Genomics via Optical Mapping (I): 0-1 Laws for Mapping with Single Molecules *

Thomas Anantharaman[1], and Bud Mishra[2,3†]
[1] BioNano Genomics, San Diego, CA 92121.
[2] Courant Institute of Mathematical Sciences, New York University,
251 Mercer Street, New York, NY, USA 10012.
[3] Cold Spring Harbor Lab, 1 Bungtown Road, Cold Spring Harbor, NY, USA 11724.

November 22, 2013

ABSTRACT: The genomic data that can be collected from a single DNA molecule by the best chemical and optical methods (e.g., using technologies from OpGen, BioNanoGenomics, NABSys, PacBio, etc.) are badly corrupted by many poorly understood noise processes. Thus, single molecule technology derives its utility through powerful probabilistic modeling, which can provide precise lower and upper bounds on various experimental parameters to create the correct map or validate sequence assembly. As an example, this analysis shows how as the number of "imaged" single molecules (i.e., coverage) is increased in the optical mapping data, the probability of successful computation of the map jumps from 0 to 1 for fairly small number of molecules.

1 Some Preliminary Remarks

Optical Mapping [AMS97, Ana+97b, AnMS99, AsMS99, Cai+98, Mishra03, Sam+95] is an approach that generates an ordered restriction map of a DNA molecule (e.g., a genome [Jing+99, Lai+99, Lim+01, lin+99, Zhou+02] or a clone [Cai+98, Giacalone+00, Sam+95, Skiadas+99]). The resulting restriction map is represented as an ordered enumeration of the restriction sites along with the estimated lengths of the restriction fragments between consecutive restriction sites and various related statistics. These statistics accurately model the errors in estimating the restriction fragment lengths as well as the

*The work reported in this paper was supported by grants from NSF's Qubic program, DARPA, HHMI biomedical support research grant, the US department of Energy, the US air force, National Institutes of Health and New York State Office of Science, Technology & Academic Research.
†To whom correspondence should be addressed. E-mail: mishra@nyu.edu

errors due to unrepresented and misrepresented restriction sites in the map. These physical maps have found applications in improving the accuracy and algorithmic efficiency of sequence assembly, validating assembled sequences, characterizing gaps in the assembly and identifying candidates for finishing steps in a sequencing project. Also, because of its inherent simplicity and scalability as well as its reliance on single molecules, optical mapping also provides a fast method for moderate resolution karyotyping and haplotyping.

The physico-chemical approach underlying optical mapping is based on immobilizing long single DNA molecules on an open glass surface, digesting the molecules on the surface and visualizing the gaps created by restriction activities using fluorescence microscopy. Thus the resulting image, in the absence of any error, would produce an ordered sequence of restriction fragments, whose masses can be measured via relative fluorescence intensity and interpreted as fragment lengths in base pairs. The corrupting effects of many independent sources of errors affect the accuracy of an optical map created from one single DNA molecule, and can *only* be tamed by combining the optical maps of many single molecules covering completely or partially the same genomic region and by incorporating accurate statistical models of the error sources. To a rough approximation, the insurmountable obstacles in the chemistry is circumvented by cleverly exploiting the statistical properties of the system through a "0-1 Law" in the parameter space. This law plays a crucial role at the heart of the entire optical mapping technology and is likely to reappear in other contexts as well, e.g., array-mapping, single-molecule sequencing and haplotyping. In this paper, we focus on one such law in the context of mapping clones and we hint at how these results are generalized to genomic mapping.

The main error sources limiting the accuracy of an optical map are due to either incorrect identification of restriction sites or incorrect estimation of the restriction fragment lengths. Since these error sources interact in a complex manner and involve resolution of the microscopy, imaging and illumination systems, surface conditions, image processing algorithm, digestion rate of the restriction enzyme and intensity distribution along the DNA molecule, statistical Bayesian approaches are used to construct a consensus map from large number of imperfect maps of single molecules. In the Bayesian approach, the main ingredients are as follows: (1) A model of the map of restriction sites (*Hypothesis*, H) and (2) A conditional probability distribution function for the single molecule map data given the hypothesis (*Conditional pdf*, $f(D|H)$). The conditional pdf models the restriction fragment sizing error in terms of a Gaussian distribution, the missing restriction site event (due to partial digestion) as a Bernoulli trial and the appearance of false restriction sites as a Poisson process. Using the Bayes' formula, the posterior conditional pdf $f(H|D) = f(D|H)\dfrac{f(H)}{f(D)}$ is computed and provides the means for searching for the best hypothetical model given the set of single molecule experimental data. Since the underlying hypothesis space is high dimensional and the distributions are multi-modal, a naïve computational search must be avoided. An efficient implementation involves approximating the modes of the posterior distribution of the parameters and accurate local search implemented using dynamic programming [AMS97, AnMS99]. The correctness of the constructed map depends crucially on the choice of the experimental parameters

(e.g., sizing error, digestion rate, number of molecules). Thus, the feasibility of the entire method can be ensured only by a proper experimental design.

This paper studies several simple models for optical mapping and explores their power and limitations when applied to the construction of maps of clones (e.g., lambdas, cosmids, BACs and YACs), by providing precise lower and upper bounds on the number of clone molecules needed to create the correct map of the clone. Our probabilistic analysis proves the existence of a 0-1 laws in the number of molecules.

The paper is organized as follows: In section 2, we formulate the problem; in sections 3, 4 and 5, we successively introduce and analyze the effects of various error sources: namely, partial digestion error, misorientation error and quantization error, respectively. We use probabilistic methods to provide upper and lower bounds on the choices of parameters that would ensure correct result with high probability. In section 6, we study the effect of sizing error and its interaction with discretization. The analysis indicates that for a reasonable choice of sizing error, the algorithms based on discretization are unlikely to work correctly with any reasonable probability.

2 Problem Formulation

The underlying bio-chemical problem concerns with the construction of an ordered restriction map of a clone (a piece of DNA of length L, where L is measured in base-pairs bps, $Kb = 10^3 bp$ or $Mb = 10^6 bp$). Typical values of L are 2–20Kb (lambda's), 20–45Kb (cosmids) 150–200Kb (BAC'S) and $\approx 1Mb$ (YAC'S). For our mathematical analysis, we will often assume that L takes some fixed value which can be arbitrarily large. These clones are sequences of length L over the alphabet { A, T, C, G }. Certain short subsequences (typically of length 6, e.g., GGATCC) can be recognized by a restriction enzyme (e.g., *BamH* I), and location of these restriction sites

$$0 < H_1 < H_2 < \cdots < H_k < L$$

in the clone is the *ordered restriction map* of the clone with respect to the given enzyme.

Let $h_i = H_i/L$ be a real number. Then the *normalized ordered restriction map* of the clone with respect to the enzyme is

$$0 < h_1 < h_2 < \cdots < h_k < 1,$$

where each h_i assumes some real value in the open unit interval $(0, 1)$.

Note that in the absence of any additional distinguishing characteristic of the clone (e.g., identification of 3' end or 5' end), we could have also taken the following as another *normalized ordered restriction map* of the same clone with respect to the same enzyme:

$$0 < h_k^R < \cdots < h_2^R < h_1^R < 1,$$

where $h_i^R = 1 - h_i$. Note that the *normalized ordered restriction map* is unique up to reversal in the absence of any additional distinguishing characteristic, and is unique if we know the orientation.

3 False Negative Errors: Partial Digestion

Let us postulate an experiment, where the desired *normalized ordered restriction map* is observed, subject to *partial digestion* error and where any particular restriction site is observed with some probability $p \leq 1$. We assume no other error sources for now; thus no other spurious sites (false restriction cuts) are included in the observation and the observed restriction map appears in the correct orientation.

Thus the result of the experiment is an ordered sequence of sites (normalized)

$$0 < s_1 < s_2 < \cdots < s_l < 1,$$

where for each s_i, there is an h_j in the true map, such that $s_i = h_j$. By assumption, for each h_j the probability

$$Pr[\text{ there exists some } s_i \text{ s.t. } h_j = s_i] = p.$$

Let us also assume that the experiment is repeated n times resulting in n observed restriction maps. Assume that the true restriction map is unknown and is to be constructed from these n observations. A straight forward algorithm for doing this would be to simply take the union of all the observed restriction sites, and output this result in sorted order.

We claim that if $n \geq \left(\frac{\ln k + c}{p}\right)$ then the result of the preceding algorithm is almost surely correct. Here c is a constant to be determined later. Note that the probability that a cut site h_j appears in at least one observation is $1 - (1-p)^n \geq 1 - e^{-c}/k$. Thus the probability that all k true cut sites show up in the final map is given by $(1 - e^{-c}/k)^k > e^{-e^{-c}} + o(1)$.

On the other hand, if $n < \left(\frac{\ln k}{p(1+p)}\right)$ $(k \geq 1$ and $0 < p < 0.69)$ then there is a high probability that the amount of data is insufficient to recover the correct map. Note again that given a true cut site h_j the probability that this cut is never observed in any of the n observations is simply $(1-p)^n > e^{-pn(1+p)} > e^{-\ln k} = 1/k$. Thus, with this value of n, the probability that we can recover all the true cut sites is simply bounded from above by $\left(1 - \frac{1}{k}\right)^k \leq \frac{1}{e} < \frac{1}{2}$.

In summary: Let ϵ be a positive constant and $c \geq \ln(1/\epsilon)$. Then for $n \geq \left(\frac{\ln k + c}{p}\right)$, with probability at least $(1-\epsilon)$, the correct ordered restriction map can be computed in $O(nk)$ time.

When $n < \left(\frac{\ln k}{p(1+p)}\right)$ $(k \geq 1$ and $0 < p < 0.69)$, there is a probability greater than half that no algorithm can compute the correct ordered restriction map.

Note, however, that since the value of k and p are not known a priori, it is impossible to use this result in a meaningful way in designing an experiment (i.e., in choosing n). The algorithm itself does not use the parameters k or p; only its success probability is determined by these parameters for a fixed set of input data.

4 Misorientation Errors

Next, let us postulate a modified experiment, where the desired *normalized ordered restriction map* is observed, subject to *partial digestion* error as well as error due to *misorientation*. Thus the result of the experiment is an ordered sequence of sites

$$0 < s_1 < s_2 < \cdots < s_l < 1,$$

where either the sequence or its reversal

$$0 < s_l^R < s_{l-1}^R < \cdots < s_1^R < 1,$$

could be assumed to be derived from the true normalized ordered restriction map

$$0 < h_1 < h_2 < \cdots < h_k < 1,$$

after partial digestion. By assumption, for each h_j and for each observation, the probability

$$Pr\left[\text{ there exists some } s_i \text{ s.t. } h_j = s_i \text{ or } h_j = s_i^R\right] = p$$

models the partial digestion.
Assumption: For the time being, we assume that the true normalized ordered restriction map has no *symmetric site*, i.e.,

$$\forall_i \, \forall_{j \neq i} \, h_i \neq h_j^R.$$

Let us also assume that the experiment is repeated n times resulting in n observed restriction maps whose orientations may be misspecified.

An algorithm to reconstruct the true map may proceed in two phases: In the first phase, all the molecules are folded by the mid-point, thus creating "folded-maps," in which the orientation of the molecule is no longer an issue. The individual "folded-maps" are combined to compute a "consensus folded-map." In the second phase, the "consensus folded-map" is unfolded back to create the restriction map where each individual site is assigned to either left half or right half, by examining the relative locations of pairs of restriction sites found in the original data (assuming that enough such information is available).

4.1 Phase 1:

Define a function

$$\begin{aligned} f &: (0,1) \to (0, \frac{1}{2}) \\ &: x \mapsto \begin{cases} x & \text{if } x \in (0, \frac{1}{2}); \\ x^R & \text{if } x \in (\frac{1}{2}, 1). \end{cases} \end{aligned}$$

In phase 1, our goal is to construct the set

$$\{f(h_1), f(h_2), \ldots, f(h_k)\},$$

which can be easily accomplished by considering the sets

$$\{f(s_{i1}), f(s_{i2}), \ldots, f(s_{il_i})\}, \quad i = 1, \ldots, n.$$

and proceeding in a manner similar to the one outlined in the preceding section. Using the arguments given earlier, we see that we will succeed in this phase with probability $e^{-e^{-c}}$, if $n \geq \left(\frac{\ln k + c}{p}\right)$.

4.2 Phase 2:

While one cannot recreate the map directly from the result of the phase 1, one can invert f correctly, if each computed site is further augmented with a sign value ($\in \{+1, -1\}$), where $+1$ denotes that the site belongs to the left half $[(0, \frac{1}{2})]$ and -1 denotes that the site belongs to the right half $[(\frac{1}{2}, 1)]$. Thus, we may define

$$\begin{aligned} \hat{f} &: (0, \frac{1}{2}) \times \{+1, -1\} \to (0, 1) \\ &: (f(h_j), \text{Sgn}\,) \mapsto \begin{cases} f(h_j) & \text{if Sgn } = +1; \\ f(h_j)^R & \text{if Sgn } = -1. \end{cases} \end{aligned}$$

We can assign the sign values correctly as follows: Define a graph $G = (V, E)$, where $V = \{f(h_1),\ f(h_2),\ \ldots,\ f(h_k)\}$ and $e = [f(h_a), f(h_b)] \in E$ if and only if

$$\exists_{s_{i,a}, s_{i,b}}\ f(s_{i,a}) = f(h_a) \text{ and } f(s_{i,b}) = f(h_b), \text{ for some } i.$$

Furthermore, label e with $+1$ if for some i, either $s_{i,a}$ and $s_{i,b} \in (0, \frac{1}{2})$ or $s_{i,a}$ and $s_{i,b} \in (\frac{1}{2}, 1)$ (*both sites belong to the same half*); and with -1 if for some i, either $s_{i,a} \in (0, \frac{1}{2})$ and $s_{i,b} \in (\frac{1}{2}, 1)$ or $s_{i,a} \in (\frac{1}{2}, 1)$ and $s_{i,b} \in (0, \frac{1}{2})$ (*two sites belong to different halves*). In other words,

$$\text{Sgn}\,(e) = \text{Sgn}\,[(\frac{1}{2} - s_{i,a})(\frac{1}{2} - s_{i,b})].$$

It is trivial to see that if the graph is *connected* then one can compute the correct vertex labels by first labeling an arbitrary vertex $+1$ (say, $f(h_j)$) and then labeling the remaining vertices by following the edge labels during a graph-search process. Thus if $f(h_i)$ and $f(h_j)$ are path connected by a simple path e_1, e_2, …, e_m then

$$\text{Sgn}\,(f(h_i)) = \text{Sgn}\,(e_1) \cdot \text{Sgn}\,(e_2) \cdots \text{Sgn}\,(e_m)\text{Sgn}\,(f(h_j)).$$

Next we compute an upper bound on the number of observations necessary for G to be connected. Let

$$n \geq \left(\frac{2 \ln k + 8c}{p^2}\right).$$

Let S_{h_1} denote the set of observations with a cut site matching $f(h_1)$. The number of such observations, $|S_{h_1}|$, follows a Binomial distribution $\sim S(n,p)$.

$$\begin{aligned} & Pr\left[S(n,p) \leq \left(\frac{\ln k + c}{p}\right)\right] \\ \leq \;& Pr\left[S(n,p) \leq (1-\epsilon_0)np\right], \quad \epsilon_0 \geq \frac{1}{2} \\ \leq \;& e^{-np/8} < e^{-c}. \end{aligned}$$

A cut corresponding to every $f(h_i)$ $[2 \leq i \leq k]$ occurs in an observation in S_{h_1}, when $|S_{h_1}| > \dfrac{\ln k + c}{p}$, with probability

$$(1-(1-p)^{|S_{h_1}|})^{k-1} \approx e^{-ke^{-p|S_{h_1}|}} \geq e^{-e^{-c}}.$$

Thus G is connected with a probability $> (1-e^{-c})e^{-e^{-c}}$ as $[f(h_1), f(h_i)]$ appears in G for all $2 \leq i \leq k$.

Summarizing: Let ϵ be a positive constant and $c \geq \ln(2/\epsilon)$. Then for

$$n \geq \max\left[\frac{\ln k + c}{p}, \frac{2\ln k + 8c}{p^2}\right],$$

with a probability at least $(1-\epsilon)$, the correct ordered restriction map can be computed in $O(nk^2)$ time. Also, see Appendix A1, for a slightly better bound when $p = O(1/k)$.

4.3 Optical Cuts

Next we shall consider the situation where we have additional spurious cuts (optical cuts) that do not correspond to any restriction sites. A sound probabilistic model for these spurious cuts can be given in terms of a Poisson process with parameters λ_f (thus the expected number of false cuts per molecule is λ_f). Hence, for any small region $[x, x+\delta x]$ in an observation,

$$\begin{aligned} Pr\left[\#\text{ false cuts} \in [x, x+\delta x] = 1\right] &= \lambda_f \delta x, \\ Pr\left[\#\text{ false cuts} \in [x, x+\delta x] \geq 2\right] &= o(\delta x). \end{aligned}$$

The probability that an observation contains exactly f spurious cuts is given by: $e^{-\lambda_f}\dfrac{\lambda_f^f}{f!}$. Typical observed values for λ_f are about 0.2 for Lambda clones, 0.5 for cosmids and 1.0 for BAC's. Thus, we expect roughly 1 false cut per $100Kb$.

Under this model, it is fairly trivial to see that the false cuts pose no serious problem. Our algorithm can be modified in a straight forward manner where Phase 1 computation needs to be somewhat more robust.

In phase 1, our goal is to construct the set

$$\{f(h_1), f(h_2), \ldots, f(h_k)\}.$$

This is accomplished by considering the observation-based sets

$$\{f(s_{i1}), f(s_{i2}), \ldots, f(s_{il_i})\}, \quad i = 1, \ldots, n.$$

and including only those $f(s_{ij})$'s that occur at least twice in the combined observations. In other words, if there exists an $i_1 \neq i_2$ such that if

$$\exists_{j_1,j_2}\ f(s_{i_1,j_1}) = f(s_{i_2,j_2}) = x,$$

then include x in the output set.

Assume that $n \geq \left(\frac{2(\ln k + c)}{p}\right)$, (with $c > 1.26$). Then if h_i is a true cut site, the probability that $f(h_i)$ is not included in the output is

$$\begin{aligned}(1-p)^n + np(1-p)^{n-1} &\leq (1 + p(n-1))e^{-p(n-1)} \\ &\leq e^{-p(n-1)/2} = \left(\frac{e^{-c}}{k}\right).\end{aligned}$$

Proceeding as before the probability that all k true sites will be included is thus bounded from below by $e^{-e^{-c}}$, Also, by the assumption regarding the distribution of spurious cuts, we see that the probability that a spurious cut is included in the final set is zero.

4.4 Symmetric Cuts

Next, assume that the true ordered restriction map consists of k asymmetric cuts and m symmetric cuts. Thus the total number of cuts is $k + 2m$. Note that a cut h_i is a symmetric cut, if both h_i and h_i^R are true cuts. Additionally, we assume that the observations are subject to the *partial digestion* errors, *misorientation* errors, *spurious cut* errors (determined by a Poisson process) and *symmetric cuts*.

In this case, we proceed as before with phase 1 from the preceding subsection, and again assuming that $n \geq \left(\frac{2(\ln k + c)}{p}\right)$, we will almost surely (with probability no smaller than $e^{-e^{-c}}$) construct a set

$$\{f(h_1), f(h_2), \ldots, f(h_k), f(h_{k+1}), \ldots, f(h_{k+m})\}.$$

Note that the $2m$ symmetric sites yield m values in the folded structure when f is applied.

However, before proceeding to phase 2, we will remove those $f(h_j)$'s from the preceding set that correspond to symmetric cuts. A simple approach we can take is to check each observation for the existence of symmetric cuts at positions s and s^R, where $f(s) = f(s^R) = f(h_j)$.

We claim that if $n \geq \left(\frac{\ln m + c}{p^2}\right)$ then the preceding steps correctly detect the symmetric cuts with probability greater than $e^{-e^{-c}}$. Note that assuming h_j to be a symmetric true cut, the probability that the above test fails in any particular observation is $(1 - p^2)$

and thus the probability that the symmetric cut h_j goes undetected in any of the n independent observations is $(1-p^2)^n$. Thus the probability that all m symmetric true cut sites are detected in the final map is given by $[1-(1-p^2)^n]^m > e^{-e^{-c}}$.

Again, by the assumption regarding the distribution of spurious cuts, we see that the probability that a spurious cut is included or symmetric cut is missed in the final set is zero.

At the end of this step, we are left with a set containing only asymmetric cuts

$$\{f(h_1), f(h_2), \ldots, f(h_k)\}.$$

At this point, we simply proceed with the phase 2 *mutatis mutandis* and claim results similar to the ones derived earlier.

4.5 Summary

Consider an ordered restriction map with $k+2m$ restriction sites, of which m are symmetric cuts. Assume that the postulated experiment observes these maps, with each observation suffering from partial digestion error ($p \leq 1$), misorientation error, spurious cuts (determined by a Poisson process with parameter λ_f), but no sizing error.

Theorem 4.1 *Let ϵ be a positive constant and $c \geq \ln(5/\epsilon)$. Then for*

$$n \geq \max\left[\frac{2(\ln(k+m)+c)}{p}, \frac{2\ln k + 8c}{p^2}, \frac{\ln m + c}{p^2}\right],$$

there is a probability of at least $(1-\epsilon)$ that the correct ordered restriction map can be computed in $O(n(L+k^2+m))$ time.

When

$$n < \frac{\ln(k+m)}{p(1+p)},$$

$(0 < p < 0.69)$, there is a probability greater than half that no algorithm can compute the correct ordered restriction map. □

5 Discretization

Next, we consider the effect of discretizing the map data by dividing it uniformly into several intervals of equal sizes. The main argument in favor of discretizing the data has been to accommodate the sizing errors that make the cut locations deviate from their true location. The main source for the sizing error has been the nonuniform attachment of the flurochromes that are necessary to visualize the DNA. We will study the effect of sizing error on discretization in a later section.

Let us assume that the clone DNA that we wish to analyze is of length L bps. Let Δ represent a small subinterval and $\delta = \Delta/L$. Thus the unit length is partitioned into $M = 1/\delta = L/\Delta$ consecutive subintervals. One assumes that it is not possible to distinguish

the restriction cuts and spurious cuts in each of these subintervals. Thus, we need to ensure that δ is significantly small so that no more than one true restriction cut location belongs to a subinterval. We now write $r = \lambda_f \delta = \Delta\lambda_f / L$ to denote the probability that we shall observe one spurious cut in a subinterval. Note that the probability that we shall observe f spurious cuts in any observation is given by

$$\binom{M}{f}(r)^f(1-r)^{M-f}, \quad \text{where } r = \frac{\lambda_f}{M} = \frac{\lambda_f \Delta}{L}.$$

Thus in the limit as $M \to \infty$ and $r \to 0$,

$$\lim_{M\to\infty} \binom{M}{f}(\lambda_f/M)^f(1-\lambda_f/M)^{M-f} = e^{-\lambda_f}\frac{\lambda_f^f}{f!},$$

the analysis given earlier holds true. Furthermore, if we are simply interested in the effect of finite M (and nonzero r), we are still able to prove that for realistic values of $r < p/27$ the earlier bounds still hold. Thus, it suffices to ensure that $\lambda_f/M < p/27$, or $M > 27\lambda_f/p$—for instance, M could be 270 and satisfy the inequality as long as $\lambda_f < 1.0$ and $p > 0.1$. (See appendix A2.)

Typical values for various clones may be as follows: for lambdas, M can range from 200 to $2,000$ and $r \approx 10^{-3}$–10^{-4}; for cosmids, M is $2,000$–$4,000$ and $r \approx 10^{-4}$; for BACs $M \approx 15,000$ and $r \approx 10^{-4}$. In general, even for significantly smaller (but still realistic) values of M, $r \ll p$.

5.1 Limit on M

It is worth noting that the discretization process makes it possible for spurious cuts to introduce a "wrong" cut site into final map. For instance, if each of the n observations contains a spurious cut in the same subinterval, then no algorithm can distinguish this spurious cut from a true cut (independent of digestion rate). Thus the probability that none of the M subintervals has a spurious cut in each of the n observations is given by

$$(1-r^n)^M.$$

Now if we assume that $n < \frac{\ln M}{\ln(1/r)}$, then the above probability is bounded from above by

$$\begin{aligned}(1-r^n)^M &< (1-r^{\ln M/\ln(1/r)})^M \\ &< \left(1-\frac{1}{M}\right)^M \\ &\leq \frac{1}{e} < \frac{1}{2}.\end{aligned}$$

Hence we must further guarantee that

$$n > \frac{\ln(L/\Delta)}{\ln(1/r)} = \frac{\ln(M)}{\ln(M/\lambda_f)},$$

since otherwise there is a probability of half or more that the computed map will be wrong.

6 Sizing Errors

Next, suppose we model the sizing error and analyze its effect. Before doing so, we need to derive some inequalities relating the size of the discretized subinterval (Δ) to several other external parameters. In particular, in order to infer the map correctly with probability greater than $1/\sqrt{2}$, we must guarantee that $\Delta \leq \left(\frac{2}{(k-1)p_E}\right)$, where k is the number of cuts and p_E denotes the probability that the restriction enzyme cuts at a site.

Assume that $\Delta > \left(\frac{2}{(k-1)p_E}\right)$. Let l denote the length of the smallest restriction fragment (piece of the molecule between two consecutive restriction sites). Note that the fragment lengths are distributed as $p_E e^{-p_E x}$, and the probability that a fragment is of length $\geq \Delta/3$ is

$$\int_{\Delta/3}^{\infty} p_E e^{-p_E x}\, dx = e^{-p_E \Delta/3}.$$

Thus the probability that the smallest of all $(k-1)$ fragments is no smaller than $\Delta/3$, is

$$e^{-(k-1)p_E \Delta/3} < e^{-2/3}.$$

Thus the probability that the smallest fragment is of length $\leq \Delta/3$ and that both ends of the fragment belong to the same subinterval is bounded by

$$\left(1 - e^{-2/3}\right)\ (1 - 1/3) > 1 - \frac{1}{\sqrt{2}}.$$

However, note that for a BAC clone, this implies that the largest value we may choose for $\Delta \leq 200bp$ (requiring M to be about 750).

Next assume that a true cut site at location h actually appears as a Gaussian distribution $\sim N(h, \sigma)$. Again, considering the complementary requirement to the one mentioned earlier, we must ensure that the observed cuts corresponding to the same true cut (at location h_i) belong to the same subinterval with high probability. As a result, we may require that

$$\forall_{1 \leq i \leq k}\ \exists_{1 \leq j \leq M}\ \ h_i \in (j\Delta + \sigma,\ (j+1)\Delta - \sigma),$$

with high probability (say, $\geq 1/\sqrt{2}$). Thus, we require that

$$\left[1 - \frac{2\sigma}{\Delta}\right]^k \approx e^{-2k\sigma/\Delta} \geq \frac{1}{\sqrt{2}}.$$

In other words, we require that $2k\sigma/\Delta \leq \ln 2/2$, and

$$\sigma \leq \frac{\ln 2}{4k}\Delta \leq \left(\frac{\ln 2}{2k(k-1)p_E}\right).$$

A simple calculation for the BAC example reveals that in order to guarantee the above inequality we need that $\sigma \leq 0.89bp$. Thus for all practical purposes, in order for the discretized algorithm to work with any degree of correctness, we must require the *observation to be free of sizing error*. As a result, one can explain why several algorithms devised to work with dicretization failed, while purely continuous versions (or some combination) have done well.

7 Experimental Verification

This section compares the performance of a program based on the maximum likelihood approach to map-computation (described in [AMS97]) with the theoretical bounds in the previous sections. At the time of this writing, AMS algorithm [AMS97] still remains the *only* algorithm that has worked successfully on raw experimental data, without access to any extraneous parameters or the final answer. In each case, when the computed map was verified with data (from sequence and gel data) derived independently and subsequent to the experiment, the algorithm was found to be remarkably successful.

For all the experiments described in this section, random data were generated using the data models of the previous sections. For each data model and assumed number of data molecules, we generated 20 random data samples and counted the fraction of these samples for which the maximum likelihood program computed the correct map. For each data model the number of data molecules is varied to obtain the fraction of cases solved correctly as a function of the number of data molecules. We show that in each case there is a fairly sharp transition from not being able to solve any of the 20 samples to being able to solve all 20 samples. Moreover this transition point lies within the theoretical bounds computed in the previous sections. Finally we examine the performance of the maximum likelihood program for the case where there is significant sizing error. In this case the discrete methods described previously fail to work altogether, whereas the maximum likelihood method continues to work, albeit requiring a larger number of data molecules as the sizing error increases.

The maximum likelihood approach described in [AMS97] is based on a continuous (non-discrete) modeling of the data. The modeling of sizing error in the model results in a singularity in the probability density when the sizing error is zero. Therefore this case was approximated by assuming a small sizing error of 10^{-5} of the total molecule size, $\sigma = 1.5$bp. Each data model is specified by providing the number (k) and value of the actual cut locations, the sizing error in the form of a standard deviation (σ), a digest rate (p) and a false cut rate (λ_f). For each model, random data is generated with the help of a random number generator in a straight forward fashion: For each of the actual cuts, we draw a random number uniformly from [0,1], and if this value is below p, the cut is assumed to be present. We then draw another random number from the standard Gaussian distribution to determine the location of the cut, thus modeling the effect of sizing error. Next, false cuts are added by first drawing a random sample from a Poisson distribution with mean λ_f to determine the number of false cuts, and then drawing the required number of random samples uniformly over [0,1] to get the false cut locations.

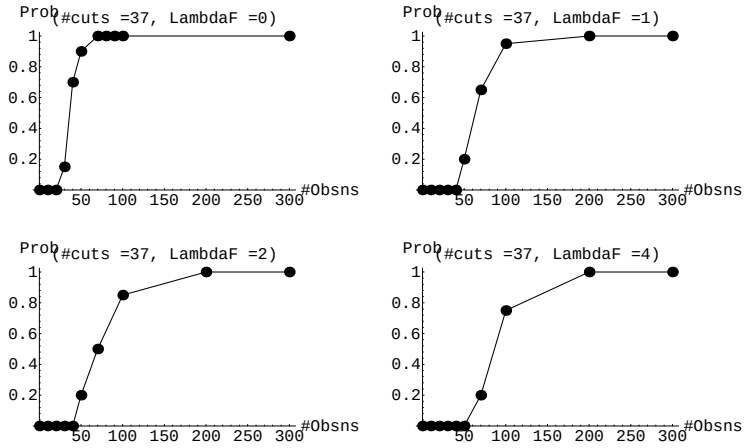

Figure 1: **Experimental Results:** #Cuts, $k = 37$, $\sigma = 1.5$bp, $p = 0.1$

This step results in the generation of one *in-silico* "molecule." This process is repeated to get the required number of molecules to make up one data set. This data set is then input as raw data to our maximum likelihood program, and the resulting map is scored a success if the number of cuts is correct, and the location of each cut is within one standard deviation (σ) of the correct location. (Note that σ is the standard deviation for the cuts of one sample molecule: the map computed by the AMS algorithm typically has a sizing error much less than that since the data from all molecules are averaged). This process is repeated for a total of 20 samples and the fraction of times the program succeeds is recorded against the data sample parameters (k, σ, p, λ_f, number of molecules). The whole process (i.e., the one generating 20 samples) was repeated for different values of the parameters. The number of cuts was varied using the values $k = 0$, 1, 2, 5, 10, 20

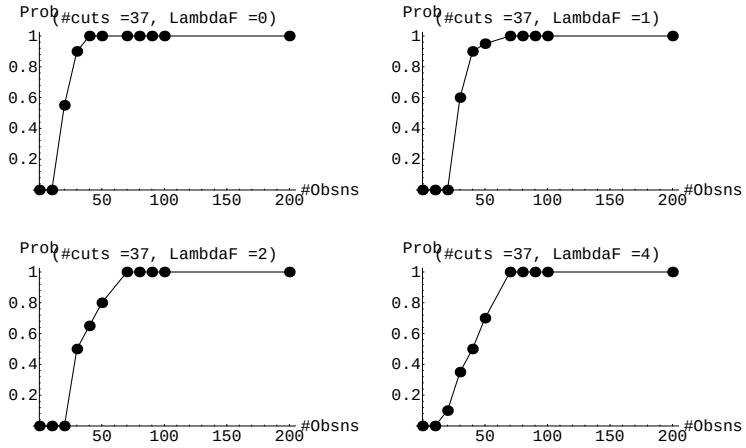

Figure 2: **Experimental Results:** #Cuts, $k = 37$, $\sigma = 1.5$bp, $p = 0.2$

and 37. The values of p tested were p=0.10 and p=0.20. The values of λ_f tested were λ_f=0 (no false cuts) and λ_f=1,2 and 4. For most experiments we selected $\sigma = 1.5$bp to approximate no sizing error, but for a small number of experiments with $k = 2$ and 37 we also tested σ= 150bp, 300bp, 750bp and 1.5Kb. Most experiments were repeated with the number of molecules set at 10, 20, 30, 40, 50, 70, 100, 200, 500 and 1000, and in a few instances 2000 or 5000.

The results are summarized in a series of graphs showing the success rate (out of 20 samples) as a function of the number of molecules used. The graph in Figure 1 shows the case for $k = 37$ and $\lambda_f = 0$, 1, 2 and 4, which corresponds to the case analyzed in Section 4. We see that for $p = 0.10$ and $\lambda_f = 1$ a sharp transition occurs when the number of molecules increases from 30 to 50. At 70 or more molecules the AMS

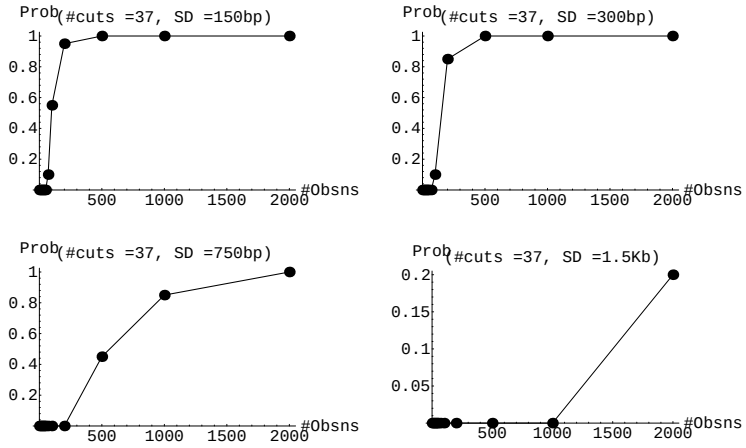

Figure 3: **Experimental Results:** #Cuts, $k = 37$, $\lambda_f = 1$, $p = 0.1$

algorithm never (out of 20 experiments) fails to find the correct map, whereas for 20 or less molecules it invariably fails to find the correct map. For $p = 0.20$ (Figure 2), the transition (from probability of near 0 to near 1) occurs at a lower value of around 20–30 molecules. Compare this with the theoretical bounds on the number of molecules required from section 4 of between 30 and 100 (lower bound and upper bound respectively).

When the number of (true) cuts in the molecules is changed to $k = 20$, 10, 5 and 2, similar graphs are obtained: Figures 4 and 5 show the results for the case $k = 20$; Figures 6 and 7, for the case $k = 10$; Figures 8 and 9, for the case $k = 5$; Figures 10 and 11, for the case $k = 2$; Figures 13 and 14, for the case $k = 1$ and Figure 15, for $k = 0$. The main trend is an increase in the number of molecules required as k is reduced down to $k = 2$: for instance, with $k = 2$ and $p = 0.1$, 500 molecules are required to find the

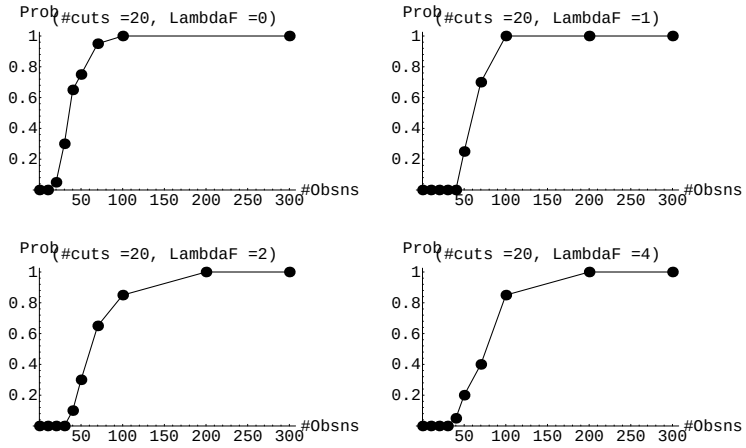

Figure 4: **Experimental Results:** #Cuts, $k = 20$, $\sigma = 1.5$bp, $p = 0.1$

correct map in every case ($\lambda_f = 0$, 1, 2 and 4), in contrast to just 200 for $k = 37$. This observation agrees with the theory from sections 4 and 5 which shows that the bounds increase slowly as k is decreased. However, the case $k = 1$, Figures 13 and 14, show that fewer molecules are required: e.g., with $p = 0.1$ and $\lambda_f = 1$, 200 molecules are sufficient to find the correct map. The reason is that orientation is less of a problem with only 1 cut.

Figure 3 shows what happens with $k = 37$ when the sizing error is increased to 150bp, 300bp, 750bp and 1.5Kb, respectively. With $p = 0.10$ and $\lambda_f = 1$, the number of molecules required to find the correct map in every case increases from 200 to about 5000 as the sizing error increases. Figure 12 shows what happens at $k = 2$ when sizing error is increased similarly. In this case the number of molecules increases from an already

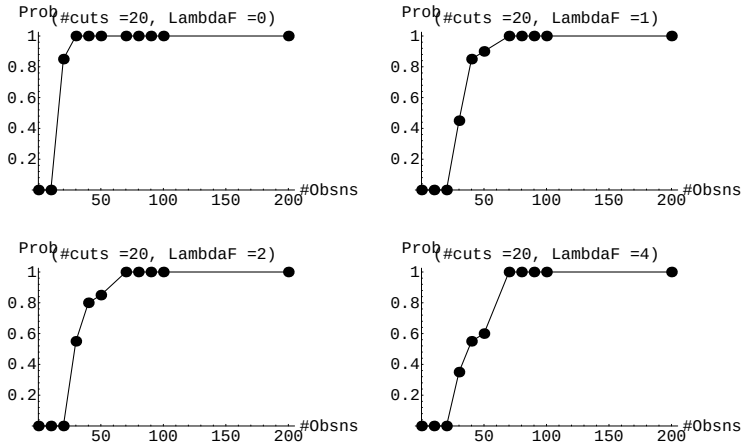

Figure 5: **Experimental Results:** #Cuts, $k = 20$, $\sigma = 1.5$bp, $p = 0.2$

larger value, but more slowly: it increases from 500 to 2000. While we do not have any theoretical bounds for this case, the intuition is that while it is harder to get the correct orientation with $k = 2$ than with $k = 37$, it is less likely that neighboring cuts will be confused with each other due to sizing errors when $k = 2$ than when $k = 37$.

8 Discussion: Genomic Mapping

The strategies for genome-wide genotype or haplotype mapping using single molecule optical maps are similar in spirit to the approaches for clone mapping, described in this paper; but there are also several differences in the details of the implementation as new and powerful heuristics need to be incorporated in order to tame the computational

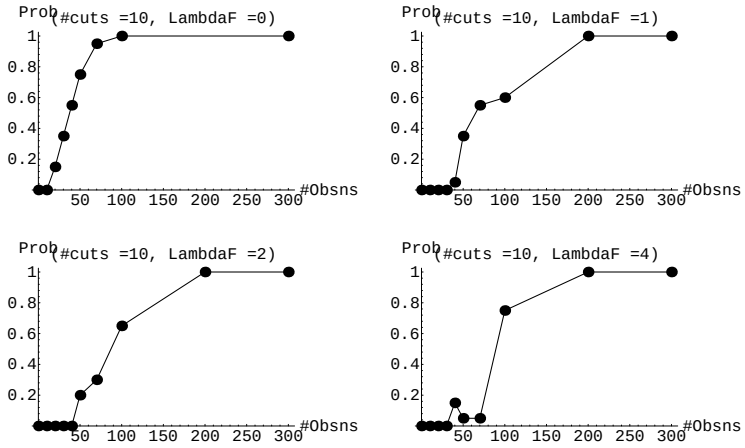

Figure 6: **Experimental Results:** #Cuts, $k = 10$, $\sigma = 1.5$bp, $p = 0.1$

complexity of searching over the hypotheses space. The details of the algorithm can be found elsewhere [AnMS99]. We summarize below a 0-1 law applicable to the experiment design in this genomic-mapping setting; the derivation of the following result is in [AM01].

Consider an optical mapping experiment for genome-wide shotgun mapping for a genome of size G and involving M molecules each of length L_d. Thus the coverage is ML_d/G. Let the a fragment of true size X have a measured size $\sim \mathcal{N}(X, \sigma^2 X)$. Let the average true fragment size be L, and the digestion rate of the restriction enzyme be p. Thus the average relative sizing error $R = \sigma\sqrt{p/L}$ and the average size of aligned fragments will be L/p^2. As usual, let θ represent the minimum "overlap threshold." Hence the expected number of aligned fragments in a valid overlap is at least $n = \theta L_d p^2/L$. Let $d = 1/p$, the inverse of the digest rate. Feasible experimental parameters are those that

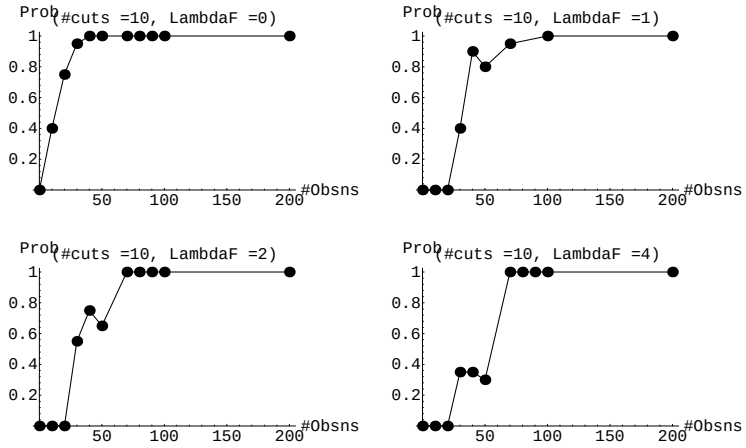

Figure 7: **Experimental Results:** #Cuts, $k = 10$, $\sigma = 1.5$bp, $p = 0.2$

result in an acceptable (e.g. $\leq 10^{-3}$) False Positive rate FPT:

$$FPT \approx 2M^2 \binom{\lceil 2nd + 2 \rceil}{\lfloor 2n(d-1) \rfloor} \frac{(R\sqrt{\frac{\pi e}{8}})^n}{\sqrt{n\pi}} e^{\frac{2(d-1)nR}{\sqrt{2\pi}}}$$

To achieve acceptable false positive rate, one needs to choose an acceptable value for the experimental parameters: p, σ, L_d and coverage. FPT exhibits a sharp phase transition in the space of experimental parameters. Thus the success of a mapping project depends extremely critically on a prudent combination of experimental errors (digestion rate, sizing), sample size (molecule length and number of molecules) and problem size (genome length). Relative sizing error can be lowered simply by increasing L with a choice of rarer-cutting enzyme and digestion rate can be improved by better chemistry.

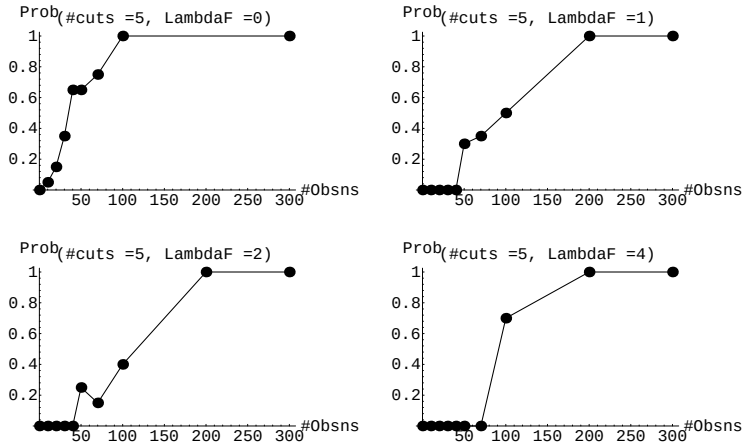

Figure 8: **Experimental Results:** #Cuts, $k = 5$, $\sigma = 1.5$bp, $p = 0.1$

As an example, for a human genome of size $G = 3,300Mb$ and a desired coverage of $6\times$, consider the following experiment. Assume a typical value of molecule length $L_d = 2Mb$. If the enzyme of choice is PAC I, the average true fragment length is about $25Kb$. Assume a minimum overlap[1] of $\theta = 30\%$. Assume that the sizing error for a fragment of $30kb$ is about $3.0kb$, and hence $\sigma^2 = 0.3kb$. With a digest rate of $p = 82\%$ we get an unacceptable $FPT \approx 0.0362$. However just increasing p to 86% results in an acceptable $FPT \approx 0.0009$. Alternately, reducing average sizing error from $3.0kb$ to $2.4kb$ while keeping $p = 82\%$ also produces an acceptable $FPT \approx 0.0007$.

Acknowledgment. Our thanks go to Naomi Silver, Rohit Parikh, Raghu Varadhan, Joel Spencer, Alan Frieze, Sylvain Cappel, Bruce Donald, Mike Wigler and Laxmi Parida

[1]This value should be selected to minimize FPT.

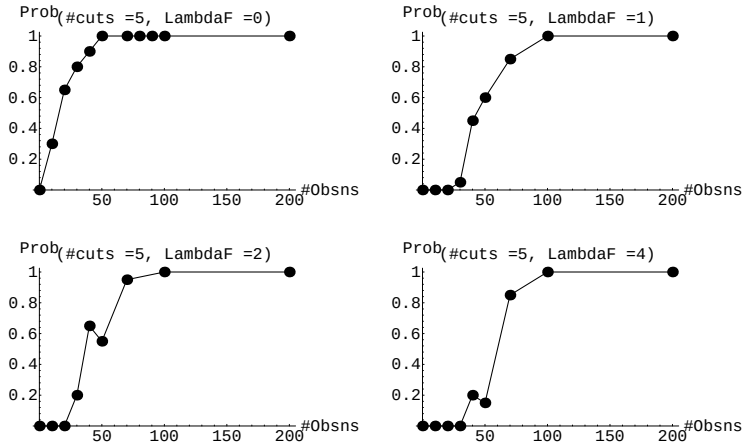

Figure 9: **Experimental Results:** #Cuts, $k = 5$, $\sigma = 1.5$bp, $p = 0.2$

for many helpful comments and encouragement.

References

[ASE92] N. Alon, J.H. Spencer and P. Erdös (1992). *The Probabilistic Method*. Wiley Interscience. (John Wiley & Sons, Inc., New York.)

[AM01] T.S. Anantharaman and B. Mishra (2001). A Probabilistic Analysis of False Positives in Optical Map Alignment and Validation. *Algorithms in Bioinformatics*, First International Workshop, WABI 2001 Proceedings **LNCS 2149**, 27–40. (Springer-Verlag, New York.)

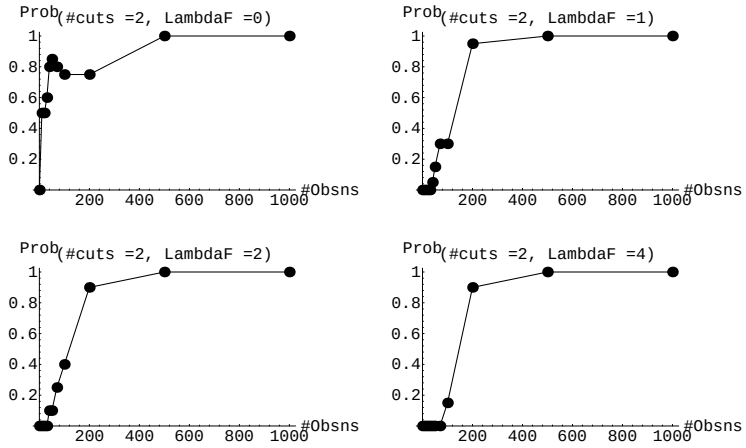

Figure 10: **Experimental Results:** #Cuts, $k = 2$, $\sigma = 1.5$bp, $p = 0.1$

[AMS97] T.S. Anantharaman, B. Mishra and D.C. Schwartz (1997). Genomics via Optical Mapping II: Ordered Restriction Maps. *Journal of Computational Biology* **4**(2), 91–118.

[Ana+97b] T.S. Anantharaman et al. (1997). Statistical Algorithms for Optical Mapping of the Human Genome. *1997 Genome Mapping and Sequencing Conference*, Cold Spring Harbor, New York.

[AnMS99] T.S. Anantharaman, B. Mishra and D.C. Schwartz (1999). Genomics via Optical Mapping III: Contiging Genomic DNA and Variations. *Proceedings 7th Intl. Cnf. on Intelligent Systems for Molecular Biology: ISMB '99* **7**, 18–27. (AAAI Press, New York).

[AsMS99] C. Aston, B. Mishra and D.C. Schwartz (1999). Optical Mapping and Its Potential for Large-Scale Sequencing Projects. *Trends in Biotechnology* **17**, 297–302.

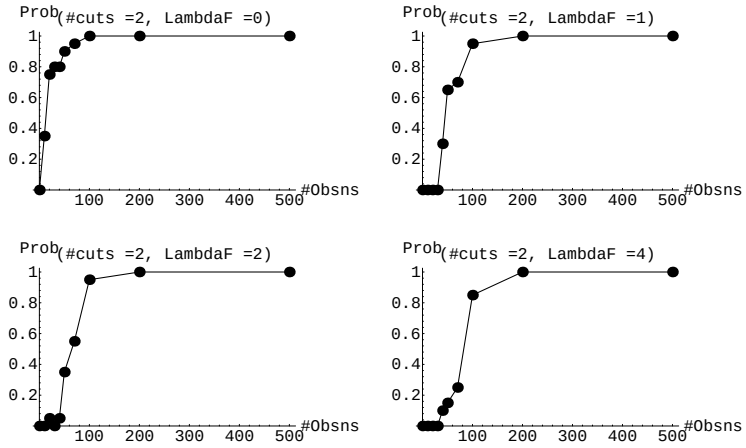

Figure 11: **Experimental Results:** #Cuts, $k = 2$, $\sigma = 1.5$bp, $p = 0.2$

[Cai+98] W. Cai et al., High Resolution Restriction Maps of Bacterial Artificial Chromosomes Constructed by Optical Mapping," *Proc. National Academy of Science*, (In Press), 1998.

[Giacalone+00] J. Giacalone et al. (2000). Optical Mapping of BAC Clones from the Human Y Chromosome DAZ Locus. *Genome Research* **10**(9), 1421–1429.

[Jing+99] J. Jing et al., "Optical Mapping of *Plasmodium falciparum* Chromosome 2," *Genome Research*, **9**:175–181, 1999.

[KS98] R. Karp and R. Shamir (2000). Algorithms for Optical Mapping. *Journal of Computational Biology* **7**(1-2), 303–316.

[Lai+99] Z. Lai et al. (1999). A Shotgun Sequence-Ready Optical Map of the Whole *Plasmodium falciparum* Genome. *Nature Genetics* **23**(3), 309–313.

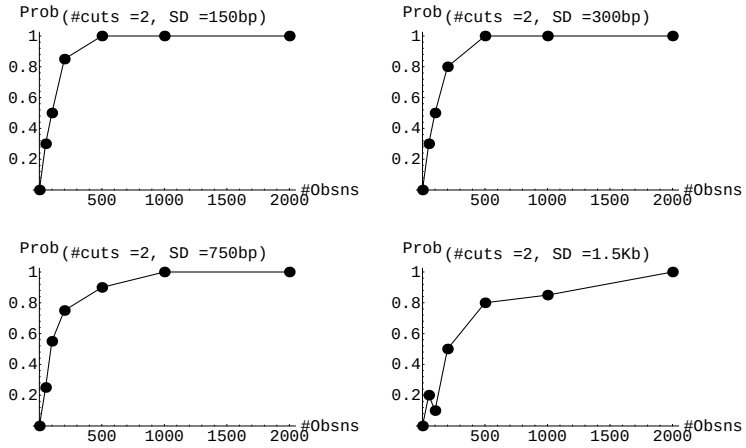

Figure 12: **Experimental Results:** #Cuts, $k = 2$, $\lambda_f = 1$, $p = 0.1$

[Lim+01] A. Lim et al. (2001). Shotgun Optical Maps of the Whole *Escherichia coli* 0157:H7 Genome. *Genome Research* **11**(9), 1584–1593.

[lin+99] J. Lin et al. (1999). Whole Genome Shotgun Optical Mapping of *Deinococcus radiodurans*. *Science* **285**(5433), 1558–1562.

[Mishra03] B. Mishra (2003). Optical Mapping. *Encyclopedia of the Human Genome*, Nature Publishing Group. (Macmillan Publishers Limited, London.)

[MP00] B. Mishra and L. Parida (2000). Partitioning Single-Molecule Maps into Multiple Populations: Algorithms And Probabilistic Analysis. *Discrete Applied Mathematics (The Computational Molecular Biology Series)* **104**(l-3), 203–227.

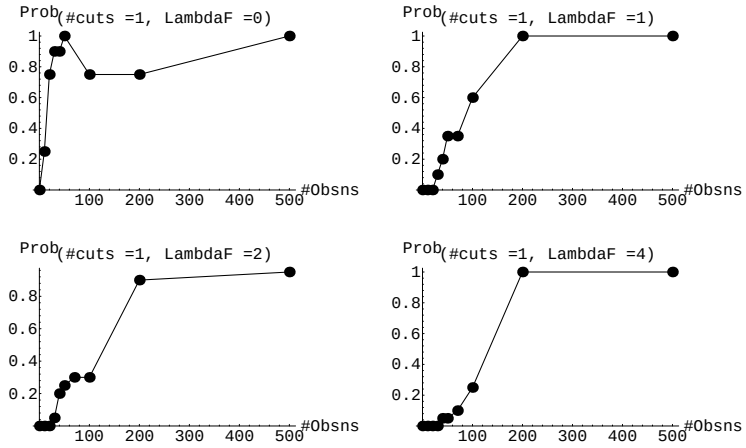

Figure 13: **Experimental Results:** #Cuts, $k = 1$, $\sigma = 1.5$bp, $p = 0.1$

[MP96] S. Muthukrishnan and L. Parida (1997). Towards Constructing Physical Maps by Optical Mapping: An Effective Simple Combinatorial Approach. In *Proceedings First Annual Conference on Computational Molecular Biology*, (RECOMB '97), ACM Press, 209–215.

[Parida98] L. Parida (1998). *Algorithmic Techniques in Computational Genomics*, PhD Thesis, Computer Science Department, New York University, New York.

[Sam+95] A. Samad et al. (1995). Mapping the Genome One Molecule At a Time—Optical Mapping. *Nature* **378**, 516–517.

[Skiadas+99] J. Skiadas et al. (1999). Optical PCR: Genomic Analysis by Long-Range PCR and Optical Mapping. *Mammalian Genome* **10**, 1005–1009.

[Spe87] J. Spencer (1987). *Ten Lectures on the Probabilistic Method*. (Society for Industrial and Applied Mathematics, Philadelphia, PA.)

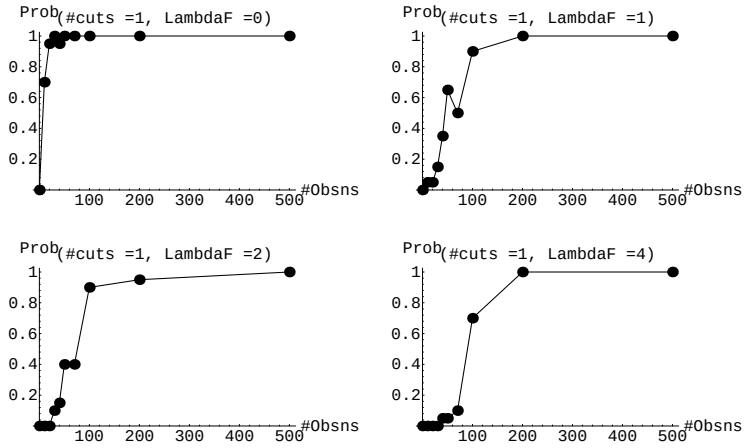

Figure 14: **Experimental Results:** #Cuts, $k = 1$, $\sigma = 1.5$bp, $p = 0.2$

[Zhou+02] S. Zhou et al. (2002). A Whole-Genome Shotgun Optical Map of *Yersinia pestis* Strain KIM. *Appl Environ Microbiol* **68**(12), 6321–6331.

A1. Bound for section 4.2

Assume the same notation as in section §4.2: We can provide a somewhat better bound when p is small, i.e., $p \approx 1/k$.

Let α be a function of p and k:

$$\alpha \equiv 1 - (1-p)^k - kp(1-p)^{k-1} \geq 1 - (1 + p(k-1))e^{-p(k-1)},$$

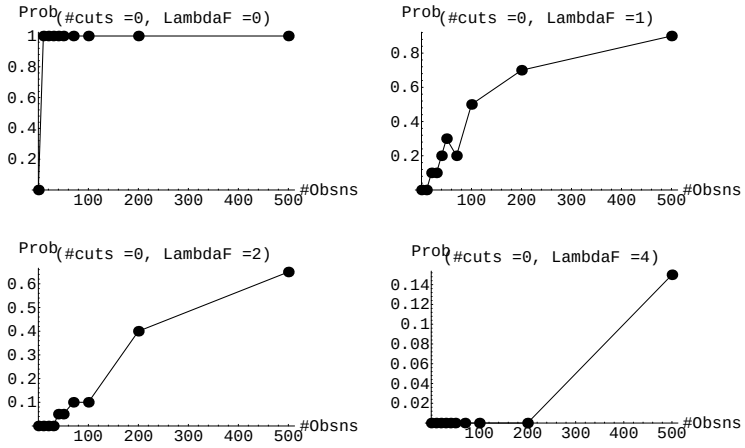

Figure 15: **Experimental Results:** #Cuts, $k = 0$

and let

$$n \geq \left(\frac{k^2}{2\alpha}\right) \ln\left(\frac{k}{k - \ln k - c}\right), \quad \text{where } k - \ln k > c.$$

Construct a random subgraph $G_R = (V, E_R)$ of G as follows: For any given observation with two or more cuts choose one edge at random from all the possible edges that the observation contributes to G. Discard those observations with fewer than two cuts. Thus with every observation, when we add an edge we do so uniformly randomly and independent of all the other edges chosen in G_R. Note that the probability that an observation has two or more cuts is α and the probability that an edge is added to G_R in an observation is $2\alpha/k(k-1) > 2\alpha/k^2$.

#Cuts, k	Digest rate, p_c	$\lambda_f = 0$	$\lambda_f = 1$	$\lambda_f = 2$	$\lambda_f = 4$
37	0.1	50	100	200	200
	0.2	40	70	70	70
20	0.1	100	100	200	200
	0.2	30	70	70	70
10	0.1	100	200	200	200
	0.2	50	70	70	70
5	0.1	100	200	200	200
	0.2	50	100	100	100
2	0.1	500	500	500	500
	0.2	100	200	200	200
1	0.1	70	200	200	200
	0.2	30	200	200	200
0	—	1	500	—	—

Table 1: Summary of Experimental Results. Number of molecules necessary as functions of the parameters: #Cuts, $k \in \{0..37\}$, Digest rate $p \in \{0.1, 0.2\}$ and $\lambda_f \in \{0, 1, 2, 4\}$.

For any pair $[f(h_i), f(h_j)]$, the probability that this edge does not occur in G_R is less than

$$\left(1 - \frac{2\alpha}{k^2}\right)^n \leq e^{-2\alpha n/k^2} \leq 1 - \frac{\ln k + c}{k},$$

and thus the "edge-probability," p_e (the probability that this edge occurs in G_R) is

$$p_e \geq \left(\frac{\ln k + c}{k}\right).$$

Thus by the well-known result on the connectivity in random graphs [Spe87], we see that with $p_e \geq \left(\frac{\ln k + c}{k}\right)$,

$$\lim_{k\to\infty} Pr\left[G_{k,p_e} \text{ is connected}\right] = e^{-e^{-c}}.$$

Note that

$$\alpha \geq \min\left[\frac{4}{5}, \frac{p^2(k-1)^2}{16}\right]$$

and if $k \gg c$ then

$$\ln\left(\frac{k}{k - \ln k - c}\right) \leq (1 + o(1))\left(\frac{\ln k + c}{k}\right).$$

Thus it suffices for our purpose to choose

$$n \geq \max\left[\left(\frac{5 + o(1)}{8}\right)k(\ln k + c), \frac{8 + o(1)}{p^2}\frac{\ln k + c}{k}\right].$$

Theorem 8.1 *Let ϵ be a positive constant and $c \geq \ln(2/\epsilon)$. Then for*

$$n \geq \max\left[\frac{\ln k + c}{p}, \left(\frac{5+o(1)}{8}\right)k(\ln k + c), \left(\frac{8+o(1)}{p^2}\right)\frac{\ln k + c}{k}\right],$$

with a probability at least $1-\epsilon$, the correct ordered restriction map can be computed in $O(nk^2)$ time. □

Furthermore, we conclude that, for $p \geq 1/k$, $n = O(k \log k)$ observations suffice to find the true map without any other prior knowledge of p.

A2. Discretization

As before, let us assume that the clone DNA is of length L bps. Let Δ represent a small subinterval and $\delta = \Delta/L$. Thus the unit length is partitioned into $M = 1/\delta = L/\Delta$ consecutive subintervals. We write $r = \lambda_f \delta = \Delta\lambda_f/L$ to denote the probability that we shall observe one spurious cut in a subinterval.

Typical values for various clones may be as follows: for lambdas, M can range from 200 to $2,000$ and $r \approx 10^{-3}$–10^{-4}; for cosmids, M is $2,000$–$4,000$ and $r \approx 10^{-4}$; for BACs $M \approx 15,000$ and $r \approx 10^{-4}$. In general, even for significantly smaller (but still realistic) values of M, $r \ll p$.

Bounds

We write $\hat{p} = p + r - pr$ to denote the probability that a subinterval contains a true or spurious cut site. We will use the following simplifying assumption:

$$27r < p.$$

More precisely, $\hat{p}/6r > 2e - 1$.

We summarize the bounds as follows:

Theorem 8.2 *Let ϵ be a positive constant and $c \geq \ln(5/\epsilon)$. Then for*

$$n \geq \frac{9}{p}\max\left[\ln(k+m) + c, \frac{2(\ln k + c)}{p}, \frac{\ln m + c}{p}, (\ln(L/2\Delta - k - m) + c)\right],$$

($L > 2\Delta$ and $r < p/27$), the probability that the correct ordered restriction map can be computed in $O(n(L + k^2 + m))$ time is at least $1-\epsilon$. □

We will now introduce two parameters $\epsilon_1 = \hat{p}/6r$ and ϵ_0, and guarantee that $\epsilon_1 > 2e-1$ and $\epsilon_0 \geq 1/2$. Furthermore, we have

$$(1+\epsilon_1) < \hat{p}/4r.$$

Phase 1 a

In phase 1 a, our goal is to construct the set

$$\{f(h_1), f(h_2), \ldots, f(h_{k+m})\},$$

by considering the observation-based sets

$$\{f(s_{i1}), f(s_{i2}), \ldots, f(s_{il_i})\}, \quad i = 1, \ldots, n.$$

and including only those $f(s_{ij})$'s that occur in *significantly large numbers* of times, determined by a threshold Th_1. Suppose that a location $f(h)$ corresponds to a true location, then the number of $f(s_{ij})$'s equal to $f(h)$ must follow a Binomial distribution $\sim S(n, \hat{p})$, if it is an asymmetric cut and $\sim S(n, 2\hat{p})$, if it is a symmetric cut. If on the other hand, $f(h)$ does not correspond to any true location, then the number of $f(s_{ij})$'s equal to this $f(h)$ must follow a Binomial distribution $\sim S(n, 2r)$.

If we set the threshold at

$$Th_1 \quad = \quad (1+\epsilon_1)2nr < \left(\frac{\hat{p}}{4r}\right) 2nr < \frac{n\hat{p}}{2}.$$

then $Th_1 = (1-\epsilon_0)n\hat{p}$, where $\epsilon_0 \geq \frac{1}{2}$.

By assumption (statement of the theorem above),

$$n > \frac{8}{\hat{p}} \max\left[\ln(k+m) + c, \ln(M/2 - k - m) + c\right].$$

Thus (using the Chernoff bound [ASE92])

$$\begin{aligned} & Pr\left[S(n,\hat{p}) \leq (1-\epsilon_0)n\hat{p}\right] \\ \leq \quad & e^{-(\epsilon_0^2/2)n\hat{p}} \\ < \quad & e^{-n\hat{p}/8} < \frac{e^{-c}}{k+m}. \end{aligned}$$

Thus the probability that all the correct cuts appear in the computed set is bounded from below by $e^{-e^{-c}}$.

Again, using the Chernoff bound [ASE92] in the other direction, we get

$$\begin{aligned} & Pr\left[S(n,2r) \geq (1+\epsilon_1)2nr\right] \\ \leq \quad & 2^{-(1+\epsilon_1)2nr} < 2^{-(\hat{p}/6r)2nr} \\ \leq \quad & e^{-((\ln 2)/3)n\hat{p}} \\ < \quad & e^{-n\hat{p}/8} < \frac{e^{-c}}{M/2 - k - m}. \end{aligned}$$

Thus the probability that no spurious cut appears in the computed set $> e^{-e^{-c}}$.

Phase 1 b

In phase 1 b, our goal is to construct the set of asymmetric cuts

$$\{f(h_1), f(h_2), \ldots, f(h_k)\},$$

by eliminating the symmetric cuts. Suppose that a location $f(h)$ corresponds to a symmetric true cut site, then the number of times an observation has sites at $s' = f(h)$ and $s'' = f(h)^R$ must follow a Binomial distribution $\sim S(n, \hat{p}^2)$. If on the other hand, $f(h)$ is not a symmetric site, then the corresponding number must follow a Binomial distribution $\sim S(n, \hat{p}r)$.

If we set the threshold at

$$Th_2 \quad = \quad (1+\epsilon_1)n\hat{p}r < \left(\frac{\hat{p}}{4r}\right) n\hat{p}r < \frac{n\hat{p}^2}{2}.$$

then $Th_2 = (1-\epsilon_0)n\hat{p}^2$, where $\epsilon_0 \geq \frac{1}{2}$.

By assumption (statement of the theorem above),

$$n > \frac{1}{\hat{p}^2} \max\left[8(\ln m + c), \left(\frac{6}{\ln 2}\right)(\ln k + c)\right].$$

Thus (using the Chernoff bound [ASE92])

$$\begin{aligned} & Pr\left[S(n, \hat{p}^2) \leq (1-\epsilon_0)n\hat{p}^2\right] \\ & \leq \quad e^{-(\epsilon_0^2/2)n\hat{p}^2} \\ & < \quad e^{-n\hat{p}^2/8} < \frac{e^{-c}}{m}. \end{aligned}$$

Thus the probability that all the symmetric cuts are correctly classified is bounded from below by $e^{-e^{-c}}$.

Again, using the Chernoff bound [ASE92] in the other direction, we get

$$\begin{aligned} & Pr\left[S(n, \hat{p}r) \geq (1+\epsilon_1)n\hat{p}r\right] \\ & \leq \quad 2^{-(1+\epsilon_1)n\hat{p}r} < 2^{-(\hat{p}/6r)n\hat{p}r} \\ & \leq \quad e^{-((\ln 2)/6)n\hat{p}^2} < \frac{e^{-c}}{k}. \end{aligned}$$

Thus the probability that no symmetric cut is misclassified is bounded from below by $e^{-e^{-c}}$.

Phase 2

The proof proceeds in a manner similar to the one given for the non-discretized case. In phase 2, our goal is to assign consistent sign labels to the asymmetric cuts

$$\{f(h_1), f(h_2), \ldots, f(h_k)\},$$

so that the final map can be constructed correctly with high probability.

Let S_{h_1} denote the set of observations containing a cut site matching $f(h_1)$, and $S^t_{h_1}$ denote the set containing a true cut site matching $f(h_1)$. Note that $|S_{h_1}| \geq |S^t_{h_1}|$ and $|S^t_{h_1}|$ follows a Binomial distribution $\sim S(n,p)$. Using the Chernoff bound, we have

$$Pr\left[S(n,p) < \frac{9}{p}(\ln k + c)\right] < e^{-np/8} < e^{-c}.$$

Let n_1 be the number cut sites matching h_1. Thus

$$n_1 \geq \frac{8(\hat{p}+r)}{\hat{p}^2+r^2}(\ln k + c) = \frac{8}{\beta_+}(\ln k + c),$$

with a probability $> 1 - e^{-c}$. Here $\beta_+ \equiv (\hat{p}^2+r^2)/(\hat{p}+r)$.

Consider a potential edge $[f(h_1), f(h_i)]$. Let n_i denote the number of times two cut sites matching $f(h_1)$ and $f(h_i)$, respectively, appear in the same half [either in $(0,1/2)$ or in $(1/2,1)$] in an observation in S_{h_1}. If the correct edge labeling is $+1$ then n_i has a Binomial distribution $\sim S(n_1,\beta_+)$, where $\beta_+ \equiv (\hat{p}^2+r^2)/(\hat{p}+r)$. If on the other hand, the correct edge labeling is -1 then n_i has a Binomial distribution $\sim S(n_1,\beta_-)$, where $\beta_- \equiv (2\hat{p}r)/(\hat{p}+r)$).

Set the threshold at

$$\begin{aligned} Th_3 &= (1+\epsilon_1)n_1\beta_- < \left(\frac{\hat{p}}{4r}\right) n_1\beta_- \\ &< \frac{n_1\hat{p}^2}{2(\hat{p}+r)} \leq \frac{n_1\beta_+}{2}. \end{aligned}$$

Thus $Th_3 = (1-\epsilon_0)n_1\beta_+$, where $\epsilon_0 \geq \frac{1}{2}$.

Thus (using the Chernoff bound [ASE92])

$$\begin{aligned} &Pr\left[S(n_1,\beta_+) \leq (1-\epsilon_0)n_1\beta_+\right] \\ &\leq e^{-(\epsilon_0^2/2)n_1\beta_+} < e^{-n_1\beta_+/8} < \frac{e^{-c}}{k}. \end{aligned}$$

Again, using the Chernoff bound [ASE92] in the other direction, we get

$$\begin{aligned} &Pr\left[S(n_1,\beta_-) \geq (1+\epsilon_1)n_1\beta_-\right] \\ &\leq 2^{-(1+\epsilon_1)n_1\beta_-} < 2^{-(1+\hat{p}/6r)n_1\beta_-} \\ &\leq e^{-((\ln 2)/3)n_1\beta_+} < \frac{e^{-c}}{k}. \end{aligned}$$

Thus it follows that the probability that all the edges receive the correct edge labeling is $> (1-e^{-c})e^{-e^{-c}}$. This concludes the proof.