequence data, two clones containing a fragment of a given length provides some positive evidence that the clones overlap, but is by no means conclusive proof.

Other types of data include combinations of the above approaches. For instance, Stallings et al. (1990) describe an approach in which fragments from restriction enzyme digestions are hybridized by repetitive sequence probes. The fact that two clones both contain a fragment of size k which contains a repetitive sequence provides more compelling evidence that the two clones overlap than either event by itself.

2.1.3 *Assembling contigs and obtaining closure*

Once we have data on individual clones, we use that data to assemble the clones into contigs. How we go about that depends to a great extent upon the type of data we have. In the case of constructing maps from unique sequence probes such as STS's, the problem is largely algorithmic. We must find an ordering of STS's and an associated configuration of clonal overlaps that is consistent with the data.

If there were no errors, this problem could be easily solved by testing whether an incidence matrix summarizing the data has the "consecutive ones property." An example will make things clear. Consider the overlap configuration in Figure 2. The horizontal lines represent clones, and the vertical arrows represent STS's. Given this configuration, we can construct an incidence matrix describing which unique sequence probes are positive for which clones. This matrix will have a 1 in row i and column j if clone i contains unique sequence j, and a 0 otherwise. Such a configuration would produce the incidence matrix on the left-hand side of (1):

		STS's			STS's
		a b c d e			*b a e d c*
	1	1 1 0 0 1		1	1 1 1 0 0
(1) Clones	2	0 1 0 0 0	$\Longrightarrow$	2	1 0 0 0 0
	3	1 0 0 1 1		3	0 1 1 1 0
	4	0 0 0 1 0		4	0 0 0 1 0
	5	0 0 1 0 0		5	0 0 0 0 1

If we permute the columns of the matrix on the left-hand side of (1) into the order (b, a, e, d, c), we will produce the matrix on the right-hand side of (1). In this permuted matrix, all the ones in each row appear consecutively. If we visualize each row of ones as a clone, this permuted matrix shows us an overlap relationship which is consistent with the data. Note that not all overlaps are captured by the matrix. For instance, in our example, clones 4 and 5 overlap, but this fact is hidden from us by a lack of unique sequence probes in the overlap region. Such *cryptic overlaps* are quite common. However, it is the best we can do with the data at hand, as the density of probes in the example is too coarse to detect all overlaps.

An incidence matrix like the incidence matrix in our example has the *consecutive ones property for rows* if its columns can be permuted so as to make

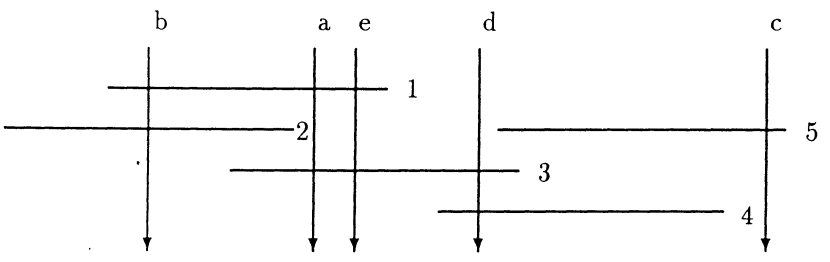

FIG. 2. *Sample overlap configuration with STS's.*

all the ones in each row appear consecutively. The matrix on the left-hand side of (1) has the consecutive ones property for rows, as seen by its permuted version on the right-hand side of (1). It is immediately clear that (a) incidence matrices corresponding to error-free probings of clones with unique sequence probes have the consecutive ones property; (b) the permutation operation corresponds to finding an ordering of STS's consistent with the data; and (c) the structure of the permuted matrix provides a consistent overlap configuration for the clones.

It is easy to determine if a matrix has the consecutive ones property. Booth and Lueker (1976) describe linear-time algorithms for determining if a matrix has the consecutive ones property. Furthermore, their algorithms also provide a compact description of all possible consistent permutations in the form of a PQ-tree. Hence, the problem is completely solvable in linear time when the data are error free.

Unfortunately, real datasets are never error free. Because of occasional experimental failures, frequent clone chimerism, mislabeling, PCR or hybridization error, measurement error and so forth, real data sets of any size will inevitably contain some erroneous entries. In this case, the relationship between consistent orderings and the consecutive ones property is destroyed, and the problem of map assembly from unique sequence data ceases to be well posed. One can continue to use approaches based on PQ-trees, but now one must use heuristics to search for a nearby matrix which has the consecutive ones property. Instead, the problem is typically recast as a combinatorial optimization problem: search through the space of all linear orderings of the objects (probes or clones) to find one that minimizes some user-defined penalty function. The underlying data structure is the same: an incidence matrix of clone–probe hits. The differences now lie in what is being ordered (rows for clones or columns for probes), how the penalty function is defined and the search technique used to generate new permutations.

In nearly all published algorithms, the penalty function is defined to be the sum of pairwise discrepancies between adjacent objects. In this case, the problem is formally identical to the traveling salesman problem (Lawler, 1985) on a complete graph. The nodes on this graph correspond to the objects to be ordered, either clones or probes. An edge between nodes a and b is weighted by a measure $d(a, b)$ of the discrepancy between objects a and b. The goal is to find a path visiting every node exactly once which minimizes the sum of the discrepancies along the edges in the path. Despite this formal similarity, most efforts to date do not exploit the considerable body of research in efficient solution techniques for the traveling salesman problem [see Newberg (1993) for an exception]. Perhaps this is because simpler heuristics have proved successful on problems encountered to date.

For example, Mott et al. (1993) order probes by defining the discrepancy between probes a and b as

$$d(a, b) = 1 - \frac{\text{Number of clones positive for } a \text{ and } b}{\text{Number of clones positive for } a \text{ or } b}.$$

They then use simple reordering rules to generate new permutations and simulated annealing to determine whether or not to accept the newly generated permutation. Once probes have been ordered, clones are then ordered with respect to the probes by an algorithm that maximizes a measure of fit between the probe data for that clone and the list of ordered probes. This approach has proved successful in constructing maps of the Huntington region (Baxendale et al., 1993) and genome of the fission yeast *S. pombe* (Hoheisel et al., 1993). In addition, they also describe a different, heuristic ordering procedure which attempts to edit the clone–probe matrix to restore consistency. In addition to the work of Mott et al. (1993), Cuticchia, Arnold and Timberlake (1992) describe efforts to construct maps from simulated clone–probe matrices via simulated annealing. In this case, clones were ordered using the ℓ_1 distance between the rows of the clone–probe matrix as a measure of discrepancy.

The combinatorial optimization approach to finding an optimal ordering described above is based on a simple, two-part strategy: compute a measure of discrepancy between all pairs of objects, and then search through the space of all orderings for a satisfactory solution. An assumption underlying this is that the data-collection process is largely independent of the map assembly process. Palazzolo et al. (1991) describe an entirely different, "directed" approach, in which the overlap information gathered at step n directs the experimental procedure at step $n + 1$. The method proceeds roughly as follows, on a library of N clones:

1. Begin with all N clones unlabeled.
2. Repeat the following until all clones are labeled:
   (a) Choose an unlabeled clone at random; call this clone a "seed" clone.
   (b) Create STS's from the two ends of the seed clone, and use them to find all clones that overlap the ends of the seed clone; call this set S.
   (c) If one or more clones in S is already labeled, merge contigs by relabeling all the clones in S, as well as the seed clone, with the smallest label among the labeled clones in S.
   (d) If no clone in S is labeled, add a new contig by labeling all the clones in S, as well as

the seed clone, with the next available contig number.

When complete, all N clones will be labeled with a contig number. Note that, as described, no mention is made of ordering probes or clones. However, evolving maps of contigs can be created and maintained as the process evolves. PQ-tree approaches can be used to edit and validate the evolving contigs. Merged contigs whose incidence matrices do not possess the consecutive ones property are symptomatic of an error, which can be localized and corrected. This approach, the success of which depends critically on the ability to do large numbers of high-accuracy probings economically, has been used successfully to map the *S. pombe* genome in cosmids by Mizukami et al. (1993).

Finally, recall that the clones in any library can be considered as a random sample of segments from some underlying genome. Hence, the map will actually consist of a collection of contigs separated by oceans of DNA not covered by any clone. As noted by Lander and Waterman (1988), for a library of N clones of average length L taken from a genome of length G, the expected proportion of the genome left uncovered is approximately $\exp(-NL/G)$. The above approaches provide an ordering, but do not unambiguously define contig boundaries.

Now let us examine how maps such as Bellarme-Chantelot et al. (1992) and Stallings et al. (1992) have been assembled from repetitive sequence data, restriction fragment data or a combination of the two. In this situation, one can follow the same basic program as above: use the data to obtain a similarity measure between clones; then order and assemble the clones into contigs using some optimization approach. However, in this situation the random, yet repetitive nature of the data lends itself to a probabilistic interpretation and to statistical approaches to detecting and evaluating potential clone configurations. Unfortunately, because of the complexity of the underlying models, most current efforts confine statistical decision making to detecting pairwise overlap and then retreat to combinatorial heuristics to assemble the map.

Statistical approaches to detecting pairwise overlap, as described by Michiels et al. (1987), Branscomb et al. (1990), Balding and Torney (1991), Fu, Timberlake and Arnold (1992) and others, all begin with a probability model for the data-generating process, given an overlap configuration between two clones. Most then compute an integrated likelihood ratio or posterior probability of overlap, which they use as a similarity measure, although Fu, Timberlake and Arnold (1992) cast the problem as a hypothesis-testing problem. The approach described by Balding and Torney (1991) has been used in Bellarme-Chantelot et al. (1992) and Stallings et al. (1992), while the approach first outlined in Branscomb et al. (1990) will be discussed in detail in Section 3.

Alizadeh et al. (1993) and Nelson, Speed and Yu (1994) describe different statistical approaches to evaluating overlap configurations involving more than two clones. Nelson, Speed and Yu (1994) recast the the problem of determining overlap as a Bayesian decision problem and examine solutions from an information-theoretic point of view. Alizadeh et al. (1993) examine the case of error-free data generated by repetitive probes against constant-length clones. They propose methods to evaluate overlap configurations among many clones. Unlike the STS case, there is no simple mapping between a configuration of clones and a permuted version of the clone–probe matrix. Probes will occur more than once along the genome, and part of the problem is to infer a probe ordering where each probe may occur more than once. In this way, the problem once again becomes one of combinatorial optimization. Alizadeh et al. (1993) describe efficient algorithms for approximating the relative likelihood of any sequence of probes, given a particular overlap configuration. They then use techniques developed for the traveling salesman problem to guide the search through the space of clone overlap configurations.

2.1.4 *Other issues*

As one can surmise from the size of the genomes and libraries involved, constructing a map of a large region or chromosome is a major, multiyear project. As a consequence, such efforts are often preceded by a considerable amount of analysis and simulation, designed to determine the feasibility, the duration, the cost and the likelihood of success of various mapping approaches.

Suppose the goal of a mapping project were to cover P percent of a chromosome in K or fewer cosmid contigs. Project managers will pose questions like the following.

"We can construct a system that will detect fifty percent overlap reliably. However, constructing a system that will detect ten percent overlap reliably will be much harder to accomplish and be more expensive per fingerprint to use. On the other hand, maintaining large libraries is also expensive. How many more clones will I need to analyze if I use a cheap, fifty percent overlap detector rather than an expensive, ten percent overlap detector?"

"After we finish we will have around K contigs. We will have to close the gaps between these contigs

by very expensive methods. How many of the gaps between contigs are actually cryptic overlaps that we will be able to close with one well-selected probe, and how many are real?"

"Suppose we have to stop after we have analyzed 5,000 clones (using whatever method we have chosen). How close to completion will we be? In other words: How many contigs will there be? What will be the average size of a contig? How much of the genome will be covered?"

Of course, the answers to these questions depend on the strategy used to construct the map. Lander and Waterman (1988) addressed these questions for libraries constructed by repetitive sequence fingerprinting of random clones. The models used were similar to those derived by Feller (1948) and Smith (1957) for analyzing the behavior of Geiger counters. Several authors, using several different approaches, have produced results allowing one to predict progress in constructing STS-content maps (Arratia et al., 1991; Barillot, Dausset and Cohen, 1991; Ewens et al., 1991; Grigoriev, 1993; Torney, 1991). Finally, Zhang and Marr (1993) and Nelson and Speed (1994) have produced results allowing one to predict progress in directed mapping projects.

2.2 Finding the Gene for Myotonic Dystrophy

To gain some appreciation of the work involved in isolating a gene, and the role that physical maps play, let us look at the sequence of events leading up to the discovery of the gene associated with myotonic dystrophy ("DM"). Myotonic dystrophy (Harper, 1979) is the most common form of adult muscular dystrophy, with prevalence estimates ranging from 2 to 14 cases per 100,000 individuals. The disease allele is strongly associated with a wide range of disorders, from myotonia and other neurological defects to cataracts. The age at onset and disease severity can also vary widely: some individuals remain asymptomatic as adults, while others present severe symptoms at birth.

As early as 1983, the locus for DM had been mapped by linkage analyses to chromosome 19. Six years later, Korneluk et al. (1989) published an analysis of markers which localized the DM gene to a 10-centimorgan region on the long or q-arm of chromosome 19, denoted by 19q, between the DNA excision repair gene ERCC1 and marker D19S50. Two years later, Tsilfidis et al. (1991) further localized the DM locus to a 2-centimorgan stretch on the band of 19q designated 19q13.3 by finding a recombination event in a pedigree which placed a polymorphic marker D19S51 just distal to the DM locus. Assuming that 1 centimorgan corresponds, on average, to 1×10^6 bases of DNA, this finding localized the DM locus to a 2×10^6 region on chromosome 19 which was flanked by known markers: ERCC1 and D19S51.

Up to this point, efforts to isolate DM had been focused on finding flanking markers close enough to make molecular analyses feasible. The next step would be to construct a high resolution physical map, consisting of (1) a set of overlapping cosmid and YAC clones which would span the region between the two flanking markers, coupled with (2) restriction maps of the region. These maps could then be used to isolate and analyze potential genes in the region.

By the end of 1991, as the result of an intensive international effort, the region was completely cloned and mapped, and a putative defect was observed (Aslanidis et al., 1992). The putative defect was a length variant observed in one of the restriction fragments among affected members. [For a detailed description of the mapping effort, see Buxton et al. (1992), Jansen et al. (1992) and Shutler et al. (1992).] Shortly thereafter, Brook et al. (1992) published evidence of the molecular basis of the length polymorphism: an unstable expansion of a three-base repeat (CTG) at the 3′ end of a new gene. Within a month Mahadevan et al. (1992) published results indicating that the unstable region was in an untranslated region of the gene. By the end of 1992, the gene containing the repeat had been identified as a type of protein kinase and completely sequenced (Mahadevan et al., 1993). Finally, by the end of 1993, diagnostic probes for DM had become commercially available, even though the mechanism by which the mutation resulted in a DM phenotype remained unknown. The mechanism remains unknown today.

Note that most of the time and effort involved in locating and characterizing DM was spent finding suitable markers bracketing a small enough stretch of DNA to make molecular analysis feasible, and much of the subsequent effort involved constructing physical maps of the localized region. Once such an infrastructure was in place, isolating and characterizing candidate genes could proceed quickly. This division of labor is not unusual: see Baxendale et al. (1993) and The Huntington's Disease Collaborative Research Group (HDCRG, 1993) for details concerning the corresponding search for the gene for Huntington's disease. Once produced, high resolution physical maps provide the infrastructure needed to enable efficient (1) discovery of new polymorphic markers that more tightly bracket the region of interest, (2) screening for genes using other libraries of candidate gene sequences called cDNA's and (3) sequencing of candidate genes.

Thus, high resolution physical maps provide a crit-

ical resource to molecular geneticists interested in understanding our genetic makeup.

3. CONSTRUCTING A PHYSICAL MAP OF CHROMOSOME 19

The previous section provided an overview of physical mapping approaches, efforts and issues. Let us now look more closely at one particular project. In this section, we describe aspects of LLNL's current effort to produce a high resolution physical map of chromosome 19. Figure 3 shows a diagram of the basic steps. Further details may be found in Carrano et al. (1989). Chromosome 19 is one of the smaller human chromosomes, containing approximately sixty million bases of DNA. The map is based on cosmids, with each insert approximately forty thousand bases long. Hence, around 1,500 clones would cover the chromosome, if laid end-to-end. However, since the cloned inserts are randomly cut from the chromosome, many more than 1,500 clones are needed to ensure nearly complete coverage of the chromosome. Consequently, the high resolution map is being constructed from more than 10,000 clones from several different cosmid libraries. Despite the fact that we are using several libraries, we will describe the process of map construction as though the clones were derived from a single library.

LLNL is taking a "bottom-up" approach to building their map. The main steps include the following:

1. creating DNA fingerprint data for each clone by restriction digestion and electrophoresis;
2. computing the posterior odds of overlap between each pair of clones, based on the similarity between the pair of fingerprints for the two clones;
3. using these posterior odds values to assemble initial contigs.

After constructing initial contigs by this method, other methods must be used to close the gaps between contigs and to associate known markers with contigs.

3.1 Creating Fingerprints

LLNL began its chromosome 19 mapping effort by spending many months analyzing concurrent efforts by other laboratories at constructing large-scale physical maps. In addition, Branscomb and coworkers performed extensive computer simulations designed to optimize their fingerprinting strategy (Branscomb et al., 1990). As a result of these simulations and analysis, they chose a bottom-up approach to fingerprinting based on restriction digestion: cutting up the clone with enzymes and measuring the sizes of the resulting fragments. The method differs from normal restriction digestion in one very important respect. In ordinary restriction digests, the fragments from the digest partition the clone. In other words, the sum of the sizes of the fragments equals the size of the clone. In LLNL's method, on the other hand, the fragments eventually visualized by a fingerprint are a subset of fragments produced by a two-step multienzyme restriction digest. To see how this happens, let us examine the experimental protocol in a little more detail.

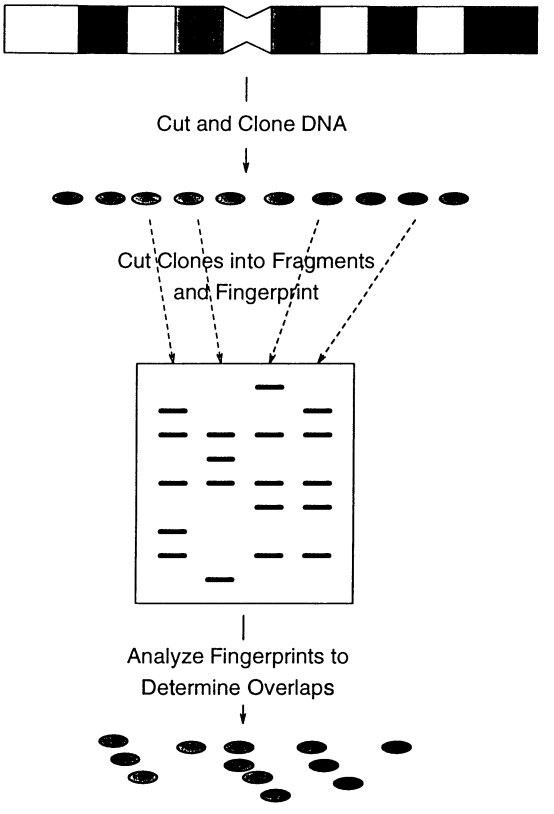

FIG. 3. *Main steps in constructing the map of chromosome* 19.

The digestion process proceeds as outlined in Figure 4. The first step is a complete double digestion with two six-base restriction enzymes, *Eco*RI and *Bgl*II. The result of this first step is a collection of fragments tagged on both ends with a fluorochrome dye. The DNA from this first digestion is then separated into three aliquots. Each of the three aliquots is further digested completely with a different four-base restriction enzyme: *Hinf*I, *Hae*III or *Dde*I. As a result of this second digestion, each fragment from step 1 has been further digested in each aliquot into a number of subfragments, only one or two of which contain color linkers and hence will be visible. The visible fragments are just those between one of the six-base recognition sites cut in step 1 and one of the four-base recognition sites cut in step 2. Of course, if a fragment from step 1 contains no four-base recognition site between its two six-base sites, the entire

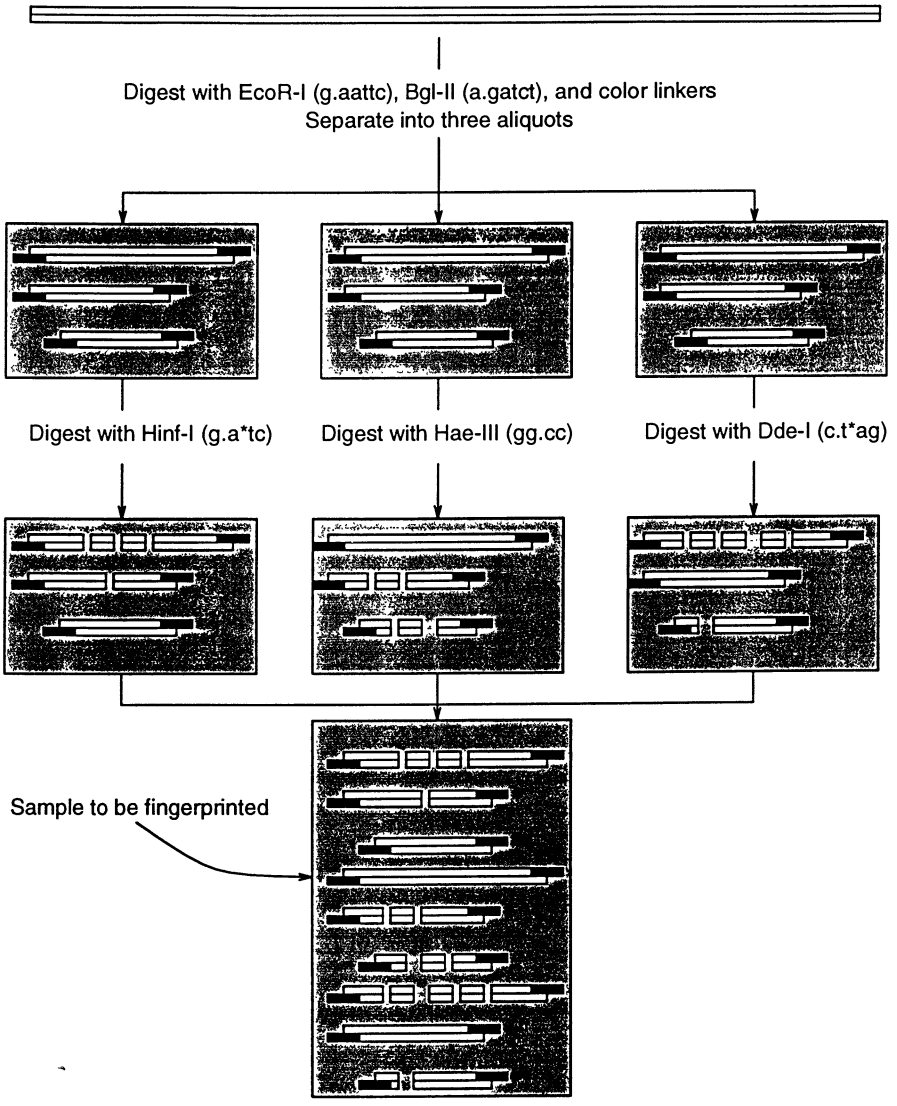

FIG. 4. *Creating fingerprints using multiple restriction digests.*

fragment will be visible. The three aliquots are then combined into a single solution called a DNA *prep* for subsequent electrophoresis. Hence, each fragment produced by step 1 results in at most six visible fragments in the final sample.

DNA fragments are currently visualized by electrophoresis on ABI373 sequencers, resulting in a list of detected fragment sizes for each clone. Because of limitations in the size standard used and the speed of electrophoresis, only fragments with sizes between 30 and 540 bases will be detected. Each sequencer run currently produces fingerprint data on up to 48 separate DNA preps: two different DNA preps can be loaded in each of 24 lanes. In addition to the two DNA preps from cloned inserts, each lane also contains a standard prep containing fragments of known sizes (a *size standard*). Detecting three different preps in a single lane is possible because the ABI373 is a four-dye-per-lane sequencer, designed to detect up to four separate sequencing reactions simultaneously in each lane, with each reaction using a different colored fluorochrome dye. Hence, if the three preps were produced with three different dyes, the reactions will be able to be captured concurrently.

Peaks are detected by a combination of locally developed software and software provided by the sequencer manufacturer, Applied Biosystems, Incorporated (ABI). The initial signal extraction software is currently provided by ABI. It produces a vector time series

$$\{(X_0(k), X_1(k), X_2(k)) | k = 1, \ldots, 6{,}000\}$$

for each lane, with approximately 10 vector samples

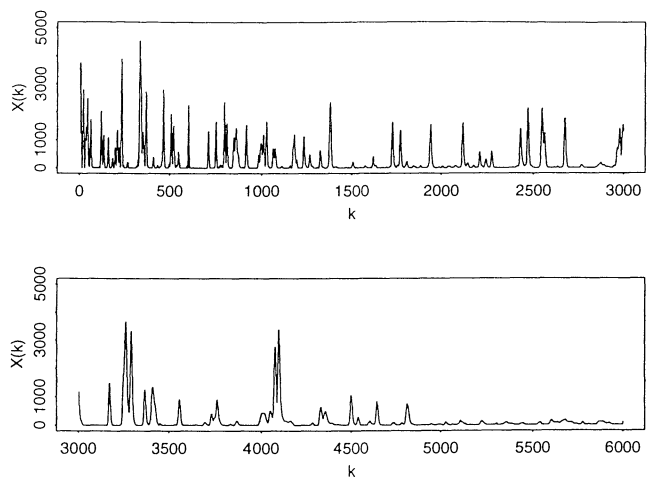

FIG. 5. *Resulting signal for one clone from an electrophoresis run.*

taken per minute. The two component series X_1 and X_2 contain sampled fluorescence intensities for the two DNA preps, while the series X_0 contains sampled fluorescence intensities for the size standard. The laser-induced fluorescence intensities are measured 25 centimeters downstream from where the DNA prep is loaded into the gel. Thus the time series measures time to passage for a fragment, rather than distance traveled in a fixed amount of time, as is the case with radioactivity-based measurement of electrophoresis. Finally, the time series has already been backgrounded and run through a linear filter to extract the three components. The graph in Figure 5 shows a typical component time series.

LLNL's current approach to peak detection breaks the analysis into two steps: finding peaks and estimating fragment length. The present software examines each component series independently to extract the locations of peaks corresponding to DNA fragments. LLNL has tried a number of approaches to extracting peak locations from signals like that shown in Figure 5. No method has proved entirely successful (Nelson et al., 1989). The current (moderately successful) approach assumes the series can be modeled as an AR(1) series contaminated by outliers (here the peaks are the outliers). It relies on robust filtering algorithms provided by S-PLUS (Statistical Sciences, Inc., 1991).

The approach proceeds in three steps. First, an AR(1) model is fitted to the series using a generalized M-estimator. Second, a model-based "filter-cleaner" is applied to the series to classify each data point as "normal" or "outlier." Finally, peak locations are found by locating large local maxima in the portion of the series classified in step 2 as "outlier." See Statistical Sciences, Inc. (1991) for details on the filtering algorithms.

The process just described produces a list of peak locations $\{k_1, k_2, \ldots\}$. To be useful for comparison between inserts, these peak locations must be translated into a standard coordinate system. The size standard in each lane is used for this coordinate transformation. LLNL's current size standard consists of a collection of fragments of known size, constructed by digesting pBR322 and SV40 with *Hae*III. The standard coordinate system is constructed by associating the peak locations of the size standard with known fragments between 30 and 540 bases in length and performing a monotone spline interpolation (Fritsch and Carlson, 1980). The result is a function which associates a standardized fragment length l_j with each peak location k_j.

Although it is known that fragments of the same length may vary in their elution time by up to 3%, the *same* fragment will migrate at approximately the same rate under standardized conditions. Indeed, experiments suggest that LLNL has been able to transform the host vector fragments to the standard coordinate system with a precision of ± 1 base out to 400 bases.

3.2 Detecting Overlap

The result of fingerprinting the library is a list of standardized fragment lengths $\{l_j\}$ for each clone in the library. The next step is to compare all pairs of clones for overlap. For a library of n clones, we compare the fingerprints of all $\binom{n}{2}$ pairs of clones, resulting in a separate comparison vector D for each pair of clones. This comparison vector details which integer fragment lengths occurred in each clone. Based on the comparison vector, we then compute the posterior odds of overlap, given the data D, by Bayes' rule, as suggested by Michiels et al. (1987) and Branscomb et al. (1990).

To create a comparison vector D from a pair of standardized fragment length lists, LLNL currently constructs a bipartite graph describing a best match between pairs of fragment lengths, one from each clone, differing by no more than one base. This graph is then used to decide which fragment lengths have occurred in each fingerprint. As mentioned above, the resulting integer fragment lengths l can range from 30 to 540 bases. We choose a subset of these potential fragment sizes to use as data for detecting overlap. For the purposes of this discussion, let us assume the fragment sizes actually used to detect overlap run from N_0 to N_1 bases, where $30 \leq N_0 < N_1 \leq 540$.

The comparison vector D, corresponding to a comparison of clones A and B, consists of a sequence $(d_{N_0}, d_{N_0+1}, \ldots, d_{N_1})$ of comparison outcomes, where

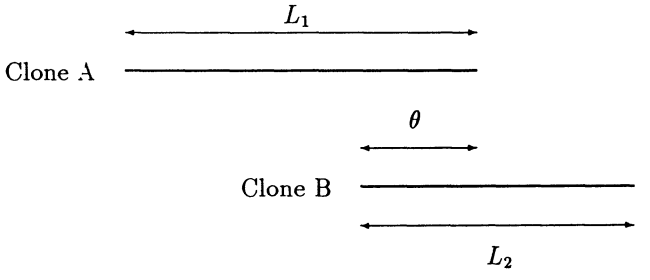

FIG. 6. *Typical configuration where two clones overlap.*

each comparison outcome is of the form

$$d_l = \begin{cases} 00, & \text{if a fragment of length } l \\ & \quad \text{is observed in neither clone,} \\ 10, & \text{if a fragment of length } l \\ & \quad \text{is observed in clone A only,} \\ 01, & \text{if a fragment of length } l \\ & \quad \text{is observed in clone B only,} \\ 11, & \text{if a fragment of length } l \\ & \quad \text{is observed in both clones.} \end{cases}$$

Here 1 and 0 represent the presence and absence of a fragment, respectively.

We can approximate the probability of the above comparison outcomes for a given fragment size l by arguing as follows. A fragment of length l occurs whenever a six-base restriction site and an appropriate four-base restriction site are l bases apart. If the restriction sites involved in the digestion process are distributed across the genome in a roughly uniform manner, then ν, the average number of fragments generated along an M-base stretch of the genome, should be roughly proportional to M; that is, $\nu \approx \alpha M$ for some α. Let us further suppose that the presence or absence on the gel of fragments of different lengths are approximately mutually independent events. Then the probability of seeing no fragments of size l in a region of length M is approximately

$$(1 - \pi_l)^{\nu} \approx \exp(-\lambda_l M),$$

where π_l is the probability that any single fragment (in the size range) is of length l, and the fragment intensity λ_l is defined by $\lambda_l = \pi_l \alpha$. Of course we do not expect the assumptions leading to this probability to be more than approximately valid, if that, but they do lead to a simple and tractable model and, furthermore, one whose utility can be evaluated.

Consider then the situation diagramed in Figure 6, in which clone A has length L_1, clone B has length L_2, the two clones overlap by an amount θ and the intensity for fragments of size l is λ_l. We immediately compute the probabilities of the above comparison outcomes in terms of $q = \exp(-\lambda_l)$ as

$$\begin{aligned} p_{00}(\theta) &= q^{L_1 + L_2 - \theta}, \\ p_{01}(\theta) &= q^{L_1}\left(1 - q^{L_2 - \theta}\right), \\ p_{10}(\theta) &= q^{L_2}\left(1 - q^{L_1 - \theta}\right), \\ p_{11}(\theta) &= 1 - q^{L_1} - q^{L_2} + q^{L_1 + L_2 - \theta}. \end{aligned} \tag{2}$$

3.2.1 *The simple trinomial model*

For the current discussion, let us assume further that all clones are the same length L, and all fragment lengths have the same intensity λ. Without (any more) loss in generality, we may also choose a coordinate system in units of G/L, where G is the length of the genome, so that all clones have length 1 and λ becomes a per-clone intensity. To simplify discussion further, let us call potential fragment sizes "probes," mimicking the vocabulary used in hybridization experiments.

Under these conditions, for any two given clones A and B, we can reduce the comparison vector D by cross-classifying each of the n probes according to which of the clones they hit, forming a fourfold table of counts:

		B: 0	B: 1	
A	0	n_{00}	n_{01}	n_{0+}
	1	n_{10}	n_{11}	n_{1+},
		n_{+0}	n_{+1}	n

(3)

where 1 means that a clone contains a probe, and 0 means that a clone does not contain a probe. The joint distribution for the cells in this table will depend upon the proportion θ of overlap between clones A and B. Since we assumed the probes were independent, for any $\theta \in [0, 1]$, the data in the fourfold table follow a multinomial distribution with probabilities

		B: 0	B: 1	
A	0	$p_{00}(\theta)$	$p_{01}(\theta)$	q
	1	$p_{10}(\theta)$	$p_{11}(\theta)$	$1-q$
		q	$1-q$	1

In this case, (2) reduces to

$$\begin{aligned} p_{00}(\theta) &= q^{2-\theta}, & p_{01}(\theta) &= q\left(1 - q^{1-\theta}\right), \\ p_{10}(\theta) &= q\left(1 - q^{1-\theta}\right), & p_{11}(\theta) &= 1 - 2q + q^{2-\theta}. \end{aligned}$$

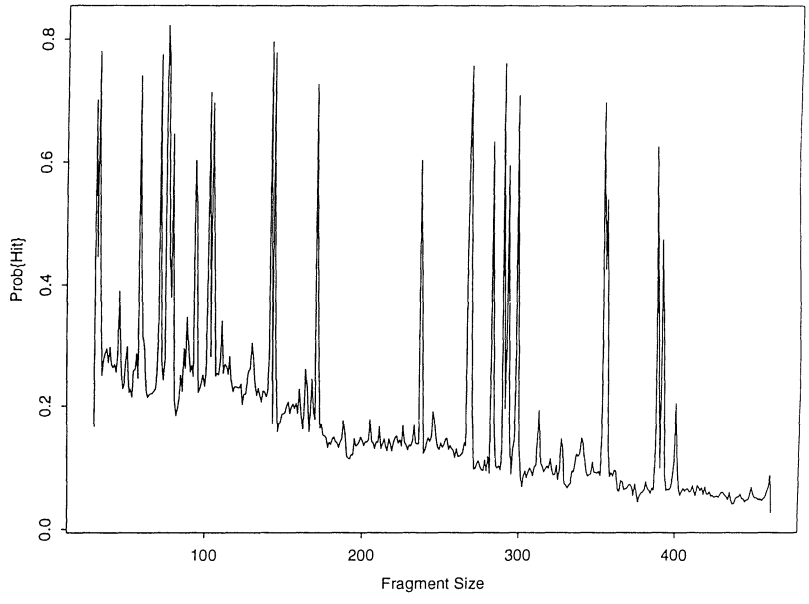

FIG. 7. *Proportion of fingerprints with positive probes as a function of fragment size l.*

We see that, for all θ, $p_{01}(\theta) = p_{10}(\theta)$. Hence, the data in the fourfold table can be reduced without loss of information to a trinomial (n_0, n_1, n_2), where $n_0 = n_{00}$, $n_1 = n_{01} + n_{10}$ and $n_2 = n_{11}$. In this case, the component probabilities reduce to

$$
\begin{aligned}
p_0(\theta) &= p_{00}(\theta) = q^{2-\theta}, \\
p_1(\theta) &= 2p_{01}(\theta) = 2q\left(1 - q^{1-\theta}\right), \\
p_2(\theta) &= p_{11}(\theta) = 1 - 2q + q^{2-\theta}.
\end{aligned} \tag{4}
$$

Furthermore, unless $p = q = \frac{1}{2}$, no further reduction is possible without some loss of information. Let us abuse notation a bit and call this trinomial vector D as well.

We need to decide whether any two clones A and B overlap based on their trinomial data D, and then use those decisions to build contigs. Our approach to deciding overlap and contig building is to consider the proportion of overlap θ to be a random variable and compute the *posterior odds* of overlap $\Pr(\theta > 0 \mid D)/\Pr(\theta = 0 \mid D)$ for every pair of clones.

Now, to compute the posterior odds of overlap for any pair of clones, we notice that

$$
\frac{\Pr(\theta > 0 \mid D)}{\Pr(\theta = 0 \mid D)} = \frac{\Pr(D \mid \theta > 0)}{\Pr(D \mid \theta = 0)} \times \frac{\Pr(\theta > 0)}{\Pr(\theta = 0)},
$$

and $\Pr(\theta > 0)/\Pr(\theta = 0)$ is the same for all pairs of clones. Therefore, we need only to compute the ratio $\Pr(D \mid \theta > 0)/\Pr(D \mid \theta = 0)$ for each pair of clones. Also, since

$$
\Pr(D \mid \theta > 0) = \int_{\theta \in (0,1]} \Pr(D \mid \theta)\, dP(\theta \mid \theta > 0)
$$

and $(\theta \mid \theta > 0) \sim \text{Uniform}(0, 1]$, the computation reduces to determining the integrated likelihood ratio

$$
L(D) = \frac{\Pr(D \mid \theta > 0)}{\Pr(D \mid \theta = 0)} = \int_0^1 \frac{\Pr(D \mid \theta = t)}{\Pr(D \mid \theta = 0)}\, dt \tag{5}
$$

for each pair of clones. Recalling that $D = (n_0, n_1, n_2)$ is distributed as a trinomial with probabilities given by (4), for this simple case (5) becomes

$$
\begin{aligned}
L(D) = \int_0^1 &\left[\frac{q^{2-\theta}}{q^2}\right]^{n_0} \left[\frac{1 - q^{1-\theta}}{1 - q}\right]^{n_1} \\
&\cdot \left[\frac{1 - 2q + q^{2-\theta}}{1 - 2q + q^2}\right]^{n_2} d\theta.
\end{aligned} \tag{6}
$$

3.2.2 *Toward more realistic assumptions*

The above description of the simple trinomial model made a number of simplifying assumptions: equal-length clones; equal intensities; and perfect fragment size detection. We may need to relax some of these assumptions to produce a usable method for overlap detection. In this paper, we will confine our attention to examining the effects on $L(D)$ of relaxing assumptions about equal intensities and perfect detection.

Different intensities. So far we have assumed that each probe is a stationary process with intensity λ, so that the probability of hitting a region of size L is $1 - \exp(-\lambda L)$. However, the graph in Figure 7, showing the proportion of fingerprints with positive probes for each fragment size between 30 and 462 bases, indicates that this value is not constant over

the range of available fragment sizes, ranging from about 0.05 to about 0.25. This graph was produced by analyzing over 10,000 fingerprints. The conspicuous spikes in the graph are the fragment sizes associated with the cloning vector, which should appear in every fingerprint, and which we ignore.

Given an estimate for the λ_l associated with each fragment size l, it is easy to incorporate differing intensities in an expression for $L(D)$. However, in this case the comparison vector D described in Section 3.2 cannot be summarized as a trinomial. The expression for $L(D)$ in (6) becomes a product of ratios over each of the potential fragment sizes:

$$L(D) = \int_0^1 \prod_{\{l \mid d_l = 00\}} \frac{q_l^{2-\theta}}{q_l^2} \prod_{\{l \mid d_l = 10 \vee 01\}} \frac{1 - q_l^{1-\theta}}{1 - q_l} \cdot \prod_{\{l \mid d_l = 11\}} \frac{1 - 2q_l + q_l^{2-\theta}}{1 - 2q_l + q_l^2} \, d\theta \tag{7}$$

where $q_l = \exp(-\lambda_l)$.

Fragment size detection error. In reality, peak detection (and hence fragment size detection) is far from perfect. The first author performed an experiment at LLNL to evaluate how often peaks were missed. We wanted to estimate two sets of parameters: π_{l0}, the probability of mistakenly finding a fragment of length l when none was there, and π_{l1}, the probability of correctly finding a fragment of length l when one actually *is* there. For this experiment, the values l ran from 30 bases to 462 bases.

We obtained a convenience sample of 40 clones, created for a prior quality analysis experiment. Each clone had been fingerprinted from three to five times. We first used a clustering algorithm to cluster peaks from different samples from the same clone into a collection of putative real peaks. This produced a data set consisting of a 433×40 matrix of counts, where each row corresponds to a fragment size from 30 to 462 bases, and each column corresponds to a clone. We modeled the distribution of the count for fragment length l for clone j, denoted by Y_{lj}, as

$$Y_{lj} \sim \begin{cases} \text{Binomial}(n_j, \pi_{l0}), & \text{if clone } j \\ \qquad \text{contains no fragment of length } l, \\ \text{Binomial}(n_j, \pi_{l1}), & \text{if clone } j \\ \qquad \text{contains a fragment of length } l, \end{cases} \tag{8}$$

where n_j is the number of samples for clone j. One can reformulate the problem as a missing data problem and estimate the $\{\pi_{l0}\}$ and $\{\pi_{l1}\}$ using the EM algorithm (Dempster, Laird and Rubin, 1977). The complete data would consist of a vector $(\mathbf{X}, \mathbf{Y}) := (X_1, X_2, \ldots, X_{40}, Y_1, Y_2, \ldots, Y_{40})$ for each fragment length l. The elements of $\mathbf{X}$ are independent, identically distributed Bernoulli random variables, where X_j takes on the value 1 whenever clone j actually has a fragment of length l. The elements of $\mathbf{Y}$ are independent of each other, and each $(Y_j \mid \mathbf{X})$ is distributed as in (8).

With this setup, the complete data are distributed as an exponential family, and we observe only $\mathbf{Y}$. Hence, the maximum likelihood estimates for our parameters are quite simple:

$$\hat{\pi}_1 = \frac{\sum_j Y_j X_j}{\sum_j n_j X_j}, \quad \hat{\pi}_0 = \frac{\sum_j Y_j (1 - X_j)}{\sum_j n_j (1 - X_j)},$$

where we have, as usual, elided the l. To complete the EM algorithm, we must compute expected values for the sufficient statistic $(\sum Y_j X_j, \sum n_j X_j)$, conditional on $\mathbf{Y}$ and the current values for π_1 and π_0:

$$\mathbb{E}\left[\sum_j Y_j X_j \mid \mathbf{Y}\right] = \Pr(X = 1 \mid Y) \sum_j Y_j,$$

$$\mathbb{E}\left[\sum_j n_j X_j \mid \mathbf{Y}\right] = \Pr(X = 1 \mid Y) \sum_j n_j;$$

and $\Pr(X = 1 \mid \mathbf{Y})$ can easily be computed by Bayes' rule, given a value for the prior probability $\Pr(X = 1)$.

We used the histogram in Figure 7 to estimate the prior probability of finding a fragment for any length l and estimated the probabilities $\{\pi_{l0}, \pi_{l1} \mid l = 30, \ldots, 462\}$ as described above. Figure 8 shows a lowess-smoothed estimate of the error rates π_{l0} and $1 - \pi_{l1}$. From this we clearly see that the false positive rate (π_{l0}) is relatively constant over l—a little higher for the shorter fragments, a little lower for the bigger fragments. However, the false negative rate $(1 - \pi_{l1})$ begins to grow after about 300 bases. In any event, it is clear that fragment size detection is far from perfect.

Perhaps the simplest way to address the problem of errors, other than ignoring it completely, is to construct a model with the following:

- a fixed probability π_0 of falsely calling a peak;
- a fixed probability $1 - \pi_1$ of dropping an existing peak;
- independent errors among different fragment sizes.

In light of the graph in Figure 8, these assumptions seem reasonable if we confine our attention to fragments less than about 300–350 bases long. In this situation, however, the expressions for $L(D)$ are no longer simple functions of q as in (6) and (7).

The new component probabilities for the comparison vector D depend upon the error probabilities $\pi = (\pi_0, \pi_1)$, as well as θ and, in the case of differing intensities, the specific intensity λ_l. These compo-

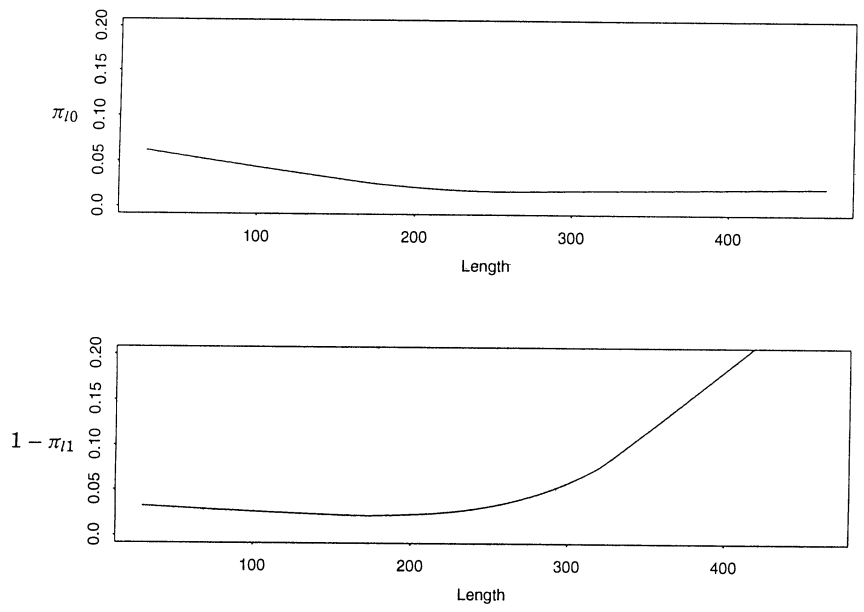

FIG. 8. *Smoothed estimates of error rates* π_{l0} *and* $1-\pi_{l1}$.

nent probabilities can be expressed as

$$\begin{bmatrix} p_0(\theta,\pi) \\ p_1(\theta,\pi) \\ p_2(\theta,\pi) \end{bmatrix} = \begin{bmatrix} (1-\pi_0)^2 & (1-\pi_0)(1-\pi_1) & (1-\pi_1)^2 \\ 2\pi_0(1-\pi_0) & \pi_0(1-\pi_1)+\pi_1(1-\pi_0) & 2\pi_1(1-\pi_1) \\ \pi_0^2 & \pi_0\pi_1 & \pi_1^2 \end{bmatrix} \cdot \begin{bmatrix} q^{2-\theta} \\ 2q\left(1-q^{1-\theta}\right) \\ 1-2q+q^{2-\theta} \end{bmatrix}. \tag{9}$$

Here the dependence of q on l in the case of differing intensities has been suppressed. Note that this reduces to (4) when $\pi_0 = 1-\pi_1 = 0$. Incorporating errors produces the following general expression for $L(D \mid \pi)$:

$$L(D \mid \pi) = \int_0^1 \prod_{\{l \mid d_l = 00\}} \frac{p_{l0}(\theta,\pi)}{p_{l0}(0,\pi)} \prod_{\{l \mid d_l = 10 \vee 01\}} \frac{p_{l1}(\theta,\pi)}{p_{l1}(0,\pi)} \cdot \prod_{\{l \mid d_l = 11\}} \frac{p_{l2}(\theta,\pi)}{p_{l2}(0,\pi)}\, d\theta, \tag{10}$$

which reduces in the obvious way for the trinomial case.

To estimate π_0 and π_1, we could perform a similar analysis on the 433×40 matrix described above, except that we would confine our attention to fragment sizes under 350 base pairs and insist that $\pi_{lk} = \pi_k$, for $l = 30, \ldots, 350$ and $k = 0, 1$. However, such an approach would be limited by, among other things, the adequacy of the clustering algorithm to identify matching peaks. Instead, we used restriction data generated during the mapping process to estimate π_0 and π_1.

Once preliminary contigs have been assembled by methods like those outlined above or in Section 3.3, they can be verified by constructing a *complete restriction map* of the contig. To build a restriction map of a contig, each clone in the contig is digested by a restriction enzyme such as *Eco*RI. The resulting digested clone is then electrophoresed on an ABI362 GeneScanner. As outlined by Lamerdin and Carrano (1993), the ABI362 can resolve fragments from about 350 bases long to over 22,000 bases long with a relative error of less than 3%. This wide range and accurate resolution ensures that, except for extraordinary situations, virtually all of the fragments produced by a six-base enzyme such as *Eco*RI will be resolved.

The result is a list of fragment sizes for each clone. The clones and their fragments are then laid out and permuted and so that identical fragments from overlapping clones are aligned vertically, as shown in Figure 9. The resulting map provides an estimate of the size of each clone in the contig, as well as the amount of overlap between any two clones in the contig. These maps complement our posterior odds data about pairs of clones, and can help us evaluate our assumptions and estimate unknown parameters. To this end, we extracted data of the form $\{(\theta, n_0, n_1, n_2)\}$ for 680 overlapping clones taken from a restriction map of 117 clones encompassing a region of chromosome 19 known as the D19S11 region, and used this data to compute maximum likelihood estimates for π.

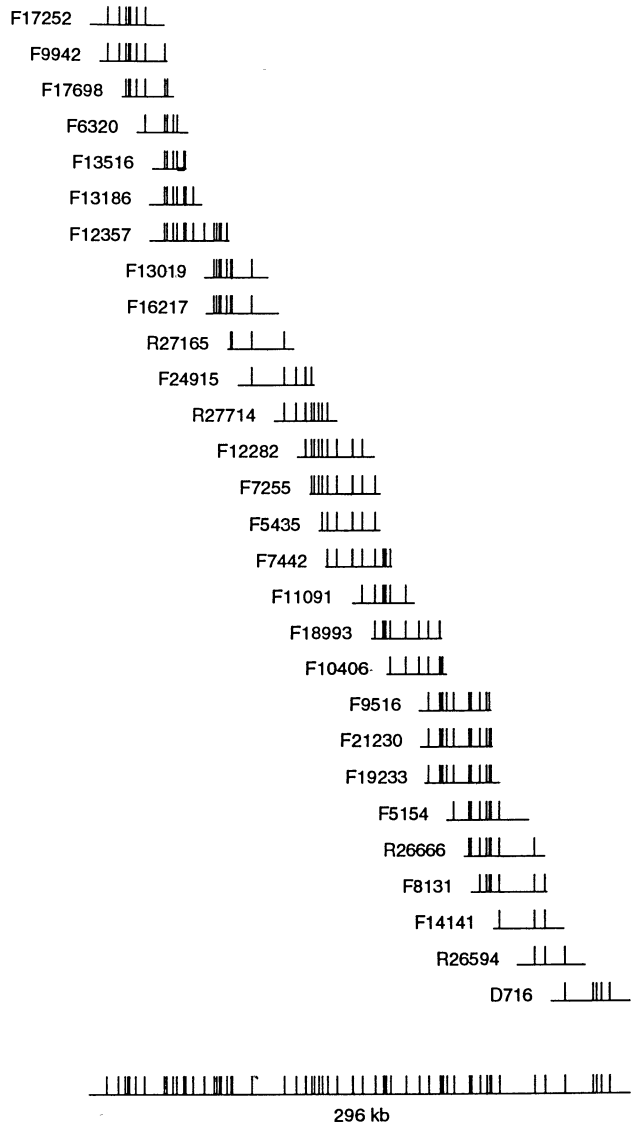

FIG. 9. *Example of a restriction map.*

The mean clone length was 40 kilobases. The θ for each overlapping pair was determined by dividing the amount of overlap by the mean clone length of all clonal inserts in the map. To determine the fragment lengths we would use as probes, we chose fragment lengths under 350 bases which had a hit probability between 0.1 and 0.3, resulting in 240 probes with an average hit probability of 0.18. Thus our data looked like $\{(\theta_i, n_{i0}, n_{i1}, n_{i2}) \mid i = 1, \ldots, 680\}$, where

$$(n_{i0}, n_{i1}, n_{i2}) \sim \text{Trinomial}\big(240, p_0(\theta_i, \pi), p_1(\theta_i, \pi), p_2(\theta_i, \pi)\big)$$

and the $p_j(\theta_i, \pi)$ are defined as in (9). The resulting maximum likelihood estimates were $\widehat{\pi}_0 \approx 0.03$ and $\widehat{\pi}_1 \approx 0.05$.

Evaluating $L(D)$. We have described several alternative expressions for the integrated likelihood ratio $L(D)$ [equations (6), (7) and (10)]. We must evaluate $L(D)$ $\binom{n}{2}$ times, where n is the size of the library. In the case of chromosome 19, we have $n \approx 10{,}000$, resulting in approximately 5×10^7 integrations, the vast majority of which will be less than 1. To speed up this evaluation process, LLNL screens all 5×10^7 pairs of clones by computing $L(D)$ for the simple trinomial model described in (6). Only those pairs with sufficiently high value for this simple model are further evaluated using the full model [equation (10)]. Even with this screening procedure, computing a value for $L(D)$ for all 5×10^7 pairs takes a weekend's worth of effort by more than 20 workstations.

3.3 Assembling Contigs

LLNL currently assembles clones into an initial set of contigs by a sequential merging algorithm designed by Thomas Slezak of the center's Informatics group. This algorithm is similar to a hierarchical clustering algorithm known as single-linkage clustering (Mardia, Kent and Bibby, 1979, page 370). Recall that, with merging hierarchical algorithms, one begins with each clone belonging to its own individual cluster. At each step in the algorithm, one merges the two clusters that are the most similar into a new, larger cluster. Differences between algorithms boil down to differences in what it means for two clusters to be similar. With single linkage clustering, the similarity between cluster A and cluster B is defined to be the maximum similarity between all pairs (a, b) of clones, where $a \in$ A, $b \in$ B. The eventual amount of clustering is defined by a similarity threshold: clustering stops when no two clusters have a similarity measure that exceeds that threshold.

LLNL's clustering algorithm uses log posterior odds as a similarity measure between clones. As in single-linkage clustering, merging stops when the log posterior odds drops below a user-defined threshold. As the log posterior odds and $\log L(D)$ differ only by a constant, the algorithm actually uses $\log L(D)$ as its similarity measure. However, Slezak has extended the basic algorithm described above in a number of ways to take advantage of the biological context in which the clustering is taking place. Most important of these extensions is the construction and maintenance of *minimal tiling paths* for each cluster. These minimal tiling paths provide a best guess of a minimal overlapping set of clones which span the cluster. When two clusters are selected to be merged, their respective tiling paths are examined. If the tiling paths of the two clusters cannot be consistently merged, then the merge of the two clusters is disallowed. On the other hand, if the two tiling paths are consistent, the two clusters are merged and the tiling path for the combined cluster is computed from the

tiling paths of the two merged clusters. Here, "consistency" is defined in a very weak sense, involving a heuristic measure of similarity between tiling path members. When the clustering stops, the algorithm has produced a set of contigs and a set of minimal tiling paths for each contig.

To a first approximation, the algorithm can be viewed as $\binom{n}{2}$ applications of a Bayesian decision rule, where for each pair of clones, we must decide whether or not $\theta > 0$, based on fingerprint data; that is, whenever the loss function is independent of actual amount of overlap, the rule "Decide overlap whenever $\log L(D) > K$" for an appropriate K is a Bayes rule for that loss function. The actual value of K will depend on the prior and the losses entailed by incorrect decisions.

In the case of contig building, the critical issue is to avoid so-called false joins. The consequences of falsely asserting that two contigs of clones should be joined into one contig are much more serious than that of failing to join two contigs that actually overlap. Thus, to a first approximation, we can analyze the decision rule in the classical sense of hypothesis testing: examine the power of the test for a given probability of falsely asserting that two clones overlap. In our case, we will be examining the behavior of the decision rule over a range of alternatives $\theta > \theta_0$. Consequently, the power we explore will be the power with respect to the marginal distribution of D, averaged over the range of alternatives. As is traditional, let us denote the probability of falsely deciding two clones overlap by α, and likewise denote the probability of falsely deciding that two clones do not overlap by β. Note that all the analyses that follow use $L(D)$ as defined by the simple trinomial model without errors [equation (6)].

What value of α is reasonable? Given a library of n clones, we expect to have $\binom{n}{2}\alpha \approx (n^2/2)\alpha$ false positives. In LLNL's case, $n \approx 10^4$, so we expect around $5 \times 10^7\alpha$ false positives. The fact that the expected number of false positives increases as the square of the library size forces us to insist upon values for α very much lower than one typically sees (say, on the order of 10^{-6}). With this harsh reality in mind, let us examine the performance of $L(D)$.

First, we examine the effect of the probability of a positive probe hit p on the power to detect any overlap: $\theta = 0$ versus $\theta > 0$. Fu, Timberlake and Arnold (1992) analyzed hybridization approaches to overlap detection from a hypothesis-testing point of view and showed results indicating that p in the range (0.4, 0.6) was optimal for detecting overlap when the test statistic was simply the total number of agreements $n_{00} + n_{11}$, which we denote by $T(D)$. In Figure 10, we compare the power of the integrated likelihood ratio $L(D)$ to that of $T(D)$ for $\alpha = 10^{-2}$, 10^{-3}, 10^{-4} and 10^{-5}. In all cases, the number of probes was 120.

Several features stand out. First, the integrated likelihood ratio decision rule $L(D)$ seems to outperform $T(D)$ for any α and p. This is not unexpected, since $L(D)$ is the Bayes rule for any "all-or-nothing" loss function. Also, the probability p that performs the best for $L(D)$ seems to be somewhat lower than that for $T(D)$: between 0.2 and 0.3. Finally, the two approaches have equal power at $p = 0.5$. This is also not unexpected, for in that situation the trinomial probabilities in (4) reduce to binomial probabilities, the number of matches $n_{00} + n_{11}$ becomes a sufficient statistic and the two approaches are identical. To summarize, then, if we use $L(D) > K$ as our decision rule, the optimal values for p seem to range between about 0.2 and 0.3.

In the above paragraphs, we examined the power to detect an overlap averaged over all $\theta > 0$. Detecting small overlaps may be too difficult. Now, let us see how the procedure detects overlaps above a certain amount: $\theta = 0$ versus $\theta > \theta_0$.

In Figure 11, we see a form of ROC curves for detecting overlaps of various sizes ($\theta > 0, 0.2, 0.5, 0.8$) and three probabilities of positive probe hits ($p = 0.1, 0.3, 0.5$). All curves are for 200 probes. Note that these curves differ from traditional ROC curves only in their vertical axis, which shows the probability of a Type II error rather than power.

In addition, we note that these graphs show that the power relationship between p and θ is not uniform. As our requirements loosen and the minimum amount of overlap we are required to detect

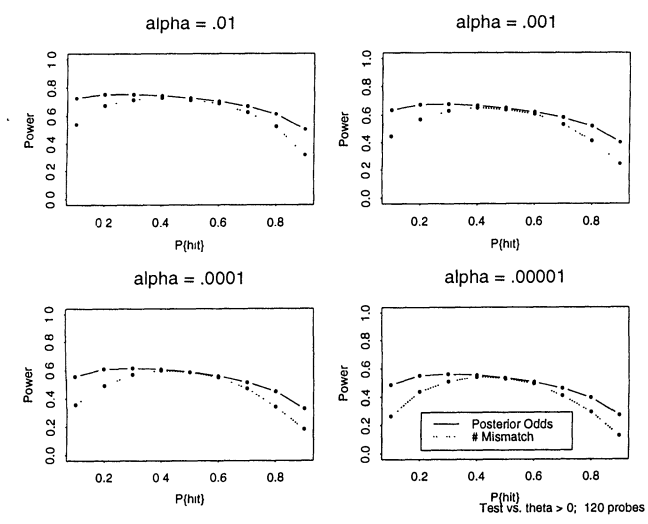

FIG. 10. *Power of posterior odds compared to number of mismatches to detect overlap.*

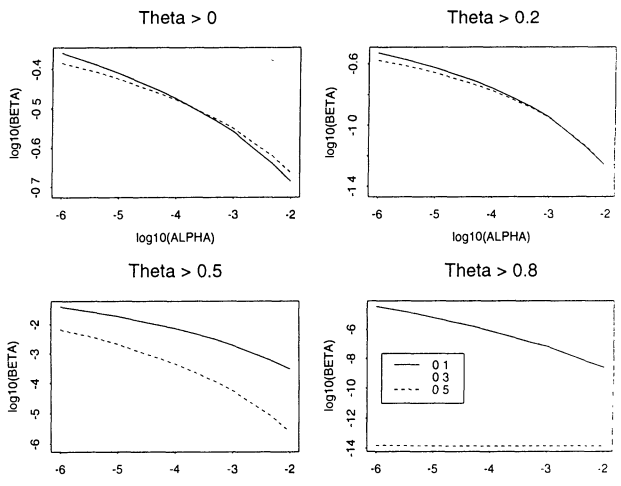

FIG. 11. *β error versus α-error for various θ and p (200 probes).*

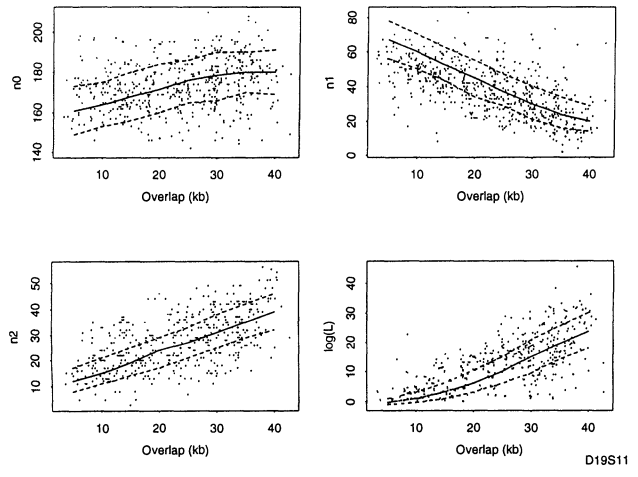

FIG. 12. *Observed versus expected for various statistics for the* D19S11 *region.*

grows larger, the higher probability probes become much more informative.

3.4 Discussion

It is instructive to examine the observed versus expected distribution of different statistics, assuming various amounts of overlap. In Figure 12, we see scatter plots of all 680 values of n_0, n_1, n_2 and $\log L(D \mid \widehat{\pi})$, plotted against the estimated amount of overlap. Superimposed upon the scatter plots are the expected 10, 50 and 90% percentiles for the statistics under the trinomial model. These percentiles were separately computed for overlaps of 5, 10, ..., 40 kilobases by simulating random draws of overlapping clones, computing the appropriate trinomial probabilities and then simulating a trinomial random variable. For each amount of overlap, 500 clone lengths were drawn as independent, identically distributed random values from a gamma distribution with mean of 40 and variance of 25, subject only to the condition that they exceeded the amount of overlap.

One feature is immediately apparent: the values of all statistics are tremendously over-dispersed, relative to their expected distributions, even after compensating for unequal clone lengths. In addition, there is a slight bias upward in $L(D)$. Some of this may be due to nonindependence between fragment sizes. We are continuing to investigate the exact source of this unexpected overdispersion. The D19S11 region is unusual in that it contains multiple copies of a repeated DNA sequence. Consequently, more fragments will match than would otherwise be the case. However, this slight bias is still evident when other restriction maps are similarly evaluated.

In addition to attempting to understand the problem with overdispersion, we are planning other future analyses aimed at further optimizing the approach. For instance, it is clear that the completely general calculation in (10) is quite involved. Is it worth the effort, or will a simple trinomial model do just as well? In addition, we are exploring likelihood calculations for configurations involving more than two clones (Nelson, Speed and Yu, 1994) to see if they can improve the contig-building process.

4. OTHER TOPICS AND OPEN PROBLEMS

The field of physical mapping is evolving extraordinarily rapidly. From less than a decade ago, when the first tentative steps toward large-scale high resolution physical mapping were being taken, map construction techniques have progressed to the point where recently Cohen, Chumakov and Weissenbach (1993) published a coarse, "first-generation" physical map in YAC's of the entire human genome. During that time, popular approaches to mapping have evolved in response to advances in technology. New cloning vectors such as YAC's made cloning larger fragments of DNA possible, and still more recent vectors like PAC's (Ioannou et al., 1994) raise the prospect of cloning large inserts without chimerism and deletion. New probing technologies like PCR have made possible the reliable detection and construction of unique segments of DNA from extremely small samples. Recently, *fluorescence in situ hybridization* (FISH) techniques (Lichter et al., 1990; Lawrence, Singer and McNeil, 1990; Brandriff, Gordon and Trask, 1991) have made possible the direct microscopic analysis of ordering and distance relationships among clones (Brandriff et al., 1994). Such advances in technology bring with them new

problems in experimental design and data analysis, and new opportunities for statistics to play a role. We now cite just a few examples.

High resolution mapping using FISH is a very recent development. In this method, two or more fluorescent probes are hybridized to chromosomes, at a particular stage in the cell cycle, in many different cells. The distance (in microns) between the fluorescent dots in each cell is measured. (Another use of FISH involves mapping a probe to a chromosomal band. In this case, only one probe is used, and the chromosome is dyed to reveal its bands.) There are many variations on this basic theme, involving the number and type of different probes, the number of different fluorescent dyes and the stage in the cell cycle. The data from a FISH experiment consists of a series of distance measurements between two or more probes. Data from multiple experiments involving multiple probes must be integrated into a set of ordering relationships among the probes. How should one model the randomness in this data? Van den Engh, Sachs and Trask (1992) have suggested that, under certain conditions, the path between probes can be modeled as a Brownian motion. However, other more recent data (Brandriff et al., 1994) strongly suggests that this model does not hold for very high resolution data. If statistical approaches are to be used to plan mapping experiments and to evaluate the resulting uncertainty in FISH maps, good analytic models for the configuration of chromosomes in nuclei must be developed.

More generally, methods are required which recognize the pervasiveness of experimental error and can quantify the resulting imprecision in generated maps. In all of the methods outlined in Section 2, statistical modeling is applied only to the simplest situation. Can one construct effective models that can be used to evaluate more complex structures such as contigs and entire maps? We would need realistic models, good approximations to the likelihood for such models and some method such as Markov chain Monte Carlo to explore the posterior distribution of contigs or maps, given the data. In ordinary hybridization experiments, for instance, the error processes are quite complex. Cosmids are typically arrayed on filters, with perhaps 10^5 cosmids per filter. The filters are exposed to a complicated experimental protocol involving the desired probe. The result of the experiment is typically a gray-scale image of the entire filter, with dark dots corresponding to cosmids which hybridized successfully to the probe, and light (or nonexistent) dots otherwise. This image is then usually reduced to a clone–probe incidence matrix for use by map assembly routines. Error models for the hybridization process would enable a suitable map assembly program to take account of the quality of the data in constructing the map. In this way, published maps could consist of not only the "best" map, but annotations indicating weaknesses in the map, as well as some indication of likely alternatives. Such additional data would be useful to subsequent researchers who need to know how much they can depend on particular attributes of a published map.

In summary, methods for physical mapping have progressed to the point where a variety of approaches have been applied with varying degrees of success to moderate-to-large regions of the human genome, at a number of different resolutions. Future issues will revolve not around feasibility, but rather designing efficient and economical mapping techniques in the face of a rapidly changing technology and emerging low resolution genomic maps. Future challenges to statisticians will more likely focus on evaluating and integrating different mapping alternatives, tailoring an approach to match the goals of the project and the strengths and interests of the laboratory. In addition to this experimental design role, statistics is uniquely suited to provide needed analytic techniques to integrate information from various maps as well as to provide map consumers with some measure of uncertainty about the maps they use.

ACKNOWLEDGMENTS

Research by D. O. Nelson was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract number W-7405-ENG-48, with additional support from NSF Grant DMS-91-13527. Research by T. P. Speed was partially supported by NSF Grant DMS-91-13527.

REFERENCES

ALIZADEH, F., KARP, R. M., NEWBERG, L. A. and WEISSER, D. K. (1993). Physical mapping of chromosomes: a combinatorial problem in molecular biology. *Algorithmica*. To appear.

ARRATIA, R., LANDER, E. S., TAVARÉ, S. and WATERMAN, M. S. (1991). Genomic mapping by anchoring random clones: a mathematical analysis. *Genomics* **11** 806–827.

ASLANIDIS, C., JANSEN, G., AMEMIYA, C., SHUTLER, G., ET AL. (1992). Cloning of the essential myotonic dystrophy region and mapping of the putative effect. *Nature* **355** 548–551.

BALDING, D. J. (1994). Design and analysis of chromosome physical mapping experiments. *Philos. Trans. Roy. Soc. London Ser. B* **334** 329–335.

BALDING, D. J. and TORNEY, D. C. (1991). Statistical analysis of DNA fingerprinting data for ordered clone physical mapping of human chromosomes. *Bull. Math. Biol.* **53** 853–879.

BARILLOT, E., DAUSSET, J. and COHEN, D. (1991). Theoretical analysis of a physical mapping strategy using random single-copy landmarks. *Proc. Nat. Acad. Sci. U.S.A.* **88** 3917–3921.

BAXENDALE, S., MACDONALD, M., MOTT, R., FRANCIS, F., ET AL. (1993). A cosmid contig and high resolution restriction map of the 2 megabase region containing the Huntington's disease gene. *Nature Genetics* **4** 181–186.

BELLARME-CHANTELOT, C., LACROIX, B., OUGEN P., BILLAULT, A., ET AL. (1992). Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* **70** 1059–1068.

BOOTH, K.S. and LUEKER, G. S. (1976). Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. System Sci.* **13** 335–379.

BRANDRIFF, B. F., GORDON, L. A., FERTITTA, A., OLSEN, A. S., ET AL. (1994). Human chromosome 19p: a fluorescence *in situ* hybridization map with genomic distance estimates for 79 intervals spanning 20 Mb. Unpublished manuscript.

BRANDRIFF, B., GORDON, L. and TRASK, B. (1991). A new system for high-resolution DNA sequence mapping in interphase pronuclei. *Genomics* **10** 75–82.

BRANSCOMB, E., SLEZAK, T., PAE, R., GALAS, D., ET AL. (1990). Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries. *Genomics* **8** 351–366.

BROOK, J. D., MCCURRACH, M. E., HURLEY, H. G., BUCKLER, A. J., ET AL. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell* **68** 799–808.

BROWN, T. A. (1990) *Gene Cloning: An Introduction*, 2nd ed. Chapman and Hall, New York.

BUXTON, J., SHELBOURNE, P., DAVIES, J., JONES, C., ET AL. (1992). Characterization of a YAC and cosmid contig containing markers tightly linked to the myotonic distrophy locus on chromosome 19. *Genomics* **13** 526–531.

CARRANO, A.V., JONG, P. D., BRANSCOMB, E., SLEZAK, T. and WATKINS, B. (1989). Constructing chromosome and region specific cosmid maps of the human genome. *Genome* **31** 1059–1065.

CHUMAKOV, I., RIGAULT, P., GUILLOU, S., OUGEN, P., ET AL. (1992). Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359** 380–387.

COHEN, D., CHUMAKOV, A. and WEISSENBACH, J. (1993). A first-generation physical map of the human genome. *Nature* **366** 698–701.

COLLINS, F. and GALAS, D. (1993). A new five-year plan for the U.S. Human Genome Project. *Science* **262** 43–46.

COULSON, A., SULSTON, J., BRENNER, S. and KARN, J.(1986). Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Nat. Acad. Sci. U.S.A* **83** 7821–7825.

CUTICCHIA, A. J., ARNOLD, J. and TIMBERLAKE W. E. (1992). The use of simulated annealing in chromosome reconstruction experiments based on binary scoring. *Genetics* **132** 591–601.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

EWENS, W. J., BELL, C. J., DONNELLY, P. J., DUINN, P., MATALLANA, E. and ECKER J. R. (1991). Genome mapping with anchored clones: theoretical aspects. *Genomics* **11** 799–805.

FELLER, W. (1948). On probability problems in the theory of counters. *Courant Anniversary Volume* 105–115.

FOOTE, S., VOLLRATH, D., HILTON, A. and PAGE, D. C. (1992). The human Y chromosome: overlapping DNA regions spanning the euchromatic region. *Science* **258** 60–66.

FRITSCH, F. N. and CARLSON, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.* **2** 238–246.

FU, Y.-X., TIMBERLAKE, W. E. and ARNOLD, J. (1992). On the design of genome mapping experiments using short synthetic oligonucleotides. *Biometrics* **48** 337–359.

GREEN, E. D. and GREEN, P. (1991). Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods and Applications* **1** 77–90.

GRIGORIEV, A. V. (1993). Theoretical predictions and experimental observations of genomic mapping by anchoring random clones. *Genomics* **15** 311–316.

HAMES, B. D. and HIGGINS, S. J. eds. (1985). *Nucleic Acid Hybridisation: A Practical Approach*. IRL Press, Oxford.

HARPER, P. (1979). *Myotonic Dystrophy*. Saunders, Philadelphia.

HOHEISEL, J. D., MAIER, E., MOTT, R. M., MCCARTHY, L., ET AL. (1993) High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast *S. pombe*. *Cell* **73** 109–120.

IOANNOU, P. A., AMEMIYA, C. T., GARNES, J., KROISEL, P. M., ET AL. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genetics* **6** 84–89.

JANSEN, G., DEJONG, P. J., AMEMIYA, C., ASLANIDIS, C., ET AL. (1992). Physical and genetic characterization of the distal segment of the myotonic dystrophy area on 19q. *Genomics* **13** 509–517.

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.

KORNELUK, R. G., MACKENZIE, A. E., NAKAMURA, Y., DUBÉ, I., ET AL. (1989). A reordering of human chromosome 19 long-arm DNA markers and identification of markers flanking the Myotonic Dystrophy locus. *Genomics* **5** 596–604.

LAMERDIN, J. E. and CARRANO, A. V. (1993). Automated fluorescence-based restriction fragment analysis. *BioTechniques* **15** 294–300.

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

LAWLER, E. L., ed. (1985). *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley, New York.

LAWRENCE, J. B., SINGER, R. H. and MCNEIL, J. A. (1992). Interphase and metaphase resolution of different distances within the human dystrophin gene. *Science* **249** 923–932.

LICHTER, P., CHANG TANG, C., CALL, K., HERMANSON, G., ET AL. (1990). High-resolution mapping of human chromosome 11 by *in situ* hybridization with cosmid clones. *Science* **247** 64–69.

MACDONALD, M. E., AMBROSE, C. M., DUYAO, M. P., MYERS, R. H., ET AL. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on the Huntington's disease chromosome. *Cell* **72** 971–983.

MAHADEVAN, M., TSILFIDIS, C., SABOURIN, L., SHUTLER, G., ET AL. (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255** 1253–1255.

MAHADEVAN, M., S., AMEMIYA, C., JANSEN, G., SABOURIN, L., ET AL. (1993). Structure and genomic sequence of the myotonic dystrophy (DM kinase) gene. *Human Molecular Genetics* **2** 299–304.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic, New York.

MICHIELS, F., CRAIG, A. G., ZEHETNER, 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.

MIZUKAMI, T., CHANG, W. I., GARKAVTSEV, I., KAPLAN, N., ET AL. (1993). A 13 kb resolution cosmid map of the 14Mb fission yeast genome by nonrandom sequence-tagged site mapping. *Cell* **73** 121–132.

MOTT, R., GRIGORIEV, A., MAIER, E., HOHEISEL, J., ET AL. (1993). Algorithms and software tools for ordering clone libraries—application to the mapping of the genome of *S. pombe*. *Nucleic Acids Research* **21** 1965–1974.

NAVIDI, W. and ARNHEIM, N. (1994). Analysis of genetic data from the polymerase chain reaction. *Statist. Sci.* **9** 320–333.

NELSON, D. O., SLEZAK, T., BRANSCOMB, E. W. and CARRANO, A. V. (1989). Robust methods for signal extraction and calibration in restriction fingerprints. In *Human Genome Program Contractor-Grantee Workshop I.*

NELSON, D. O. and SPEED, T. P. (1994). Predicting progress in directed mapping projects. *Genomics.* In press.

NELSON, D. O., SPEED, T. P. and YU, B. (1994). A decision problem in physical mapping. Unpublished manuscript.

NEWBERG, L. A. (1993). Finding, evaluating, and counting DNA physical maps. Ph.D. dissertation, Dept. Electrical Engineering and Computer Science, Univ. California, Berkeley.

OLSON, M. V. (1993). The human genome project. *Proc. Nat. Acad. Sci. U.S.A.* **90** 4338–4344.

OLSON, M. V., DUTCHIK, J. E., GRAHAM, M. Y., BRODEUR, G. M., ET AL. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Nat. Acad. Sci. U.S.A.* **83** 7826–7830.

OLSON, M., HOOD, L., CANTOR, C. R. and BOTSTEIN, D. (1989). A common language for physical mapping of the human genome. *Science* **245** 1434–1435.

PALAZZOLO, M. J., SAWYER, S. A., MARTIN, C. H., SMOLLER, D. A. and HARTL, D. L. (1991). Optimized strategies for sequence-tagged-site selection in genome mapping. *Proc. Nat. Acad. Sci. U.S.A.* **88** 8034–8038.

SHUTLER, G., KORNELUK, R. G., TSILFIDIS, C., MAHADEVAN, M., ET AL. (1992). Physical mapping and cloning of the proximal segment of the myotonic dystrophy gene region. *Genomics* **13** 518–525.

SMITH, W. L. (1957). On renewal theory, counter problems, and quasi-poisson processes. *Proceedings of the Cambridge Philosophical Society* **53** 175–193.

STALLINGS, R. L., DOGGETT, N. A., CALLEN, D., APOSTOLOU, S., ET AL. (1992). Evaluation of a cosmid contig physical map of human chromosome 16. *Genomics* **13** 1030–1039.

STALLINGS, R. L., TORNEY, D. C., HILDEBRAND, C. E., LONGMIRE, J. L., ET AL. (1990). Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Nat. Acad. Sci. U.S.A.* **87** 6218–6222.

STATISTICAL SCIENCES, INC. (1991). *S-PLUS User's Manual.* Statistical Sciences, Inc., Seattle, WA.

TORNEY, D. C. (1991.) Mapping using unique sequences. *Journal of Molecular Biology* **217** 259–264.

TSILFIDIS, C., MACKENZIE, A. E., SHUTLER, G., LEBLOND, S., ET AL. (1991). D19S51 is closely linked with and maps distal to the Myotonic Dystrophy locus on 19q. *American Journal of Human Genetics* **49** 961–965.

VAN DEN ENGH, G., SACHS, R. and TRASK, B. J. (1992). Estimating genomic distance from DNA sequence location in cell nucleii by a random walk model. *Science* **257** 1410–1412.

WATSON, J. D., HOPKINS, N. H., ROBERTS, J. W., STEITZ, J. A., ET AL. (1987). *Molecular Biology of the Gene,* 4th ed. Benjamin/Cummings, Menlo Park, CA .

ZHANG, M. Q. and MARR, T. G. (1993). Genome mapping by nonrandom anchoring: a discrete analysis. *Proc. Nat. Acad. Sci. U.S.A.* **90** 600–604.