On the Consistency of a Physical Mapping Method to Reconstruct a Chromosome in Vitro

Momaio Xiong,∗ Hubert J. Chen,† Rolf A. Prade,‡ Yuhong Wang,§ James Griffith,¶ William E. Timberlake#, and Jonathan Arnold$∗$

∗Department of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089, †Department of Statistics, University of Georgia, Athens, Georgia 30602, ‡Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, Oklahoma 74078, §Department of Molecular Biology, Jilin University, Chang Chun 130023, ¶Department of Genetics, University of Georgia, Athens, Georgia 30602, and $Myco Pharmaceuticals, Inc., Cambridge, Massachusetts 02139

Manuscript received February 6, 1995
Accepted for publication September 15, 1995

ABSTRACT

During recent years considerable effort has been invested in creating physical maps for a variety of organisms as part of the Human Genome Project and in creating various methods for physical mapping. The statistical consistency of a physical mapping method to reconstruct a chromosome, however, has not been investigated. In this paper, we first establish that a model of physical mapping by binary fingerprinting of DNA fragments is identifiable using the key assumption—for a large randomly generated recombinant DNA library, there exists a staircase of DNA fragments across the chromosomal region of interest. Then we briefly introduce epi-convergence theory of variational analysis and transform the physical mapping problem into a constrained stochastic optimization problem. By doing so, we prove epi-convergence of the physical mapping model and epi-convergence of the physical mapping method. Combining the identifiability of our physical mapping model and the epi-convergence of a physical mapping method, finally we establish strong consistency of a physical mapping method.

One goal of the Human Genome Project is to create detailed maps of the human genome and of several other model organisms (Collins and Galas 1993) as a prelude to determining their entire DNA sequence. The size of these genomes vary from a few million base pairs of DNA in a bacterial genome to over 3 billion base pairs in a mammalian genome. In the last several years considerable progress has been made in mapping a diverse array of organisms ranging from bacteria to humans (Coulson et al. 1986; Olson et al. 1986; Kohara et al. 1987; Azevedo et al. 1993; Cohen et al. 1993; Eiglmair et al. 1993; Hoheisel et al. 1993; Mizukami et al. 1993; Cai et al. 1994). This progress has been made possible by a variety of new recombinant DNA methodologies (Davies and Tilghman 1990; Billings et al. 1991), which include cloning and automated sequencing of DNA fragments. The power of this technology for generating mapping and sequencing data has fundamentally shifted the problem from one of collecting the mapping data to the inference problem of assembling the maps. It is the latter problem, which is the focus of this article.

A central statistical challenge of the Human Genome Project is assembling a physical map of a whole chromosome from a large number of DNA markers along a chromosome. A physical map represents a partial ordering of distinguishable DNA fragments by their position along a chromosome. While recombinant DNA technology provides a wealth of markers to distinguish DNA fragments for mapping, physical maps have existed since nearly the beginning of genetics. One of the simplest and oldest examples is a cytological map in fruit flies.

There is a wide variety of physical maps including cytological maps, radiation hybrid maps, STS content maps, ordered clone collections (i.e., “contig maps”), restriction maps, and the DNA sequence of an entire chromosome, reflecting the diversity of experimental approaches used to generate them. The focus of this report is on assembling one particular kind of physical map, an ordered clone collection. Even within this narrower class of physical maps, a rich variety of fingerprinting methodologies exist for distinguishing clones in a library for the purpose of ordering this library. Some of these fingerprinting methods include the use of (1) restriction enzymes (Coulson et al. 1986; Olson et al. 1986); (2) restriction enzymes supplemented with Southern hybridization to synthetic L1 and Alu sequences (Stallings et al. 1990; Bellane-Chantelot et al. 1992); (3) sequence-tagged sites (STSs) (Green and Olson 1990a,b; Foote et al. 1992); (4) synthetic oligonucleotides (Craig et al. 1990; Hoheisel et al. 1991, 1993); and (5) low copy number DNA probes (Mizukami et al. 1993; Wang et al. 1994a). The focus of this paper will be on creating an ordered clone collection with fingerprinting methods that utilize “low copy

*Address correspondence to: Jonathan Arnold, Genetics Department, University of Georgia, Athens, GA 30602.

E-mail: arnold@bscr.uga.edu

number” DNA probes that occur once or a few times in a genome [i.e., methods (3) and (5)] because these two methods are being used to create a high resolution physical map of the human genome over the next 5 years. These methods are also being used to create physical maps of a variety of other organisms including Arabidopsis thaliana, Aspergillus fumigatus, Candida albicans, Cryptococcus neoformans, Neurospora crassa, Staphylococcus aureus, and Enterococcus faecium.

Modeling physical mapping experiments statistically and estimating their parameters, including the original order of the clones in the library, are important for planning a physical mapping project, evaluating inference tools for map assembly, and for understanding the data from mapping experiments. In physical mapping by fingerprinting methods (1)–(5) there is a stochastic process by which clones in a library are sampled as well as a stochastic process by which the DNA within each clone is sampled in order to fingerprint that clone. No matter which method of fingerprinting is used, procedures for physical mapping consist mainly of ordering random clones by their fingerprints in much the same way that “call numbers” might be used to order books in a real library. A mathematical description and statistical analysis of these processes have been proposed and used for designing physical mapping experiments, for identifying overlaps between clones, and for ordering clones (Lander and Waterman 1988; Grigoriev and Mironov 1990; Arratia et al. 1991; Balding and Torney 1991; Barillett et al. 1991; Karlin and Macken 1991; Palazzo et al. 1991; Torney 1991; Cuticchia et al. 1992a; Fu et al. 1992; Zhang and Marr 1993; Wang et al. 1994a).

For any particular fingerprinting method, it is not an easy task to order clones in a library. For n clones, the total number of possible orders is n!. Goldstein and Waterman (1987) considered the efficacy of a simulated annealing algorithm in assembling a restriction map. Unfortunately, they demonstrate that under a certain probability model there is an exponentially increasing number of solutions as a function of the length of chromosome segment being mapped with probability 1.

Cuticchia et al. (1992a) examine another approach to ordering clones in a library. Each clone is scored for the presence or absence of target DNA sequences by hybridization to a panel of probes, thereby assigning a digital call number to each clone. The number of differences between a pair of call numbers establishes a distance between clones. They formulated the clonal-ordering problem in terms of minimizing the sum of linking distances between successive clones as a function of their order along a reconstructed chromosome, i.e., the traveling salesman problem (TSP) (Garey and Johnson 1979) and computed a solution to the minimization problem using simulated annealing. Whether or not this estimator of the original order of the clones based on the data from physical mapping experiments is consistent remains an open question and may depend on the identifiability of an underlying physical mapping model.

The objective of this paper is to prove the consistency of an ordering procedure based on a physical mapping criterion, such as total linking distance (Cuticchia et al. 1992a). First, a description of the physical mapping data generated by methods (3) and (5) and a real example of an inferred physical map of an entire chromosome are given in the example section. Second, a statistical model for the physical mapping process and a statistical method for reconstructing the order of clones in a library is developed in the model section. A principal ingredient in the proof of consistency is the “identifiability” of this model. The definition of identifiability of a model and its proof for the physical mapping model are also included in the model section. A nontechnical statement of a consistency result for a physical mapping method is given together with a sketch of its proof and a statement of its biological implications in the consistency section. The problem of ordering clones is then formulated as a stochastic optimization problem under methods, and the problem of proving consistency of clonal ordering methods is translated into showing convergence of an optimal solution of a stochastic optimization problem to its true solution. Under methods a basic concept and a theorem of stochastic optimization are also briefly introduced. This theorem is then used to prove the consistency of a particular estimator of a library’s true clonal order under results. Finally, in the discussion some comments are made, and some directions for future research, indicated.

**EXAMPLE**

Several steps are involved in generating a physical map of a whole chromosome by fingerprinting methods involving the use of low copy number probes [methods (3) and (5)]: (1) intact chromosomal DNA is isolated (Brody et al. 1991); (2) isolated chromosomal DNA is physically sheared or cut by restriction fragments into smaller fragments; (3) these smaller fragments are size-selected by gel electrophoresis into 40 kilobase (kb) fragments (or larger, depending on the size of the target genome); (4) size-selected DNA fragments are inserted into a cloning vector (e.g., Wahl et al. 1987) (i.e., a plasmid, phage, cosmid, or artificial chromosome). The result is a chromosome-specific library of DNA fragments. The number of clones n in a library is designed to be large enough so that with high probability each base pair of chromosomal DNA is represented at least once in the library (i.e., the expected coverage probability is near 1) (Fu et al. 1992). Choice of library size is related to the genome size (N), the cloning vector’s insert size (M), and the expected coverage probability. The resulting library is a sample of size n from the chromosome.

Clones in the library are then fingerprinted to order them. Methods (3) and (5) involve the selection of nearly unique probe sequences scattered at random along the chromosome and hybridizing them to clones in the library (Foot et al. 1992; Hoheisel et al. 1993;
### Physical Mapping Consistently

<table>
<thead>
<tr>
<th>Raw data</th>
<th>Redundant order of clones</th>
<th>Minimal order of probes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probes</td>
<td>Probes</td>
<td>Probes</td>
</tr>
</tbody>
</table>
| NDEKAJFMBGHILC | 13 01 02 03 04 05 06 07 08 09 10 11 12 13 20 21 22 23 24 25 26 27 | A B C D E F G H I J K L M N |}

**Figure 1:** Ordering clones and then probes converts the data matrix into a physical map. The physical map is displayed as a two-way layout with clones down the rows in their order along the chromosome and probes across the columns in their order along the chromosome. A 1 indicates clone/probe hybridization or equivalently, detectable overlap. A period indicates no clone/probe hybridization or equivalently, no detectable overlap.

---

Mizukami et al. 1993; Wang et al. 1994a). The fingerprinting process then involves taking a sample of \( m \) probes and hybridizing them to \( n \) clones in a library. Selecting the number of probes \( m \) is discussed in Fu et al. (1992). The result is a \( n \times m \) binary data matrix (Figure 1), in which 1 indicates hybridization and period, no hybridization. The probes can be randomly sampled from the primary library used for mapping (Hoheisel et al. 1993; Mizukami et al. 1993; Wang et al. 1994a) or from a secondary library with smaller inserts (Foote et al. 1992). In the example of this section probes are sampled from the primary library (Wang et al. 1994a). The resulting collection of \( m \) probes is a means to sample the DNA from each clone.

The resulting data are extremely simple, and the methods above lend themselves to robotic data collection and to computer analysis. Each clone in the library corresponds to a row of the binary data matrix. Each probe sampled corresponds to a column of the data matrix. Contiguous blocks of clones or “contigs” are linked together by their common hybridization pattern to a panel of probes. The physical mapping problem is then to uncover a permutation of the rows and columns that reveals the true ordering of clones and probes along the chromosome. This process is diagrammed in Figure 1. This process is analogous to the job of a librarian, namely examining a clone’s binary call number (i.e., a row in Figure 1) and placing the clone next to other clones with similar call numbers. The distance between a pair of clones can be measured by the number of differences in their digital call numbers. One method of ordering the clones is simply to minimize the sum of the linking distances between adjacent clones in the data matrix by randomly permuting the rows (Cutticchia et al. 1992a). The same process can be applied to the transpose of the binary data matrix to order the columns. The end product of this process is a physical map of a whole chromosome (Figure 2, Wang et al. 1994a).

In this example a panel of 115 probes is hybridized to 593 clones from chromosome IV of a fungus, *Aspergillus nidulans*. This ordered clone collection represents an *in vitro* reconstruction of a 2.9 megabase (Mb) chromosome. The horizontal lines demarcate 31 *contigs*, a sequence of 1 or more clones, each overlapping with its neighbor(s). The vertical lines demarcate 12 *cells*, a sequence of 1 or probes, each overlapping with its neighbors(s). This visualization of a physical map as a two-way layout permits the researcher to see how the
FIGURE 2.—Physical map of chromosome IV in *A. nidulans*. The chromosomal map was constructed by minimizing the total linking distance $D$ with the random cost algorithm (Wang et al. 1994a). Clone names are given in the margin of the map in their inferred order along the chromosome down the rows. Probes are assigned to columns. Clone/probe hybridization is indicated by a 1 on the interior of the matrix and no hybridization, by a period. Each row is the digital call number of a particular clone, indicating hybridization or no hybridization with a particular probe. A contig is defined as a contiguous block of one or more clones, in which each clone overlaps with its neighbor(s). Contigs (and isolated clones) are demarcated by horizontal lines. A cell is defined as a contiguous block of one or more probes, in which each probe is linked to its neighbor(s) by intervening clones. Cells are demarcated by vertical lines. The number of differences in the digital call numbers between cosmid $c_a$ and the next neighboring cosmid $c_{a+1}$ in the map, *i.e.*, the pairwise Hamming distance $d(c_a, c_{a+1})$, is also given in the margin. An electronic version of the map is available by email request to arnold@bscr.uga.edu.

probes (*i.e.*, the hybridization data) support the physical ordering of clones.

MODEL

Here a framework for statistical analysis of physical mapping experiments by binary fingerprinting of clones is presented. The problem of ordering clones in a chromosome-specific library by minimizing the total linking distance is investigated. The identifiability of the model obtained by minimizing the total linking distance is also investigated. The identifiability of the model for a physical mapping experiment is defined and proven. A mathematically equivalent formulation can be given for physical...
mapping with oligonucleotide probes \([i.e., \text{fingerprinting method (4)}]\) in \text{Hohise} et al. (1993).\

**Model for physical mapping by binary fingerprinting:** A simple model from Fu et al. (1992) and Arratia et al. (1991) is now described for the above physical mapping experiment. There are three assumptions in this model.

**Assumption 1:** Each clone has a constant length \(M\).

The number of clones \(n\) is modeled as a random variable. Each possible chromosomal DNA fragment has the same probability of being included in the clonal library and is cut independently from the chromosome.

The first assumption is tantamount to assuming the library is a random sample of DNA fragments of the same size from the chromosome. There are a number of experimental limitations that may compromise this assumption. For example, certain short DNA sequences in the target genome may be unclonable or may be unstably maintained in a particular cloning vector. Other sequences may encode genes that are lethal to the bacterial or fungal host when they are expressed in high copy number. One experimental solution to these "cloning biases" is to use multiple cloning vectors (Brody et al. 1991) and to vary the insert size. Assumption 1 can be tested statistically and experimentally from Figure 2 and is the subject of another paper (Prade et al. 1995). This assumption does not appear problematical in Figure 2.

The auxiliary assumption of constant insert size \(M\) for a cloning vector is a good one for plasmids, phage, and cosmid cloning vectors because the fragments are size-selected and because there are precise packaging requirements placed upon the insert to enter a phage head, for example. This assumption becomes more problematical with use of artificial chromosomes (Burke et al. 1987; Pierce et al. 1992; Shizuya et al. 1992; Ioannou et al. 1994). Even for these artificial chromosome vectors, it is possible to size-select their inserts (e.g., Cai et al. 1994).

**Assumption 2:** The number of target sites for hybridization by a particular probe along the chromosome is a Poisson process with intensity parameter \(\lambda\).

The number of hybridization sites along a chromosome are rare, although multiple hybridization sites do occur because of the presence of repeated sequences within an insert of a cloning vector. These hybridization sites are also likely to be separated by many kilobases. In a number of organisms dependencies between bases along a DNA sequence appear weak and appear to extend only over short distances of a few base pairs (Phillips et al. 1987; Arnold et al. 1988; Cuticchia et al. 1992b). So, if two DNA fragments are nonoverlapping, it is reasonable to suppose hybridization to each fragment is independent.

There are a number of experimental limitations that may compromise the rarity and independence components of Assumption 2. For example, a probe may con-
tain a repeated sequence, which elevates its frequency of hybridization. These repeated sequences are often associated with structural features of the chromosome, like centromeres, and may be nonuniform in their distribution across a chromosome (Alberts et al. 1994). Chromosomal sequences may be tandemly repeated, as in rDNA genes (Brody et al. 1991), leading to a clustering of hybridization sites. There are a number of experimental steps that can be taken to address these limitations. Libraries can be prescreened with probes of known repeated sequences, and chromosome-specific probes (i.e., probes that uniquely hybridize to one chromosome) can be selected (Pra de et al. 1995).

The intensity parameter can be estimated directly from Figure 2 by counting how often a probe hybridizes to clones in the library. The most problematic aspect of Assumption 2 is the constancy of the Poisson intensity parameter along the chromosome (Karlin and Macken 1991) and across probes. Repeated sequences may be nonrandomly distributed along the chromosome (Brody et al. 1991) or be enriched in certain chromosomal bands, like heterochromatin (Alberts et al. 1994). A probe containing such repeated sequences will not have uniformly distributed hybridization sites or have the same number of hybridization sites. These kinds of inhomogeneities can be tested in part by examining the frequency of hybridizations by each probe in Figure 2. The inhomogeneity of probe hybridization sites can be controlled experimentally in part by use of chromosome-specific probes and prescreening the list of potential probes for repeated sequences. Inhomogeneity in the intensity of hybridization sites can also be experimentally controlled by reducing the size of the hybridization site by only probing with the ends of a clone (Mizukami et al. 1993).

Assumption 3: The hybridization of one probe to any clone is independent of the hybridization of another distinct probe to any clone.

This assumption asserts that the columns of the binary hybridization matrix are independent. Whether or not this assumption is valid will depend in part on the design of the physical mapping experiment (Arratia et al. 1991; Palazzolo et al. 1991; Fu et al. 1992; Zhang and Marr 1992). Some probe sampling schemes involve choosing probes by "sampling without replacement" so that the probes are nonoverlapping, leading to weak dependencies in probe hybridization. Here it is assumed that probes are randomly sampled "with replacement."

There are several experimental limitations that may come into play to compromise this assumption as well. For example, probes may share common hybridization
sites through shared repeated sequences that may lead to a correlated pattern of hybridization. This can be checked by testing the association of the columns in

Figure 2 (Prade et al. 1995) and by examining runs of positives off the diagonal in Figure 2. Experimental steps to insure the validity of this assumption include...
the use of chromosome-specific libraries and using only the ends of clones as probes (MIZUKAMI et al. 1993).

A statistical study of physical mapping begins with the distributional properties of the binary hybridization matrix. In the process of creating physical maps, the binary fingerprints (rows of the data matrix) are com-
pared and a distance computed between clonal fingerprints. A random distance $D_{ij}$ is now defined between clone $i$ and $i'$. Let $m$ be the number of probes, and let $P$ be the length of a probe in base pairs (bp). Define an indicator function for whether or not a probe $j$ hybridizes to clone $i$ as follows:

$$X_{ij} = \begin{cases} 1, & \text{if probe } j \text{ hybridizes to clone } i, \\ 0, & \text{if probe } j \text{ does not hybridize to clone } i. \end{cases}$$

When clones overlap, their pattern of hybridization will be similar because they will share sites of hybridization. The similarity of clones is then measured by

$$S_{ij} = \sum_{j=1}^{m} X_{ij},$$

where

$$S_{ij} = \begin{cases} 1, & \text{if } X_{ij} = X_{i'j}, \\ 0, & \text{otherwise.} \end{cases}$$

The count $S_{ij}$ (summing over probes) is referred to as the pair's similarity score. Define the quantity $D_{ij} = m - S_{ij}$ as the distance between clone $i$ and $i'$. Hence,

$$D_{ij} = \sum_{j=1}^{m} D_{ij},$$

where

$$D_{ij} = \begin{cases} 1, & \text{if } X_{ij} \neq X_{i'j}, \\ 0, & \text{if } X_{ij} = X_{i'j}. \end{cases}$$
The similarity score and distance measure the degree to which two clones overlap.

It is clear that the distance $D_r$ is a random variable. Before calculating the expectation $E(D_r)$, we should find the distribution of the distance $D_r$ between clones in order to study the asymptotic behavior of the physical mapping criterion in (15). There are two cases to consider: (i) clone $i$ and $i'$ do not overlap or (ii) clones $i$ and $i'$ do overlap. These two cases are discussed separately.

(i) Nonoverlapping pair of clones: By assumption 2 the number of hybridization sites within a clone of length $M$ has a Poisson distribution with mean $\lambda M$. Since the effective length of a clone for which detecting the hybridization of a probe with a clone is possible, is $(M - P + 1)_+$, where $X_+ = \max(0, X)$, the probabilities of hybridization or no hybridization are given by:

$$P[X(j) = 0] = \exp(-\lambda (M - P + 1)_+)$$

$$P[X(j) = 1] = 1 - \exp(-\lambda (M - P + 1)_+) = f(M).$$

(6)

Since the two clones are nonoverlapping, the probe then hybridizes (or does not hybridize) independently to a least one site within each clone, and the event that the hybridization is the same or different between clones occurs with constant probability. The probability that hybridization is the same ($\rho_0$) or different ($\rho_1$) is given by:

$$p_0 = P[S_r(j) = 1] = f(M)^2 + (1 - f(M))^2$$

$$p_1 = P[S_r(j) = 0] = 2f(M)(1 - f(M)).$$

(7)

By assumption 2, each probe hybridizes independently so that the distance $D_r$ has a binomial distribution:

$$P[D_r = d] = \binom{m}{d} p_0^d p_1^{m-d}, d = 0, 1, \ldots, m.$$  

(ii) Overlapping pair of clones: The degree of overlap between clone $i$ and $i'$ is a random variable $K$ measured in base pairs (bp) with realized value $k$. Suppose that the two clones $i$ and $i'$ overlap by $k$ (≥ 0) bp so that the overlap $k$ is given and so that the minimal detectable overlap is defined as $k_0 = (k + P + 1)_+$. Then the conditional probability that hybridization is the same ($\rho_0$) or different ($\rho_1$), given $K = k$, is as shown:

$$p_k = P[S_r(j) = 1 | K = k] = (1 - f(k))$$

$$\times [(1 - f(M - k_0)^2 + f(M - k_0)^2] + f(k)$$

and

$$q_k = P[S_r(j) = 0 | K = k] = 2(1 - f(k))(1 - f(M - k_0))f(M - k_0).$$

(9)

Again because of the Poisson distribution assumption and the independence of probe hybridizations, the distance between a pair of clones has a binomial distribution, $B(m, p_k)$, with changed probability parameter $p_k$:

$$P[D_r = d | K = k] = \binom{m}{d} q_k^d p_k^{m-d}, d = 0, 1, \ldots, m.$$  

(10)

Hence, under the assumption that there is an overlap between clone $i$ and $i'$, the probability density of $D_r$ is a mixture of binomial distributions:

$$P[D_r = d | K \geq 1] = E[P[D_r = d | K = k]].$$

(11)

$$P[D_r = d | K \geq 1] = \sum_{k=0}^{K} P[D_r = d | K = k] P[K = k].$$

(12)

From a statistical point of view, our goal is to estimate...
the original order of clones in the library based on the observed distances \(d_{il}, i, i' = 1, \ldots, n\) between clones. These distances give us information about the overlaps \((k_{il}, i, i' = 1, \ldots, n - 1)\) between pairs of clones via (7)-(14), enabling us to place clones relative to each other on the chromosome. In some cases the ordering information will be partial when clones belong to different contigs, i.e., a contiguous block of clones, in which each clone overlaps with its neighbor(s). If the overlap \(k_{il}\) between two clones is less than the probe length \(P\), i.e., \(k_{il} \leq P - 1\), clones \(i\) and \(i'\) will be considered nonoverlapping, and we will set \(k_{il} = P - 1\) in the analysis below.

One method for estimating the order of the clones is by minimizing the total linking distance \(D\) across the collection of clones (Cuticchia et al. 1992a). Minimum linking distance estimation is analogous to least squares in statistics. Clones are assigned ID numbers \(i\) or \(i' = 1, \ldots, n\), which may indicate, for example, a grid location on a particular plate in a clonal library stored in a freezer. (The ID number provides no information about a clone’s physical location on the chromosome.) In addition, clones have an inferred order along the chromosome, \(t = 1, \ldots, n\). The clone in position \(t\) in the order has ID, \(i\). For example, clones \(c_1, \ldots, c_n\) might be in the following order along the chromosome, \((c_1, \ldots, c_n) = (c_2, c_3, c_5, c_4, c_1)\). Assume that the inferred order of clones is \([i_1, \ldots, i_n]\). Then the total linking distance \(D\) is defined as

\[
D = \sum_{l=1}^{n-1} D_{i_l i_{l+1}}. 
\]  

Our goal is to find an order \([i_1^0, \ldots, i_n^0]\) such that

\[
E[D^0] = \sum_{l=1}^{n} E[D_{i_l i_{l+1}^0}] = \text{Min} \sum_{l=1}^{n-1} E[D_{i_l i_{l+1}}]. 
\]  

This is a stochastic traveling salesman problem, which can be solved by resorting to stochastic approximation. The rationale for this criterion is that two clones, \(i_1\) and \(i_{l+1}\), with smaller expected distance \(E[D_{i_l i_{l+1}}]\) between them will have a greater overlap \(k_{i_l i_{l+1}}\) between them. The criterion in (16) is used to select an order so that adjacent clones have maximum overlap.

**Identiﬁability:** Identiﬁability of a model is an extremely important property and principal ingredient in a consistency proof. Goldstein and Waterman (1987) pointed out that under a certain probability model, the multiple digest problem for restriction mapping has an exponentially growing number of solutions as a function of the segment being mapped. In general the model for restriction mapping by multiple digests is nonidentifiable. It is natural to ask whether or not there are any restrictions on the physical mapping problem by fingerprinting which ensure a unique solution, i.e., whether or not the physical mapping probe by fingerprinting is identifiable. The following theorem will answer this question with Assumption 4.

**Assumption 4. (Monotonicity of overlaps with physical distance):** Denote the overlap between clones \(i\) and \(i'\) by \(k_{ii}\). If \(k_{ii} \leq P - 1\), then clones \(i\) and \(i'\) are viewed as nonoverlapping, and let \(k_{ii} = P - 1\). Assume that the original order of clones which form the chromosome are either, as shown in Figure 3, or as shown in Figure 4, and

\[
k_{12} > k_{13} > \cdots > k_{1w}, \quad k_{25} > k_{24} > \cdots > k_{2m}
\]

\[
k_{n-2, n-1} > k_{n-2, w}
\]

This assumption is a direct consequence of the linearity of DNA. For a large recombinant DNA library generated from randomly sheared genomic DNA, there is a staircase of cloned DNA fragments that span a chromosomal region of interest. The biological justification for this assumption rests with the requirement that an in vitro reconstruction of a chromosome, i.e., a physical map, should be a faithful model of the original chromosome in vivo and logically consistent with current theories about the structure of DNA (Alberts et al. 1994). Assumption 4 is tantamount to assuming that a physical map look like our conception of a real chromosome.

In practice there are a number of experimental limitations that may compromise this staircase or ladder property. The most common one is clones may have identical digital fingerprints, and in this case there will be no information about the degree of overlap of clones with “identical call numbers” next to each other on a “contig map.” Even a pair of adjacent clones with nonidentical fingerprints may not satisfy this monotonicity property, if an adjacent clone shares an identical call number with another clone on the contig map. However, clones that do not extend the physical map maximally up the ladder are “redundant,” and a variety of experimental strategies are used to strike them from the clone collection to create a “minimum tiling of the chromosome.” One strategy involves the estimation of overlap between pairs of clones and the elimination of clones that are “redundant” and do not conform to the monotonicity condition in Assumption 4.

For example, there are new cloning vectors, like the pDUAL family (Strausbaugh et al. 1990; Wang et al. 1993a,b) that allow the estimation of overlap between adjacent cosmid clones on a contig map. These new

\[
\begin{array}{c}
1 \\
2 \\
\vdots \\
n \\
\end{array}
\]

This assumption is a direct consequence of the linearity of DNA. For a large recombinant DNA library generated from randomly sheared genomic DNA, there is a staircase of cloned DNA fragments that span a chromosomal region of interest. The biological justification for this assumption rests with the requirement that an in vitro reconstruction of a chromosome, i.e., a physical map, should be a faithful model of the original chromosome in vivo and logically consistent with current theories about the structure of DNA (Alberts et al. 1994). Assumption 4 is tantamount to assuming that a physical map look like our conception of a real chromosome.

In practice there are a number of experimental limitations that may compromise this staircase or ladder property. The most common one is clones may have identical digital fingerprints, and in this case there will be no information about the degree of overlap of clones with “identical call numbers” next to each other on a “contig map.” Even a pair of adjacent clones with nonidentical fingerprints may not satisfy this monotonicity property, if an adjacent clone shares an identical call number with another clone on the contig map. However, clones that do not extend the physical map maximally up the ladder are “redundant,” and a variety of experimental strategies are used to strike them from the clone collection to create a “minimum tiling of the chromosome.” One strategy involves the estimation of overlap between pairs of clones and the elimination of clones that are “redundant” and do not conform to the monotonicity condition in Assumption 4.

For example, there are new cloning vectors, like the pDUAL family (Strausbaugh et al. 1990; Wang et al. 1993a,b) that allow the estimation of overlap between adjacent cosmid clones on a contig map. These new

\[
\begin{array}{c}
1 \\
2 \\
\vdots \\
n \\
\end{array}
\]
vectors involve the use of transposons. Transposon γδ (and its engineered derivatives) moves more efficiently and more randomly than other known elements in the host bacterium Escherichia coli. The ability of γδ to form adjacent deletions by intramolecular transposition enables a simple means to carry out outer estimation. Members of the "pDUAL" family of cosmid cloning vectors use an engineered γδ transposon to generate nested deletions along the cosmid insert (Wang et al. 1993a,b; Berg et al. 1995). The deletions are easy to select, and evidence to date indicates that the γδ site of insertion is only weakly dependent on sequence context (Strausbaugh et al. 1990; Wang et al. 1993a,b; Berg et al. 1993a,b, 1995). The deletion endpoints are easily selected or mapped by sizing deletion constructs on an agarose gel. Probing an adjacent clone with the ends of these size-selected deletion constructs produces an experimental estimate of the overlap between two adjacent clones on the contig map. This is one of several strategies used to select for the desired "ladder property" for an ordered clone collection.

With this fourth assumption the identifiability of the physical mapping model can be established in the following theorem.

**Theorem 1. (Identifiability theorem):** Under Assumptions 1, 2, 3, and 4, and the further assumption that the length of overlap between two clones is larger than the length of a probe, the model of physical mapping is identifiable as defined in (16), i.e., there exist unique orders [i1, ..., iL] and [i1, ..., iL] such that

\[ E[D] = \sum_{l=1}^{L-1} E[D_{i_l i_{l+1}}] = \min_{[i_1, ..., i_L]} \sum_{l=1}^{L-1} E[D_{i_l i_{l+1}}]. \]  

(17)

The proof of this theorem can be found in Xiong (1993) as well as on the World Wide Web at address http://fungus.genetics.uga.edu:5080.

An order estimation problem (16) based on the total linking distance is a stochastic optimization problem. The distribution of the linking distance \( D_{xy} \) will be approximated by the empirical distribution of \( D_{xy} \). Suppose that the physical mapping experiment by fingerprinting \( n \) clones is replicated \( L \) times and that the observed distances between clones are \( [D_{xy}^{(l)}, i = 1, \ldots, n, j = 1, \ldots, n, l = 1, \ldots, L] \). In the case of chromosome IV in Figure 2 the physical mapping experiment will have been replicated nearly \( L = 3 \) times. Then problem (16) can be approximated by

\[ \min_{[i_1, \ldots, i_L]} \frac{1}{L} \sum_{l=1}^{L} \sum_{j=1}^{n} d_{ij}^{(l)}. \]  

(18)

Sometimes we will denote the sample mean, i.e., \( \bar{D} \), as an expectation, i.e., \( E[D] \), with respect to the empirical distribution. An alternate but equivalent view of this replication process is that the number of probes utilized is being sequentially expanded in batches of \( m \) probes to examine the large sample behavior of the physical mapping method.

Another key step in proving the consistency of the physical mapping algorithm is to transform problems (17) and (18) into a constrained stochastic optimization problem. The convergence of a sequence of optimal solutions will ensure the consistency of a physical mapping method. The transformation of (17) and (18) begins with the introduction of two indices \( x \) and \( i \). Index \( x \) is the ID number of clone \( x \), and index \( i \) is the \( i \)th position of clone \( x \) in the order of clones across the chromosome. We also introduce the indicator variables \( V_{xy} \) \( (x = 1, \ldots, n, i = 1, \ldots, n) \) for clone \( x \) appearing in the \( i \)th position in the ordering. The indicator variables \( V_{xy} \) can take values of 1 or 0 only, depending on whether or not clone \( x \) appears in the \( i \)th position in an ordered library. Denote these indicators collectively by \( V = \{V_{xy}\} \). Since each clone has only one position along the chromosome and each position along the chromosome can be occupied by one and only one clone, the indicator variables must satisfy the following constraints:

\[ \sum_{i=1}^{n} \sum_{j=1}^{n} V_{xy} = 0, \quad \sum_{i=1}^{n} \sum_{j=1}^{n} V_{xy} = 0, \quad \text{and} \quad \left( \sum_{j=1}^{n} V_{xy} - n \right)^2 = 0 \]  

(19)

Thus, the minimization problem in (17) can be transformed into the following stochastic optimization problem:

\[ \min_{V} E[D] = \frac{1}{2} \sum_{x=1}^{n} \sum_{y=x}^{n} |E[D_{xy}]V_{xy}(V_{xy} + V_{xy}) \]

\[ + E[D_{xy}]V_{xy} + E[D_{xy}]V_{xy} \]  

such that

\[ \sum_{i=1}^{n} \sum_{j=1}^{n} V_{xy} = 0, \quad \sum_{i=1}^{n} \sum_{j=1}^{n} V_{xy} = 0, \quad \text{and} \quad \left( \sum_{j=1}^{n} V_{xy} - n \right)^2 = 0 \]  

\[ 0 \leq V_{xy} \leq 1, \quad x = 1, \ldots, n; \quad i = 1, \ldots, n. \]  

(21)

In computing solutions to this minimization problem it may prove useful to entertain values for \( V_{xy} \) between 0 and 1 (Hopfield and Tank 1985). The expectation of the total linking distance \( D \) in (17) equals the criterion in (21). The order of clones along the chromosome is determined by the array of indicator variables \( V \), and the original true ordering of clones along the chromosome is determined by the array of indicator variables, \( V' = \{V'_{xy} \}_{x=1}^{n} \). In the case of chromosome IV the physical mapping experiment was replicated nearly \( L = 3 \) times. Then problem (16) can be approximated by

\[ \min_{[i_1, \ldots, i_L]} \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{n} d_{ij}^{(l)}. \]  

(18)

Similarly, the empirical approximation to the original physical mapping problem in (18) can also be rewritten in terms of the indicator variables:

\[ \min_{V} E[D] = \frac{1}{2} \sum_{x=1}^{n} \sum_{y=x}^{n} |d_{xy}^{(l)}V_{xy}(V_{xy} + V_{xy}) \]

\[ + d_{xy}^{(l)}V_{xy} + d_{xy}^{(l)}V_{xy} \]  

\[ \min_{V} E[D] = \frac{1}{2} \sum_{x=1}^{n} \sum_{y=x}^{n} |d_{xy}^{(l)}V_{xy}(V_{xy} + V_{xy}) \]

\[ + d_{xy}^{(l)}V_{xy} + d_{xy}^{(l)}V_{xy} \]  

(21)

In computing solutions to this minimization problem it may prove useful to entertain values for \( V_{xy} \) between 0 and 1 (Hopfield and Tank 1985). The expectation of the total linking distance \( D \) in (17) equals the criterion in (21). The order of clones along the chromosome is determined by the array of indicator variables \( V \), and the original true ordering of clones along the chromosome is determined by the array of indicator variables, \( V' = \{V'_{xy} \}_{x=1}^{n} \).
such that
\[ \sum_{i=1}^{n} \sum_{x=1}^{m} V_{x \cdot i} V_{x \cdot y} = 0, \quad \sum_{i=1}^{n} \sum_{x=1}^{m} V_{x \cdot i} V_{y \cdot j} = 0, \text{ and} \]
\[ \left( \sum_{i=1}^{n} \sum_{x=1}^{m} V_{x \cdot i} - n \right)^2 = 0 \]
\[ 0 \leq V_{x \cdot i} \leq 1, \quad x = 1, \ldots, n; \quad i = 1, \ldots, n. \quad (22) \]

Now the problem is showing that the optimal ordering of clones given by \( V \) solving the minimization problem in (22) converges to the true ordering given by \( V^0 \).

**CONSISTENCY**

There are two fundamental but distinct questions in developing mathematical methods to assemble a physical map: (1) why is a physical mapping criterion, like the total linking distance in (15), reasonable? (2) how is the physical map that optimizes the physical mapping criterion found? While many researchers (Cuticchia et al. 1992; Churchill et al. 1993; Parsons et al. 1993; Soderlund and Burks 1994; Zhang et al. 1994; Wang et al. 1994a) have addressed the second question, no one to our knowledge has provided a mathematical justification for one of the many physical mapping criteria in the literature. One reasonable property that any sensible physical mapping criterion should have is that as the mapping data set becomes large, then the physical mapping criterion should recover the true underlying physical map. This property is known as a **consistency result**. This is the most fundamental property for an estimator (of a physical map). For example, the variance of an estimator alone does not make sense unless the estimator is consistent. The property of consistency then becomes a tool by which to sort through which physical mapping criteria are reasonable and which are unreasonable. The consistency of an estimator for the true physical map can be addressed independently of how the estimator is computed.

What is shown under **RESULTS** is that a clonal ordering derived by minimizing the total linking distance in (15) is consistent when (1a) insert size in the cloning vector is constant; (1b) the genomic library is random; (2) probe hybridization sites along a chromosome are random and rare; (3) probes are selected randomly from the library; and (4) there is a staircase of inserts spanning the chromosomal region of interest. The first three assumptions are the standard ones in Arratia et al. (1991) and Fu et al. (1992). The last assumption is introduced to insur that the resulting physical map looks like a real chromosome.

A sketch of the proof is now presented. Imagine repeating a physical mapping experiment with \( n \) clones and \( m \) probes, \( L \) times. In each replicate the same \( n \) clones are used. If you were to compute an ordering of the \( n \) clones for each replicate, a slightly different ordering and final minimum total linking distance in (16) would be obtained because the distances between clones in (4) are random. The effect of replicating the physical mapping experiment \( L \) times is equivalent to increasing the number of probes in blocks of \( m \) probes to a total of \( mL \) probes. If the estimation procedure were consistent, the sample mean of the \( L \) total linking distances in (18) might be expected to converge to the expected total linking distance \( E(D) \) by the Law of Large Numbers. If there were only one ordering of clones that minimized \( E(D) \), namely the true ordering, we might also hope that the clonal ordering minimizing the sample mean of the \( L \) total linking distances in (18) might converge to the true ordering. We established this latter fact under **RESULTS**.

The problem of estimating the physical map is fundamentally different from the ones usually encountered in statistics (Dupacova and Wets 1988) because (1) the estimation criterion in (16) is a nonsmooth function of the parameter, the clonal ordering; (2) the minimization problem is constrained in (22); and (3) the physical mapping problem may have more than one solution (Goldstein and Waterman 1987). This means the usual methods utilizing the concept of pointwise convergence of random variables and the Strong Law of Large Numbers (Chow and Teicher 1988) are not sufficient to establish the consistency of an estimator associated with (16). A novel approach is required, and a new concept of convergence of a sequence of random variables, called **epi-convergence**, is introduced under **METHODS**.

Under **RESULTS** we begin by showing that the empirical mean of the total linking distance in (18) epi-converges and pointwise converges to the same limit in Theorem 3. Consideration then turns to the sequence of orderings that minimize the sequence of minimization problems in (18) as \( L \) (i.e., the number of probes) gets large. Using a result from variational analysis due to Wets (1991), we establish that this sequence of optimal orderings found by minimizing (18) converges to one that minimizes the expected total linking distance \( E(D) \). Since there is one and only clonal ordering that minimizes the expected total linking distance \( E(D) \) by Theorem 1 and since this clonal ordering is the true ordering, we were able to prove the consistency of the physical mapping method (16) in Theorem 4.

The biological implications of a consistency result are far reaching. One of the main challenges (Cox et al. 1994) for the human genome project is assessing the statistical reliability of its maps. Calculating something like a variance or confidence value for an ordering, however, presupposes the existence of a consistent estimator of the clonal ordering. Such reliability measures are critical in integration of distinct physical maps of the same genome, comparing maps of different genomes for their similarities and dissimilarities, reconciliation of a community resource, namely an entire physical map, with the local results of individual laboratories, and ultimately determining whether or not a mapping project is complete on the basis of map reliability. Not only does a consistency result provide a logical basis for establishing the statistical reliability of a physical map, but the tools
used under RESULTS can be used to develop reliability measures as the number of probes grows large.

METHODS

A unified approach to statistical estimation of stochastic optimization problems with or without constraints and with differentiable or nondifferentiable criteria has been proposed (Duponcova and Wets 1988). The problem of estimating the order of clones falls into the following general framework for a constrained optimization problem. Introduce \((\Theta, A, \Pr)\) as a probability space, where \(\Theta\) is the support of \(\Pr\) (a closed subset of a Polish Space \(X\)), where \(A\) is the Borel \(\sigma\)-field relative to \(\Theta\), and where \(\Pr\) is a probability measure. Consider a function \(f : R^n \times \Theta \rightarrow R\), a set \(S \subset R^n\), and the associated stochastic programming problem:

\[
\min_{V \in S} \phi(V) = \min_{V \in S} E[f_0(V, \xi)]. \quad (23)
\]

(The set \(S\) embodies the constraints placed on the parameters \(V \in R^n\).)

Let \(\xi_1, \ldots, \xi_m\) be a sample of independent random variables with values in \(\Theta\) having the common probability distribution \(\Pr\), and consider the mathematical programming problem,

\[
\min_{V \in R^n} \Psi(V) = \min_{V \in R^n} \frac{1}{L} \sum_{i=1}^L f_0(V, \xi_i) = \min_{V \in R^n} E[f_0(V)]. \quad (24)
\]

Our aim is to study the asymptotic behavior of the optimal solution to \((24)\),

\[
V^*_L = \arg \inf \{\Psi(V) : V \in S\}, \quad (25)
\]

as the sample size \(L\) tends to infinity.

It is convenient to study an extended real-valued objective function which incorporates the constraints implied by \(V \in S\):

\[
f(V, \xi) = \begin{cases} f_0(V, \xi), & \text{if } V \in S, \\ \infty, & \text{otherwise}. \end{cases} \quad (26)
\]

Thus, problems \((23)\) and \((24)\) are transformed into the following two unconstrained problems by the introduction of extended real-valued functions:

\[
\min_{V \in R^n} \phi^*(V) = \min_{V \in R^n} E[f(V, \xi)] \quad \text{and} \quad (27)
\]

\[
\min_{V \in R^n} \Psi^*(V) = \min_{V \in R^n} E[f(V)], \quad \text{respectively}. \quad (28)
\]

Thus, for theoretical purposes, the study of optimization problems, their general properties, as well as their classification, can be undertaken in the framework of extended, real-valued functions defined on \(R^n\). However, the traditional approach to functional analysis is no longer quite appropriate for extended, real-valued functions. For example, the concept of pointwise convergence is replaced by the concept of epi-convergence (defined shortly in this section). In the following section we briefly introduce some concepts and a result from variational analysis.

For statistical estimation and stochastic optimization problems, the objective function depends on a set of parameters, i.e., the clone ordering \(V\) in this example. When the parameters are varied, a whole family of problems having the same structure is generated. Since the parameter values are not known with certainty, the study of one instance of a problem entails the study of other related instances of the same problem. It has been proposed that the concept of pointwise convergence is not appropriate for the study of convergence of optimal solutions of constrained optimization problems. Here we introduce some theory about epi-convergence which is particularly suited for studying the convergence of optimal solutions to the physical mapping problem.

**Definition 1. (epi-convergence):** A sequence of functions \(h \in R^n \rightarrow R, l = 1, \ldots\) is said to epi-converge to \(h \in R^n \rightarrow R\) if for all \(V \in R^n\), we have

\[
\liminf_{l \rightarrow \infty} h'(V) \geq h(V), \quad \text{for all sequences } l \rightarrow \infty
\]

and is denoted by

\[
h = \text{epi-lim } h'. \quad (31)
\]

Although closely connected to the notion of pointwise convergence, epi-convergence is neither stronger nor weaker. In fact, certain sequences of functions have different pointwise and epi-limits.

Now we are in a position to discuss the main results about convergence of approximate solutions to stochastic optimization problems, such as the physical mapping problem. We wish to know when do approximate solutions to stochastic optimization problems converge to the optimal solution of the original optimization problem. The following theorem due to Wets (1991) answers this question and allows us to prove the first consistency result for a physical mapping method.

**Theorem 2. [Wets (1991)], (convergence of approximate solutions to a stochastic optimization problem):** Suppose that \(\{h_l \in R^n \rightarrow R, l = 1, \ldots\} \) is a collection of functions such that

\[
h = \text{epi-lim } h'.
\]

Then

\[
\limsup_{l \rightarrow \infty} \inf h_l \leq \inf h,
\]

and if \(V^* \in \arg \min h\), for some subsequence \(\{l\}\) and \(V = \lim_{l \rightarrow \infty} V_l\), it follows that

\[
V \in \arg \min h,
\]

and

\[
\lim_{l \rightarrow \infty} h(V) = \inf h. \quad (35)
\]

Thus, if we have

\[
h = \text{epi-lim } h'
\]

and if there exists a bounded set \(D \subset R^n\) such that for some subsequences \(\{l\}\),

\[
\arg \min h_l \cap D \neq \emptyset,
\]

then the minimum of \(h\) is attained at some point in the closure of \(D\).

**RESULTS**

The approach here for proving consistency is classical and consists of two steps: (a) proving identifiability of the physical mapping problem; (b) proving the convergence of an approximate solution of the physical mapping problem to its true optimal solution. The identifiability of a physical mapping problem (by fingerprinting) based on minimizing total linking distance
Physical Mapping Consistently

has been established in the MODEL section. Now we will prove the convergence of the approximating optimal solution in (22) to the solution of (17). From the theory developed under METHODS, it follows that proving consistency of map estimators might first begin with establishing that empirical expectations of the physical mapping criterion $h_t = E^t[f(V, \xi)]$ epi-converge to $h = E[f]$. Then it will be possible to establish when epi-convergence of $h_t = E^t[f(V, \xi)]$ to $h = E[f]$ is equivalent to pointwise convergence.

Let $V = \{V_{x,t}: x = 1, \ldots, n; i = 1, \ldots, n\}$ be the ordering of $n$ clones along a chromosome, and define

$$S = \left\{ V: \sum_{i=1}^{n} \sum_{x=1}^{n} \sum_{y=x}^{n} V_{x,t}V_{y,t} = 0, \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{x=1}^{n} V_{x,t}V_{j,t} = 0, \left( \sum_{x=1}^{n} \sum_{i=1}^{n} V_{x,t} - n \right)^2 = 0, 0 \leq V_{x,t} \leq 1 \right\}$$

as the set in which constraints on the parameters $V$ are satisfied. Let

$$f(V, \xi) = \begin{cases} \frac{1}{2} \sum_{x=1}^{n} \sum_{y=x}^{n} \sum_{i=1}^{n-2} [D_{x,y}V_{x,t-1} + V_{y,t+i}] + D_{x,y}V_{x,t} + D_{y,x}V_{y,t,n-1}, & V \in S \\
+\infty, & \text{otherwise} \end{cases}$$

where the random variables $\xi = [D_{11}, \ldots, D_{nn}]^T$ are the pairwise distances between all clones. The physical mapping criterion to be minimized is the expectation of the total linking distance, $E[f(V)] = E[f(V, \xi)]$.

Let $(Z, F, \mu)$ be a sample space for which $(F^t, L = 1, 2, \cdots)$ is an increasing sequence of $\sigma$-fields in $F$, $\mu$ is a measure, and let $\xi$ be a sample, e.g., $\xi = [\xi^1, \xi^2, \cdots]$ obtained by independent sampling of the values of $\xi$ (a distance matrix between clones).

Define the empirical expectation of the physical mapping criterion as

$$E^t[f(V)] = E^t[f(V, \xi)] = \frac{1}{L} \sum_{l=1}^{L} f(V, \xi^l), \quad (36)$$

and the optimal solution for the physical mapping problem as

$$V^\ast = \arg\min E[f(V)].$$

Our aim is to show that for any empirical approximation to this solution,

$$V^{\ast l} = \arg\min E^l[f(V)], \quad \text{we have } \lim_{l \to \infty} V^{\ast l} = V^\ast.$$

First we prove the epi-convergence of the approximate solutions $V^{\ast l}$ of the physical mapping problem. Since the physical mapping criterion $\xi \to f(V, \xi)$ is not continuous on $\Theta$, which violates the assumptions listed in (DUPACOVA and WETS 1988), theorems developed in (DUPACOVA and WETS 1988) cannot be directly applied to the physical mapping problem. However, under a general framework developed in (DUPACOVA and WETS 1988), we can still prove the epi-convergence of the approximate solution $V^{\ast l}$ to the true solution $V^\ast$. For the physical mapping problem we can show an empirical approximation to the criterion converges to the true mean of the criterion.

**Theorem 3 (a strong law of large numbers for the physical mapping criterion):** Under the assumptions of the Identifiability Theorem 1, $\mu$ a.s.

$$E[f] = \text{epi-lim } E^t[f] = \text{ptwse-lim } E^t[f], \quad (37)$$

where ptwse-lim$_{t \to \infty} E^t[f]$ denotes the pointwise limit.

The proof can be found in XIÖNG (1993) and on the World Wide Web at address http://fungus.genetics.uga.edu:5080.

Now we are ready to prove the consistency of a physical mapping method in the same spirit as Theorem 3.9 of DUPACOVA and WETS (1988).

**Theorem 4 (consistency of a physical mapping method):** Under the assumptions of the Identifiability Theorem 1, the method of physical mapping based on minimizing total linking distance is strongly consistent.

**Proof:** From Theorem 3, it follows that there exists a set $Z_0 \in F$ with $\mu(Z \setminus Z_0) = 0$ such that for any $\xi \in Z_0$,

$$\text{epi-lim } E^t[f] = E[f]. \quad (38)$$

Since $0 \leq V_{x,t} \leq 1, \forall x = 1, \ldots, n; i = 1, \ldots, n$, there exists a compact set $D$ such that $V \in D$ for all $V$. Thus,

$$\arg\min E^t[f] \subseteq D. \quad (39)$$

Let $V^l \in \arg\min E^t[f]$. Then there exists a subsequence $\{V^l_k, k = 1, 2, \cdots\}$ such that

$$\lim_{k \to \infty} V^l_k = V^\ast. \quad (40)$$

It follows from Theorem 2 that

$$V^\ast \in \arg\min E[f]. \quad (41)$$

and

$$\lim_{k \to \infty} (\inf E^l[f]) = \inf E[f]. \quad (42)$$

By Theorem 1, $V^\ast = V^0$ and is unique. Since $V^l \subseteq D$ and every subsequence $\{V^{l_k}, k = 1, 2, \cdots\}$ will converge to the same $V^0$, then

$$\lim_{l \to \infty} V^l = V^0. \quad (43)$$

From the discussion at the end of the MODEL section, we know that $V^l$ corresponds to the order determined by the physical mapping method (18) and that the true order $V^0$ corresponds to the original order of clones along the chromosome; therefore, with probability one, the order of clones produced by minimum linking distance estimation converges to the original order of clones along the chromosome.
DISCUSSION

Genetic and physical mapping data are currently being acquired and stored at a tremendous rate. Once the experimental methods for generating Figure 2 were worked out, this binary data matrix was generated over a period of 22 weeks. The time taken to analyze this data became the bottleneck and took more time than generating the mapping data. In most physical mapping projects, data analysis and data management have become rate limiting steps. The physical maps from these projects will enable important biomedical, agricultural, and biological applications. The human physical map is a case in point (COHEN et al. 1993). Since at least one third of all human disease has a genetic basis, it is likely that a physical map of the human genome (COHEN et al. 1993) will be the single most important medical diagnostic tool in the next century. There are a number of statistical problems in building such a resource, including the creation of new statistical models for physical mapping experiments (KARLIN and MACKEN 1991), design of physical mapping experiments (FU et al. 1992; ZHANG and MARR 1993), creating new inference tools for assembling physical maps (CUTICCHIA et al. 1992a), creating new inference tools for testing conformity of physical maps (e.g., Figure 2) to physical mapping models, and performing data analysis in mapping experiments (PRADE et al. 1995).

The scope for new problems in the statistical analysis of physical mapping experiments is quite large. The particular analysis described here is for one of six kinds of physical maps mentioned in the introduction, ranging from cytological maps to the entire DNA sequence of a chromosome. The questions raised and answered in this article remain to be addressed for other kinds of physical maps and mathematical methods for assembling them. In particular, with large scale genomic sequencing now underway it would be desirable to know whether or not any of the existing sequence assembly methods were consistent. Within one class of physical maps considered here, i.e., ordered clone collections, at least five fingerprinting protocols for creating ordered clone collections exist. Two of these fingerprinting protocols await a statistical analysis of associated map reconstruction methods.

Lastly, it is likely that other statistical geneticists will invent better physical mapping methods than minimum linking distance estimation for the two fingerprinting methods considered here. Minimum linking distance estimation is an example of one of several distance-based physical mapping methods that might be considered. Another example might be constructed from multiplying probabilities of overlap between successive pairs of clones as in (11) along a clonal ordering, and it is now natural to ask whether or not this criterion produces a consistent estimator. Another distinct class of physical mapping methods are character-based physical mapping methods, which utilize directly the clone/probe data matrix (in Figure 2). An example of this class of methods would be a maximum likelihood estimator (MLE) of the physical map derived from the likelihood of the data in Figure 2. One natural line of inquiry will be into the development of the method of maximum likelihood for the problem described here and the establishment of the MLE's consistency.

Within the realm of creating ordered clone collections from binary fingerprinting data with low copy number probes there still remains a number of important problems. The most fundamental one (COHEN et al. 1994) is the development of new statistical methods to assess the statistical reliability of a physical map. A consistency result for an estimator of a physical map is a logical precursor to assessing an estimator’s statistical reliability. For example, the rate of convergence of the inferred physical map to the true physical map might be used to establish asymptotic confidence levels for links between pairs of clones (or probes) on the physical map. Variational analysis will provide the necessary tools for an asymptotic analysis, and the bootstrap (EFRON 1982; WANG et al. 1994b) might in practice provide a tool for estimating confidence in links on the map, ground truthed by an asymptotic analysis.

The model in the MODEL section is one of the simplest and most widely used (ARRATIA et al. 1988; FU et al. 1992). This model is the ideal, against which the success of an experimental design for physical mapping is measured. If a physical mapping tool cannot perform well in this ideal setting, it is unlikely to succeed in application. This is one reason to prove a consistency result in this context. Nonetheless, it will prove useful to analyze physical mapping methods in more complicated models, which more closely fit data like Figure 2 to understand the exact behavior of physical mapping methods like (16) in practice and to generate improved physical mapping methods.

The most vulnerable assumption in the MODEL section is the homogeneity of the Poisson process for probe hybridization sites along the chromosome (i.e., Assumption 2). This assumption is likely to fail in two ways (PRADE et al. 1995). One, the Poisson intensity parameter may vary across probes, so it would be useful to consider a generalization of the model in the MODEL section, in which each probe has its own intensity parameter. Two, even for a single probe, the occurrence of hybridization sites along the chromosome may be nonuniform. KARLIN and MACKEN (1991) suggest some alternative models, which need to be fitted to data like Figure 2. For a revised model that empirically accounts for inhomogeneity of probe hybridization sites, it will be necessary to ask whether or not map estimators satisfying generalizations to Theorems 5.2 and 5.3 can be established.

One of the key reasons that hybridization sites may be nonuniform is the presence of repetitive DNA. One way that repetitive DNA reveals itself in the physical map in Figure 2 is the presence of multiple hybridization
signals for an individual clone. While usually positive hybridization signals are on the main diagonal in Figure 2, even in *Aspergillus nidulans* there are some cross-hybridization signals off the diagonal. These repeated sequences will lead to ambiguities in the correct placement of clones. In simpler eukaryotic genomes these ambiguities can be resolved by anchoring clones with repeats by flanking clones with no repeats. This strategy will fail only if there is a long stretch of repeats greater than 40 kilobases, as in the case of the rDNA cluster on chromosome V of *A. nidulans* (Brody et al. 1991).

While extensive efforts were made to remove errors in the data (false-positive and false-negatives in DNA/DNA hybridization) in Figure 2, such errors could also lead to some of the hybridization signals off the main diagonal. While Cuticchia et al. (1992a) showed that a map estimator based on minimizing total linking distance performed reasonably well in the presence of errors, such errors were also shown to have a serious impact on the map. These kinds of errors can be controlled in part by having multiple probes supporting links between clones on the map (Wang et al. 1994b). Errors in the data were not explicitly modeled in the MODEL section, but what is striking about the variational analysis used in the METHODS and RESULTS sections is that these tools were developed for establishing the validity of robust estimation methods (Dupacova and Wets 1988). The tools in this paper will provide a basis for evaluating robust methods of map assembly in large samples.

A key assumption, which permitted Theorem 4 to go through, is Assumption 4, the monotonicity of overlaps with distance along the physical map. This assumption has a dual justification. One, the assumption is based on current theory about the structure of chromosomes in *vivo*, which physical maps are supposed to reflect in *vitro*. The second justification is a utilitarian one—physical maps that are a "minimum tiling" (i.e., Figure 4), have little redundancy and are easier to work with. While it is possible to generalize Identifiability Theorem 1 in various ways, it captures an essential element of the linear structure of DNA as well as what molecular geneticists consider a "good" physical map. Perhaps a more fundamental question (than questions about generalizations of Theorem 1) is whether or not our notion of consistency should be revised in the light of new mathematical tools for constrained, nonsmooth estimation problems (Dupacova and Wets 1988) with multiple solutions. It may be that conclusion (42) from Theorem 2 should be considered sufficient. That is, in complicated combinatorial inference problems, like the Stochastic Traveling Salesman Problem, it may be enough that a statistical inference tool produces at least one true solution among many "equally good" solutions (Goldstein and Waterman 1987) when the sample (of probes) gets large. Rather than being bound by classical notions of consistency based on identifiability, we might consider other definitions of statistical consistency (Dupacova and Wets 1988; Wets 1991) for statistical estimation tools that do not require identifiability.

In this article a flavor for the importance of statistical tools for creating physical maps in genetics is given, and many open problems that statistical geneticists are uniquely equipped to address are described. A few statistical methods for inferring physical maps have been rigorously justified here. Once the identifiability of a statistical model is established, the consistency of one physical mapping method, minimum linking distance estimation, is proven using epi-convergence theory of variational analysis. It is hoped that these theorems and the methods used to prove them will act as a catalyst for others to examine these kinds of physical mapping problems. The possible consequences of considering such applications include a broadening of classical statistical concepts like consistency and an understanding of the structure, function, and evolution of genomes.

We thank Bob Taylor, Isinwar Basawa, Bill McCormick, Jaxn Reeves, and Bruce Weir for their comments on earlier drafts of this manuscript. We also thank the reviewers for an opportunity to strengthen the manuscript. This work was supported by National Institutes of Health grant GM-52925 and National Science Foundation BIR-9422896.

LITERATURE CITED


Cai, H., P. Kirkei, J. Yee and I. Duncan, 1995 A yeast artificial


DAWES, EIGLMEIR, K., N. HONORE, S. A. WOODS, B. CAUDRON and DUPAMNA, J., and R. WETS, 1988 Asymptotic behavior of statistical...