An Algorithm Combining Discrete and Continuous Methods for Optical Mapping*

Richard M. Karp[†] Itsik Pe'er[‡] Ron Shamir[§]

Keywords: Optical mapping, EM algorithm, Discrete optimization, Maximum likelihood, Restriction maps

Abstract

Optical mapping is a novel technique for generating the restriction map of a DNA molecule by observing many single, partially digested copies of it, using fluorescence microscopy. The real-life problem is complicated by numerous factors: false positive and false negative cut observations, inaccurate location measurements, unknown orientations and faulty molecules. We present an algorithm for solving the real-life problem. The algorithm combines continuous optimization and combinatorial algorithms, applied to a non-uniform discretization of the data. We present encouraging results on real experimental data, and on simulated data.

1 Introduction

Even in the era of whole genome methods, the mapping of restriction sites still plays an important role in genomic analysis and motivates further development of mapping procedures [4, 10]. Optical mapping is a novel technique for obtaining restriction

*Portions of this paper were presented at the 7th international conference on Intelligent Systems for Molecular Biology (ISMB '99), Heidelberg, Germany [15]

†Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720-1776 +510-642-5799 / FAX: +510-642-5775 E-mail: `karp@cs.berkeley.edu`.

‡School of Mathematics, Tel Aviv University, Tel Aviv, 69978, Israel, Tel: +972-3-6408979, Fax: +972-3-6409357, E-mail: `izik@math.tau.ac.il`. Supported by the Clore foundation scholarship.

§School of Mathematics, Tel Aviv University, Tel Aviv, 69978, Israel, Tel: +972-3-6409356, Fax: +972-3-6409357, E-mail: `shamir@math.tau.ac.il`. Supported in part by an infrastructure grant from the Israeli Ministry of Science

maps [16, 3, 18, 11, 17, 7, 4, 10]. In an optical mapping experiment, many copies of the target DNA molecule are elongated and attached to a glass surface. Restriction endonuclease enzymes are applied to the molecules, partially digesting them. At the cleaved cut sites the two cleaved ends of the molecule coil away from each other due to the elasticity of DNA. The molecules are stained and photographed. Molecules appear in the image as lines separated by gaps at cut locations. For each molecule the locations of these cuts along the DNA are recorded. The goal is to deduce from this data the correct locations of all restriction sites.

Optical mapping has several advantages over traditional mapping methods: It can provide high resolution maps which determine physical landmarks in the DNA; it is insensitive to repetitions in the sequence, and can be automated to a considerable extent (see also [19]). Moreover, since the data preserves the linear order of the sites, it is more informative than the data from traditional gel-based restriction mapping methods. Currently, BAC sized clones are routinely mapped, and it is possible to map molecules of up to several hundred Kb. A variant of this technique in which cut sites in subfragments of the target are recorded was successfully used recently to obtain a 3.2Mb map of Deinococcus radiodurans [10].

In this paper we address the basic optical mapping technique, where cuts are recorded with respect to the complete target molecule. Even in that case, the deduction of the true restriction sites is not straightforward, due to the following factors:

- An observed cut may not correspond to a restriction site. We call such a cut *false*.
- No cut may be observed at a true restriction site, due to partial digestion.
- The orientation of each molecule is unknown, *i.e.*, observed cut locations are known only up to complete reversal of the molecule.
- There is a sizing error when measuring observed cut locations.
- There are molecules on which the observed cuts do not correspond to real sites in any orientation, and instead produce "random" results. We call these molecules *faulty*, and the other molecules *proper*.

Optical mapping uses a number of copies of the target molecule ranging between several dozen and several hundred. The digestion rate, false cut rate, and the sizing error vary considerably, even when comparing restriction sites in the same experiment [3].

Various combinatorial formulations of the optical mapping problem have been proven NP-complete [2, 14, 12] and even hard to approximate [13]. Bayesian approaches using global optimization have been applied in [2] and [9], where a probabilistic model of the experiment is set up, and global optimization techniques are used to find the most probable parameters of this model, which comprise the desired solution. The implementation reported in [2] is successfully used in practice for analysis of real laboratory data. Other studies [8, 6, 12] discretize the input, and devise combinatorial solution methods. The algorithm of [8] has a proven performance guarantee under a simple probabilistic model of the data.

In this paper we describe a new strategy combining the combinatorial approach with the global optimization approach. We first determine the orientations of the molecules and then determine the restriction sites by applying continuous optimization in the space of certain model parameters, as done in [2] and [9]. In order to orient the molecules we adapt a combinatorial algorithm from [8]. However, instead of using a fixed, uniform discretization of the data as in [8], we use a nonuniform discretization based on identifying certain informative intervals derived from the data itself. The chief innovation in this paper is the use of continuous and discrete algorithmic methods to enhance each other.

Our method was applied in a blind test to eight examples provided by the lab of David Schwartz at New York University. The results show that we determine the orientation of the molecules correctly, enabling us to identify almost all the restriction sites. Occasional misestimation of the number of restriction sites can probably be alleviated by finer tuning of the algorithm. We also present results on simulated data, for several different choices of simulation parameters.

The paper is organized as follows: Section 2 outlines the general strategy of our algorithm. In Section 3 we present our probabilistic model. Sections 4, 5, and 6 describe the three major stages in our algorithm. Section 7 gives results on experimental and simulated data.

2 General Strategy

2.1 Disadvantages of the Discrete Approach

We sketch the *signature method* that was suggested in [8] as a method of determining the locations of restriction sites in a discretized version of the problem. We assume that the data is scaled so that the target molecule and each of its copies extends over the interval $[0, 1]$. This interval is partitioned into n equal sections, each of which may contain a single restriction site (typically, $n = 200$). Sections $k \leq \frac{n}{2}$ and $n + 1 - k$ are called *conjugate*. Thus each conjugate pair contains a section from the left half of the target molecule and a symmetrically placed section from the right half of the molecule. For each conjugate pair, the numbers of molecules with no observed cuts, one observed cut and two observed cuts are determined. On the basis of this information, the conjugate pairs are divided into three types: those likely to contain no restriction sites, those likely to contain two restriction sites, and those likely to contain one restriction site. The conjugate pairs of the first two types are set aside, and the remaining conjugate pairs are then divided into two classes such that two pairs are in the same class if their restriction sites appear to lie in the same half of the target molecule, and in the opposite classes otherwise. The details will be omitted here, but the idea is that two conjugate pairs should be placed in the same class if, for those copies of the molecule in which each of the two conjugate pairs contains one observed cut, the observed cuts tend to occur in the same half of the molecule. We refer to this process as *resolving* the conjugate pairs. In [8] the signature method was directly used to determine the restriction sites. In this paper a refinement of the signature method is used to orient the molecules, in preparation for a later stage in which the restriction sites are determined.

The signature method (and other algorithms of [8]) were applied to uniformly discretized real data and failed. The main reason is that the sizing errors (and hence, the errors in measured cut locations) are too large to conform with the uniform discretization. Probabilistic analysis of the error under uniform discretization [1] illuminates the limitations of such discretization from a theoretical perspective, and motivates finding a better way to apply combinatorial methods. The fact that the algorithms in [8] disregard faulty molecules adds to the problem by assuming errors occur independently in the data.

2.2 Disadvantages of the Global Optimization Approach

In the continuous approach, one formulates a probabilistic model with many parameters (*e.g.*, molecule orientations, site locations, cut intensities, noise, etc.) and attempts to find, by global optimization methods, the most likely parameter values given the data. This approach has been demonstrated to perform well on real data [2, 9]. However, there is room for improvement in some important respects, as detailed below.

The score (likelihood) function takes into account both orientations of each molecule. This is done by averaging two probability functions, one for each orientation, so one of them just adds noise to the computation. Clearly, if all other factors are unchanged, a score based on correctly oriented data would give better optimization results.

Furthermore, among the parameters to be optimized, and probably the most important ones, are the locations of the restriction sites. These locations are determined by global optimization of a joint likelihood function which associates, with each finite set S of points along the molecule, the likelihood that S is the set of locations of the restriction sites. Usually there are many short intervals in which a restriction site is likely to have occurred, separated by intervals in which the occurrence of a restriction site is unlikely. Thus the joint likelihood function tends to have many local optima, corresponding to selections of points from different combinations of these short intervals. This multi-modality makes the task of finding the global optimum quite difficult for standard techniques. Our methods alleviate this difficulty in two ways. First, since the molecules have been oriented before global optimization is undertaken, the number of short intervals likely to contain a restriction site is reduced, leading to a great reduction in the number of local maxima of the joint likelihood function. Secondly, we initialize the search with a reasonably good estimation of the locations of the restriction sites. If the initialization is good enough, then the global optimum point will lie within a small neighborhood of the initialization point, and within this neighborhood the likelihood function will be unimodal, and therefore easy to maximize.

2.3 Our Approach

Our approach uses elements of both the continuous and the discrete approaches, attempting to remedy the shortcomings of each: We reduce the main error source in the discrete procedure for orienting the molecules (*i.e.* the uniform discretization) by a more subtle consideration of the continuous data. The results of the refined orientation procedure eliminate the need to optimize over orientation parameters for individual molecules, and thus help the continuous global search heuristic avoid local optima. We now intuitively sketch our algorithm, deferring formal description to later sections.

We assume a continuous probabilistic model of the data, similar to the model used in [2] and [9]. To orient the molecules we use a variant of the signature algorithm that is less sensitive to sizing errors. This variant depends on the concept of an *informative interval*, which we now define. The *folding operation* maps each point $0 \leq x \leq 1$ to $\hat{x} = \min\{x, 1-x\}$. Two intervals along the target molecule are called *conjugate* if their folded images coincide. This generalizes the definition of conjugate pairs from [8, 12], since it does not specify the sizes of these intervals or restrict their possible endpoints to a predefined discrete set. This generalization enhances the performance of the signature algorithm, allowing it to overcome sizing errors. Under this folding operation each conjugate pair of intervals maps to a single *folded interval.* A folded interval is called an *informative interval* if it has a substantial density of observed cuts and passes a statistical test indicating that all the restriction sites within it come from one half of the target molecule; *i.e.*, from one member of the corresponding conjugate pair.

Using dynamic programming we identify a set S of disjoint informative intervals. Using the signature algorithm we can resolve S into two classes, such that two folded intervals in S are in the same class if their restriction sites appear to lie in the same half of the target molecule, and in opposite classes otherwise. We then choose a standard orientation of the target sequence in which the restriction sites occurring in intervals from the first class are placed in the left half, and those occurring in intervals from the second class are placed in the right half. In this way the restriction sites from each folded interval I in S are assigned to one of the two conjugate intervals that map onto I. This process is called *resolving the informative intervals.* Finally, we orient each molecule so as to maximize the number of observed cuts in it that lie within folded intervals from S and occur in the "correct" members of the corresponding conjugate pairs.

Once the molecules have been oriented we apply likelihood optimization to determine the restriction sites. To improve the search for restriction sites, we initialize it with a good approximation of the site locations, which is obtained by identification of *good intervals.* Informally, an interval is good if its density of observed cuts is high, and most of the observed cuts within it can be attributed to restriction sites within the interval itself. The process of identifying good intervals makes use of the fact that the molecules have been oriented. We also screen out molecules suspected to be faulty, first in a pre-processing step, and again after orienting the molecules. The general scheme of our algorithm is as follows:

1. Screen out faulty molecules from the unoriented data.
2. Identify informative intervals.
3. Apply the signature algorithm to resolve the informative intervals and orient the molecules.
4. Screen out more faulty molecules from the oriented data.
5. Identify good intervals.
6. Determine the restriction site locations.

Stages 1 and 4 are described in Section 4. Stages 2 and 5 are described in Section 5. Stage 6 is described in Section 6.

3 Model and Terminology

We now define our probabilistic model of the problem. Similar models were used in [2, 9]. Each of the N_{mol} molecules is faulty with (independent) probability p_{faulty}, in which case, it displays (false) cuts with Poisson rate λ_{faulty}. In each proper molecule false cuts are Poisson distributed with rate λ. We denote $\lambda_{av} = p_{faulty}\lambda_{faulty} + (1 - p_{faulty})\lambda$.

We assume there is some unknown number t of (true) restriction sites, with the i-th site R_i located at position μ_i along the molecule. The input data D is a set of N_{mol} lists, $D_1, \ldots, D_{N_{mol}}$. The list D_m of the m-th molecule contains $N_{cut}(m)$ entries (observed cuts), at positions $c_{m,1}, \ldots, c_{m,N_{cut}(m)}$. Further define N_{cuts} to be $\sum_m N_{cut}(m)$. In each proper molecule, the cut R_i is actually *observed* (registers as a cut) with (independent) probability p_i. Its actual observed position is normally (and independently) distributed around μ_i with variance σ_i^2. Additionally, each molecule is independently oriented as straight or reverse with equal probability. Our problem is to determine the restriction sites $\{\mu_i\}$ from the data D. In the course of doing this we shall also determine the orientations of the molecules and the other parameters of the probabilistic model.

4 Screening Out Faulty Molecules

4.1 Screening Oriented Data

We begin by describing how to screen out faulty molecules from oriented data, and then, in Section 4.2, present the slightly more complicated analysis, for unoriented data. This order of description is the reverse of the chronological order of these stages in our algorithm.

Denote by $\vec{D} = \{\vec{c}_{m,j}\}$ the data after the molecules have been oriented. We define f, the probability density of observing a cut at x, to be the probability of observing a cut in a short interval centered at x, per molecule, per unit length of the interval. Formally, $f(x) : [0,1] \mapsto \mathcal{R}$ is:

$$f(x) \equiv \lim_{\epsilon \to 0} \frac{\left|\left\{(m,j) : |\vec{c}_{m,j} - x| < \frac{\epsilon}{2}\right\}\right|}{\epsilon N_{mol}} \tag{1}$$

Cuts in proper molecules tend to be observed near restriction sites, while cuts in faulty molecules occur at random locations. Thus, the observed cuts in proper molecules should tend to occur at points of higher probability density than the observed cuts in faulty molecules. This is the basis for our screening procedure, which we now detail more formally.

According to our assumptions,

$$f(x) = \lambda_{av} + (1 - p_{faulty}) \sum_i p_i Normal(x, \mu_i, \sigma_i^2)$$

where $Normal(x, \mu, \sigma^2)$ is the probability density, at point x, of a normal random variable with mean μ and variance σ^2, *i.e.*,

$$Normal(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The overall average value of f, denoted f_{av}, is

$$f_{av} = \lambda_{av} + (1 - p_{faulty}) \sum_i p_i$$

However, depending upon whether x is near restriction sites, $f(x)$ behaves differently. If x is not near any μ_i, then the probability of an observation in this interval is:

$$f(x) \simeq \lambda_{av}$$

In contrast, if x is close to some μ_i, and the μ_i-s are sufficiently separated, then

$$f(x) \simeq \lambda_{av} + (1 - p_{faulty}) p_i Normal(x, \mu_i, \sigma_i^2)$$

We now describe how we can estimate $f(x)$ from the data, and how we can use this estimation. To this end, define $X_I(D)$, for an interval I, as the number of cuts observed in D, inside I. Note that $X_I(D)$ is a random variable. Let $I(x, \epsilon)$ denote the interval of length ϵ centered at x. Then for a small ϵ, we may estimate $f(x)$ by:

$$\phi_{\epsilon,\vec{D}}(x) = \frac{X_{I(x,\epsilon)}(\vec{D})}{\epsilon N_{mol}}$$

We compute $\phi_{\epsilon,\vec{D}}(x)$ for every cut position $x = \vec{c}_{m,j}$. In practice, we choose $\epsilon = \epsilon_x$ so that $X_{I(x,\epsilon)}(\vec{D})$ will be some predetermined constant. This way we have a large enough sample size when $\phi_{\epsilon,\vec{D}}(x)$ is small, while concentrating on a small interval around x for better accuracy, when $\phi_{\epsilon,\vec{D}}(x)$ is large. This is, in fact, the k-th-nearest-neighbor method for density estimation (cf. [20]).

For each molecule m, we compute ϕ_m, the average of the estimated density in all the molecule's observed cut positions:

$$\phi_m = \frac{\sum_{j=1}^{N_{cut}(m)} \phi_{\epsilon,\vec{D}}(\vec{c}_{m,j})}{N_{cut}(m)}$$

The expected value of ϕ_m for a faulty molecule m is f_{av}, which in turn, can be estimated as

$$f_{av} \simeq \frac{\sum_m N_{cut}(m)}{N_{mol}}$$

In contrast, for a proper molecule m the expected value of $\phi_{\epsilon,\vec{D}}(\vec{c}_{m,j})$ is larger, and thus, the expected value of ϕ_m is larger.

The molecules with the smallest ϕ_m are the ones most likely to be faulty, and should therefore be discarded. We repeatedly find the molecule m with minimal ϕ_m (the molecule most likely to be faulty), designate it as faulty, filter it out of our data set and recompute the quantities ϕ_m on the basis of the remaining molecules. We repeat this process N_{faulty} times, where N_{faulty} is an input number. We note that a more subtle analysis might be able to estimate N_{faulty} from the observed distribution of ϕ_m.

4.2 Screening Unoriented Data

We now refer to the original unoriented data set D. We fold the molecule in half to create a *folded* data set $\widehat{D} = \{\widehat{c_{m,1}}, \ldots, \widehat{c_{m,N_{cut}(m)}}\}_{m=1}^{N_{mol}}$, using the folding operation $\widehat{x} = \min\{x, 1-x\}$ introduced in Section 2.3. Note, that if the orientation of a molecule D_j is unknown, then the folded molecule $\widehat{D_j}$ data conveys all the information obtainable from D_j on each cut separately. We can now screen out faulty molecules from the folded (unoriented) data using the same idea as in Section 4.1. We present the formulae for completeness:

We define $\hat{f}$, the probability density, of observing a (folded) cut at x, to be the probability of an observed cut being in a short interval centered at x, relative to the length of this interval. Formally, $\hat{f}(x) : [0, \frac{1}{2}] \mapsto \mathcal{R}$, is defined:

$$\hat{f}(x) \equiv \lim_{\epsilon \to 0} \frac{\left|\left\{(m,j) : |\widehat{c_{m,j}} - x| < \frac{\epsilon}{2}\right\}\right|}{\epsilon N_{mol}} \tag{2}$$

where our assumptions imply:

$$\hat{f}(x) = 2\lambda_{av} + (1 - p_{faulty}) \sum_i p_i(Normal(x, \mu_i, \sigma_i^2) + Normal(x, 1-\mu_i, \sigma_i^2))$$

The expected value of $\hat{f}(x)$ is $\hat{f_{av}} = 2f_{av}$, which is estimated by $2\frac{\sum_m N_{cut}(m)}{N_{mol}}$. Again,

$$\hat{f}(x) \simeq \begin{cases} 2\lambda_{av} + (1 - p_{faulty}) p_i Normal(x, \widehat{\mu_i}, \sigma_i^2) & \text{if } x \text{ is near } \widehat{\mu_i} \\ 2\lambda_{av} & \text{otherwise} \end{cases}$$

An unbiased estimator of $\hat{f}(x)$ is

$$\phi_{\epsilon,\hat{D}}(x) = \frac{X_{I(x,\epsilon)}(\hat{D})}{\epsilon}$$

and we can compute the average of the estimated folded density in all the molecule's observed cut positions:

$$\hat{\phi}_m = \frac{\sum_{j=1}^{N_{cut}(m)} \phi_{\epsilon,\hat{D}}(\widehat{c_{m,j}})}{N_{cut}(m)}$$

We observe that the expected value of this average estimated density is, for a faulty m, $\hat{\phi}_m = \hat{f}_{av}$; for a proper m, $\hat{\phi}_m$ is larger. We can therefore perform a screening procedure similar to the one described in Section 4.2, this time using $\hat{\phi}_m$, instead of ϕ_m.

From this point on, we shall assume there are no more faulty molecules, and N_{mol} will denote the number of proper molecules.

5 Good Intervals and Informative Intervals

5.1 Good Intervals

Once the molecules have been oriented we seek to determine the restriction sites by an iterative maximum likelihood computation. Since the method we use (the EM algorithm) is only guaranteed to converge to a local maximum, it is important to start the iteration with a good estimate of the restriction site locations. For this purpose we attempt to localize the restriction site positions to a set of disjoint *good intervals*. Informally, an interval is good if its density of observed cuts is high and most of the observed cuts within it can be attributed to restriction sites which indeed reside in this interval. In this subsection we describe the process of finding the good intervals.

5.1.1 Goal and Compromise

We would like to partition the observed cuts according to their true restriction sites, and to separately handle all the observed cuts originating from a single restriction site. More formally, we want to find intervals $I_1, \ldots, I_t$ contained in $(0,1)$, so that I_i contains only observations originating from R_i, and all of the observations originating from R_i are contained in I_i. However, this is not always possible, because the sizing error may be large relative to the distance between restriction sites. In such cases, there are true restriction sites R_i and R_{i+1} such that the observed cut sites originating from R_i are interleaved with those originating from R_{i+1}.

We therefore settle for finding disjoint intervals $I_1, \ldots, I_k$, and a partition of the sites $R_1, \ldots, R_t$ into k disjoint subsets, with I_i containing mostly true observed cuts originating from the restriction sites of the i-th subset. We call such intervals *good intervals.*

5.1.2 Initial Filtering

We restrict attention to the $O(N_{cut}^2)$ intervals having observed cuts as their end points. We also reject unreasonably long intervals. More precisely, we discard intervals in which at least one of our molecules exhibits more than k cuts (we use $k = 2$). For each such interval I, we estimate the average value of the density $f(x)$ within I by $\frac{X_I(\vec{D})}{|I|N_{mol}}$. We further eliminate those intervals for which this value is smaller than a chosen threshold, as they are unlikely to be good. On the remaining intervals we perform a more subtle analysis, as explained in the next section.

5.1.3 Finding Good Intervals

We consider the oriented data set $\vec{D}$. For each interval $I = (a, b) \subset (0, 1)$ we wish to estimate the number of restriction sites within I and the probability that I contains at least one restriction site. We base this estimation on the random variables $X_I(\vec{m})$, where $\vec{m}$ ranges over the molecules in the probability space $\vec{D}$. Assuming there are j true restriction sites in I, and that a negligible number of the observed cuts within I arise from restriction sites outside I, the distribution of $X_I(\vec{m})$ is a function of the following parameters:

1. The false cut rate λ.
2. For each restriction site R_i within I, the probability p'_i of a true cut originating from R_i to be observed in I. For simplicity, we further approximate all the p'_i values by a single p'_I.

We define:

$$
\begin{aligned}
Poisson(h, \alpha) &= \frac{e^{-\alpha}\alpha^h}{h!} \\
Bin(i, j, p) &= \binom{j}{i} p^i (1-p)^{j-i}
\end{aligned}
$$

$$
\begin{aligned}
P(k,j,\lambda,p'_I) &= Prob\{X_I(\vec{m}) = k|j \text{ restriction sites in } I, \lambda, p'_I\} \\
&= \sum_{h+i=k, i\leq j} Poisson(h,\lambda|I|)Bin(i,j,p'_I)
\end{aligned}
$$

Each molecule in $\vec{D}$ gives rise to an independent sample $X_I(\vec{m})$ from this distribution. From these samples we can obtain the empirical frequency count:

$$
\mathcal{X}_k(\vec{D}) = |\{\vec{m} \in \vec{D} | X_I(\vec{m}) = k\}|
$$

Hence:

$$
\begin{aligned}
Likelihood(j,\lambda,p'_I) &= Prob(\vec{D}|j,\lambda,p'_I) \\
&= \prod_{\vec{m}\in\vec{D}} P(X_I(\vec{m}),j,\lambda,p'_I) \\
&= \prod_k P(k,j,\lambda,p'_I)^{\mathcal{X}_k(\vec{D})}
\end{aligned}
$$

Using standard numerical maximization techniques, for each such j, we can get the most likely parameters, λ and p'_I, given the observed values of $\mathcal{X}_k(\vec{D})$. We optimize these parameters, and denote the log of the maximum likelihood by $L^j(I)$. Let $L(I) = \max_{j>0} L^j(I) - L^0(I)$ be the log-likelihood of the most likely such assumption, compared to the null hypothesis of no restriction sites at all. In practice, it is enough to consider only small values of j, *i.e.*, $j \leq 2$. The higher $L(I)$ is, the more we consider I a good interval.

We extend this measure to any set S of non-overlapping intervals:

$$
L(S) = \sum_{I\in S} L(I)
$$

It is possible to find the set S maximizing $L(S)$ by dynamic programming, as follows. We denote the ordered set of observed cut locations by $x_1, \ldots, x_{N_{cuts}}$, and set $x_0 = 0$. Let $OPT(k)$ be an optimal set of non-overlapping good intervals in $[0, x_k]$ and let $F(k)$ be the corresponding optimal value. Then $F(0) \equiv 0$, and we compute for $k = 1, 2, \ldots, N_{cuts}$:

$$
F(k) = \max\{F(k-1), \max_{j<k-1}\{F(j) + L([x_{j+1}, x_k])\}\} \tag{3}
$$

and save $OPT(k)$, a set of intervals attaining that optimum. $OPT(N_{cuts})$ is the desired solution.

5.2 Informative Intervals

5.2.1 Naive Discretization of the Data

We now discuss the original, unoriented data set D. We naturally "fold the molecule in half" to create a *folded* data set $\widehat{D} = \{\widehat{c_{m,1}}, \ldots, \widehat{c_{m,N_{cut(m)}}}\}_{m=1}^{N_{mol}}$. The straightforward way to discretize the input data is to uniformly partition the (folded) molecule m into k segments of equal length, replacing it with a vector $v = (v_1, \ldots, v_k) \in \{0,1\}^k$, such that v_j is an indicator variable, indicating whether a cut was observed in the j-th segment. However, real life sizing error is rather large: In many cases, it is of the same order of magnitude as the distance between nearby restriction sites (1-3 percent of the molecule length). This causes naive discretization to fail.

5.2.2 Finding Informative Intervals

We now devise a discretization method that non-uniformly slices the interval $[0, \frac{1}{2}]$ corresponding to the folded molecule in a manner more favorable for our purposes. This method aims at finding those intervals whose associated conjugate pairs supply much information on the orientation of the data. In order for such a pair, $(I, \bar{I})$, to be informative, there must be strong evidence that either I contains at least one restriction site and I' does not, or vice versa. We measure the informativeness of a folded interval by a likelihood calculation analogous to the one given in Section 5.1. The added complications are that we now consider the original, unfolded data set, D, and examine the two dimensional random variable $Y_I(m) = (X_I(m), X_{\bar{I}}(m))$, for a single molecule $m \in D$. Clearly, its distribution is symmetric, *i.e.*, $Prob(Y_I(m) = (k,l)) = Prob(Y_I(m) = (l,k))$. We would use the observations we have on this variable to learn its distribution, and estimate the probability of having true sites at I.

Assuming there are i true restriction sites in I, and j restriction sites in $\bar{I}$, this distribution can be approximated as a function of the following instance parameters:

1. The noise (false cuts) rate λ.
2. For each restriction site R_x of the $i+j$ sites, the probability $\widehat{p'_x}$ of a true cut originating from R_x to appear in I or in $\bar{I}$. As in Section 5.1, we approximate all the $\widehat{p'_x}$ values by a single $\widehat{p'_I}$.

We can define:

$$\widehat{P}((k,l), i, j, \lambda, \widehat{p'_I}) \quad = \quad Prob\{Y_I(m) = (k,l) | i, j \text{ restriction sites}, \lambda, \widehat{p'_I}\}$$

$$= \sum_{g+g'=k,h+h'=l} Poisson(g+h, 2\lambda|I|)Bin(g', i, \widehat{p'_I})Bin(h', j, \widehat{p'_I})$$

The molecules in D are N_{mol} independent random observations of this random variable, from which we obtain the empirical frequency count:

$$\mathcal{Y}_{k,l}(D) = |\{m \in D | Y_I(m) = (k,l) \text{ or } Y_I(m) = (l,k)\}|$$

Hence:

$$\begin{aligned} Likelihood(i, j, \lambda, \widehat{p'_I}) &= Prob(D|i, j, \lambda, \widehat{p'_I}) \\ &= \prod_{m \in D} \widehat{P}(Y_I(m), i, j, \lambda, \widehat{p'_I}) \\ &= \prod_{k,l} \widehat{P}((k,l), i, j, \lambda, \widehat{p'_I})^{\mathcal{Y}_{k,l}(D)} \end{aligned}$$

Using standard numerical maximization techniques, we find, for each such i, j, the most likely parameters, given the observed values of $\mathcal{Y}_{k,l}(D)$. We optimize these parameters, and denote the log of this likelihood, by $\hat{L}^{i,j}(I)$. We define:

$$\hat{L}(I) = \max_{i>0} \hat{L}^{i,0}(I) - \max_{i,j \neq 0 \text{ or } i=j=0} \hat{L}^{i,j}(I)$$

In other words, $\hat{L}(I)$ is the likelihood of the most likely such informative assumption, compared to the null-hypothesis of the most likely non-informative assumption. The higher $\hat{L}(I)$ is, the more informative interval I is. In practice, it is enough to consider only small values of i, j, *i.e.*, $i + j \leq 2$. Furthermore, we stress that although the computation of $\hat{L}^{i,j}(I)$ is analogous to that of $L^j(I)$ in Section 5.1, the analogy does not carry over to $\hat{L}(I)$ versus $L(I)$, as the criteria for informative intervals differ from those for good intervals.

After computing $\hat{L}(I)$ for each candidate interval, we employ a dynamic program, analogous to the one presented in Equation 3, in order to find a set of non overlapping intervals $I_1, \ldots, I_k$, with maximal $\sum_i \hat{L}(I_i)$. This is a set of informative intervals, and we use it to discretize the data to a $k \times N_{mol}$ binary matrix, on which the signature methods of [8] are applicable.

6 Determining Restriction Site Locations

6.1 Probabilistic Model

After stage 5 of the algorithm, we have a set of good intervals, $I_1, \ldots, I_k$ in $(0, 1)$. For each I_i, and for each value of j, we know $L(I_i, j)$, the likelihood of the data within I_i

(for the best values of λ and p'_I) assuming there are j restriction sites in I_i and that no observed cuts within I_i arise from restriction sites outside I_i. In practice, j is no more than 2 (for larger values, this likelihood is negligible). For any vector $J = (j_1, \ldots, j_k)$, we can estimate the likelihood of the event that, for each i, I_i contains exactly j_i restriction sites, by $\prod_{i=1}^{k} L(I_i, j_i)$. This formula is a good approximation to the actual likelihood, provided that for each good interval I_i, restriction sites outside I_i do not give rise to a significant number of observed cuts within I_i. For each vector J with significantly high likelihood, we generate initial values for λ and the set of triplets (p_r, μ_r, σ_r), where r ranges from 1 to $\sum_{i=1}^{k} j_i$, as follows: for an interval I_i containing j_i restriction sites, a different site is placed at the center of each disjoint sub-interval of size $\frac{|I_i|}{j_i}$, and for each such site, p_r is set to p'_{I_i}. λ is set to $\frac{N_{cuts}}{N_{mol}} - \sum p_r$. It turns out that there are not too many likely values for the vector J, and we perform a heuristic likelihood maximization from the starting solution associated with each such vector. Although our starting solutions have all their restriction sites within good intervals, this property is not required to hold at later iterations.

6.2 Approximate Likelihood Score

We now describe the score which we optimize. Let $\psi = (\lambda, \{(p_r, \mu_r, \sigma_r)\}_{r=1}^{k})$ be a set of assumed parameters. We need to compute the likelihood score $s(\psi)$, *i.e.*, the probability of the data given ψ:

$$s(\psi) = Prob(\vec{D}|\psi) = \prod_m Prob(\vec{D_m}|\psi) \tag{4}$$

For a molecule m with the observed cuts $\vec{c}_{m,1}, \ldots, \vec{c}_{m,N_{cut}(m)}$, we do not know which of these observed cuts originated from which of the true restriction sites in ψ, and which are due to background noise. A matching between the observed cuts $\vec{c}_{m,1}, \ldots, \vec{c}_{m,N_{cut}(m)}$ and the assumed restriction sites (or noise) is called an *alignment* between m and ψ. We can therefore write:

$$s(\psi) = \prod_m \sum_a Prob\left(\begin{array}{l} \vec{D_m}\psi, \text{ and the alignment } a \\ \text{between } m \text{ and } \psi \end{array} \middle| \psi \right) \tag{5}$$

The inner summation is done over all possible alignments between the restriction sites assumed by ψ, and the cuts observed in $\vec{D_m}$.

We call an alignment *order preserving* if for every two observed cuts c, c', which are matched to restriction sites r, r', respectively, $c < c'$ iff $\mu_r < \mu_{r'}$. Other alignments are highly unlikely. In practice, we therefore perform the summation of Equation 5 only over the order preserving alignments. We calculate the logarithm of the required probability by dynamic programming.

Typically, the score of the optimal alignment a^* is considerably higher than the score of any other alignment. Therefore calculating the probability of a^* is a reasonable approximation to the true likelihood score. We denote this score by $s^*(\psi)$. $s^*(\psi)$ can be computed by a recurrence relation similar to the dynamic program used to compute $s(\psi)$, the only difference being that we compute maximum instead of summation, replacing Equation 5 with:

$$s^*(\psi) \cong \prod_m \max_a Prob\left(\begin{array}{l} \vec{D_m}\psi, \text{ and the alignment } a \\ \text{between } m \text{ and } \psi \end{array} \middle| \psi \right) \tag{6}$$

We have found the difference between these two scores to be small.

6.3 EM Optimization

We use a variant of the EM (Expectation Maximization) heuristic for this optimization, as detailed below. We remark that [5], [2] and [9] use EM and other heuristics (gradient descent and Monte Carlo Markov Chain simulation) to optimize a related score. For completeness, we include a description of this stage in our algorithm, although it is not unique to this work. In contrast to prior art, we have the advantage of working with oriented data using a good starting solution.

At each iteration of the optimization procedure, we assume a parameter set $\tilde{\psi}$, and reassign observed cuts to the postulated restriction sites. Throughout this procedure, we maintain, for each restriction site r, the set $L(r)$ of locations of cuts, observed across all the molecules in the data set, that were assigned to r by the optimal alignments between each of the molecules and $\tilde{\psi}$; $L(noise)$ is the set of all observed cut locations that optimal alignments attribute to background noise. Define $N(r) = |L(r)|$, $N(noise) = |L(noise)|$.

In order to find the maximum of this score function, we iterate as follows:

1. **Expectation:** For each restriction site r, estimate its parameters given the set $L(r)$ of all observed cut locations $\{l_j\}_{j=1}^{N(r)}$ aligned with r. This statistical estimation is performed as follows:
   - $\tilde{p_r} = \frac{N(r)}{N_{mol}}$ is a maximum likelihood estimator for p_r.
   - Similarly, $\tilde{\lambda} = \frac{N(noise)}{N_{mol}}$ is a maximum likelihood estimator for λ.
   - Assuming that the locations in $L(r)$ are normally distributed with expectation μ_r, the term $\tilde{\mu_r} = \frac{\sum l_j}{N(r)}$ is a maximum likelihood estimator for μ_r.
   - Assuming that the locations in $L(r)$ are normally distributed with variance σ_r^2, the term $\tilde{\sigma_r^2} = \frac{\sum l_j^2}{N(r)} - \tilde{\mu_r}^2$ is a maximum likelihood estimator for σ_r^2.

2. **Maximization:** Given the current set $\tilde{\psi} = (\tilde{\lambda}, \{(\tilde{p_r}, \tilde{\mu_r}, \tilde{\sigma_r})\}_{r=1}^k$ of estimated parameters, for each molecule m, find the optimal alignment $a^*(m)$ between $\tilde{\psi}$ and m. Adjust the sets $L(r)$ for each r, to contain the observed cut locations that were matched to r by the new optimal alignments $\{a^*(m)\}_m$.

Note, however, that even if our probabilistic model is correct, and the locations of the observed cuts originating from the restriction site r indeed have the distribution $Normal(\mu_r, \sigma_r^2)$, the locations in $L(r)$ do not have the same distribution. Rather, the distribution of these locations is a doubly truncated Gaussian. Luckily, it seems from our experiments with real data that the difference between these distributions is negligible.

In practice, we have observed that the score s^* occasionally splits a site into two nearby sites with lower p_i-s. In order to make sure this only happens when the evidence for such a pair is solid, we do the following: For each pair of nearby sites $r, r+1$, we count the number $M(r, r+1)$ of molecules m, for which both sites are matched by $a^*(m)$. We multiply the score, s^*, by the probability of observing $M(r, r+1)$, given that there are two independent cuts $r, r+1$. This modification seems to solve the splitting problem in practice, provided there are no chains of such nearby sites.

7 Results

7.1 Simulations

The latter stage of our algorithm is actually a hard-decision EM algorithm, iteratively improving s^*. Our main contribution is being able to perform this search on the reduced space of oriented molecules. After orientation, any existing method for tackling this problem can, in principle, be employed, with the advantage of knowing the correct orientation. We therefore gathered statistics indicating our success in orienting the molecules, presenting the percentage of mis-oriented molecules as a figure of merit.

We have applied our algorithm to many data sets, generated randomly according to a range of statistical scenarios: we altered the number of restriction sites, digestion rate, magnitude of sizing error, and false cut rate. For each such set of parameters, data sets of different size (*i.e.*, number of molecules) were analyzed. The results are presented in Figure 7.1. They show that for standard values of the model parameters errors in orientation are quite rare, dropping further, of course, as more molecules are examined. Furthermore, these results are stable with respect to alterations of instance parameters.

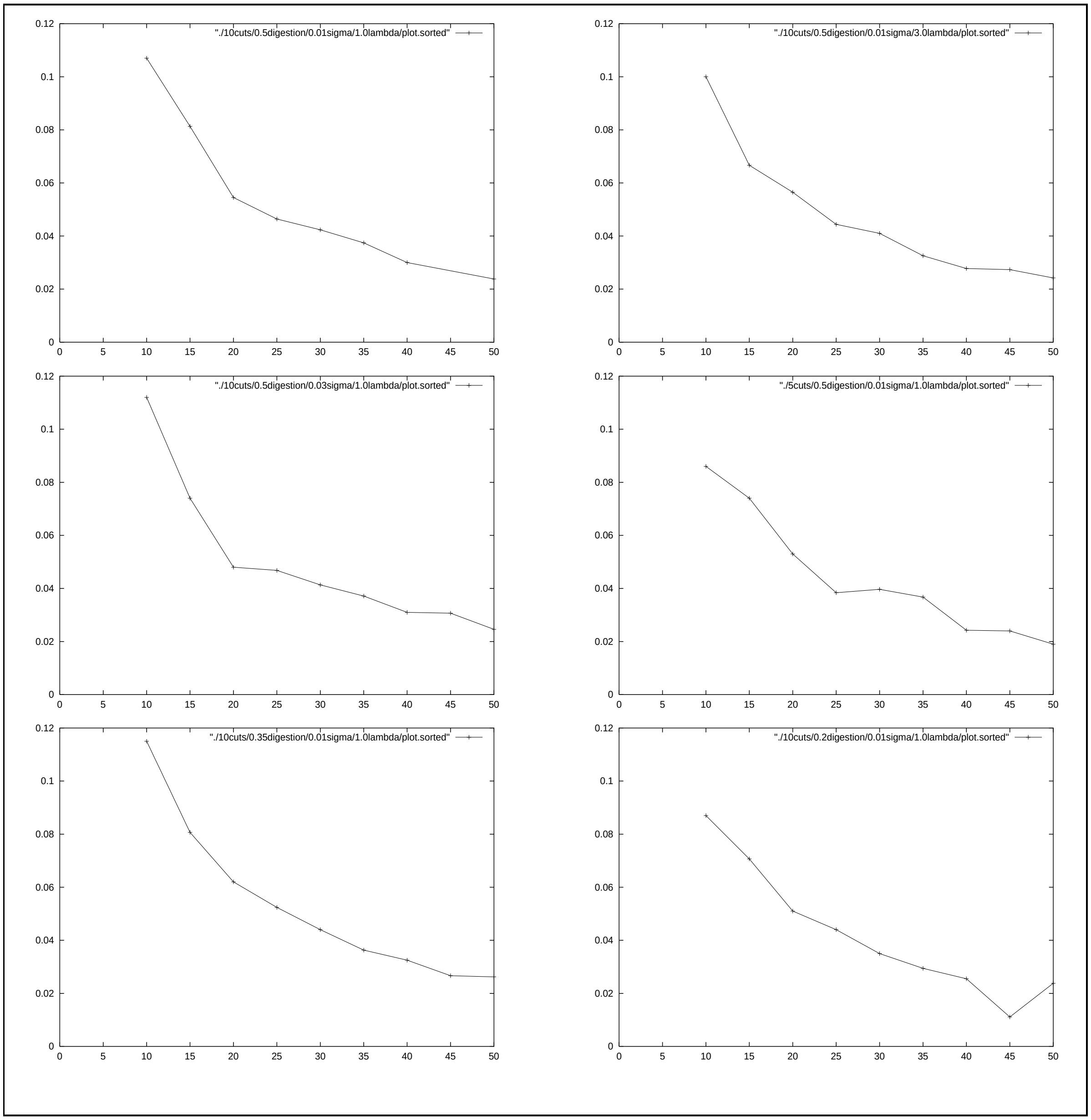


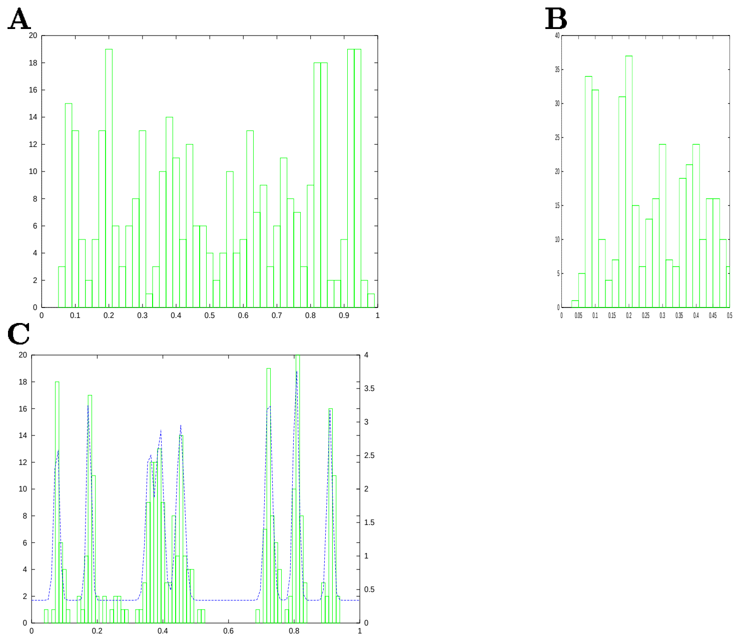
Figure 1: Results on simulated data. Each graph presents results with a specific set of simulation parameters (number of true cuts, digestion rate, standard sizing error, and false cut rate). The x-axis is the number of molecules, while the y-axis is the fraction of the molecules that are incorrectly oriented by the algorithm, averaged across 50 simulated data sets with the appropriate parameters and number of molecules. The basic scenario (top left) is for 10 restriction sites, digestion rate of 0.5, sizing error of standard deviation 0.01, and false cuts rate of 1.0. The other cases shown are for larger (3.0) false cut rate (top right), larger (0.03) sizing error (middle left), fewer (5) restriction sites (middle right), and lower (0.35 and 0.2) digestion rates (bottom left and right, respectively).

7.2 Real Data

Our algorithm was implemented in a blind test on real biological data provided by D. Schwartz's laboratory. There were eight data sets, with cosmids (50Kb) and BAC (130-180Kb) molecules. We encountered at most a dozen restriction sites in each data set, with the digestion rate varying widely from ~ 0.1 to ~ 0.9 between data sets. There were differences of up to 0.3 in the digestion rate between sites in the same data set. The false cut rate varied too, sometimes exceeding one false cut per molecule, on the average. The average sizing error was on the order of magnitude of $\frac{1}{100}$ to $\frac{1}{20}$ of the molecule length. Examples of the results are given in Figures 2, 3 and 4. Figures 2 and 4 describe two experiments with disjoint data-sets for the same molecule, demonstrating application of our algorithm for different digestion rates.

Bud Mishra and Thomas Anantharaman of N.Y.U. kindly examined our results and then provided us with the true restriction sites, as determined directly from sequence data or indirectly by inference from pulsed field gel electrophoresis data combined with the optical mapping data. In several cases where two restriction sites were separated by less than 1000 bases, our algorithm reported only one site. This difficulty was also encountered in [2] and may be an inherent problem due to the inability of the imaging system to detect small restriction fragments. Apart from this, our results were correct on four of the eight examples; these include the examples shown here. On two further examples the results were correct except that one restriction fragment with a low digestion rate was missed. More careful inspection showed that almost all (95%) of the molecules were correctly oriented for these two examples, with only the last stage of the algorithm, EM optimization, being inaccurate. On two additional examples where the data was of low quality (as measured by the method reported in [2]) the results were less good. In one of these examples two restriction sites were missed, despite the fact that they were evident by inspection from the histogram of observed cut locations in the oriented molecules. In the other example our program failed, missing two true restriction sites and introducing three false ones. On the positive side, we provided the correct solution for the data-set in Figure 4, which was classified by [2] as the hardest among the eight data-sets we have considered.

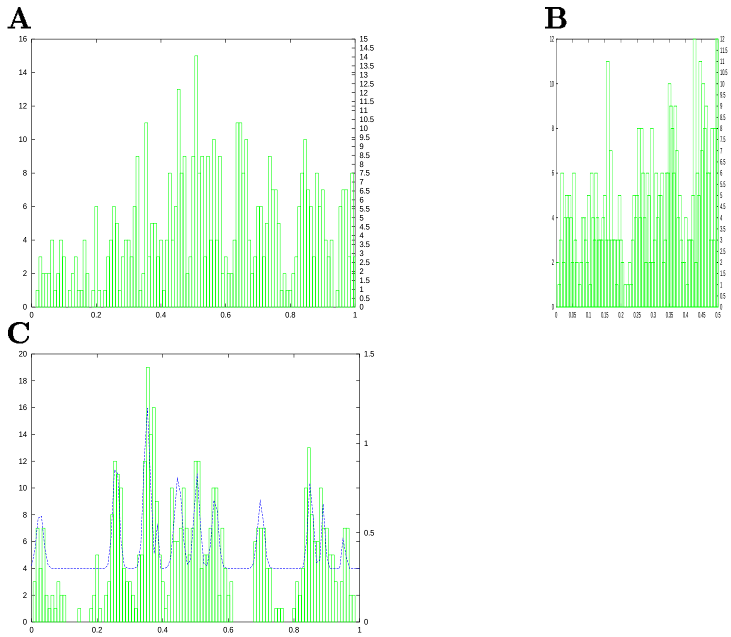
It appears that our program tends to determine the orientations of the molecules correctly, thereby substantially reducing the parameter space for any subsequent maximum likelihood method. The occasional misestimation of the number of restriction sites can be alleviated by further tuning of the algorithm for determining the restriction site data from the oriented molecules.



D

True	Computed		
μ_i	μ_i	p_i	σ_i
0.074	0.076	0.680	0.011
0.174	0.174	0.787	0.009
0.363 0.369	0.358	0.702	0.014
0.397 0.398	0.391	0.808	0.015
0.464	0.453	0.851	0.016
0.727	0.722	0.978	0.014
0.808	0.805	0.978	0.011
0.914	0.908	0.766	0.011
False Cut Rate:			0.340

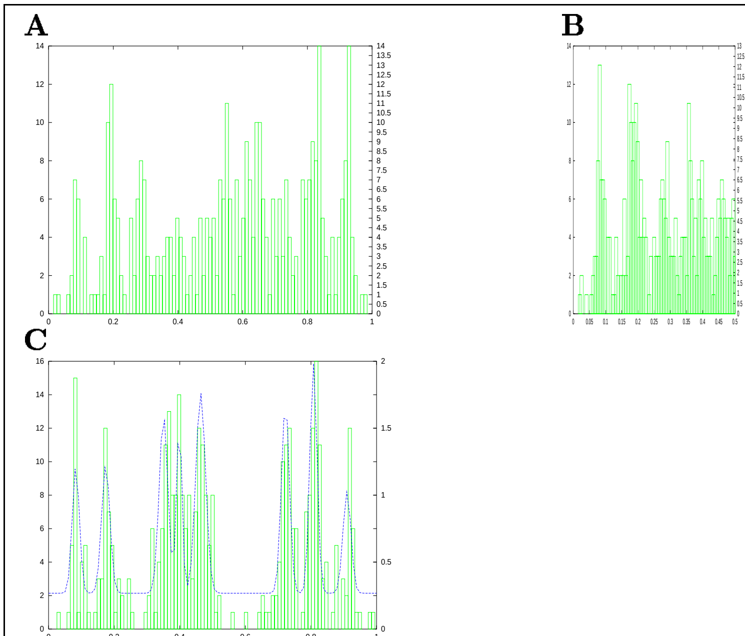
Figure 2: Test example 6262-1: Instance parameters: 54 molecules, 370 observed cuts in total. 7 molecules screened out. A: Histogram of observed cuts in equalized sub-intervals. B: Histogram of the 'folded' data. C: Histogram of the oriented, cleaned data superimposed with the optimal density function. D: The suggested solution. Pairs of sites in the same row of the "True" column indicate close-by sites that cannot be distinguished.



D

True	Computed		
μ_i	μ_i	p_i	σ_i
0.008	short: reported as 0		
0.024 0.031	0.025	0.097	0.014
0.254	0.257	0.180	0.013
0.346	0.352	0.266	0.013
0.361	0.381	0.050	0.004
0.448	0.447	0.162	0.015
0.515	0.504	0.147	0.012
0.562	0.559	0.119	0.013
0.697	0.698	0.111	0.013
0.837 0.854	0.850	0.129	0.011
0.889	0.889	0.079	0.007
0.951 0.953	0.953	0.050	0.005
False Cut Rate:			0.300

Figure 3: Test example 6401: Instance parameters: 280 molecules, 494 observed cuts in total. 7 molecules screened out. See Figure 2 for legend.



D

True	Computed		
μ_i	μ_i	p_i	σ_i
0.074	0.082	0.294	0.015
0.174	0.174	0.321	0.017
0.363 0.369	0.350	0.455	0.018
0.397 0.398	0.398	0.348	0.011
0.464	0.463	0.544	0.021
0.727	0.722	0.464	0.017
0.808	0.807	0.535	0.015
0.914	0.909	0.250	0.017
False Cut Rate:			0.267

Figure 4: Test example 6262-0: Instance parameters: 114 molecules, 396 observed cuts in total. 7 molecules screened out. See Figure 2 for legend.

The results presented here are the first blind test of a novel algorithm. The performance of our algorithm does not match up, at this point, to that of [2], which has been employed for three years now with considerable success. The results illustrate that a refined discretization procedure, combined with the signature method, greatly help the global optimization in cases where the likelihood landscape is unfavorable for optimizing over unoriented data. Since the orientation phase of our algorithm seems to be very successful, one natural suggestion for future work is to combine that phase with a refined global optimization procedure for finding the cut sites on the more favorable landscape of oriented data.

Acknowledgments

We thank Thomas Anantharaman, Bud Mishra and David Schwartz for providing us with experimental data and analyzing our results. We also thank Roded Sharan for implementing the signature method. We further thank an anonymous referee for helpful comments.

References

[1] T. Anantharaman and B. Mishra. Genomics via optical mapping (i): Probabilistic analysis of optical mapping models. Technical Report TR1998-770, Computer Science Department, New York University, 1998.

[2] T. S. Anantharaman, B. Mishra, and D. C. Schwartz. Genomics via optical mapping ii: Ordered restriction maps. *Journal of Computational Biology*, 4(2):91–118, 1997.

[3] W. Cai, H. Aburatani, V. P. Stanton, D. E. Housman, Y. K. Wang, and D. C. Schwartz. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proceedings of the National Academy Science U.S.A*, 92:5164–8, 1995.

[4] W. Cai, J. Jing, B. Irvin, L. Ohler, E. Rose, H. Shizuya, U. J. Kim, M. Simon, T. A. T, B. Mishra, and D. C. Schwartz. High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proceedings of the National Academy Science U.S.A*, 95(7):3390–3395, 1998.

[5] V. Dančík and M. S. Waterman. Simple maximum likelihood methods for the optical mapping problem. In *Proceedings of the Workshop on Genome Informatics (GIW '97)*, 1997.

[6] V. Dančík, S. Hannenhalli, and S. Muthukrishnan. Hardness of flip-cut problems from optical mapping. *Journal of Computational Biology*, 4:119–125, 1997.

[7] J. Jing, J. Reed, J. Huang, X. Hu, V. Clarke, J. Edington, D. Housman, T. S. Anantharaman, E. J. Huff, B. Mishra, B. Porter, A. Shenker, E. Wolfson, C. Hiort, R. Kantor, C. Aston, and D. C. Schwartz. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proceedings of the National Academy Science U.S.A*, 95(14):8046–8051, 1998.

[8] R. M. Karp and R. Shamir. Algorithms for optical mapping. In *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology, New York, 1998*, pages 117–124. ACM Press, 1998. to appear in Journal of Computational Biology.

[9] J. K. Lee, V. Dančík, and M. S. Waterman. Estimation for restriction sites observed by optical mapping using reversible-jump markov chain monte carlo. In *Proc. RECOMB 98*, pages 147–152, 1998.

[10] J. Lin, R. Qi, C. Aston, J. Jing, T. Anantharaman, B. Mishra, O. White, M. Daly, K. Minton, J. Venter, and D. Schwartz. Whole-genome shotgun optical mapping of deinococcus radiodurans. *Science*, 285:1558–1562, 1999.

[11] X. Meng, K. Benson, K. Chada, E. J. Huff, and D. C. Schwartz. Optical mapping of lambda bacteriophage clones using restriction endonucleases. *Nature Genetics*, 9:432–438, 1995.

[12] S. Muthukrishnan and L. Parida. Towards constructing physical maps by optical mapping: An effective, simple, combinatorial approach. In *Proc. RECOMB 1997*, pages 209–219. ACM Press, 1997.

[13] L. Parida. On the approximability of physical map problems using single molecule methods. In *Procceedings of Discrete Mathematics and Theoretical Computer Science*, 1999.

[14] L. Parida and B. Mishra. Partitioning k clones: Hardness results and practical algorithms for the k-populations problem. In *Proc. RECOMB 98*, pages 192–201, 1998. To appear in Discrete Applied Math.

[15] I. P. R.M. Karp and R. Shamir. Algorithms for optical mapping of dna combining discrete and continuous methods. In *Proceedings of the Seventh international Conference on Intelligent Systems for Molecular Biology (ISMB'99)*, pages 159–168, 1999.

[16] A. Samad, E. J. Huff, and D. C. Schwartz. Optical mapping: A novel, single-molecule approach to genomic analysis. *Genome Rsearch*, 5(1), 1995.

[17] H. Samad, W. W. Cai, X. Hu, B. Irvin, J. Jing, J. Reed, X. Meng, J. Huang, E. Huff, B. Porter, et al. Mapping the genome one molecule at a time–optical mapping. *Nature*, 378:516–517, 1995.

[18] D. C. Schwartz, X. Li, L. I. Hernandez, S. P. Ramnarain, E. J. Huff, and Y. K. Wang. Ordered restriction maps of saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262:110–114, 1993.

[19] D. C. Schwartz and A. Samad. Optical mapping approaches to molecular genomics. *Current Opinion. in Biotechnology*, 8(1):70–74, 1997.

[20] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.