Simple Maximum Likelihood Methods for the Optical Mapping Problem

Vlado Dančík **Michael S. Waterman**

dancik@hto.usc.edu msw@hto.usc.edu

Department of Mathematics, University of Southern California
Los Angeles, U.S.A.

Abstract

Recently a new method for obtaining restriction maps was developed by David Schwartz at NYU. Using this method restriction maps are created from fluorescent images of individual molecules obtained using a microscope. For every individual observed molecule, image processing methods are used to generate a list of the approximate locations of the sites where the molecule is cut by the restriction enzyme. Our task is to find the location of all restriction sites given the observed cutting sites. This is also complicated by the fact that an orientation of the molecules is unknown, i.e. for a cut-site x we do not know whether x or $1 - x$ corresponds to a restriction site in a unit length molecule.

First we consider the case that the orientation of all molecules and the number c of restriction sites are known. We suppose that for each restriction site location y_j the corresponding measured cut-sites follow the normal distribution with the density function $g(x; \theta_j, \sigma_j)$ for some σ_j. (This means the measurement is unbiased with mean θ_j.) The observed cut-sites locations $x_1, \dots, x_n$ then follow the mixture distribution $f(x; \mathbf{p}, \theta, \sigma) = \sum_{j=1}^{k} p_j g(x; \theta_j, \sigma_j)$, where $\sum p_j = 1$. Using the likelihood principle we wish to find parameters $\mathbf{p}, \theta, \sigma$ that achieve the maximum of the likelihood function $\prod_{i=1}^{n} f(x_i; \mathbf{p}, \theta, \sigma)$. In our case it is natural to assume that $p_1 = \cdots = p_k = 1/k$ and $\sigma_1 = \cdots = \sigma_k = \sigma$ for a constant σ.

Frequently in the Optical Mapping there appear "false" cuts, i.e. cuts corresponding to no restriction site. In our model we accommodate false cuts by using an uniform component in the mixture distribution. We use EM algorithm and Bayes theorem for computing the maximum likelihood estimate and compare our results for the different variants of our model.

We explore how the change of the orientation of some molecules influences the maximum likelihood estimate and show that the orientation question can be in our case answered for each molecule separately. Finally we present few ideas for specifying the orientation of molecules without investigating the positions of restriction sites.

1 Introduction

There is a group of enzymes known as restriction endonucleases (or restriction enzymes) that are able to cleave (cut) DNA molecules. The *restriction sites* – the positions where DNA molecule is cleaved is usually specified by a short sequence of nucleotides. For given restriction enzyme(s) a DNA molecule exhibits a typical pattern of restriction sites called *restriction map*. Restriction maps are frequently used in molecular biology from genetic engineering to genome mapping. The standard way for constructing maps is by sizing the restriction fragments using gel electrophoresis. *Optical Mapping* (OM) is a new single-molecule approach to constructing restriction maps developed by D. Schwartz at the W.M. Keck Laboratory for Biomolecular Imaging, Department of Chemistry, New York University[2, 7, 8, 11]. It has already been used in constructing restriction maps for medium-sized molecules effectively and has a potential for highly effective automated creation of restriction maps.

Here is an overview of the Optical Mapping technique. Fluorescently stained DNA molecules are elongated and attached to a glass surface so that biochemical activity is preserved. This can be achieved in a couple of ways, the most recent technique uses the fluid flows within drying droplets. The molecules are then exposed to a restriction enzyme and after digestion microscope images of cleaved molecules are taken. Restriction sites appear as gaps in the image of a molecule and fragment lengths can be computed based on fluorescent intensity of the fragments.

In the idealized experiment we would expect restriction maps of individual molecules to be almost identical, however due to various experimental imprecision there are errors in the detection of restriction sites. False negative errors, when molecules are not cleaved at all restriction sites, are mostly due to the the fact that restriction enzymes cannot cleave the DNA molecule at the places where molecule is attached to the glass surface. False negative errors can be easily eliminated by increasing the number of scanned images. It is more difficult to eliminate the false positive errors – when there is a cleavage detected not at the restriction site. It is suspected that false positive errors are mostly due to imperfection of machine vision, namely 1) misidentification of spurious data, 2) identification of multiple molecules as one, 3) identification of partial molecules as complete, 4) errors in the size estimation, 5) missing fragments.

Given the restriction maps of the individual molecules, the major computational challenge is to derive consensus locations of restriction sites. Another issue involved with the current system is that it may not produce the exact orientation information on individual molecules, i.e. the real ordering of the sites may be the reverse of what we observe. The orientation problem can be relaxed by attaching a marker to one end of the DNA molecule. This can make it easier to find a multiple alignment of restriction maps, but it still remains a chalenging algorithmic and statistical problem.

A model similar to our has been presented in [1] and its implementation is used in the Schwartz's laboratories at NYU. Our aim has been to explore certain simplifications of that model in hopes of having faster algorithms that remain reliable. For example, we only include false cuts bat not "bad" molecules as does [1]. Also we employ certain heuristics.

2 Known Orientation

Let $\theta = \theta_1, \ldots, \theta_k$, $\theta_l \in (0,1)$ be the restriction sites of the unit length DNA molecule. We assume that the number k of restriction sites is known. We have got images of M different copies of the DNA molecule, for i-th copy of the molecule we have observed m_i positions where the molecule is cleaved. We will call these positions *cut sites* and denote them $X_i = \{x_{i,1}, \cdots, x_{i,m_i}\}$, $x_{i,j} \in (0,1)$ for $i = 1, \ldots, M$ and $j = 1, \ldots, m_i$. With each $x_{i,j}$ we can associate an unobservable zero-one indicator variable $z_{i,j,l}$, where value of $z_{i,j,l}$ is one or zero depending on whether cut site $x_{i,j}$ comes as observation of the restriction site θ_l or does not. The knowledge of $z_{i,j,l}$ would allow us to estimate θ_l with

$$\hat{\theta}_l = \frac{\sum_{i=1}^{M} \sum_{j=1}^{m_i} z_{i,j,l} x_{i,j}}{\sum_{i=1}^{M} \sum_{j=1}^{m_i} z_{i,j,l}}. \tag{1}$$

We will simplify our statistical model by considering each cut site to be an independent observation. This is true for cut sites from different molecules and we believe that dependences among cut-sites within a molecule are weak enough to justify our simplification. More complex models have been studied [1] and it seems that this simplification does lead to comparable results.

Let $X = x_1, \ldots, x_n = x_{1,1}, \ldots, x_{M,m_M}$ be the collection of all cut sites and let $z_{i,l}$ be the corresponding unobservable variables. Let $Y_l = \{x_i : z_{i,l} = 1\}$ be the collection of cut sites that arise from a restriction site θ_l. We assume that each cut site from Y_l is distributed according to a normal distribution with mean θ_l and some variance σ_l^2. Unfortunately we also observe "false cut sites"

$Y_0 = \{x_i : z_{i,l} = 0, 1 \leq l \leq k\}$. We can extend the definition of $z_{i,l}$ for $l = 0$, we put $z_{i,0} = 1$ when $x_i \in Y_0$. We assume that cut sites from Y_0 are distributed according to a uniform distribution on interval $(0, 1)$. Therefore cut sites from X are distributed according to a mixture of uniform $U(0, 1)$ and k normal $N(\theta_1, \sigma_1^2), \dots, N(\theta_k, \sigma_k^2)$ distributions. The probability density function for this mixture is

$$f(x; p_0, \dots, p_k, \theta_1, \dots, \theta_k, \sigma_1^2, \dots, \sigma_k^2) = p_0 + \sum_{l=1}^{k} p_l g(x; \theta_l, \sigma_l^2)\,,$$

where $g(x; \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\theta)^2}{2\sigma^2}$ is the normal probability density function.

We will make one more simplifying assumption, we will consider only the case when mixing proportions of the normals are the same and variances are the same too. So we have $p_1 = \cdots = p_k = (1 - p_0)/k$, $\sigma_1^2, \dots, \sigma_k^2 = \sigma^2$ and the probability density function simplifies to

$$f(x; p, \theta_1, \dots, \theta_k, \sigma^2) = p + \frac{1-p}{k} \sum_{l=1}^{k} g(x; \theta_l, \sigma^2)\,.$$

Given data $X = x_1, \dots, x_n$, the best estimate of the positions of restriction sites is $\theta = \theta_1, \dots, \theta_k$ that maximizes the likelihood function

$$L(X; p, \theta, \sigma^2) = \prod_{i=1}^{n} f(x_i; p, \theta, \sigma^2)\,.$$

This is the same as maximizing the log-likelihood function

$$l(X; p, \theta, \sigma^2) = \log L(X; p, \theta, \sigma^2) = \sum_{i=1}^{n} \log f(x_i; p, \theta, \sigma^2)\,. \tag{2}$$

We use the EM (expectation-maximization) algorithm to find the maximum likelihood estimate (MLE). The EM algorithm is an iterative algorithm, in each iteration we compute a new estimate of parameters based on the estimate of parameters from the previous iteration (the question of starting values will be discussed later). It can been shown that iterative estimates of parameters obtained by the EM algorithm converge to the MLE [5, 9].

Every iteration of the EM algorithm consists of an E-step and an M-step. In the E-step we compute the estimate of unobservable data $z_{i,l}$ from the values of parameters $p, \theta_1, \dots, \theta_l, \sigma^2$ using the following expressions.

$$\begin{aligned} \hat{z}_{i,0} &= \frac{p}{f(x_i, p, \theta, \sigma^2)} \\ \hat{z}_{i,l} &= \frac{1-p}{k} \cdot \frac{g(x_i; \theta_l, \sigma^2)}{f(x_i, p, \theta, \sigma^2)}\,, \quad 1 \leq l \leq k\,. \end{aligned}$$

Note that while the indicator variables $z_{i,l}$ can have only zero-one values, the estimate $\hat{z}_{i,l}$ is the conditional probability that observation x_i belong to the l-th component and can have any value from $[0, 1]$.

In the M-step the new estimates of the parameters are computed from $\hat{z}_{i,l}$. The estimate for θ_l is similar to (1), we have

$$\hat{\theta}_l = \frac{\sum\limits_{i=1}^{n} \hat{z}_{i,l} x_i}{\sum\limits_{i=1}^{n} \hat{z}_{i,j,l}}\,.$$

The estimate for p is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{i,0}$$

and for σ^2 we have

$$\hat{\sigma}^2 = \frac{\sum_{l=1}^{k}\sum_{i=1}^{n} \hat{z}_{i,l}(x_i - \hat{\theta}_l)^2}{\sum_{l=1}^{k}\sum_{i=1}^{n} \hat{z}_{i,l}} .$$

Equations for $\hat{p}, \hat{\theta}, \hat{\sigma}^2$ can be justified in the sense that if p, θ, σ^2 are convergence points such that $\hat{p}, \hat{\theta}, \hat{\sigma}^2 = p, \theta, \sigma^2$, then $\partial l / \partial \xi = 0$, for $\xi = p, \theta_1, \ldots, \theta_k, \sigma^2$ and the convergence point is a local maximum (we can use augmentation technique to avoid being stuck in a stationary point thats not a maximum).

3 Unknown Orientation

In the case of unknown orientation we can introduce a new set of unobserved (rather that unobservable) variables f_i, $i = 1, \ldots, M$. We set $f_i = 1$ when orientation of X_i corresponds to the orientation of Y and $f_i = -1$ when orientation of $X_i^{-1} = \{1 - x_{i,m_i}, \cdots, 1 - x_{i,1}\}$ corresponds to orientation of Y.

There are 2^M possible choices for orientation of molecules, however we can incorporate orientation variables into the likelihood model in such a way that the orientation question for each molecule can be decided independently.

Let $l(X_i^{f_i}; p, \theta, \sigma^2)$ be the contribution of molecule X_i to the log-likelihood function,

$$l(X_i^{f_i}; p, \theta, \sigma^2) = \begin{cases} \sum_{j=1}^{m_i} \log f(x_{i,j}; p, \theta, \sigma^2) & \text{if } f_i = 1, \\ \sum_{j=1}^{m_i} \log f(1 - x_{i,j}; p, \theta, \sigma^2) & \text{if } f_i = -1. \end{cases}$$

For given $\mathbf{f} = f_1, \ldots, f_M$ the log-likelihood function (2) has form

$$l(X, \mathbf{f}; p, \theta, \sigma^2) = \sum_{i=1}^{M} l(X_i^{f_i}; p, \theta, \sigma^2)$$

and the MLE in this setting is the set parameters p, θ, σ^2 that maximizes

$$\begin{aligned} l(X; p, \theta, \sigma^2) &= \max\{l(X, \mathbf{f}; p, \theta, \sigma^2) : \mathbf{f} \in \{-1, 1\}^M\} \\ &= \sum_{i=1}^{M} \max\{l(X_i; p, \theta, \sigma^2), l(X_i^{-1}; p, \theta, \sigma^2)\} . \end{aligned}$$

We will extend E-step of EM-algorithm to estimate the orientation of molecules. Given p, θ, σ^2 we set $\hat{f}_i = 1$ if $l(X_i; p, \theta, \sigma^2) \geq l(X_i^{-1}; p, \theta, \sigma^2)$ and $\hat{f}_1 = -1$ otherwise. Clearly $\hat{\mathbf{f}} = \hat{f}_1, \ldots, \hat{f}_M$ is the orientation of molecules that for given p, θ, σ^2 maximizes $l(X, \mathbf{f}; p, \theta, \sigma^2)$. We then compute $\hat{\mathbf{z}}, \hat{p}, \hat{\theta}, \hat{\sigma}^2$ assuming the orientation of molecules is given by $\hat{\mathbf{f}}$.

Another approach we have tried was to incorporate unobservable f_i in the same manner as $z_{i,l}$ and to compute conditional probabilities $\hat{f}_i$ of the orientations in the E-step. This required to evaluate two sets of $z_{i,l}$, for each of two possible orientations. In the M-step the new parameter estimates are then computed based on the values of conditional probabilities f_i and $z_{i,l}$. Unfortunately this approach did not lead to satisfactory results.

4 Initial Values of Parameters and Independent Flipping

The major drawback of the EM-algorithms is the dependence of the outcome on the initial values of parameters. This is the consequence of the multimodality of the likelihood function. We are searching for the global maximum, but the EM-algorithm is only able to find a local maxima. A straitforward but time consuming approach is described in [1], to generate many starting points and to use the maximizing procedure on the most promising starting points.

We describe a heuristic approach, which allows us to find an orientation of molecules without the knoledge of the estimates $\hat{p}, \hat{\theta}, \hat{\sigma}^2$ and even without the knoledge of the number of restriction sites k. Our heuristic is based on the "voting (majority)" principle – the decision whether two molecules have the same orientation or do not is based on how these two molecules compare to all remaining molecules.

First, every two molecules X_i, X_j are assigned *orientation score* $os_{i,j}$ expressing how likely the molecules are to have the same orientation. To get the orientation score we investigate how well X_i aligns with X_j and X_j^{-1}. The problem of aligning restriction maps is discussed in [6]. Here We use very simple scoring scheme. For two aligned cut sites $x_i \in X_i$ and $x_j \in X_j$ the score is $1 - |x_i - x_j|/w$ (w is a fixed parameter, only cut sites within distance w are considered aligned). The score of the alignment is the sum of scores of aligned pairs. The alignment score $as(X_i, X_j)$ is the score of the highest scoring alignment.

We define orientation score by

$$os_{i,j} = os(X_i, X_j) = \frac{as(X_i, X_j)}{as(X_i, X_j) + as(X_i, X_j^{-1})},$$

where we set $os_{i,j} = 1/2$ when $as(X_i, X_j) + as(X_i, X_j^{-1}) = 0$. The orientation score $os_{i,j}$ can be seen as an estimate of $\mathbf{1}(f_i = f_j)$ (for Boolean expression E the indicator value $\mathbf{1}(E)$ is 1 when E is true and 0 otherwise). Our aim is to find the orientation $\mathbf{f}$ of molecules such that $|\mathbf{1}(f_i = f_j) - os_{i,j}|$ is minimal[1].

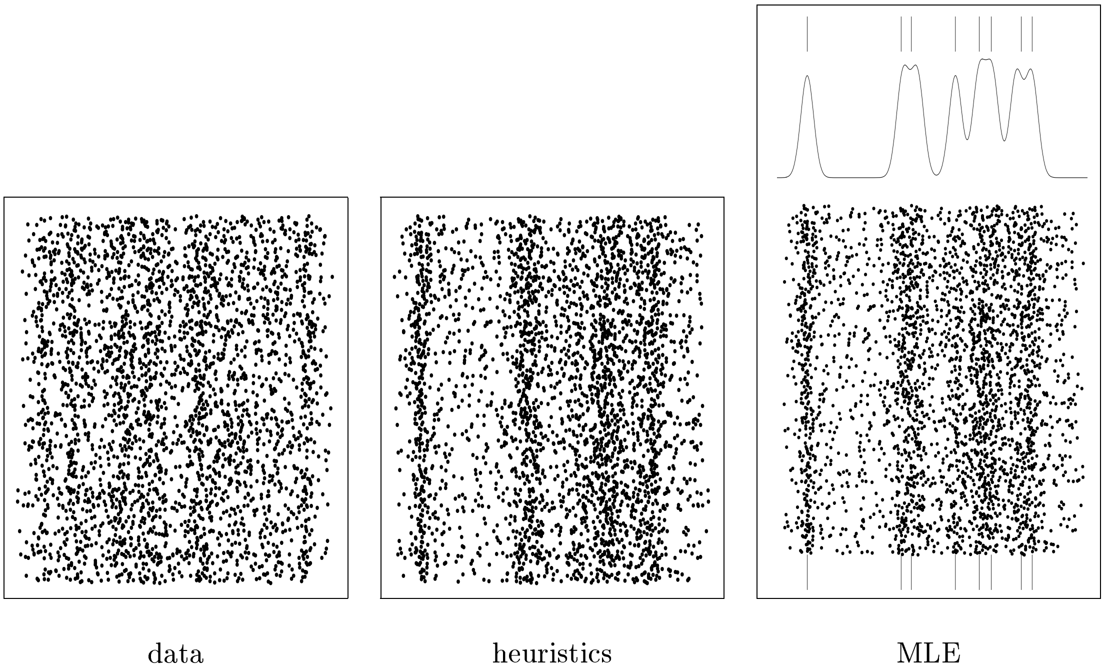
Consider two molecules X_i and X_j, the cut sites of X_i can correspond to different restriction sites then do the cut sites of X_j thus making score $os_{i,j}$ small even if $f_i = f_j$ (or having $os_{i,j}$ large when $f_i \neq f_j$). We can avoid this by looking at two corresponding rows of orientation scores $os(i) = os_{i,l}$ and $os(j) = os_{j,l}$, $1 \le l \le M$. If X_i and X_j have the same orientation, we should see some agreement between rows $os(i)$ and $os(j)$, and on the contrary, if X_i and X_j have the different orientation, we should see some disagreement between rows $os(i)$ and $os(j)$. In general, we consider two values from (0,1) to agree when they are either both larger than 0.5 or both smaller that 0.5. The new estimate $\widehat{os}_{i,j}$ thus is

$$\widehat{os}_{i,j} = \frac{1}{M} \sum_{l=1}^{M} \mathbf{1}((os_{i,l} > 0.5 \wedge os_{j,l} > 0.5) \vee (os_{i,l} < 0.5 \wedge os_{j,l} < 0.5)).$$

We can iterate this process and get more and more accurate estimates. We continue till convergence to a (zero-one) matrix F is achieved. Vector f such that $F = \{\mathbf{1}(f_i = f_j)\}$ specifies the orientation of molecules. The outcome of this heuristic is dependent on parameter w, when w is selected very small then most orientation scores are 1/2, the resulting matrix consists of all 1's. We will not consider such values of parameter w. For some data sets, especially when resulting restriction map is quite symmetric, we do not get 0-1 matrix as the output of the heuristic. Than we have more then one answer to the orientation problem. Also in such case we can try to find w that yields 0-1 matrix. We have observed, that the most appropriate parameter is the smallest w that yields to 0-1 matrix.

Even if the orientation of molecules is known (specified), the EM-algorithm remains sensitive to the initial values of the parameters, however to a much less extent. Again we can generate many sets of starting values and continue from the most promising points. However, having specified the orientation and knowing the number of restriction sites we can use following simple heuristic. We order the observed cut sites and divide them according to their order into k bins, each containing n/k

[1] This problem can be shown to be NP-hard by reduction from the EBFC problem ([3, 10])



data heuristics MLE

Figure 1: λ DNA, *Ava* I Enzyme.

cut-sites. If we assume that the digestion rate for every restriction site is the same, we can assume that the cut sites in the l-th bin are mostly cut sites corresponding to the l-th restriction site. Therefore a good starting value for θ_l might be the average value of cut sites in the l-th bin. And the initial value for σ^2 would be the average variance.

There are two factors that make the described heuristic for starting point imprecise. We do not have an appropriate starting value for p and the false cut sites artificially increase the variance. Therefore we include one extra bin (say bin 0), in which we will capture cut sites that appear to be false cuts. To specify the potential false cut sites we use the near neighbor technique [12]. For each point x_i we determine the distance $d_m(x_i)$ between x_i and its m-th nearest neighbor, i.e. $d_m(x_i)$ is such that $|\{x \in X : |x_i - x| < d_m(x_i)\}| < m$ and $|\{x \in X : |x_i - x| \leq d_m(x_i)\}| > m$. Naturally the cut sites in the dense populated areas have small m-th nearest neighbor distance and vice versa. Therefore we can expect cut sites with large $d_m(x_i)$ to be false cuts. Given threshold t we put all cut sites x_i with $d_m(x_i) > t$ into bin 0, order the remaining cut sites and distribute them into bins $1, \ldots, k$. The initial value for p then is the size of bin 0 divided by n, the initial value for θ_l is the average value of cut sites in the l-th bin and the initial value for σ^2 is the average variance.

5 Experimental Results

We have implemented the EM algorithm for maximum likelihood estimate and the heuristics for orientations and starting point. The algorithms will be accessible through the USC Computational Biology server "`http://www-hto.usc.edu/software/`". The performance of the algorithms is shown in Fig. 1 – 3. Data we used were provided by D. Schwartz, Laboratory for Biomolecular Imaging, Department of Chemistry, NYU.

The first columns show data as we obtained it. The second columns show the outcome of the orientation heuristic. The third columns show the outcome of the EM-algorithm. The vertical bars are actual real restriction sites obtained from the sequence of the DNA. Also there is shown the density

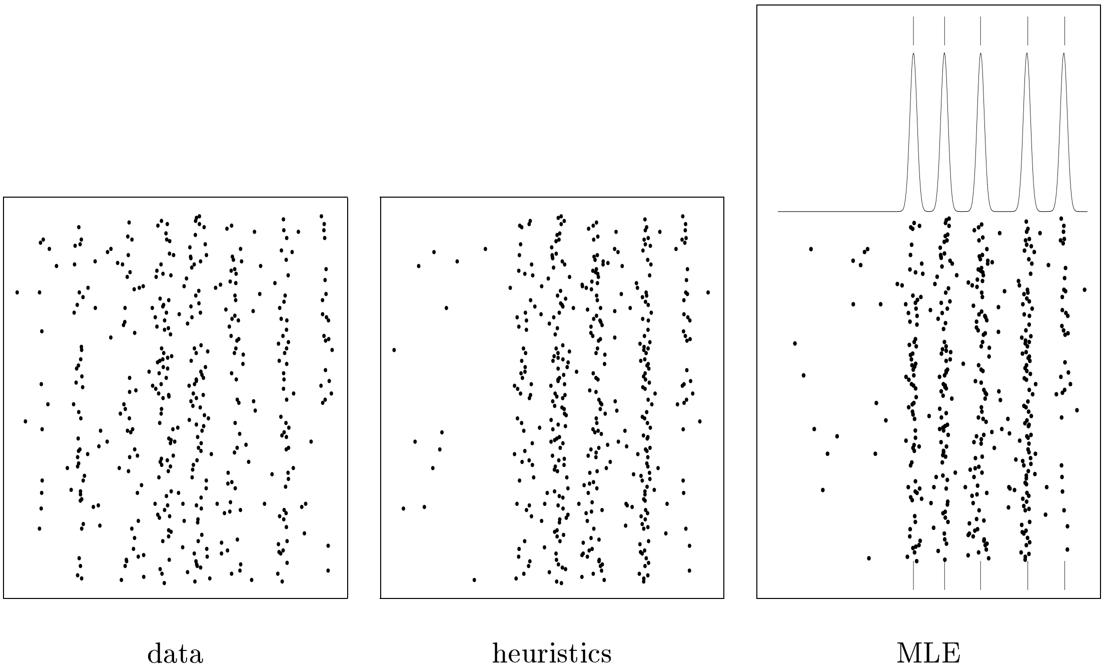


Figure 2: λ DNA, *EcoR* I Enzyme.

function corresponding to the maximum likelihood estimate of the parameters.

6 Conclusion

We have described a simple maximum likelihood approach for solving the multiple restriction map alignment problem from Optical Mapping. The major drawback of the maximum likelihood methods is the dependence of the outcome on the starting point, caused by the multimodality of the likelihood surface. To overcome this obstacle we have designed heuristic algorithms to find plausible orientations of the molecules and to suggest appropriate initial values for the parameters. Unfortunately, these techniques are not able to specify the number of restriction sites and we plan to use more sophisticated approaches for this problem [4].

Acknowledgments

The authors wish to thank David Schwartz for the motivation and for providing us with data. We also thank Bud Mishra, Larry Goldstein, Sridhar Hannenhalli, George Komatsoulis and Jae K. Lee for helpful discussions. This work is supported by NIH grant GM36230.

References

[1] T. Ananthraman, B. Mishra B, D. Schwartz. Genomics via optical mapping II: Ordered restriction maps, *Journal of Computational Biology*, 4, 91–118, 1997.

[2] W. Cai, H. Aburatani, D. Housman, Y. Wang, D. C. Schwartz. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces, *Proc. Nat. Acad. Sci.*, 92, 5164-5168, 1995.

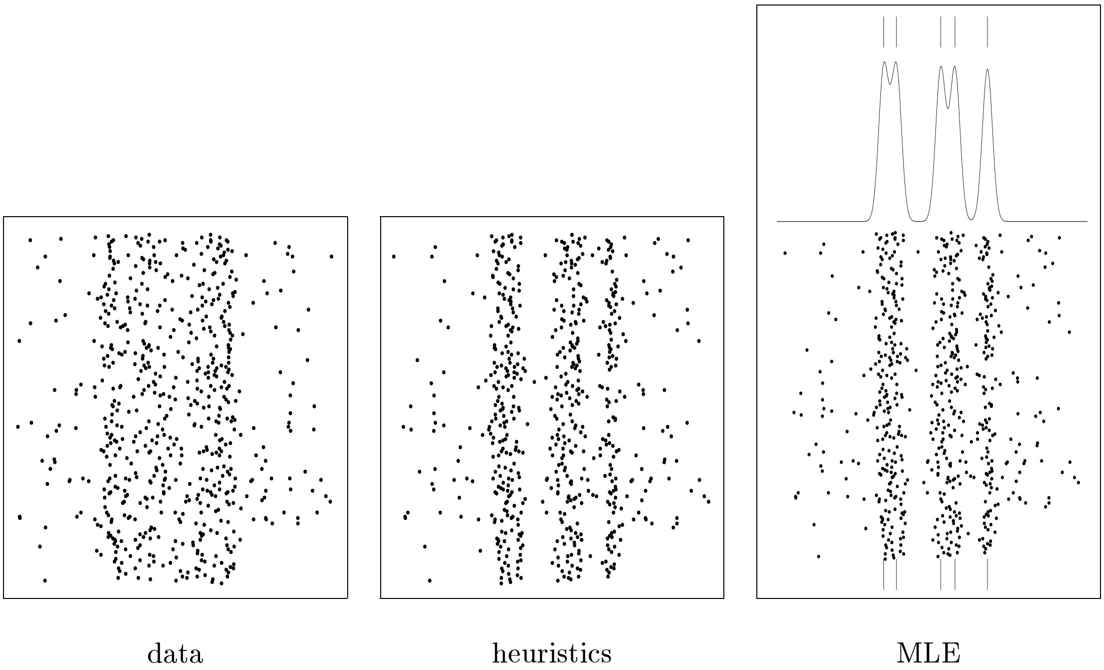


Figure 3: λ DNA, *Sca* I Enzyme.

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

[4] V. Dančík, J. K. Lee, and Michael S. Waterman. Estimation for restriction sites observed by optical mapping using Markov Chain Monte Carlo. *Unpublished manuscript.*

[5] B. S. Everitt and D. J. Hand. Finite Mixture Distributions. Chapman and Hall, London 1981.

[6] X. Huang, M. S. Waterman. Dynamic programming algorithms for restriction map comparison. *CABIOS* 8, 511-520, 1992.

[7] Y. K. Wang, E. J. Huff, D. C. Schwartz. Optical mapping of site-directed cleavages on single DNA molecules by the RecA-assisted restriction endonuclease technique, *Proc. Nat. Acad. Sci.*, 92, 165-169, January 1995.

[8] X. Meng, K. Benson, K. Chada, E. Huff, D. C. Schwartz. Optical mapping of lambda bacteriophage clones using restriction endonucleases. *Nature Genetics*, 9, 432-438, April 1995.

[9] G. J. McLachlan and T. Krishnan. The EM Algorithm and Extensions. John Wiley & Sons, New York 1997.

[10] S. Muthukrishnan and L. Parida. On constructing physical maps by optical mapping: A simple, highly effective, combinatorial approach. *Proc. of the First ACM Conference on Computational Molecular Biology* (RECOMB), 209–219, Santa Fe, January 1997.

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

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