An Algorithm Combining Discrete and Continuous Methods for Optical Mapping

R.M. Karp
Department of Computer Science & Engineering
University of Washington
Box 352350, Seattle, WA 98195-2350, USA
karp@cs.washington.edu

I. Pe'er   R. Shamir
School of Mathematics
Tel Aviv University
Tel Aviv, 69978, Israel
izik+shamir@math.tau.ac.il

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.

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 [Cai *et al.* 1998]. Optical mapping is a novel technique for obtaining restriction maps [Samad, Huff, & Schwartz 1995],[Cai *et al.* 1995],[Schwartz *et al.* 1993],[Meng *et al.* 1995], [Samad *et al.* 1995],[Jing *et al.* 1998],[Cai *et al.* 1998]. 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 [Schwartz & Samad 1997]. 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 Kilobases.

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.
- 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 within the same experiment [Cai *et al.* 1995].

Various combinatorial formulations of the optical mapping problem have been proven NP-complete [Anantharaman, Mishra, & Schwartz 1997], [Parida & Mishra 1998], [Muthukrishnan & Parida] and even hard to approximate [Par 1999]. Bayesian approaches using global optimization have been applied in [Anantharaman, Mishra, & Schwartz 1997] and [Lee, Dančík, & Waterman 1998]. 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 [Anantharaman, Mishra, & Schwartz 1997] is successfully used in practice for analysis of real laboratory data. Other studies [Dančík, Hannenhalli, & Muthukrishnan 1997], [Muthukrishnan & Parida], [Karp & Shamir 1998] discretize the input, and devise combinatorial solution methods. The algorithm of [Karp & Shamir 1998] 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 [Anantharaman, Mishra, & Schwartz 1997] and [Lee, Dančík, & Waterman 1998]. In order to orient the molecules we adapt a combinatorial algorithm from [Karp & Shamir 1998]. However, instead of using a fixed, uniform discretization of the data as in [Karp & Shamir 1998], 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 real data sets provided by David Schwartz's lab 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 further tuning of the algorithm.

The paper is organized as follows: The next section outlines the general strategy of our algorithm. In the subsequent section we present our probabilistic model. The three following sections describe the major stages in our algorithm. The last section gives experimental results.

General Strategy

We sketch the *signature method* that was suggested in [Karp & Shamir 1998] 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 extend 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 [Karp & Shamir 1998] the signature method was used to actually 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 [Karp & Shamir 1998] were applied to uniformly discretized real data and failed. The main reason is that the sizing errors (and hence the errors in cut observations) are too large to conform with the uniform discretization. Probabilistic analysis of the error under uniform discretization [Anantharaman & Mishra 1998] illuminates the limitations of such discretization from a theoretic perspective, and motivates finding a better way to apply combinatorial methods. The fact that the algorithms in [Karp & Shamir 1998] disregard faulty molecules adds to the problem.

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 [Anantharaman, Mishra, & Schwartz 1997],[Lee, Dančík, & Waterman 1998]. However, there is room for improvement in some important respects: 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 a score based on correctly oriented data would give better optimization results. Moreover, the likelihood function for restriction site locations has many local optima, making it quite hard to determine the global optimum unless one starts the search with a solution that is close to the global optimum.

Our approach uses elements of both the continuous approach and the discrete approach, attempting to remedy the shortcomings of each: We reduce the main error source in the discrete procedure for orienting the molecules by a more subtle consideration of the continuous data. The results of the refined orientation procedure eliminate the need for orientation parameters for individual molecules, and thus help the continuous global search heuristic avoid local optima.

We assume a continuous probabilistic model of the data, similar to [Anantharaman, Mishra, & Schwartz 1997] and [Lee, Dančík, & Waterman 1998]. 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 prepare to 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 [Karp & Shamir 1998],[Muthukrishnan & Parida], since it does not specify the sizes of these intervals or restrict their possible end points to a predefined discrete set. This generalization enhances the performance of the signature algorithm, allowing it to overcome the sizing error. 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 partition 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 maximum 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.

Model and Terminology

We now define our probabilistic model of the problem. Similar models were used in [Anantharaman, Mishra, & Schwartz 1997],[Lee, Dančík, & Waterman 1998]. Each of the N_{mol} molecules is faulty with (independent) probability p_{faulty}. In each proper molecule false cuts are Poisson distributed with rate λ.

We assume there is some unknown number t of (true) restriction sites, with the i-th site R_i located at positicn 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 $N_{cuts}(m)$ entries (observed cuts), at positions $c_{m,1}, \ldots, c_{m,N_{cuts}(m)}$. In each proper molecule 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 μ_i from the data D. In the course of doing this we will also determine the orientations of the molecules and the other parameters of the probabilistic model.

Define N_{cuts} to be $\sum_m N_{cuts}(m)$. For an interval I, define $X_I(D)$ to be the number of cuts observed in D, inside the interval I. Note, that since the data is a function of our probability space, $X_I(D)$ is a random variable.

Screening Out Faulty Molecules

We describe how to screen out faulty molecules from oriented data (step 4 in the algorithm). The procedure for unoriented data (step 1) is analogous, and is omitted due to lack of space.

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:

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

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.

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 when $\phi_{\epsilon,\vec{D}}(x)$ is large.

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

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

The molecules with the smallest ϕ_m are the ones most likely to be faulty, and should therefore be discarded. We 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.

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

Good Intervals and Informative Intervals

Good Intervals

We will eventually determine the restriction sites by an iterative maximum likelihood computation. Since the method we use (the E-M 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 the interval can be attributed to restriction sites within the interval itself. In this subsection we describe the process of finding the good intervals.

We restrict attention to the $O(N_{cuts}^2)$ intervals having observed cuts as their end points. 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 eliminate those intervals for which this value is smaller than a chosen threshold, as well as those intervals that are unreasonably long.

We consider the oriented data set $\vec{D}$. For each interval $I = (a,b) \subset (0,1)$, we examine the random variable $X_I(\vec{m})$, for a single molecule $\vec{m} \in \vec{D}$ in our probability space. 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 instance parameters:

1. The false cut rate λ
2. For each restriction site R_i within I, the probability p'_i of a true cut 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:

$$\begin{aligned}
Psn(h,\alpha) &= \frac{e^{-\alpha}\alpha^h}{h!} \\
Bin(i,j,p) &= \binom{j}{i} p^i(1-p)^{j-i} \\
P(k,j,\lambda,p'_I) &= Pr\{\, X_I(\vec{m}) = k | \\
&\qquad j \text{ restriction sites}, \lambda, p'_I \,\} \\
&= \sum_{h+i=k, i\le j} Psn(h,\lambda|I|)Bin(i,j,p'_I)
\end{aligned}$$

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:

$$\chi_k(\vec{D}) = |\{\vec{m} \in \vec{D} | X_I(\vec{m}) = k\}|$$

Hence:

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

Using standard numerical maximization techniques, for each such j, we can get the most likely parameters, given the observed values of $\chi_k(\vec{D})$. We optimize these parameters, and denote the log of this 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.* $j \le 2$. The higher $L(I)$ is, the more we consider I to be 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: Let $x_1,\ldots,x_{N_{cuts}}$ be the ordered set of observed cut locations, and let $x_0 = 0$. 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 $k = 1,2,\ldots,N_{cuts}$:

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

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

Informative Intervals

We now discuss the original, unoriented data set D. We "fold the molecule in half" to create a *folded* data set $\widehat{D} = \{\widehat{c_{m,1}},\ldots,\widehat{c_{m,N}}_{N_{cuts}(m)}\}_{m=1}^{N_{mol}}$. In order for a folded interval to be informative there must be strong evidence that, of the two conjugate intervals associated with the folded interval, one contains at least one restriction site and the other does not. We measure the informativeness of a folded interval by a likelihood calculation similar to the one given in the previous sub-section, with the added complication that we consider the original, unfolded, data set D, and examine the two dimensional random variable $Y_I(m) = (X_I(m), X_{\bar{I}}(m))$, with $\bar{I} = (1-b, 1-a)$ being the conjugate interval of $I = (a,b)$. The computation is an easy extension of the likelihood computation presented in the previous sub-section, and is omitted due to lack of space. Once this measure of informativeness has been calculated, an optimal disjoint set of informative intervals is easily found by dynamic programming.

Determining Restriction Site Locations

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 all the observed cuts within I_i arise from restriction sites within I_i. In practice, j is no more than 2 (for larger values, this likelihood is practically 0). For any vector $\bar{j} = (j_1,\ldots,j_k)$, 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 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 $\bar{j}$ 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 site is placed at the center of each sub-interval of size $\frac{|I_i|}{j_i}$, and for each such site

p_r is set to p'_{I_i}. λ is fixed to $\frac{N_{cuts}}{N_{mol}} - \sum p_r$. It turns out that there are not too many likely values for the vector $\vec{j}$, 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.

We use a variant of the EM (Expectation Maximization) heuristic for this optimization, as detailed below. We remark that [Dančík & Waterman 1997], [Anantharaman, Mishra, & Schwartz 1997] and [Lee, Dančík, & Waterman 1998] use EM and other heuristics (gradient descent and Monte Carlo Markov Chain simulation) to optimize a related score. We have the advantage of working with oriented data using a good starting solution.

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

$$s(\psi) = Pr(\vec{D}|\psi) = \prod_m Pr(\vec{D_m}|\psi) \qquad (2)$$

For the molecule m with the observed cuts $\vec{c}_{m,1}, \ldots, \vec{c}_{m,N_{cuts}(m)}$, 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 $\vec{c}_{m,1}, \ldots, \vec{c}_{m,N_{cuts}(m)}$ and the true restriction sites (or noise) is called an *alignment* between m and ψ. We can therefore write:

$$s(\psi) = \prod_m \sum_a Pr\left(\vec{D_m} \;\middle|\; \begin{array}{c} \psi, \text{ and the} \\ \text{alignment } a \\ \text{between } m \text{ and } \psi \end{array}\right) \cdot Pr(a) \qquad (3)$$

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. We therefore perform the summation in equation 3 only over the order preserving alignments. Since $Pr(a)$ depends only on $N_{cuts}(m)$ and J, it only multiplies the total score by a constant factor, and we omit it when assuming the same number J of sites:

$$s(\psi) \cong \prod_m \sum_a Pr\left(\vec{D_m} \;\middle|\; \begin{array}{c} \psi, \text{ and the alignment } a \\ \text{between } m \text{ and } \psi \end{array}\right) \qquad (4)$$

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 $s^*(\psi)$. $s^*(\psi)$ is computable by a dynamic programming recurrence similar to the one used to compute $s(\psi)$, but taking maximum instead of summation, replacing equation 4 with:

$$s^*(\psi) \cong \prod_m \max_a Pr\left(\vec{D_m} \;\middle|\; \begin{array}{c} \psi, \text{ and the alignment } a \\ \text{between } m \text{ and } \psi \end{array}\right) \qquad (5)$$

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

Throughout the optimization procedure, we maintain, for each restriction site $r = \tilde{R}_i$, the set $L(r)$ of locations of observed cuts that were assigned to r by the optimal alignments between each of the molecules and ψ. 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 $r = \tilde{R}_i$, estimate its parameters, given the set $L(r)$ of all observed cut locations $\{l_j\}_{j=1}^{N(r)}$ aligned with the site. The statistical estimation of the parameters of the restriction site r using $L(r)$ is done as follows:
   - $\bar{p_r} = \frac{N(r)}{N_{mol}}$ is a maximum likelihood estimator for p_r.
   - Assuming that the locations in $L(r)$ are normally distributed with expectation μ_r, the term $\bar{\mu_r} = \frac{\sum l_j}{N(r)}$ 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 $\bar{\sigma_r^2} = \frac{\sum l_j^2}{N(r)} - \bar{\mu_r}^2$ is a maximum likelihood estimator for σ_r^2.
2. **Maximization:** Given the current estimated parameters, ψ, for each molecule m, find the optimal alignment $a^*(m)$ between ψ 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 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 of molecules m, for which both sites are matched by $a^*(m)$. We multiply the score s^* by the probability of observing this number, 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.

Results

The above algorithm was implemented in a blind test on real biological data provided by D. Schwartz's laboratory. 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. We encountered 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's length. Examples of the results are given in figures 1,2 and 3. Figures 2 and 3 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 (in one data set,) 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 [Anantharaman, Mishra, & Schwartz 1997] and may be an inherent problem due to the inability of the imaging system to detect very 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. On two additional examples where the data was of low quality (as measured by the program reported in [Anantharaman, Mishra, & Schwartz 1997]) 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 criented molecules. In the other example our program failed, missing two restriction sites and introducing three false restriction sites.

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.

The results presented here are the first blind test of a novel algorithm. Obviously, the performance of our algorithm does not match up, at this point, to that of [Anantharaman, Mishra, & Schwartz 1997], which has been employed for two 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. For example, our algorithm gives a correct solution for the data set in figure 3, which was classified by [Anantharaman, Mishra, & Schwartz 1997] as the hardest global optimization challenge among the 8 data sets we have considered.

Acknowledgments

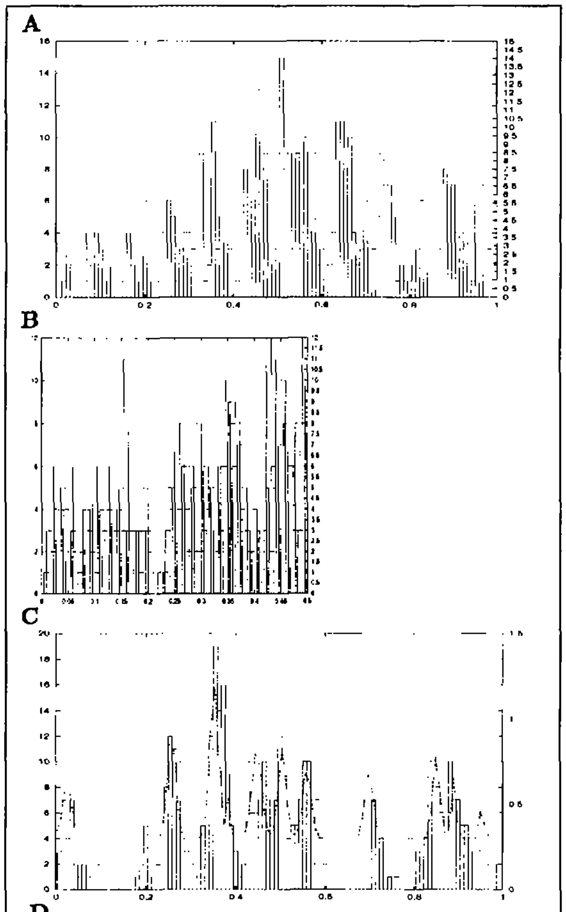
We thank Thomas Anantharaman, Bud Mishra and David Schwartz for providing us with experimental data and analyzing our results. We also thank Roded Sharan for implementing the signature method. I.P. thanks the Clore foundation, for partially supporting this research. Research of R.S. was supported by an infrastructure grant from the Israeli Ministry of Science.

References

Anantharaman, T., and Mishra, B. 1998. Genomics via optical mapping (i): Probabilistic analysis of optical mapping models. Technical Report TR1998-770, Courant Institute of Mathematical Sciences, New York University.

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

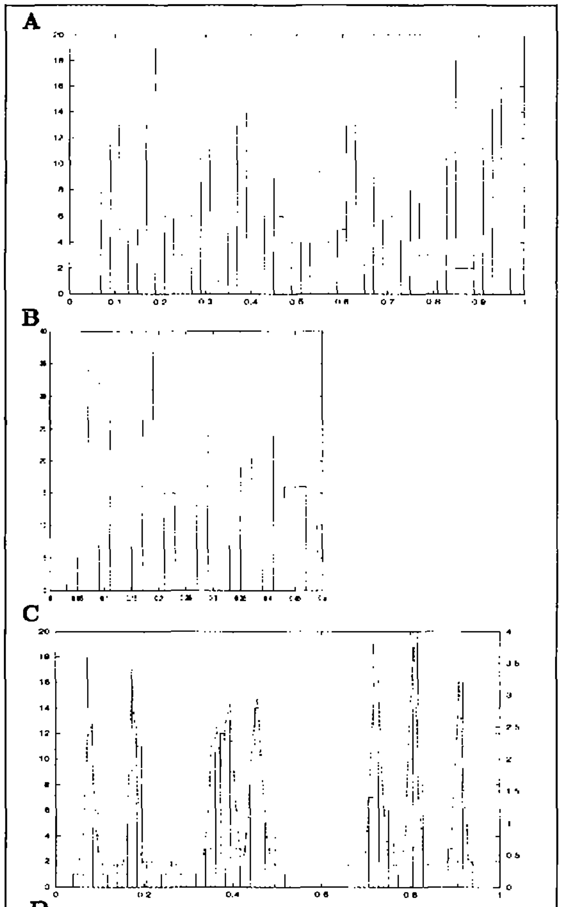
Cai, W.; Aburatani, H.; Stanton, V. P.; Housman, D. E.; Wang, Y. K.; and Schwartz, D. C. 1995. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proceedings of the National Academy Science U.S.A* 92:5164–8.



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

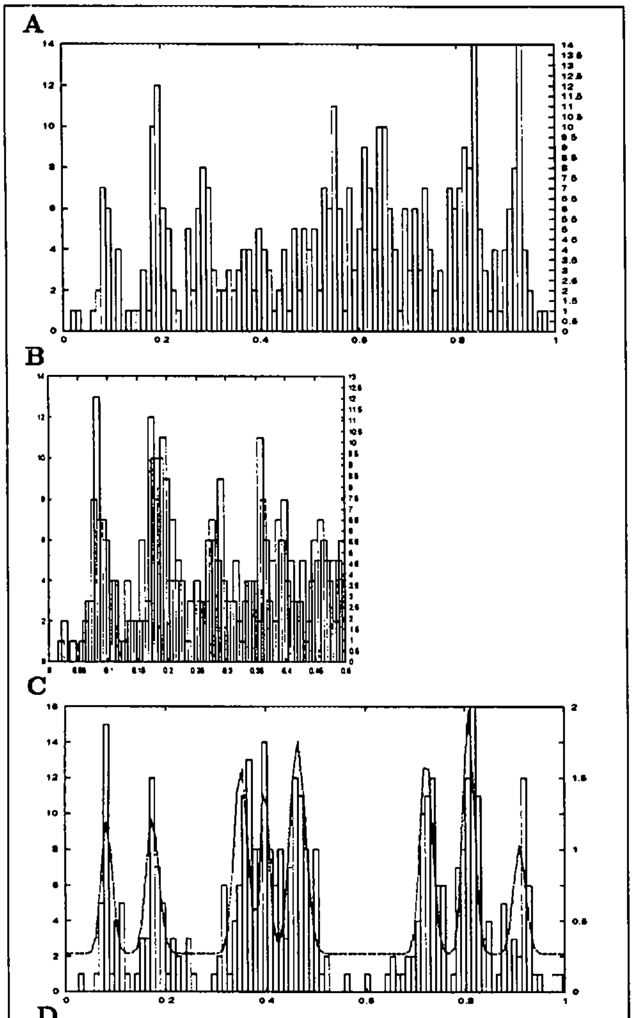
Figure 1: Test example 6401: Instance parameters: 280 molecules, 494 observed cuts in total. 7 molecules screened out. A: Histogram of observed cuts in equal-sized sub-intervals. B: Histogram of the 'folded' data. C: Histogram of the oriented, cleaned data superimposed with the optimal density function. 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
0.074	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. See figure 1 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 3: Test example 6262-0: Instance parameters: 114 molecules, 396 observed cuts in total. 7 molecules screened out. See figure 1 for legend.

Cai, W.; Jing, J.; Irvin, B.; Ohler, L.; Rose, E.; Shizuya, H.; Kim, U. J.; Simon, M.; T, T. A.; Mishra, B.; and Schwartz, D. C. 1998. 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.

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

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

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

Karp, R. M., and Shamir, R. 1998. Algorithms for optical mapping. In *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology (RECOMB), New York, 1998*, 117–124. ACM Press.

Lee, J. K.; Dančík, V.; and Waterman, M. S. 1998. Estimation for restriction sites observed by optical mapping using reversible-jump markov chain monte carlo. In *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology (RECOMB),New York*, 147–152.

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

Muthukrishnan, S., and Parida, L. Towards constructing physical maps by optical mapping: An effective, simple, combinatorial approach.

1999. On the approximability of physical map problems using single molecule methods. In *Procceedings of Discrete Mathematics and Theoretical Computer Science (DMTCS), Auckland*, 310–328.

Parida, L., and Mishra, B. 1998. Partitioning k clones: Hardness results and practical algorithms for the k-populations problem. In *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology (RECOMB),New York*, 192–201.

Samad, H.; Cai, W. W.; Hu, X.; Irvin, B.; Jing, J.; Reed, J.; Meng, X.; Huang, J.; Huff, E.; Porter, B.; Shenker, A.; Anantharaman, T.; Mishra, B.; Clarke, V.; Dimolata, E.; Edington, J.; Hiort, C.; Rabbah, R.; Siada, J.; and Schwartz, D. 1995. Mapping the genome one molecule at a time–optical mapping. *Nature* 378:516–517.

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

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

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