Placing Probes along the Genome Using Pairwise Distance Data

Will Casey[1], Bud Mishra[1], and Mike Wigler[2]

[1] Courant Institute, New York University
251 Mercer St., New York, NY 10012, USA
wcasey@cims.nyu.edu,mishra@nyu.edu
[2] Cold Spring Harbor Laboratory
P.O. Box 100, 1 Bungtown Rd.
Cold Spring Harbor, NY 11724, USA
wigler@cshl.org

Abstract. We describe the theoretical basis of an approach using microarrays of probes and libraries of BACs to construct maps of the probes, by assigning relative locations to the probes along the genome. The method depends on several hybridization experiments: in each experiment, we sample (with replacement) a large library of BACs to select a small collection of BACs for hybridization with the probe arrays. The resulting data can be used to assign a local distance metric relating the arrayed probes, and then to position the probes with respect to each other. The method is shown to be capable of achieving surprisingly high accuracy within individual contigs and with less than 100 microarray hybridization experiments even when the probes and clones number about 10^5, thus involving potentially around 10^{10} individual hybridizations.

This approach is not dependent upon existing BAC contig information, and so should be particularly useful in the application to previously uncharacterized genomes. Nevertheless, the method may be used to independently validate a BAC contig map or a minimal tiling path obtained by intensive genomic sequence determination.

We provide a detailed probabilistic analysis to characterize the outcome of a single hybridization experiment and what information can be garnered about the physical distance between any pair of probes. This analysis then leads to a formulation of a likelihood optimization problem whose solution leads to the relative probe locations. After reformulating the optimization problem in a graph-theoretic setting and by exploiting the underlying probabilistic structure, we develop an efficient approximation algorithm for our original problem. We have implemented the algorithm and conducted several experiments for varied sets of parameters. Our empirical results are highly promising and are reported here as well. We also explore how the probabilistic analysis and algorithmic efficiency issues affect the design of the underlying biochemical experiments.

1 Introduction

Genetics depends upon genomic maps. The ultimate maps are complete nucleotide sequences of the organism together with a description of the transcription units. Such maps in various degrees of completion exist for many of the microbial organisms,

O. Gascuel and B.M.E. Moret (Eds.): WABI 2001, LNCS 2149, pp. 52–68, 2001.

yeasts, worms, flies, and now humans. Short of this, genetically or physically mapped collections of objects derived from the genome under study are still of immense utility, and are often precursors to the development of complete sequence maps. These objects may be markers of any sort, DNA probes, and genomic inserts in cloning vectors.

We have been exploring the use of microarrays to assist in the development of genomic maps. We report here one such mapping algorithm, and explore its foundation using computer simulations and mathematical treatment. The algorithm uses unordered probes that are microarrayed and hybridized to an organized sampling of arrayed but unordered members of libraries of large insert genomic clones.

In the foregoing we assume some knowledge of genome organization, DNA hybridization, repetitive DNA, gene duplication, and the common forms of microarray experiments. In the proposed experimental setting, one sample at a time is hybridized to microarrayed probes, and hybridization is measured as an absolute quantity. We assume probes are of zero dimension, that is, of negligible length compared to the length of the large genomic insert clones. Most importantly, we assume that hybridization signal of a probe reflects its inclusion in one or more large genomic insert clones present in the sample, and negligible background hybridization. Our analysis is general enough to include the effects of other sources of error. *The novelty of the results reported here is in their ability to deal with ambiguities, an inevitable consequence of the use of massive parallelism in microarrays involving many probes and many clones.* Similar algorithms are reported in the literature [7], but assume only the knowledge of clone-probe inclusion information for every such combination and suggest different algorithms that do not exploit the underlying statistical structure.

One important application of our method is in measuring gene copy number in genomic DNA [10]. Such techniques will eventually have direct application to the analysis of somatic mutations in tumors and inherited spontaneous germline mutations in organisms when those mutations result in gene amplification or deletion. In contrast, low signal-to-noise ratios, due to the high complexity of genomic DNA, make other approaches such as the direct application of standard DNA microarray methods highly problematic.

2 Related Literature

The problem of physical mapping and sequencing using hybridization is relatively well studied. As shown in [6], the general problem of physical mapping is NP-complete. An approach based on traveling salesman problem (TSP) in the absence of statistics is given in [1]. The problem formalism used in this paper will be similar to the foundational work in [1–3, 5, 7, 11, 12, 14]. Our method extends the previous results by devising efficient algorithms as well as biochemical experiments capable of achieving higher resolution of probe placement within contigs. In [8] the MATRIX-TO-LINE problem is suggested as the model problem for determining Radiation Hybrid maps. Probe mapping using BACs is slightly different in that pairwise distances for probes far away cannot be resolved directly using BACs. Our design is general in that the inputs are modeled as random variables with known statistics determined a priori by our experimental designs chosen appropriately for a range of applications. Also we provide estimated probabilities of correctness for the map we produce. In this sense, this paper invokes an an experimental optimization as recommended in [2].

3 Mathematical Definitions

Given a set of P probes listed as $\{p_1, p_2, \dots, p_P\}$ and contained in some contiguous segment of the genome we define a *probe map* to be a pair of sequences, **ordering** $= \{p_{\pi(1)}, p_{\pi(2)}, \dots, p_{\pi(P)}\}$ and **position** $= \{x_1, x_2, \dots, x_P\}$. The position sequence infers the positions of the probes and the ordering sequence is determined by the permutation[1] $\pi \in S_P$ that sorts the given list of probes by position.

However the underlying correct position of each probe remains unknown. We infer probe maps approximating the correct positions as best as possible from an experimental set of data which is stochastic. Experimental data sets are represented by graphs; given a set of probes $\{p_1, p_2, \dots, p_P\}$, let V be the set of indices. Then a *pairwise distance graph* is an undirected graph $\mathcal{G} = \langle V, E \rangle$, $E \subset V \times V$ where each edge $e_{i,j}$ maps to a distance $d_{i,j}$ between probe i and probe j.

We model various experimental errors arising from the hybridization experiment used to measure probe to probe distance. With the model we can understand the distribution of pairwise distance graphs as a random variable. Under certain parameters we can implement Bayes formula to build a Maximum Likelihood Estimator (MLE) for probe map reconstruction. With the MLE established we attempt to optimize the computation involved for practical implementation.

3.1 Experimental Procedure

Consider a genome represented by the interval $[0, G]$. Take P random short sub-strings (about 200bps) which appear on the genome uniquely. Represent these strings as points $\{x_1, \dots, x_P\}$. Assume that the probes are i.i.d. with uniform random distribution over the interval $[0, G]$. Let S be a collection of intervals of the genome, each of length L (usually ranging from few 100kbs to Mbs). Suppose the left-hand points of the intervals of S are i.i.d. uniform random variables over the interval $[0, G]$. Take a small, even in number sized subset of intervals $S' \subset S$, chosen randomly from S. Divide S' randomly into two equal-size disjoint subsets $S' = S'_R \cup S'_G$, where R indicates a red color class and G indicates a green color class. Now specify any point x in $[0, G]$ and consider the possible associations between x, and the intervals in S':

- x is not covered by any interval in S'.
- x is covered by at least one interval of S'_R but no intervals of S'_G.
- x is covered by at least one interval of S'_G but no intervals of S'_R.
- x is covered by at least one interval of S'_R and at least one interval of S'_G.

If we perform a sequence of M such experiments then for each x we get a sequence of M outcomes represented as a color vector of length M. We are interested in observing sequences of such outcomes on the set $\{x_1, \dots, x_P\}$.

For DNA the short sub-strings can be produced with the use of restriction enzymes, or synthesized as oligoes. The collection of covering intervals may be provided by a BAC or YAC clone library. The division of a random sample taken from the clone library may be done with phosphorescent molecules added to the DNA and visible

[1] We denote the permutation group on P indices as S_P.

with a laser scanner. Hybridization microarrays allow us to observe such an outcome sequence for each of the 100,000 probes in a constant amount of time.

Consider an example with human. To make a set of Human Oligoe Probes we may use restriction enzymes to cut out P probe substrings of size 200bp to 1200bp from the genome and choose a low complexity representation (LCR) as discussed in [10, 9]. We may arrange for a sequence of M random samples from the BAC library, suppose each sample has K BACs and coverage $c = \frac{KL}{G}$. Samples are then randomly partitioned into two color classes $\Sigma = \{R, G\}$, and then hybridized to a microarray, arrayed with P probes. If we pick one probe p_i, then the possible outcomes for one experiment are:

- p_i hybridizes to zero BACs. We say the outcome is 'B' (blank).
- p_i hybridizes to at least one red BAC and zero green BACs. We say the outcome is 'R' (red).
- p_i hybridizes to at least one green BAC and zero red BACs. We say the outcome is 'G' (green).
- p_i hybridizes to at least one green BAC and at least one red BAC. We say the outcome is 'Y' (yellow).

We call these events i_B, i_R, i_G, and i_Y respectively. We use M random samples to complete the full experiment. The parameter domain for the full experiment is $\langle P, L, K, M\rangle$, where P is the number of probes, L is the average length of the genomic material used (for BACs, $L = 160$kb), K is the sampling size, and M is the number of samples. The output is a color sequence for each probe. The sequence corresponding to probe p_j is $\mathrm{s}_j = \langle s_{j,k}\rangle_{k=1}^{M}$ with $s_{j,k} \in \{B, R, G, Y\}$.

How the Distances Are Measured. With the resulting color sequences s_j we can compute the pairwise Hamming distance. Let

$$\begin{aligned} H_{i,j} &= \#\text{ places where } \mathrm{s}_i \text{ and } \mathrm{s}_j \text{ differ },\\ C_{i,j} &= \#\text{ places where } \mathrm{s}_i \text{ and } \mathrm{s}_j \text{ are the same but } \mathrm{s}_i \neq B,\\ T_{i,j} &= \#\text{ places where } \mathrm{s}_i \text{ and } \mathrm{s}_j \text{ are } B. \end{aligned}$$

The Hamming distance defines a distance metric on the set of probes.

Lemma 31. *Consider an experiment with parameters $\langle P, L, K, M\rangle$, and $c = \frac{KL}{G}$. Let i and j be arbitrary indices from the clone set and x_{ij} is the actual distance (in number of bases) separating probe p_i from probe p_j on the genome. Let $\hat{x}_{ij} = \min\{x_{ij}, L\}$. Then:*

$$\begin{aligned} H_{i,j} &\sim \mathit{Bin}\,(M, \frac{2ce^{(\frac{-c}{2})}\hat{x}_{ij}}{L} + O((\hat{x}_{ij})^2))\\ C_{i,j} &\sim \mathit{Bin}\,(M, 1 - e^{-c} + \frac{c}{2}(e^{-c} - 2e^{-\frac{c}{2}})\hat{x}_{ij} + O((\hat{x}_{ij})^2))\\ T_{i,j} &\sim \mathit{Bin}\,(M, (\mathrm{e}^{-\mathrm{c}(1+\hat{x}_{ij})})) \end{aligned}$$

Proof. See appendix.

These computations for small x lead to an accurate estimator:

Corollary 32. *The estimator of* x_{ij} *given by* $\tilde{x}_{ij} = H_{i,j}\frac{e^{\frac{c}{2}}L}{2cM}$ *is good in the sense that there are values of* c *so that:*

$$f(\tilde{x}_{ij} = d|x_{ij}) \to \begin{cases} \frac{1}{\sqrt{2\pi}\sigma\sqrt{x_{ij}}} e^{-\frac{(d-x_{ij})^2}{2\sigma^2 x_{ij}}} & \text{if } x_{ij} < L; \\ \frac{1}{\sqrt{2\pi}\sigma\sqrt{L}} e^{-\frac{(d-x_{ij})^2}{2\sigma^2 L}} & \text{if } x_{ij} \geq L; \end{cases} \quad \text{as } M \to \infty.$$

with $\sigma^2 = \left(\frac{e^{\frac{c}{2}}}{2c}\right)$.

Proof. It is based on a standard approximation. We have developed Chernoff bounds to analyze tradeoff between parameters K (determining c), and M. For $x < \frac{L}{2}$ one can show that for nearly any value of c the above convergence in distribution is significantly rapid w.r.t. M. □

We have developed an estimator of x_{ij} given by $\tilde{x}_{ij} = \frac{H_{i,j}}{H_{i,j}+2C_{i,j}} e^{\frac{H_{i,j}+2C_{i,j}}{4M}} L$ this estimator takes into account the variation of sample coverage over the genome.

Lemma 33. *The distribution for distance* d *is a function of* x *and is approximated by*

$$f(d|x) = \mathbb{I}_{0 \leq x < L} \frac{e^{-(d-x)^2/2\sigma^2 x}}{\sqrt{2\pi x}\sigma} + \mathbb{I}_{L \leq x \leq G} \frac{e^{-(d-L)^2/2\sigma^2 L}}{\sqrt{2\pi L}\sigma}.$$

Proof. Simple restatement of corollary 2.2 □

Since we have assumed that any given probe is distributed uniformly randomly over the genome, the density function for the probe's position is:

$$f(x) = \frac{1}{G}$$

Our next lemma is an application of Bayes' formula to compute $f(x|d)$ from $f(x)$ and $f(d|x)$ computed above.

Lemma 34. *If* $f(d|x) = \mathbb{I}_{0 \leq x < L} \frac{e^{-(d-x)^2/2\sigma^2 x}}{\sqrt{2\pi x}\sigma} + \mathbb{I}_{L \leq x \leq G} \frac{e^{-(d-L)^2/2\sigma^2 L}}{\sqrt{2\pi L}\sigma}$. *Then*

$$f(x|d) \approx \mathbb{I}_{d<L} \frac{e^{-(x-d)^2/2\sigma^2 d}}{\sqrt{2\pi d}\sigma} + \mathbb{I}_{d \geq L}\, \mathbb{I}_{L \leq x \leq G} \left(\frac{1}{G-L}\right).$$

Proof. See appendix □.

With conditional $f(x|d)$ we can now define the Maximum Likelihood Estimation problem:

Given an arbitrary pair-wise distance edge weighted complete graph $\mathcal{G}$ of P vertices, representing probes, and each edge (i,j) labeled with $d_{i,j}$, a sampled value of a random variable with the distribution $f(d||x_i - x_j|)$, we would like to choose an embedding of $\mathcal{G}$ (or more precisely, an embedding of the vertices of $\mathcal{G}$) into the real line:

$$\{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_P\} \subset [0, G],$$

that maximizes a likelihood function $F(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_P | d_{ij} : i, j \in [1, P])$. By ignoring the week dependencies, we approximate F as:

$$\prod_{1 \leq i,j \leq P} f(|\tilde{x_i} - \tilde{x_j}||d_{ij}).$$

Hence, we can minimize a related cost function

$$\sum_{1 \leq i,j \leq P} - \ln f(|\tilde{x_i} - \tilde{x_j}||d_{ij}).$$

Lemma 35. *The Optimization problem of finding $\tilde{x}_j$ to minimize $f(\tilde{x}_j | \{\tilde{x}_i : i < j\}, \{d_{i,j} : i < j\})$ is approximated by solving the following optimization problem:*

$$\textit{minimize} \sum_{1 \leq i < j \leq P} W_{ij}(|\tilde{x}_i - \tilde{x}_j| - d_{ij})^2,$$

where W_{ij}'s are positive real valued weight functions:

$$W_{ij} = \begin{cases} \dfrac{1}{2\sigma^2 d_{ij}} & \textit{if } d_{ij} < L; \\ \epsilon & \textit{otherwise,} \end{cases}$$

and $\epsilon = O\left(\frac{1}{(G-L)^2}\right)$.

Proof.

$$- \ln f(x|d) \approx \begin{cases} \dfrac{(x-d)^2}{2\sigma^2 d} + \ln(\sqrt{2\pi d}\sigma) & \text{if } d < L; \\ \ln(G-L) - \ln \mathbb{I}_{L \leq x \leq G} & \text{otherwise.} \end{cases}$$

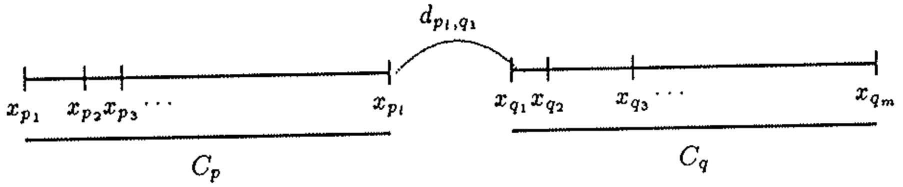
Hence

$$\sum_{1 \leq i,j \leq P} - \ln f(|\tilde{x_i} - \tilde{x_j}||d_{ij}) = \sum_{1 \leq i < j \leq P} W_{ij}(|\tilde{x}_i - \tilde{x}_j| - d_{ij})^2.$$

Note that $\epsilon = \frac{1}{2\sigma_M^2 d_{ij}} \leq \frac{1}{2\sigma_M^2 L} \leq \frac{1}{2(G-L)^2 L}$ as σ_M being the maximum variance is bounded by $(G-L)$ □.

3.2 Simple Algorithm

The simplest algorithm to place probes proceeds as follows: Initially, every probe occurs in just one singleton contig, and the relative position of a probe $\tilde{x}_i$ in contig C_i is at the position 0. At any moment, two contigs $C_p = [\tilde{x}_{p_1}, \tilde{x}_{p_2}, \ldots, \tilde{x}_{p_l}]$ and $C_q = [\tilde{x}_{q_1}, \tilde{x}_{q_2}, \ldots, \tilde{x}_{q_m}]$ may be considered for a "join" operation: the result is either a failure to join the contigs C_p and C_q or a new contig C_r containing the probes from the constituent contigs. Without loss of generality, assume that $|C_p| \geq |C_q|$, and that the probe corresponding to the right end of the first contig (x_{p_l}) is closet to the left end of the other contig (x_{q_1}). That is the estimated distance d_{p_l,q_1} is smaller than all other estimated distances: d_{p_1,q_1}, d_{p_1,q_m} and d_{p_l,q_m}.



Let $0 < \theta \leq 1$ be a parameter to be explored further later, and $L' = L\theta \leq L$. If $d_{p_l,q_1} \geq L'$ then the join operation fails. Otherwise, the join operation succeeds with the probes of C_p placed to the left of the probes of C_q, with all the relative positions of the probes of each contig left undisturbed. We will estimate the distance between the probes in C_p and the probe x_{q_1} by minimizing the function:

$$\text{minimize} \sum_{i \in \{p_1,\ldots,p_l\}:d_{i,q_1}<L'} \frac{(\tilde{x}_{q_1} - \tilde{x}_i - d_{i,q_1})^2}{2\sigma^2 d_{i,q_1}},$$

where $\tilde{x}_i$'s ($i \in \{p_1, \ldots, p_l\}$) are fixed by the locations assigned in the contig C_p. Thus taking a derivative of the expression above with respect to $\tilde{x}_{q_1}$ and equating it to zero, we see that the optimal location for x_{q_1} in C_r is

$$d^* = \max \left[\tilde{x}_{p_l}, \frac{\sum_{i \in \{p_1,\ldots,p_l\}:d_{i,q_1}<L'} (\tilde{x}_i + d_{i,q_1}) / \sigma^2 d_{i,q_1}}{\sum_{i \in \{p_1,\ldots,p_l\}:d_{i,q_1}<L'} 1/\sigma^2 d_{i,q_1}} \right].$$

Once the location of x_{q_1} is determined in C_r at d^*, the locations of all other probes of C_q in the new contig C_r are computed by shifting them by the value d^*. Thus

$$C_r = [\tilde{x}_{r_1}, \ldots, \tilde{x}_{r_l}, \tilde{x}_{r_{l+1}}, \ldots, \tilde{x}_{r_{l+m}}],$$

where $r_i = p_i$ and $\tilde{x}_{r_i} = \tilde{x}_{p_i}$, for $1 \leq i \leq l$; $r_{l+i} = q_i$ and $\tilde{x}_{r_{l+i}} = d^* + \tilde{x}_{q_i}$, for $1 \leq i \leq m$. Note that when the join succeeds, the distance between the pair of consecutive probes $\tilde{x}_{r_l}$ and $\tilde{x}_{r_{l+1}}$ is

$$0 \leq \tilde{x}_{r_{l+1}} - \tilde{x}_{r_l} \leq L',$$

and the distances between all other consecutive pairs are exactly the same as what they were in the original constituent contigs. Thus, in any contig, the distance between every pair of consecutive probes takes a value between 0 and L'. Note that one may further simplify the distance computation by simply considering the k nearest neighbors of $\tilde{x}_{q_1}$ from the contig C_p: namely, $\tilde{x}_{p_{l-k+1}}, \ldots, \tilde{x}_{p_l}$.

$$d_k^* = \max \left[\tilde{x}_{p_l}, \frac{\sum_{i \in \{p_{l-k+1},\ldots,p_l\}:d_{i,q_1}<L'} (\tilde{x}_i + d_{i,q_1}) / \sigma^2 d_{i,q_1}}{\sum_{i \in \{p_{l-k+1},\ldots,p_l\}:d_{i,q_1}<L'} 1/\sigma^2 d_{i,q_1}} \right].$$

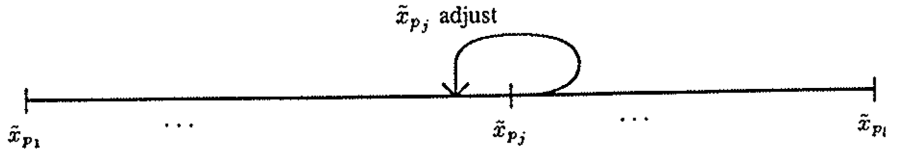
In the greediest version of the algorithm $k = 1$ and

$$d_1^* = \tilde{x}_{p_l} + d_{p_l,q_1},$$

as one ignores all other distance measurements.

At any point we can also improve the distances in a contig, by running an "adjust" operation on a contig C_p with respect to a probe $\tilde{x}_{p_j}$, where

$$C_p = [\tilde{x}_{p_1}, \dots, \tilde{x}_{p_{j-1}}, \tilde{x}_{p_j}, \tilde{x}_{p_{j+1}}, \dots, \tilde{x}_{p_l}.]$$



We achieve this by minimizing the following cost function:

$$\text{minimize} \sum_{i\in\{p_1,\dots,p_l\}\setminus\{p_j\}:d_{i,p_j}<L'} \frac{(|\tilde{x}_{p_j} - \tilde{x}_i| - d_{i,p_j})^2}{2\sigma^2 d_{i,p_j}},$$

where $\tilde{x}_i$'s ($i \in \{p_1, \dots, p_l\}\setminus\{p_j\}$) are fixed by the locations assigned in the contig C_p.

Let:

$$I_1 = \{i_1 \in \{p_1, \dots, p_{j-1}\} : d_{i_1,p_j} < L'\}$$
$$I_2 = \{i_2 \in \{p_{j+1}, \dots, p_l\} : d_{i_2,p_j} < L'\}$$
$$x^* = \frac{\sum_{i_1\in I_1} (\tilde{x}_{i_1} + d_{i_1,p_j}) / \sigma^2 d_{i_1,p_j} + \sum_{i_2\in I_2} (\tilde{x}_{i_2} - d_{i_2,p_j}) / \sigma^2 d_{i_2,p_j}}{\sum_{i_1\in I_1} 1/\sigma^2 d_{i_1,q_1} + \sum_{i_2\in I_2} 1/\sigma^2 d_{i_2,p_j}}.$$

At this point, if $x^* \neq \tilde{x}_{p_j}$, then the new position of the probe $\tilde{x}_{p_j}$ in the contig C_p is x^*. As before, one can use various approximate version of the update rule, where only k probes from the left and k probes from the right are considered and in the greediest version only the two nearest neighbors are considered. Note that the "adjust" operation always improves the quadratic cost function of the contig locally and since it is positive valued and bounded away from zero, the iterative improvement operations terminate.

4 Implementation of the k-Neighbor Algorithm

INPUT

The input domain is a probe set V, and a symmetric positive real-valued distance weight matrix $D \in \mathbb{R}_+^{P\times P}$, where $P = |V|$.

PRE-PROCESS

Construct a graph $\mathcal{G}' = \langle V, E'\rangle$, where $E' = \{e_k = (x_i, x_y) | d_{i,j} < L'\}$. The edge set of the graph $\mathcal{G}'$ is sorted into an increasing order as follows: $e_1, e_2, \dots, e_Q$, with $Q = |E'|$ such that for any two edges $e_{k_1} = [x_{i_1}, x_{j_1}]$ and $e_{k_2} = [x_{i_2}, x_{j_2}]$, if $k_1 < k_2$ then $d_{i_1,j_1} \leq d_{i_2,j_2}$. $\mathcal{G}'$ can be constructed in $O(|V|^2)$ time, and its edges can be sorted in $O(|E'|\log(|V|))$ time. In a simpler version of the algorithm it will suffice to sort the edges into an "approximate" increasing order by a parameter $H_{i,j}$ (related to $d_{i,j}$) that takes values between 0 and M. Such a simplification would result in an algorithm with $O(|E'|\log M)$ runtime.

MAIN ALGORITHM

Data-structure: Contigs are maintained in a modified union-find structure designed to encode a collection of disjoint unordered sets of probes which may be merged at any time. Union-find supports two operations, *union* and *find* [13], union merges two sets into one larger set, find identifies the set an element is in. At any instant, a is represented by the following:

- Doubly linked list of probes giving left and right neighbor with estimated consecutive neighbor distances.
- Boundary probes: each contig has a reference to left and right most probes.

In the kth step of the algorithm consider edge $e_k = [x_i, x_j]$: if find(x_i) and find(x_j) are in distinct contigs C_p and C_q, then join C_p and C_q, and update a single distance to neighbor entry in one of the contigs.

At the termination of this phase of the algorithm, one may repeatedly choose a random probe in a randomly chosen contig and apply an "adjust" operation.

OUTPUT

A collection of probe contigs with probe positions relative to the anchoring probe for that contig.

4.1 Time Complexity

First we estimate the time complexity of the main algorithm implementing the k−neighbor version: For each $e \in E'$ there are two find operations. The number of union operations cannot exceed the number of probes $P = |V|$, as every successful join operation leading to a union operation involves a boundary vertex of a contig. Any vertex during its life time can appear at most twice as a boundary vertex of a contig, taking part in a successful join operation. The time cost of a single find operation is at most $\gamma(P)$, where γ is the inverse of Ackermann's function. Hence the time cost of all union-find operations is at most $O(|E'|\gamma(P))$. The join operation on the other hand requires running the k−neighbor optimization routine which is done at a cost $O(k)$. Thus the main algorithm has a worst case time complexity of:

$$O\Big(|E'|\gamma(|V|) + k|V|\Big)$$

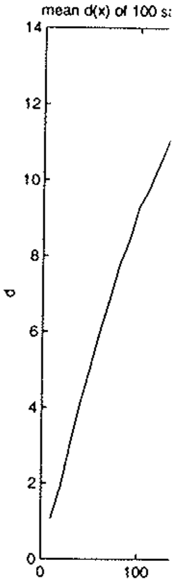
The Full Algorithm including preprocessing is:

$$O\Big(|E'|\log(|V|) + |V|^2\Big)$$

In a slightly more robust version the contigs may be represented by a dynamic balanced binary search tree which admit find and implant operations. Each operation has worst

case time complexity of
worst case runtime for t

and for the full algorithn



5 What Do the Sin

5.1 Simulation: Obser

The sample mean and va
simulation done in-silico
place them randomly on
In this experiment 100 ra
compute the Hamming di
on the Genome. Color se
chosen BACs of which ha
nar

5.2 Simulation: Full E

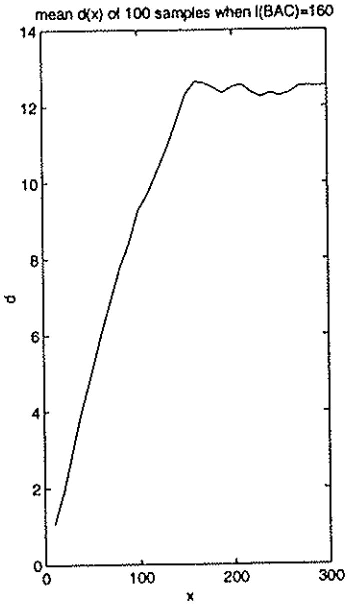
Below we describe an in-s
of size 5,000 Kb we rand
the function in the probe

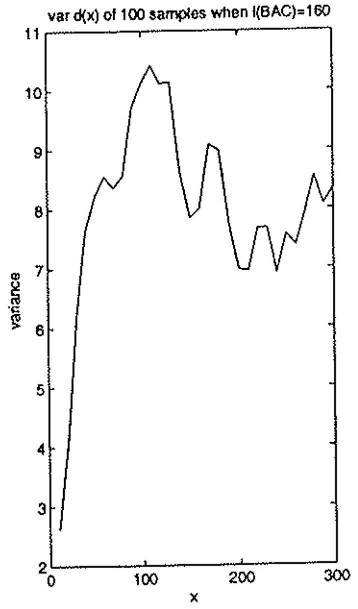
case time complexity of $O(\log(|V|))$. Thus after summing over all $|E'|$ operations the worst case runtime for the main algorithm is:

$$O\Big(|E'|\log(|V|) + k|V|\Big)$$

and for the full algorithm is:

$$O\Big(|E'|\log(|V|) + |V|^2\Big)$$





5 What Do the Simulations Tell?

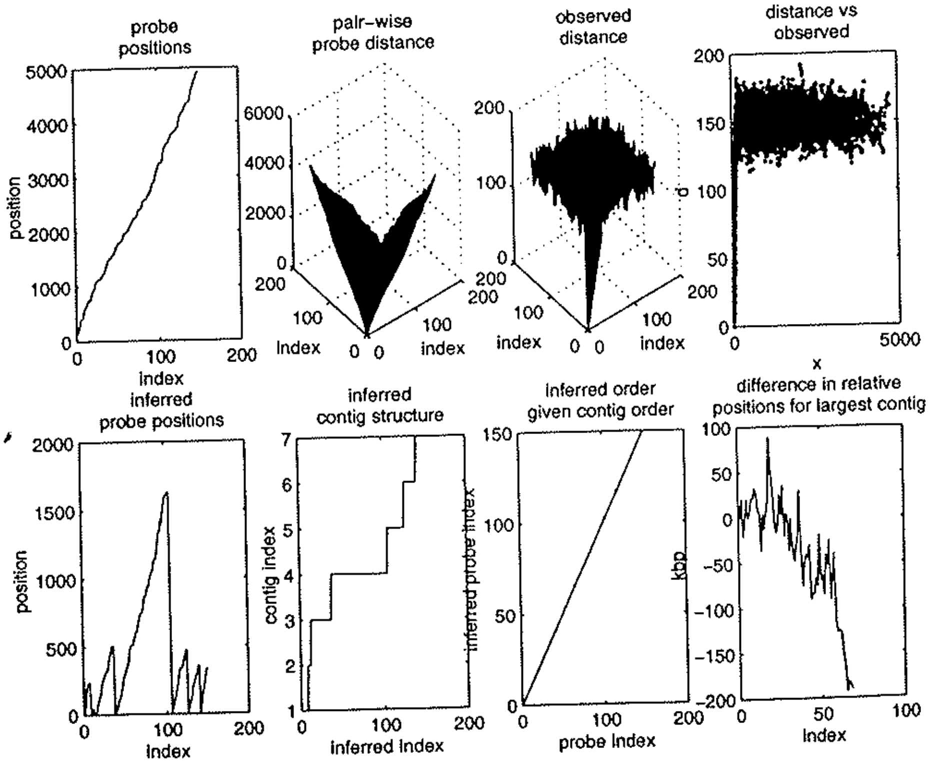
5.1 Simulation: Observed Distance

The sample mean and variation of the distance function are computed with a simple simulation done in-silico. BACs are 160Kb in length, we generate 1,200 BACs and place them randomly on a genome of size $G = 32,000$Kb, This gives a 6× BAC set. In this experiment 100 random points are chosen on the genome and for each point we compute the Hamming distance compared to points $10, 20, 30, \ldots 300$ Kb to the right on the Genome. Color sequences are computed by using 20 samples of 130 randomly chosen BACs of which half are likely to be red and the other half green.

5.2 Simulation: Full Experiment

Below we describe an in-silico experiment for a problem with 150 probes. On a Genome of size 5,000 Kb we randomly place 150 probes, there positions are graphed as a monotone function in the probe index. Next we construct a population of 500 randomly placed

BACs. From the population we repeat a sampling experiment using a sample size of 32 BACS 16 are colored red, and 16 are colored green. Each sample is hybridized in-silico to the probe set. Here we assume a perfect hybridization so there are no cross hybridizations or failures in hybridizations associated with the experiment. We repeat the sample experiment 130 times. This produces the observed distance matrix, whose distribution we modeled earlier. This is the input for the algorithm presented in this paper. In the distance vs. observed data plot we see that using a large $M = 130$ (suggested by the Chernoff bounds) has its benefits in cutting down the rate of the false positives. The observed distance matrix is input into the (10−neighbor, $\theta = \frac{11}{16}$) algorithm without the use of the adjust operation, the result is 7 contigs. The order within contigs had five mistakes. We look at the the 4th contig and plot the relative error in probe placement.



6 Future Work

The more robust variation of the algorithm based on a dynamically balanced binary search tree will be studied in more detail. A comparison with Traveling Salesman heuristics and an investigation of an underlying relation to the heat equation will show why this algorithm works well. We will work on a probabilistic analysis for the statistics of contigs. A model incorporating failure in hybridization and cross hybridization will be developed. We are able to prove that, if errors are not systematic, then a slight modification of the Chernoff bounds presented here can be applied to ensure the same

results. We shall also con
ror and a choice of meltin
noise. A set of experimen
extensive simulations, an
appears to be a promising

Acknowledgments

The research presented h
Challenge Grant, and an N

Appendix

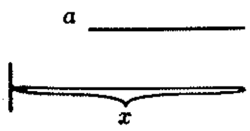
A Proof of Lemma

Lemma A1. $H_{i,j} \sim$ *Bin*
$2e^{-\frac{c}{2}})x + O(x^2)), T_{i,j}$
above and $c = \frac{KL}{G}$, i, j
distance as number of ba

Proof. Since the M samp
when $M = 1$ the probabi
events $T = (i_B \wedge j_B), C$

Given a set of K BA
interval of length l is $(1 -$

Shown below is a dia
C, H, T when $x < L$. Th
covers probe p_i but not p_j
p_i and p_j; finally, the bar



Hence we derive:

$P(T|x \leq L) = \text{ex}$
$(i_R \wedge j_R|x < L) = e^{-}$

results. We shall also consider the choice of probes to limit the cross-hybridization error and a choice of melting points to further add to the goal of decreasing experimental noise. A set of experimental designs will be presented for the working biologists. More extensive simulations, and results on real experiments shall report the progress of what appears to be a promising algorithm.

Acknowledgments

The research presented here was supported in part by a DOE Grant, an NYU Research Challenge Grant, and an NIH Grant.

Appendix

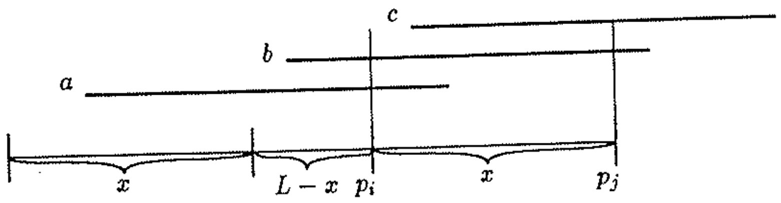
A Proof of Lemma 2.1

Lemma A1. $H_{i,j} \sim$ *Bin* $(M, \frac{2c\exp(\frac{-c}{2})x}{L} + O(x^2)), C_{i,j} \sim$ *Bin* $(M, 1-e^{-c}+\frac{c}{2}(e^{-c} - 2e^{-\frac{c}{2}})x + O(x^2)), T_{i,j} \sim$ *Bin* $(M, (e^{-c(1+\frac{x}{L})}))$ *with parameters* $\langle P, L, K, M\rangle$ *as above and* $c = \frac{KL}{G}$, i, j *are arbitrary indices from the Clone Set and* x *is the actual distance as number of bases separating the probe positions on the Genome.*

Proof. Since the M samples are done independently the proof reduces to showing that when $M = 1$ the probabilities are Bernoulli with respective parameters. Let us define events $T = (i_B \wedge j_B)$, $C = ((i_R \wedge j_R) \vee (i_G \wedge j_G) \vee (i_Y \wedge j_Y))$, and $H = (\neg T \wedge \neg C)$.

Given a set of K BACs on a genome $[0, G]$ the probability that none start in an interval of length l is $(1-\alpha)^l \approx e^{-\alpha l}$ where $\alpha = \frac{K}{G}$.

Shown below is a diagram that is helpful in computing the probabilities for events C, H, T when $x < L$. The heavy dark bar labeled a represents a set of BACs which covers probe p_i but not p_j; the bar labeled b represents a set of BACs that covers probe p_i and p_j; finally, the bar labeled c represents a set of BACs that covers p_j but not p_i.



Hence we derive:

$$P(T|x \leq L) = \exp(-(\alpha_R + \alpha_G)(L + x))$$
$$P(i_R \wedge j_R|x < L) = e^{-\alpha_G(L+x)}\{(1 - e^{-\alpha_R(L-x)})$$

$$
\begin{aligned}
&+(1-e^{-\alpha_R x})(e^{-\alpha_R(L-x)})(1-e^{-\alpha_R x})\} \\
&= e^{-\alpha_G(L+x)}\{1-2e^{-\alpha_R L}+e^{-\alpha_R(L+x)}\} \\
P(i_G \wedge j_G|x \leq L) &= e^{-\alpha_R(L+x)}\{1-2e^{-\alpha_G L}+e^{-\alpha_G(L+x)}\} \\
P(i_Y \wedge j_Y|x \leq L) &= \left(1-2e^{-\alpha_R L}+e^{-\alpha_R(L+x)}\right)\left(1-2e^{-\alpha_G L}+e^{-\alpha_G(L+x)}\right) \\
P(C|x \leq L) &= P(i_R \wedge j_R|x \leq L)+P(i_G \wedge j_G|x \leq L)+P(i_Y \wedge j_Y|x \leq L) \\
P(H|x \leq L) &= 1-[P(T|x \leq L)+P(C|x \leq L)]
\end{aligned}
$$

When $x \geq L$ the probabilities are:

$$
\begin{aligned}
P(T|x \geq L) &= \exp(-(\alpha_R+\alpha_G)(2L)) \\
P(i_R \wedge j_R|x \geq L) &= e^{-\alpha_G(2L)}\{(1-e^{-\alpha_R L})^2\} \\
P(i_G \wedge j_G|x \geq L) &= e^{-\alpha_R(2L)}\{(1-e^{-\alpha_G L})^2\} \\
P(i_Y \wedge j_Y|x \geq L) &= (1-e^{-\alpha_R L})^2(1-e^{-\alpha_G L})^2 \\
P(C|x \geq L) &= P(i_R \wedge j_R|x \geq L)+P(i_G \wedge j_G|x \geq L)+P(i_Y \wedge j_Y|x \geq L) \\
P(H|x \geq L) &= 1-[P(T|x \geq L)+P(C|x \geq L)]
\end{aligned}
$$

Because $\alpha_R = \alpha_G$, $\alpha_R L = \alpha_G L = \frac{c}{2} = \frac{KL}{2G}$. Let $q = q(x) = P(H)$ and $p = p(x) = P(C)$. In general $q(x)$ and $p(x)$ are complicated function of x, below we derive a first order approximation of $x(q)$ to be used as a biased estimator.

$$
\begin{aligned}
P(H) &= (1-(1-2e^{\frac{-c}{2}}+2e^{\frac{-c}{2}(1+\frac{x}{L})})^2) = \frac{2c\exp(\frac{-c}{2})x}{L}+O(x^2) \\
P(T) &= \left(e^{-c(1+\frac{x}{L})}\right) \\
P(C) &= 1-e^{-c}+\frac{c}{2}(e^{-c}-2e^{-\frac{c}{2}})x+O(x^2)
\end{aligned}
$$

With independent sampling:

$$
\begin{aligned}
P(H_{i,j}) &\sim \text{ Bin}\,(M, \frac{2c\exp(\frac{-c}{2})x}{L}+O(x^2)) \\
P(C_{i,j}) &\sim \text{ Bin}\,(M, 1-e^{-c}+\frac{c}{2}(e^{-c}-2e^{-\frac{c}{2}})x+O(x^2)) \\
P(T_{i,j}) &\sim \text{ Bin}\,(M, (e^{-c(1+\frac{x}{L})})) \qquad \square
\end{aligned}
$$

B Proof of Lemma 2.3 Using Bayes' Formula

Lemma B1. *If* $f(d|x) = \mathbb{I}_{0 \leq x < L}\frac{e^{-(d-x)^2/2\sigma^2 x}}{\sqrt{2\pi x\sigma}} + \mathbb{I}_{L \leq x \leq G}\frac{e^{-(d-L)^2/2\sigma^2 L}}{\sqrt{2\pi L\sigma}}$. *Then*

$$
f(x|d) \approx \mathbb{I}_{d<L}\frac{e^{-(x-d)^2/2\sigma^2 d}}{\sqrt{2\pi d\sigma}} + \mathbb{I}_{d \geq L}\,\mathbb{I}_{L \leq x \leq G}\ \left(\frac{1}{G-L}\right),
$$

Proof.

$$f(x|d) = \frac{f(d|x)f(x)}{\int_0^G f(d|x)f(x)\;dx}$$
$$= \frac{\frac{1}{G}\left(\mathbb{I}_{0\leq x<L}\frac{e^{-(d-x)^2/2\sigma^2 x}}{\sqrt{2\pi x\sigma}} + \mathbb{I}_{L\leq x\leq G}\frac{e^{-(d-L)^2/2\sigma^2 L}}{\sqrt{2\pi L\sigma}}\right)}{\frac{1}{G}\int_0^G\left(\mathbb{I}_{0\leq x<L}\frac{e^{-(d-x)^2/2\sigma^2 x}}{\sqrt{2\pi x\sigma}} + \mathbb{I}_{L\leq x\leq G}\frac{e^{-(d-L)^2/2\sigma^2 L}}{\sqrt{2\pi L\sigma}}\right)\;dx}$$

For small values of σ^2 the denominator in the above expression can be approximated as follows[2]:

$$f(d) = \left(\frac{1}{G}\right)\int_0^L \frac{e^{-(d-x)^2/2\sigma^2 x}}{\sqrt{2\pi x\sigma}}\;dx + \left(\frac{G-L}{G}\right)\frac{e^{-(d-L)^2/2\sigma^2 L}}{\sqrt{2\pi L\sigma}}$$
$$\approx \frac{1}{G}\mathbb{I}_{d<L} + \left(1-\frac{L}{G}\right)\delta_{d=L}.$$

Thus, we make further simplifying assumptions and choose the following likelihood function:

$$f(x|d) \approx \mathbb{I}_{d<L}\frac{e^{-(x-d)^2/2\sigma^2 d}}{\sqrt{2\pi d\sigma}} + \mathbb{I}_{d\geq L}\;\mathbb{I}_{L\leq x\leq G}\;\left(\frac{1}{G-L}\right), \qquad \square$$

C How Good Are the Results?

C.1 False Positives, False Negatives

We treat the problem of false positives, and false negatives with Chernoff's tail bounds. We find upper bounds on the probability of getting a false positive or false negative in terms of the parameters $\theta, M, c = \frac{KL}{G}, 0 \leq \theta \leq 1, L' = L\theta \leq L$.

A *false positive* is a pair of probes that appear to be close by the Hamming distance but are actually far apart on the genome. We denote the event as:

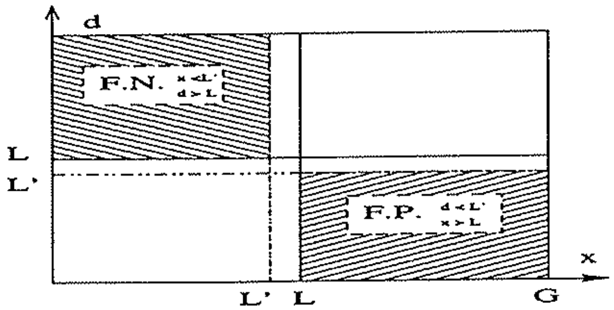
$$\text{F.P.} = (d < L') \wedge (x > L)$$

A *false negative* is a pair of probes that appear to be far by the Hamming distance but are actually close on the genome. We denote the event as:

$$\text{F.N.} = (x < L') \wedge (d > L)$$

In the following picture the volume of data which are false positives and false negatives are indicated by the squares noted F.P. and F.N. respectively.

[2] The Dirac Delta Function is distribution defined by the equations $\left\{\begin{array}{ll}\delta_{x=0} = 0 & \text{if } x \neq 0\\ \int_x \delta_{x=0}dx = 1 & \end{array}\right\}$.



We develop a Chernoff bound to bound the probability that the volume of false positive data is greater than a specified size.

The Chernoff bounds for a binomial distribution with parameters (M, q) are given by:

$$P(H > (1+v)Mq) < \left(\frac{e^{v}}{(1+v)^{(1+v)}}\right)^{Mq} \quad \text{with } v > 0$$
$$P(H < \theta Mq) < e^{\frac{-Mq(1-\theta)^2}{2}} \quad \text{with } 0 \leq \theta < 1$$

Let $H(M)$ be the Hamming distance when M phases are complete. Let $q(L) = P(H|x \geq L) \approx \frac{2\alpha L}{e^{\frac{L}{2}}} = \frac{2c}{e^{\frac{L}{2}}}$ We start by noting equivalent events:

$$\begin{aligned}(d < \theta L|x > L) &= (\sigma^2 H(M) < \theta L|x > L)\\ &= (H(M) < \theta\frac{L}{\sigma^2}|x > L)\\ &\subset (H(M) < \theta\frac{2cM}{e^{\frac{c}{2}}})\\ &= (H(M) < \theta Mq(L))\end{aligned}$$

Using the Chernoff bound we have:

$$P(d < \theta L|x > L) \leq P(H(M) < \theta Mq_L) < e^{\frac{-Mc(1-\theta)^2}{e^{\frac{L}{2}}}}$$

For the False Negatives we begin by noting that:

$$\begin{aligned}(d > L|x \leq \theta L) &= (\sigma^2 H(x) > L|x < L')\\ &= (\sigma^2 H(x) > (1+v)L'|x < L') \text{ where } v = (\frac{1}{\theta} - 1)\\ &= (H(x) > \frac{(1+v)L'}{\sigma^2}|x < L')\\ &\subset (H(x) > (1+v)Mq(x))\end{aligned}$$

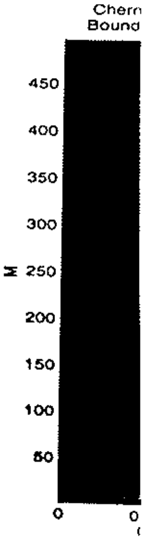
The last eve

Applying th

$P($ F.N. $)$

Chernofl

The Cher



References

1. F. Alizadel
Using Unic
2. E. Barillot,
Using Ran
A185, 1995.

The last event inclusion is because:

$$(x \leq L') \Rightarrow (\frac{2cMx}{e^{\frac{c}{2}}L} \leq \frac{2cML'}{e^{\frac{c}{2}}L}) \Rightarrow (Mq(x) \leq \frac{1}{\sigma^2}L')$$

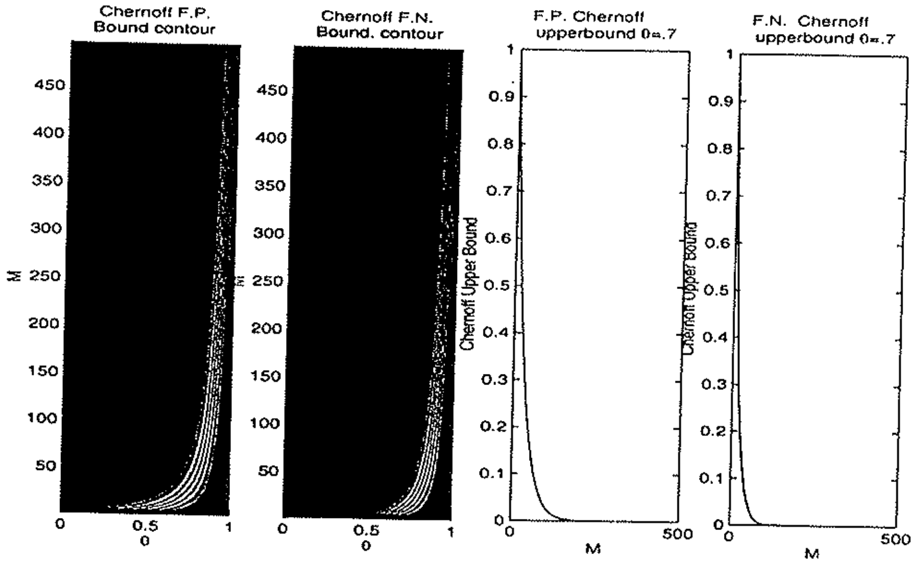
Applying the Chernoff bound we get:

$$P(\text{ F.N. }) \leq P(H > (1+v)Mq(x)) < \left(\frac{e^v}{(1+v)^{(1+v)}}\right)^{Mq(x)} < (e^{(\frac{1}{\theta}-1)}\theta^{\frac{1}{\theta}})^{Mq_L}$$
$$= (e^{(\frac{1}{\theta}-1)}\theta^{\frac{1}{\theta}})^{M\frac{2c}{e^{\frac{c}{2}}}}$$

Chernoff bounds are:

$$P(\text{ F.P. }) < e^{\frac{-Mc(1-\theta)^2}{e^{\frac{c}{2}}}}$$
$$P(\text{ F.N. }) < (e^{(\frac{1}{\theta}-1)}\theta^{\frac{1}{\theta}})^{M\frac{2c}{e^{\frac{c}{2}}}}$$

The Chernoff bounds for typical parameters are shown below.



References

1. F. Alizadeh,R.M. Karp,D.K. Weisser, and G. Zweig. "Physical Mapping of Chromosomes Using Unique Probes," **Journal of Computational Biology, 2(2)**:159–185, 1995.
2. E. Barillot, J. Dausset, and D. Cohen. "Theoretical Analysis of a Physical Mapping Strategy Using Random Single-Copy Landmarks," **Journal of Computational Biology, 2(2)**:159–185, 1995.

3. A. Ben-Dor and B. Chor. "On constructing radiation hybrid maps," **Proceedings of the First International Conference on Computational Molecular Biology,** 17–26, 1997
4. H. Chernoff. "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," **Annals of Mathematical Statistics, 23**:483–509,1952.
5. R. Drmanac *et al.*, "DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing," **Science, 163(5147)**:596, Feb 4, 1994.
6. P.W. Goldberg, M.C. Golumbic, H. Kaplan, and R. Shamir. "Four Strikes Against Physical Mapping of DNA," **Journal of Computational Biology, 2(1)**:139–152, 1995.
7. D. Greenberg and S. Istrail. "Physical mapping by STS hybridization: Algorithmic strategies and the challenge of software evaluation," **Journal of Computational Biology, 2(2)**:219–273, 1995.
8. J. Håstad, L. Ivansson, J. Lagergren, "Fitting Points on the Real Line and its Application to RH Mapping," **Lecture Notes in Computer Science, 1461**:465–467, 1998.
9. N. Lisitsyn and M. Wigler, "Cloning the differences between two complex genomes," **Science, 258**:946–951,1993.
10. R. Lucito, J. West, A. Reiner, J. Alexander, D. Esposito, B. Mishra, S. Powers, L. Norton, and M. Wigler, "Detecting Gene Copy Number Fluctuations in Tumor Cells by Microarray Analysis of Genomic Representations," **Genome Research, 10(11)**: 1726–1736, 2000.
11. M.J. Palazzolo, S.A. Sawyer, C.H. Martin, D.A. Smoller, and D.L. Hartl "Optimized Strategies for Sequence-Tagged-Site Selection in Genome Mapping," **Proc. Natl. Acad. Sci. USA, 88(18)**:8034–8038, 1991.
12. D. Slonim, L. Kruglyak, L. Stein, and E. Lander, "Building human genome maps with radiation hybrids," **Journal of Computational Biology, 4(4)**:487–504, 1997.
13. R. E. Tarjan. **Data Structures and Network Algorithms**, CBMS 44
SIAM, Philadelphia, 1983.
14. D.C. Torney, "Mapping Using Unique Sequences," **J Mol Biol, 217(2)**:259–264, 1991.