# Chapter 6

# Clone Maps

> *"Nobody's honest. Scientists are not honest. And people usually believe that they are. That makes it worse. By honest I don't mean that you only tell what's true. But you make clear the entire situation. You make clear all the information that is required for somebody else who is intelligent to make up their mind."*
>
> —Richard Feynman, *The Meaning of It All*, pp. 106.

## 6.1 Some Preliminary Remarks

We study several simple models for optical mapping and explore their power and limitations when applied to the construction of maps of clones (e.g., lambdas, cosmids, BACs and YACs). We provide precise lower and upper bounds on the number of clone molecules needed to create the correct map of the clone. Our probabilistic analysis shows that as the number of clone molecules 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 (for typical values of the parameters, the transition point is around 70 molecules).

## 6.2 Problem Formulation

The underlying bio-chemical problem concerns with the construction of ordered restriction map of a clone (a piece of DNA of length $L$). Typical values of $L$ are 2–20$Kb$ (lambda's), 20–45$Kb$ (cosmids) 150–200$Kb$ (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.

# 6.3 False Negative Errors: Partial Digestion

Let us postulate an experiment, where the desired *normalized ordered restriction map* is observed, subject to *partial digestion* error 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 $n$-many 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 straightforward 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 $n \geq \frac{c}{p} + \frac{\ln k}{p}$ ($k \geq 1$ and $c \geq 1$) then the result of the preceding algorithm is correct with probability greater than $e^{-e^{-c}} e^{-(e^{-2c})/2k}$. Note that the probability that a cut site $h_j$ does not appear in any particular observation is $(1-p)$ and thus the probability that $h_j$ does not appear in any of the $n$ independent observations is $(1-p)^n$. Thus, we see that the probability that the cut site $h_j$ appears in the final result is $[1-(1-p)^n]$. Note that

$$\begin{aligned}(1-p)^n &\leq e^{-pn}\\ &\leq e^{-c-\ln k} = \frac{e^{-c}}{k},\end{aligned}$$

and
$$1-(1-p)^n \geq 1-\frac{e^{-c}}{k}.$$
Thus the probability that all $k$ true cut sites show up in the final map is given by
$$\begin{aligned} & [1-(1-p)^n]^k \\ \geq & \left(1-\frac{e^{-c}}{k}\right)^k \\ = & \left(e^{-1-\varepsilon_{k,c}}\right)^{e^{-c}} \\ = & e^{-e^{-c}} e^{-\varepsilon_{k,c}e^{-c}} \\ > & e^{-e^{-c}}\, e^{-(e^{-2c})/2k}, \end{aligned}$$
since $\varepsilon_{k,c} < 1/(ke^c)$ ($k \geq 1$ and $c \geq 1$ and $ke^c > 1.45$).

On the other hand, if $n < \frac{\ln k}{p(1+p)}$ ($k \geq 1$ and $0 < p < 0.69$) then it is easy to argue that the amount of data is insufficient to recover the correct map with high probability. 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 $(1-p)^n > e^{-pn(1+p)} > e^{-\ln k} = \frac{1}{k}$. Thus, with $n$ data the probability that we can recover all the true cut sites is simply bounded from above by
$$\left(1-\frac{1}{k}\right)^k \leq \frac{1}{e} < \frac{1}{2}.$$
Thus with probability half or higher any algorithm will fail to produce the correct ordered restriction map.

**Theorem 6.3.1** *Let $\epsilon$ be a positive constant and $c \geq 1$ be so chosen that $1-e^{-e^{-c}} = \epsilon$. Then for $n \geq \frac{c(1+e^{-c})}{p} + \frac{\ln k}{p}$ ($k \geq 1$), with probability at least $1-\epsilon$, the correct ordered restriction map can be computed in $O(nk)$ time.*

*When $n < \frac{\ln k}{p(1+p)}$ ($k \geq 1$ and $0 < p < 0.69$), no algorithm can compute the correct ordered restriction map with probability better than half.* □

For example, for a BAC clone digested by a 6-cutter enzyme, $k \approx 37$ (expected value), with a partial digestion rate $p \geq 0.1$, if we compute an ordered restriction map from $n = 250$ observations then the probability that we have a correct map is at least $1 - 2 \times 10^{-9}$. Similarly, for $n = 100$ (with all other parameters unchanged), the same probability is at least $1 - 2.5 \times 10^{-3}$. In contrast, for the same values $k \approx 37$ and $p = 0.1$, but with $n < 30$ observations, the probability of obtaining a correct map drops to less than half.

*Note*, however, that since the value of $k$ and $p$ are not known a priori, it is really impossible to use this result in a meaningful way in designing an experiment (i.e., in choosing $n$).

## 6.4 Misorientation Errors

Now, 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

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

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 $n$-many times resulting in $n$ observed restriction maps whose orientations may be misspecified.

An algorithm to reconstruct the true map may proceed in two phases:

### 6.4.1 Phase 1:

Define a map

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

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 as given earlier, we see that we will succeed in this phase with probability $e^{-e^{-c}}$, if $n \geq \frac{c(1+e^{-c})}{p} + \frac{\ln k}{p}$.

### 6.4.2 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 $[(0,1/2)]$ and $-1$ denotes that the site belongs to the right half $[(1/2,1)]$. Thus, we may define

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

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 $e = [f(h_i), f(h_j)] \in E$ if and only if

$$\exists_{s_{i'}, s_{j'}} f(s_{i'}) = f(h_i) \text{ and } f(s_{j'}) = f(h_j).$$

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

$$\operatorname{sgn}(e) = \operatorname{sgn}[(1/2 - s_{i'})(1/2 - s_{j'})].$$

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 $+1$ 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 $e_1, e_2, \ldots, e_m$ then

$$\operatorname{sgn}(f(h_i)) = \operatorname{sgn}(e_1) \cdot \operatorname{sgn}(e_2) \cdots \operatorname{sgn}(e_m) \operatorname{sgn}(f(h_j)).$$

Let us assume that $n \geq \frac{1}{p^2} \ln\left(\frac{k}{k - \ln k - c}\right)$.

With partial digestion rate $p > 1$, for any pair $[f(h_i), f(h_j)]$, the probability that this edge does not occur is

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

and

$$p_e = 1 - (1 - p^2)^n \geq \frac{\ln k}{k} + \frac{c}{k}.$$

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

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

On the other hand, if $n < \frac{1}{p^2(1+p^2)} \ln \frac{k}{k-1}$ ($k > 1$ and $0 < p < 0.83$) then it is easy to argue that the amount of data is insufficient to recover the correct map with high probability. Note that, since

$$(1-p^2)^n > e^{-np^2(1+p^2)} > e^{-\ln(k/k-1)} = 1 - \frac{1}{k},$$

we have $p_e < \frac{1}{k}$. In this case the graph is almost surely not connected (the largest connected component has size $o(\ln k)$), and the final map cannot be computed uniquely (correctly). Thus with probability half or higher any algorithm will fail to produce the correct ordered restriction map.

**Theorem 6.4.1** *Let $\epsilon$ be a positive constant and $c \geq 1$ be so chosen that $1 - e^{-2e^{-c}} = \epsilon$. Then for*

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

*($k > c + \ln k$), with probability at least $1-\epsilon$, the correct ordered restriction map can be computed in $O(nk^2)$ time.*

*When*

$$n < \max\left[\frac{\ln k}{p(1+p)}, \frac{1}{p^2(1+p^2)}\ln\frac{k}{k-1}\right],$$

*($k > 1$ and $0 < p < 0.69$), no algorithm can compute the correct ordered restriction map with probability better than half.* □

Considering the earlier example, for a BAC clone digested by a 6-cutter enzyme, $k \approx 37$ (expected value), with a partial digestion rate $p \geq 0.1$, if we compute an ordered restriction map from $n = 250$ observations then the probability that we have a correct map is at least $1 - 6 \times 10^{-9}$. Similarly, for $n = 100$ (with all other parameters unchanged), the same probability is at least $1 - 7.5 \times 10^{-3}$. In contrast, for the same values $k \approx 37$ and $p = 0.1$, but with $n < 30$ observations, the probability of obtaining a correct map drops to less than half. Thus, the effect of the misorientation is dominated by the partial digestion error for $p \geq 0.1$.

In fact, comparing the two terms in the earlier theorem, we see that as long as

$$p \geq \frac{\ln(k/(k-\ln k - c))}{\ln k + c} = \frac{-\ln(1-(\ln k + c)/k)}{\ln k + c} \approx \frac{(\ln k + c)/k}{\ln k + c} = \frac{1}{k},$$

then only the term due to partial digestion dominates. For instance, if we assume that every observation contains at least one true restriction site, then $p \geq 1/k$ and $n = O(k \log k)$ observations will suffice to find the true map without any other a priori knowledge of $p$.

### 6.4.3 Optical Cuts

Next we shall consider the situation when 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 $\lambda_f$ (thus the expected number of false cuts per molecule is $\lambda_f$). Hence, for any small region $[x, x+\delta x]$ in an observation,

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

The probability that an observation contains exactly $f$ spurious cuts is given by: $e^{-\lambda_f}\frac{\lambda_f^f}{f!}$. Typical observed values for $\lambda_f$ 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 straightforward 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 exist $i_1 \neq i_2$ 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 $n > \frac{2c}{p} + \frac{2\ln k}{p}$, (with $c > 1.26$) then if $h_i$ is a true cut site then the probability that $f(h_i)$ is not included in the output is

$$(1-p)^n + np(1-p)^{n-1} \leq (1+p(n-1))e^{-p(n-1)} \leq e^{-p(n-1)/2} = \frac{e^{-c}}{k}.$$

Proceeding as before the probability that all $k$ true sites will be included is thus bounded from below by

$$e^{-e^{-c}}\ e^{-(e^{-2c})/2k},$$

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.

### 6.4.4 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 with the phase 1, as in the preceding subsection, and again assuming that $n > \frac{2c}{p} + \frac{2\ln k}{p}$ ($c > 1.26$), we will almost surely (with probability no smaller than $e^{-e^{-c}} e^{-(e^{-2c})/2k}$) construct a set

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

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 $n \geq \frac{c}{p^2} + \frac{\ln m}{p^2}$ ($m \geq 1$ and $c \geq 1$) then the preceding steps correctly detect the symmetric cuts with probability greater than $e^{-e^{-c}} e^{-(e^{-2c})/2m}$. Note that the probability that, assuming $h_j$ to be a symmetric true cut, 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 $(1 - p^2)^n$. Observe that

$$\begin{aligned} (1 - p^2)^n &\leq e^{-p^2 n} \\ &\leq e^{-c - \ln m} = \frac{e^{-c}}{m}, \end{aligned}$$

and

$$1 - (1 - p^2)^n \geq 1 - \frac{e^{-c}}{m}.$$

Thus the probability that all $m$ symmetric true cut sites are detected in the final map is given by

$$\begin{aligned} & [1 - (1 - p^2)^n]^m \\ \geq & \left(1 - \frac{e^{-c}}{m}\right)^m \\ > & e^{-e^{-c}}\, e^{-(e^{-2c})/2m}, \end{aligned}$$

since $me^c > 1.45$.

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.

Using an argument, similar to the one presented in section 1, we are led to the conclusion that if $n < \frac{\ln m}{p^2(1+p^2)}$ ($m \geq 1$ and $0 < p < 0.83$) then we will fail to detect at least one symmetric cut (and hence fail to create the correct final map) with probability half or higher.

At the end of this step, we are left with a set only corresponding to 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.

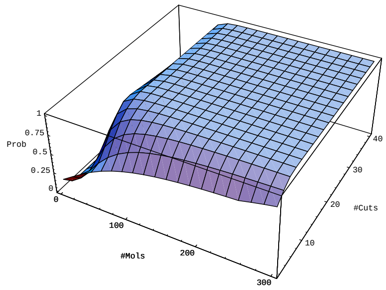


Figure 6.1: **Theoretical Results.** The probability of successfully computing the correct restriction map as a function of the number of cuts in the map and the number of molecules used in creating the map. The individual maps for each single molecule is assumed to be subject to several sources of error: partial digestion (10%), unknown orientation and false cut (with a rate of 1 false cut in 100Kb) error. Effects of the other error sources are not explicitly accounted for. Note that, in almost all cases (number of cuts exceeding 4), 70–100 molecules suffice to find the correct map. Our experimental results with significantly better digestion rate agree with the theoretical analysis with remarkable fidelity. Notice the sharp transition in the probability of successfully computing the map (from a probability of near-0 value to near-1 value) as the number of molecules used in the construction is increased from 30 to 90.

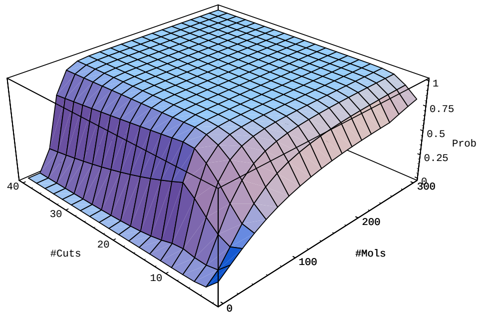


Figure 6.2: **Theoretical Results.** Another view of the earlier plot.

### 6.4.5 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 $\lambda_f$), but no sizing error.

**Theorem 6.4.2** *Let $\epsilon$ be a positive constant and $c \geq 1.26$ be so chosen that $1 - e^{-3e^{-c}} = \epsilon$. Then for*

$$n \geq \max\left[\frac{2c(1+e^{-c})}{p} + \frac{2\ln(k+m)}{p}, \frac{1}{p^2}\max\left[c(1+e^{-c}) + \ln m, \ln\left(\frac{k}{k-\ln k - c}\right)\right]\right],$$

*($k > c + \ln k$ and $m \geq 1$), with probability at least $1 - \epsilon$, the correct ordered restriction map can be computed in $O(n(L+k^2+m))$ time.*

*When*

$$n < \max\left[\frac{\ln(k+m)}{p(1+p)}, \frac{1}{p^2(1+p^2)} \max\left[\ln m, \ln\frac{k}{k-1}\right]\right],$$

*($k > 1$, $m > 1$ and $0 < p < 0.69$), no algorithm can compute the correct ordered restriction map with probability better than half.* □

## 6.5 Discretization

There now remain to introduce two more significant effects in order to make the observation model somewhat more realistic. Firstly, we need to study the effect of the underlying discrete model, as one may argue that at the most basic level the maps can only be presented with fragment lengths in base pairs. (Although this itself is not a realistic model, as what one observes are randomly attached flurochromes filtered by various optical and image processing steps.) The good news here is that our analysis so far holds with minor modification; the detailed analysis only makes the effect of spurious cuts a lot more obvious. Secondly, we need to model the errors in fragment sizes. The bad news are two fold: a) the analysis tools we have used so far simply do not apply; b) the obvious discretization algorithms that we have relied on simply fails. The last fact is rather important as it clearly explains why at least three of the algorithms that have appeared in the literature have failed to produce maps for any data set.

### 6.5.1 Base Pair Accuracy

Let us assume that the clone DNA that we wish to analyze is of length $L$ bps and the restriction enzyme used is a 6-cutter. Let $\Delta = 12$ and $\delta = \Delta/L$. In the normalized ordered restriction map, we may assume that the computational processes cannot distinguish between two locations if they are only $\delta$ apart. More formally, we say that

$$\begin{aligned} x &\approx_\delta y, \quad \text{if } y-\delta \le x \le y+\delta \\ x &<_\delta y, \quad \text{if } x < y-\delta \\ x &>_\delta y, \quad \text{if } x > y+\delta. \end{aligned}$$

We can thus imagine that the unit length is partitioned into $M = 1/\delta = L/\Delta$ consecutive subintervals and it is not possible to distinguish the restriction cuts and spurious cuts in each of these subintervals. Thus, we need to ensure that $\delta$ 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. Here, we are simply interested in the effect of finite $M$ (and nonzero $r$).

To give some idea of the numbers involved, we see that 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 $M$, $r \ll p$. We will use the following simplifying assumption:

$$26r < p.$$

More precisely, $(12e-7)r < p$, thus implying that $(p+r)/6r > 2e-1$. While it is interesting to analyze the case when $p$ and $r$ are arbitrarily close, the analysis only produces unrealistic and pessimistic results, and differs widely from experimentally observed results.

### 6.5.2 Limit on $M$

The discretization process, now, 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

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

Hence we must guarantee that

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

since otherwise the computed map will be wrong with probability half or more. Note that, in order for the above expression to have any discernible effect, $L$ has to be astronomically large, and of course, none of the single molecule approaches is ever planned to be used for molecules much longer than few Mbs.

**Theorem 6.5.1** *When*

$$n < \max\left[\frac{\ln(k+m)}{p(1+p)}, \frac{1}{p^2(1+p^2)} \max\left[\ln m, \ln\frac{k}{k-1}\right], \frac{\ln(L/\Delta)}{\ln(L/\lambda_f\Delta)}\right],$$

*($k > 1$, $m > 1$, $L > \Delta$ and $0 < p < 0.69$), no algorithm can compute the correct ordered restriction map with probability better than half.* □

### 6.5.3 Statistical Analysis

Next we shall provide a simple statistical analysis for the success of the phases of the earlier algorithm, which need to be adapted to the new case. As mentioned earlier, we shall make use of the following inequality in the analysis:

$$26r < p,$$

although in general we expect $r \ll p$ ($r \approx 10^{-4}$ and $p > 0.1$).

#### 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 *significantly large number* of times, determined by a threshold. 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 $\sim S(n, p+2r)$. 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 $\sim S(n, 2r)$. Let

$$\epsilon_1 = \frac{p+r}{6r} \geq \frac{p}{6r}, \quad \epsilon_1 > 2e - 1,$$

and the threshold be

$$\begin{aligned}
(1+\epsilon_1)2nr &= \left(1 + \frac{p+r}{6r}\right) 2nr = \frac{n(p+r)}{3} + 2nr \\
&< \frac{n(p+r)}{3} + \frac{2n(p+r)}{27} \\
&< \frac{n(p+2r)}{2}.
\end{aligned}$$

By assumption (since $26r < p$), this threshold is less than

$$\frac{n(p+2r)}{2} = (1-\epsilon_0)n(p+2r),$$

where

$$\epsilon_0 \geq \frac{1}{2}.$$

Assume that

$$n > \frac{8}{p} \max\left[c + \ln(k+m), \frac{3}{8\ln 2}(c + \ln(M/2 - k - m))\right].$$

Now we can use the Chernoff's bound to note that

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

Thus the probability that all the correct cuts appear in the computed set is bounded from below by

$$\left(1 - \frac{e^{-c}}{k+m}\right)^{k+m} > e^{-e^{-c}} e^{-(e^{-2c})/2(k+m)}.$$

Again, using the Chernoff's bound in the other direction, we get

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

Thus the probability that no spurious cut appears in the computed set is bounded from below by

$$\left(1 - \frac{e^{-c}}{M/2-k-m}\right)^{M/2-k-m} > e^{-e^{-c}} e^{-(e^{-2c})/M}.$$

### 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 $\sim S(n, (p+r)^2)$. If on the other hand, $f(h)$ is not a symmetric site, then the corresponding number must follow a Binomial distribution $\sim S(n, 2(p+r)r)$. Let

$$\epsilon_1 = \frac{p+r}{6r} \geq \frac{p}{6r},$$

and the threshold be

$$(1+\epsilon_1)2n(p+r)r = \left(1 + \frac{p+r}{6r}\right) 2n(p+r)r = \frac{n(p+r)^2}{3} + 2n(p+r)r,$$

where

$$\epsilon_1 > 2e - 1.$$

By assumption (since $26r < p$), this threshold is less than

$$\frac{n(p+r)^2}{2} = (1-\epsilon_0)n(p+r)^2,$$

where

$$\epsilon_0 \geq \frac{1}{2}.$$

Assume that

$$n > \frac{8}{p^2} \max\left[c + \ln m, \frac{3}{8\ln 2}(c + \ln k)\right].$$

Now we can use the Chernoff's bound to note that

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

Thus the probability that all the symmetric cuts are detected is bounded from below by

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

Again, using the Chernoff's bound in the other direction, we get

$$\begin{aligned} & Pr\Big[S(n, 2(p+r)r) \geq (1+\epsilon_1)2n(p+r)r\Big] \\ \leq\ & 2^{-(1+\epsilon_1)2n(p+r)r} \\ <\ & e^{-(\ln 2/3)np^2} < e^{-c-\ln k} = \frac{e^{-c}}{k}. \end{aligned}$$

Thus the probability that no asymmetric cut is mistakenly eliminated is bounded from below by

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

### Phase 2

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. Consider a potential edge $[f(h_i), f(h_j)]$, then the number of times an observation has sites at $s_{i'}$ and $s_{j'}$ such that $f(h_i) = f(s_{i'})$ and $f(h_j) = f(s_{j'})$ assigning the correct edge labeling must follow a Binomial distribution $\sim S(n, (p+r)^2)$. If on the other hand, the edge labeling is incorrect, then the corresponding number must follow a Binomial distribution $\sim S(n, 2(p+r)r)$. Let

$$\epsilon_1 = \frac{p+r}{6r} \geq \frac{p}{6r},$$

and the threshold be

$$(1+\epsilon_1)2n(p+r)r = \left(1 + \frac{p+r}{6r}\right) 2n(p+r)r = \frac{n(p+r)^2}{3} + 2n(p+r)r,$$

where

$$\epsilon_1 > 2e - 1.$$

By assumption (since $26r < p$), this threshold is less than

$$\frac{n(p+r)^2}{2} = (1-\epsilon_0)n(p+r)^2,$$

where

$$\epsilon_0 \geq \frac{1}{2}.$$

Assume that

$$n > \frac{8}{p^2} \max \left[ \ln \left( \frac{k}{k - \ln k - c} \right), \frac{3}{8 \ln 2}(c + 2\ln k) \right].$$

Now we can use the Chernoff's bound to note that

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

Thus the probability, $p_e$ that any edge is correctly labeled is bounded from below by

$$\frac{\ln k}{k} + \frac{c}{k},$$

and the resulting random graph $G_{k,p_e}$ is connected with probability higher than $e^{-e^{-c}}$.

Again, using the Chernoff's bound in the other direction, we get

$$\begin{aligned} &Pr\Big[S(n, 2(p+r)r) \geq (1+\epsilon_1)2n(p+r)r\Big] \\ &\leq \; 2^{-(1+\epsilon_1)2n(p+r)r} \\ &< \; e^{-(\ln 2/3)np^2} < e^{-c-2\ln k} = \frac{e^{-c}}{k^2}. \end{aligned}$$

Thus the probability that no non-edge is included with incorrect label is bounded from below by

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

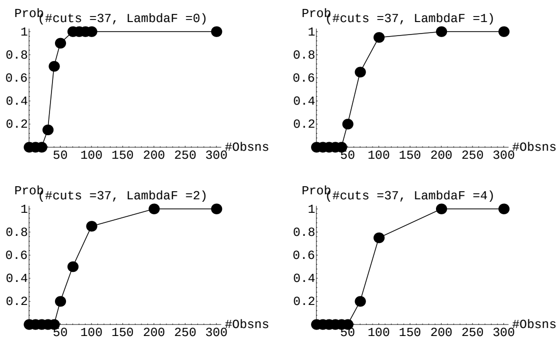


Figure 6.3: **Experimental Results:** #Cuts, $k = 37$, $\sigma = 1.5$bp, $p = 0.1$

### 6.5.4 Summary

**Theorem 6.5.2** *Let $\epsilon$ be a positive constant and $c \geq 1$ be so chosen that $1 - e^{-12e^{-c}} = \epsilon$. Then for*

$$n \geq \frac{8(1+e^{-c})}{p} \max\left[c + \ln(k+m), \frac{c + \ln m}{p}, \frac{1}{p} \ln\left(\frac{k}{k - \ln k - c}\right),\right.$$

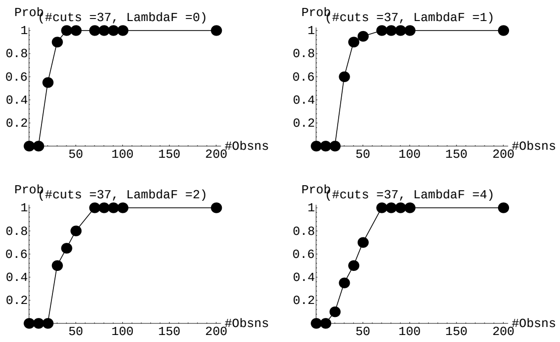


Figure 6.4: **Experimental Results:** #Cuts, $k = 37$, $\sigma = 1.5$bp, $p = 0.2$

$$\left(c + \ln(L/2\Delta - k - m)\right), \frac{c + 2\ln k}{p}\Bigg],$$

*($k > c + \ln k$, $m \geq 1$, $L > 2\Delta$ and $r < p/26$), with probability at least $1 - \epsilon$, the correct ordered restriction map can be computed in $O(n(L + k^2 + m))$ time.* $\square$

The previous result can be tightened further, if we generalize our assumption to $(2^\alpha \theta e)\ r < p$, where $\alpha \geq 1$ and $\theta \geq 3$. For instance, $\alpha = 6$ and $\theta = 9/2$, requires that $r < p/785$ as opposed to $1/26$. In this case, we can show that the appropriate choice

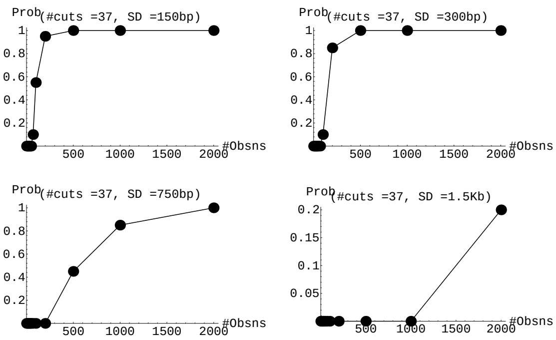


Figure 6.5: **Experimental Results:** #Cuts, $k = 37$, $\lambda_f = 1$, $p = 0.1$

of $n$ in the preceding theorem could be changed to:

$$
\begin{aligned}
n \;\geq\; & (1+e^{-c}) \max\Bigg[ \\
& \frac{2}{(1-2/\theta(1+2^{-\alpha}/e))^2 p} \max\left(c+\ln(k+m), \frac{c+\ln m}{p}, \frac{1}{p}\ln\left(\frac{k}{k-\ln k - c}\right)\right), \\
& \qquad \frac{\theta}{(2\ln 2\alpha)\; p^2} \max\left((c+\ln(L/2\Delta - k - m)), \frac{c+2\ln k}{p}\right)\Bigg]
\end{aligned}
$$

*Note*: Use essentially the same analysis as before: Simply take $\epsilon_1$ to be $(p+r)/(\theta r) > 2^{\alpha}e - 1$, and note that $\epsilon_0$ is then bounded by $(1 - 2/\theta(1+2^{-\alpha}/e))$.

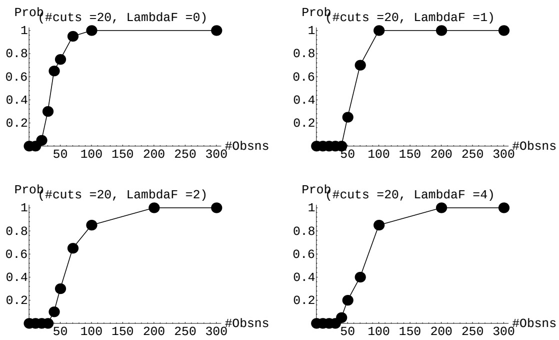


Figure 6.6: **Experimental Results:** #Cuts, $k = 20$, $\sigma = 1.5$bp, $p = 0.1$

For example, for a BAC clone digested by a 6-cutter enzyme, $k \approx 37$ (expected value), with a partial digestion rate $p \geq 0.1$, with the false optical cut rate $r = 10^{-4}$ and $M = 15,000$ if we compute an ordered restriction map from $n = 1050$ observations then the probability that we have a correct map is at least $1 - 7 \times 10^{-5}$. In contrast, for the same values $k \approx 37$, $p = 0.1$ and $r \approx 10^{-4}$, but with $n < 30$ observations, the probability of obtaining a correct map drops to less than half.

Let $p_E$ denote the probability that the restriction enzyme cuts at a site. For instance, the probability $p_E = 1/4^6 \approx 1/4,000$ for 6-cutters. Also note that assuming that $c \geq 4$ and $Lp_E \geq 8c$, we see that since the total number of cuts $(k + 2m) \sim S(L, p_E)$, the

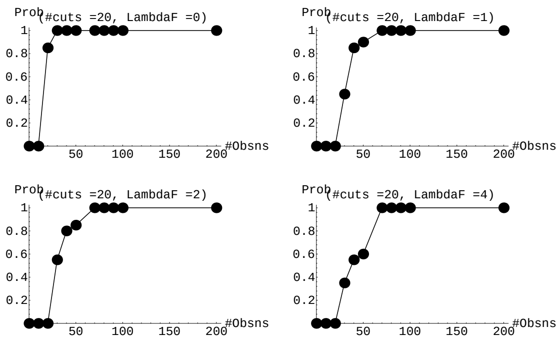


Figure 6.7: **Experimental Results:** #Cuts, $k = 20$, $\sigma = 1.5$bp, $p = 0.2$

probability that a random clone of length $L$ has total number of cuts ranging between $2c$ and $(1+c)Lp_E$ is close to one:

$$Pr\Big[S(L, p_E) \in [2c, (1+c)Lp_E]\Big] \geq 1 - 2e^{-2c}.$$

Using the previous arguments, we see that in general if we use

$$n > \frac{8}{p^2}(c' + \ln L), \quad c' = c + 2\ln(1+c),$$

observations (with $r < p/26$), we will compute the correct map almost surely for almost all clones.

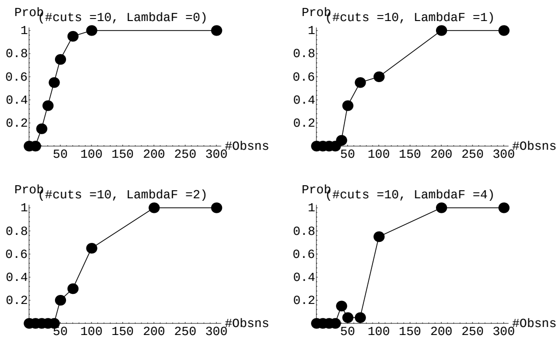


Figure 6.8: **Experimental Results:** #Cuts, $k = 10$, $\sigma = 1.5$bp, $p = 0.1$

## 6.6 Sizing Errors

However, we need to model the sizing error and analyze its effect. Before doing so, we need to derive some inequalities relating the size of the discretized subinterval ($\Delta$) 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 $\Delta \leq \frac{2}{(k-1)p_E}$, where $k$ is the number of cuts and $p_E$, as usual, denotes the probability that the restriction enzyme cuts at a site.

Assume that $\Delta > \frac{2}{(k-1)p_E}$. 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

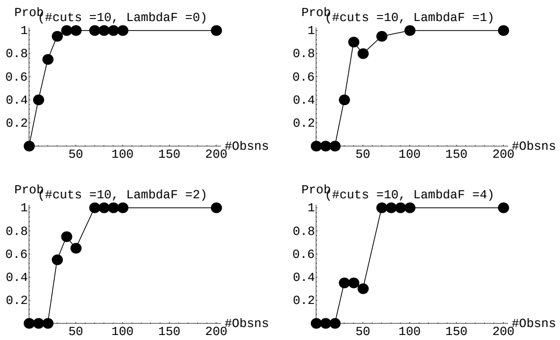


Figure 6.9: **Experimental Results:** #Cuts, $k = 10$, $\sigma = 1.5$bp, $p = 0.2$

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

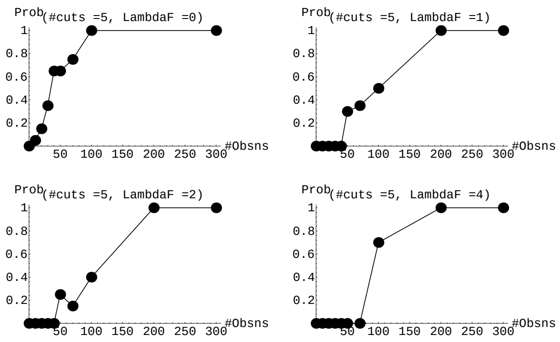


Figure 6.10: **Experimental Results:** #Cuts, $k = 5$, $\sigma = 1.5$bp, $p = 0.1$

subinterval is bounded by

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

However, note that for the running BAC example, 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

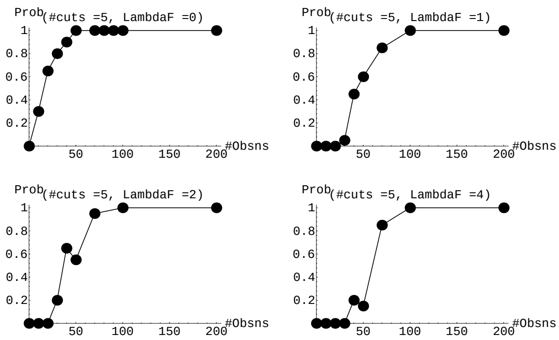


Figure 6.11: **Experimental Results:** #Cuts, $k = 5$, $\sigma = 1.5$bp, $p = 0.2$

probability. As a result, we may require that

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

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

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

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

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

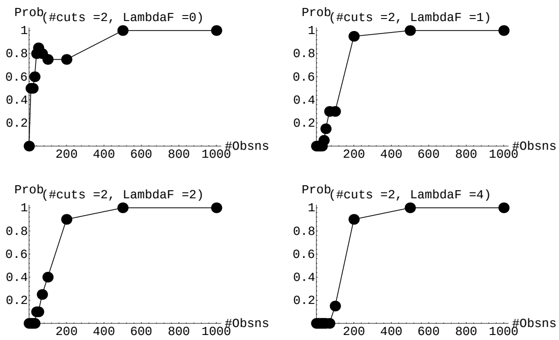


Figure 6.12: **Experimental Results:** #Cuts, $k = 2$, $\sigma = 1.5$bp, $p = 0.1$

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.

It is somewhat interesting to note that if we are considering an example involving lambda's or very small cosmids (with $k \approx 5$), and willing to accept a rather small probability of success, then a $\sigma \approx 60bp$ is feasible. The experimentally observed value for $\sigma$ has ranged around $700bps$, and it is inconceivable that even in this case such an approach could be made to work except for few lucky instances.

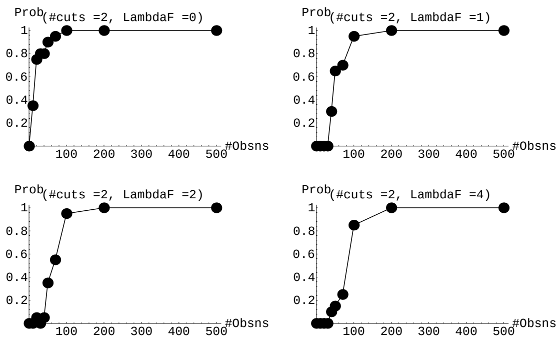


Figure 6.13: **Experimental Results:** #Cuts, $k = 2$, $\sigma = 1.5$bp, $p = 0.2$

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 ($\sigma$), a digest rate ($p$) and a false cut rate ($\lambda_f$). For each model random data is generated with the help of a random number generator in a straightforward 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. Then another random number is drawn from the standard Gaussian distribution to determine the location of the cut with sizing error. Next false cuts are added by first drawing a random sample from a Poisson distribution with mean $\lambda_f$ to determine the number of false cuts, then drawing the required

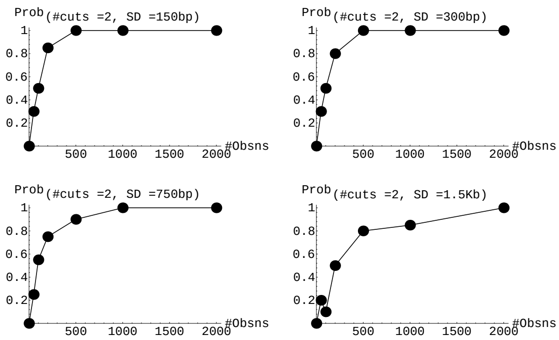


Figure 6.14: **Experimental Results:** #Cuts, $k = 2$, $\lambda_f = 1$, $p = 0.1$

number of random samples uniformly over [0,1] to get the false cut locations. This 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 scored a success if the number of cuts is correct and their location is within one standard deviation ($\sigma$) of the correct location. (Note that $\sigma$ is the standard deviation for the cuts of one sample molecule: the map computed by the AMS algorithm typically has 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

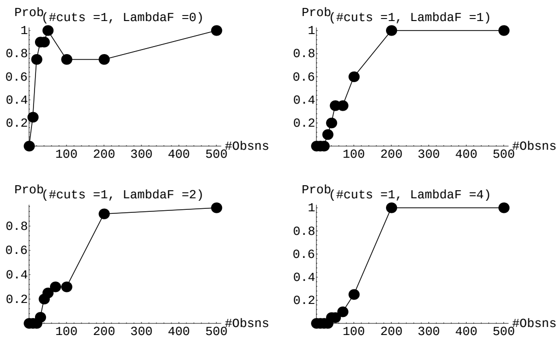


Figure 6.15: **Experimental Results:** #Cuts, $k = 1$, $\sigma = 1.5$bp, $p = 0.1$

succeeds is recorded against the data sample parameters ($k$, $\sigma$, $p$, $\lambda_f$, number of molecules). The whole process was repeated for different values of the parameters. The number of cuts was varied using the values $k = 0$, 1, 2, 5, 10, 20 and 37. The values of $p$ tested were $p$=0.10 and $p$=0.20. The values of $\lambda_f$ tested were $\lambda_f$=0 (no false cuts) and $\lambda_f$=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 $\sigma$= 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 few instances 2000 or 5000.

The results are summarized in a series of graphs showing the

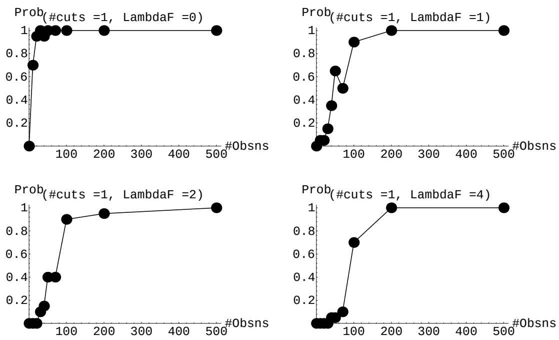


Figure 6.16: **Experimental Results:** #Cuts, $k = 1$, $\sigma = 1.5$bp, $p = 0.2$

success rate (out of 20 samples) as a function of the number of molecules used. The graph in Figure 6.3 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 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 6.4), 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

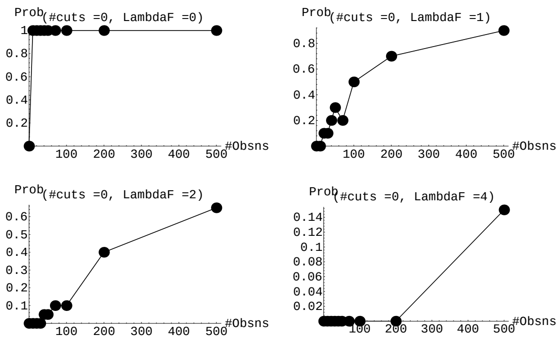


Figure 6.17: **Experimental Results:** #Cuts, $k = 0$

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 6.6 and 6.7 show the results for the case $k = 20$; Figures 6.8 and 6.9, for the case $k = 10$; Figures 6.10 and 6.11, for the case $k = 5$; Figures 6.12 and 6.13, for the case $k = 2$; Figures 6.15 and 6.16, for the case $k = 1$ and Figure 6.17, 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 correct map in every case ($\lambda_f = 0$, 1, 2 and 4), in contrast to just 200 for $k = 37$. This agrees with the theory from sections 4 and 5 which shows that the bounds increase slowly as $k$ is decreased. The case $k = 1$

| #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 6.1: 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 $\lambda_f \in \{0, 1, 2, 4\}$.

in Figures 6.15 and 6.16, however, 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 6.5 shows what happens with $k = 37$ when 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 6.14 shows what happens at $k = 2$ when sizing error is increased similarly. In this case the number of molecules increases from an already larger value, but more slowly: it increase 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 orientation right 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 $k = 37$.