Towards Constructing Physical Maps by Optical Mapping: An Effective, Simple, Combinatorial Approach

(Extended Abstract)

S. Muthukrishnan* Laxmi Parida†

Abstract

We initiate the complexity study of physical mapping with the emerging technology of Optical Mapping (OM) pioneered by the team lead by David Schwartz at the W. M. Keck Laboratory for Biomolecular Imaging, Dept of Chemistry, NYU. In currently popular electrophoretic approaches, information about the relative ordering of the fragments comprising the DNA molecule is lost, thus leading to difficult computational problems of composing the fragments in to a physical map depicting their relative order. In contrast, the relative ordering of the pieces is readily obtained in OM. However, OM faces serious technological challenges as it has low resolution and is fault-prone.

We take a combinatorial approach to the problem of constructing physical maps from the erroneous data generated by OM. We identify two abstract problems in this context, namely, the Exclusive Binary Flip-Cut and Exclusive Weighted Flip-Cut problems. For both, we present polynomial time approximation schemes. However, our main contribution here is an extremely simple heuristic algorithm that rapidly and accurately (with in 3% error) constructs the physical map from input data with immense experimental errors and imprecision (even with only 10% expression of a restriction site in the molecules).

Our strong experimental results, while being preliminary, seem to indicate that although OM has immense experimental imprecision, the errors appear to be "local" and hence more easily manageable than the ones in other approaches where the errors appear "global". Also, although OM may not be suitable for producing physical maps at the resolution of few base pairs, our results indicate that it may be appropriate for rapidly generating accurate physical maps at the resolution of a few 100's of base pairs.

*Bell Labs, Lucent Technologies. Work partly done while at Dept. of Computer Science, Univ. of Warwick, UK, and while visiting DIMACS partly supported by the NATO grant CRG 960215.

†Dept. of Computer Science, NYU, USA, parida@cs.nyu.edu.

RECOMB 97, Santa Fe New Mexico USA

1 Introduction

A step towards the ultimate goal of many efforts in Molecular Biology (including the Human Genome Project), namely to determine the entire sequence of Human DNA and to extract the genetic information from it, is to build *physical maps* of portions of the DNA [9, 5]. A physical map merely specifies the location of some identifiable markers (restriction sites of up to 20 base pairs) along a DNA molecule. Physical maps provide useful information about the arrangement of the DNA, and they serve as recognizable posts to help search it. In this paper, we propose and study the complexity of a combinatorial approach to constructing physical maps of medium sized molecules (20K - 40K base pairs long) using an emerging technology, called *Optical Mapping* [15].

There are several known technological approaches to building physical maps with their associated computational problems [16, 1, 10, 12, 13, 7]; most of these use restriction enzymes. A *restriction enzyme* is an enzyme that recognizes a unique sequence of nucleotides and it cleaves every occurrence (called a *restriction site*) of that sequence in a DNA molecule. In a well-established approach to physical mapping, a restriction enzyme is applied to cleave the molecule at these restriction sites producing pieces of the molecule. In this process, the information about their relative positioning is lost. Thus we are faced with the problem of assembling these pieces into their relative order: this leads to difficult combinatorial and computational problems (such as the *partial digest problem*, *probed-partial digest problem* etc.) most of which are NP-hard, and many of which have been exten-

sively studied from the point of workable heuristics (See [9, 7, 1] etc. and Section 3 of [14] for several open problems in this area).

An alternative approach to physical mapping is based on a new technology pioneered by David Schwartz at the W. M. Keck Laboratory for Biomolecular Imaging, Dept. of Chemistry, NYU, called the *Optical Mapping* (OM) technology [8, 11, 15]. At a very high level, here is an overview of that method. Single strand of a DNA molecule is attached to the surface of a slide by electrostatic forces. Then it is treated in a controlled manner with a restriction enzyme. The molecule still remains attached to the slide although the restriction sites get digested by the enzyme. Now by applying appropriate fluorescent dyes, the molecule may be viewed under a microscope or recorded by a camera as an image on a Computer. For a more detailed description of this complex process, see [8, 11, 15].

As it is clear from our overview of OM, the relative order of the pieces is not lost. In fact, the image itself is a physical map (although perhaps not at desirable levels of resolution, and not in a form compatible with genomic data we handle now). In this sense, this technology seems to cut through the Gordian Knot of physical mapping[1] described above faced by current technologies, such as gel electrophoresis. However, OM too faces severe difficulties: at the core, the technological process is highly error-prone. Some such issues are: (*i*) poor digestion of restriction sites and physical factors such as the coiling of DNA and fragments getting washed away, leading to high rate of false negatives and false positives, (*ii*) noise and lack of precision in capturing and processing images, and (*iii*) crude measures of parameters such as intensity, length etc.. Thus the problem of physical mapping is not immediately solved by OM. Nevertheless, it is a promising technology that is being made more robust (See the second generation versions in [11]).

In this paper, we consider the computational problem of constructing physical maps from the OM technology. For exposition in this section, consider the following idealized version of the problem. The image processing software, after analyzing the image obtained from OM, generates a discretized binary string of the molecule indicating the presence of restriction sites along it. This resolution is not at the level of base pairs (bps)[2]. If the technology were perfect, that will suffice as a physical map (modulo the resolution). However, because of poor digestion rates, not all sites are represented in that string. In order to get all the sites, several experiments (100's) are done on the same molecule (but with different sample molecules) and the same restriction enzyme. Thus the restriction sites will be those obtained by *consensus* from these experiments. Of course now there are basic technological problems getting the discretized strings with reasonably consistent alignment of the string positions. However, the major conceptual problem is that the different samples are not necessarily laid down along the same direction on the slide. Specifically, each sample is laid down along one of two anti-parallel directions. There are exponentially many alignments of the string positions depending on how each sample is laid and informally the problem we study here is to decode the direction for each molecule and isolate the consensus restriction sites. Formally we study an optimization version of this problem which we call the *Binary Flip-Cut* (BFC) problem. Handling real data is considerably harder. In particular, the positioning of the restriction sites as reported by the imaging software may not be accurate. For this case, we generalize the version above to the *Weighted Flip-Cut* (WFC) problem and study that as well. (See Sections 2 and 3 for the precise definition of the problems).

Our contributions are as follows. First, we initiate the study of the computational complexity of physical mapping by OM. In particular, we take a combinatorial approach and formulate two novel problems, namely, the BFC and WFC problems. In solving these problems, we reduce them to certain dense, hard optimization versions that we call the *exclusive* BFC and *exclusive* WFC problems respectively. Our main technical contribution is theoretical, and more importantly, efficient practical results for exclusive versions of BFC and WFC problems. Our theoretical result is a strong approximation result: a polynomial time approximation scheme for them (that is, a polynomial time algorithm that for any fixed fraction ϵ, produces a solution that is at least $1 - \epsilon$ of the maximum (optimal) solution).

The bulk of what we consider our contribution comes from our simple heuristic algorithm for the exclusive BFC and WFC problems (the core of BFC and WFC problems that is hard). It is an appropriate greedy algorithm that may be viewed as doing limited backtracking; as primitives, it merely uses sorting and bookkeeping. We do not prove any thing nontrivial for this heuristic (it is a 0.5 approximation algorithm but that is trivial for exclusive BFC and WFC problems). But this algorithm is extremely accurate in predicting the direction of each sample molecule and the consensus restriction sites. See 4 for detailed descriptions and figures. To sum, we claim: *our simple heuristic, running on Sun Sparc Station 2, rapidly (< 1 min) and accurately (gross overestimate of* 3% *or* 1000 *bps error) computes the physical map of medium sized molecules (40K) from real data with immense experimental and image processing error (most restriction sites having only* 10% *expression in molecules)*[3]

[1] It was a knot tied by Gordius, king of Phrygia, held to be capable of being untied only by the future ruler of Asia; it was unceremoniously cut by Alexander the Great with his sword! Now the phrase "cut the Gordian Knot" is used to mean solving an intricate problem in a surprisingly different, highly effective manner.

[2] For a molecule of 20000 base pairs, the discretized string has usually 200 positions.

[3] As a digression, consider the compromise in the quality due to the error in our algorithm. The gross upper bound of

We discuss three further points. First, why does our simple heuristic algorithm perform as well as it does? (In contrast, for physical mapping arising from other technological approaches, sophisticated heuristics such as Lin-Kernighan heuristic, or Hamming metric TSP were used [1]). We believe the explanation lies in the strength of the OM technique. Although it measures lengths and other parameters coarsely (in contrast to say gel electrophoresis) the errors are *local* such as in boundary of the restriction sites etc (in contrast, in currently prevalent approached to physical mapping, interaction between clones far apart can affect the quality of data and therefore errors appear "global"). Therefore, local search methods such as ours will tend to work well.

Second, does our result bring any insight to OM technology? Following from the point above, perhaps it is true that although OM is more error-prone, the errors are computationally more manageable since they have a "local" nature. Also, from our experimental evidence, we believe that the data from OM can be rapidly analyzed to obtain fairly accurate physical maps although not refined to the level of bps. Combined with the potential for automating the entire process, this might be the strength of OM (as opposed to generating data for very high quality physical maps given more computational resources). David Schwartz, the pioneer of OM, expressed this intuition in a personal communication, before we began work on this problem.

Finally, how far is the goal of physical mapping by OM resolved by our work? There are several other combinatorial formulations and cost functions we can envisage. Non-combinatorial approaches (eg., probabilistic, maximum likelihood) are relevant as well, and some of these are currently under investigation [2]. It remains to be seen how these formulations and solutions compare with ours. Also, OM is an evolving technology. Therefore, new technical problems arise with changes in the laboratory procedures. In this paper, we have tackled only one version for which we obtained the real data from the lab.

Map. In Section 2, we describe our results for the BFC problem. We sketch the modifications to handle real data in Section 3 using WFC. In Section 4, we present a small sample of our experimental results with the real data.

2 The BFC problem

In this section, we consider the binary flip-cut problem (BFC) informally stated below. Given n binary molecules each with m sites, determine a subset of sites (called the *cuts*) and an assignment of *flip* or *no-flip* to each of the molecules so that the number of *consensus* cut sites is minimized; a cut site is a consensus one under an assignment of flips to the molecules if at least cn 1's line up on that site when the molecules are flipped accordingly, for some small constant parameter c. A *flip* of a molecule is its reversal. In reality, c depends on various experimental parameters such as the false positive and false negative rates, enzyme digestion rate etc. Although there is no inherent reason to look for minimizing consensus cut sites, in the absence of additional discriminatory evidence, seeking such "minimal" explanation for the input data seems suitable. Throughout the paper, the *conjugate* of column i is the column $m+1-i$; we denote the conjugate of i by $\bar{i}$.

Even though we formalized the problem combinatorially as above, in our approach to its solution we kept the spirit of the underlying problem in mind. Specifically, our approach to solving this problem is the following two step process. In the first step, called the *elimination step*, we eliminate sites and their conjugates in pairs as described below. Eliminating the sites in conjugate pairs means that positions which might map on to each other because of flips continue to be able to do so. Thus this elimination is a reduction that does not affect the optimization criteria on the remaining sites owing to the molecule flips. In the second step, we solve a more specific problem, namely the *exclusive* BFC problem which is the original BFC problem except that for each conjugate pair $i, \bar{i}$, precisely one of them may be a cut site. For a collection of molecules, this *fixes* the number of cut sites and therefore we need alternate optimization criteria for this problem. We chose the total number of 1's in the cut sites as this measure.

Therefore, formally, the exclusive BFC problem is as follows. Given n binary molecules of m sites each, determine the flip for each molecule and an assignment of either i or $\bar{i}$ as a cut (but not both) for i, $1, \leq i \leq m/2$, such that the total number of 1's in the cut sites is maximized. Note that we can assume without loss of generality that m is even since otherwise, we can remove the middle site, that is, the site $(m+1)/2$, and the problem remains unchanged.

Step 1. In the elimination step, we remove two types of sites (in conjugate pairs) from consideration. First, we remove those sites i and $\bar{i}$ that have fewer than $n\tau_p$ of 1's each; here τ_p (say, 1/50) is a parameter we set from the knowledge of the error parameters in the experimental set up. We look upon these as sites where there is no underlying cut, but some molecules display the cut owing to false positive errors. Second, we remove all those sites i and $\bar{i}$ where the *sum* of the number of 1's in them exceeds $n\tau_n$, for a parameter τ_n (say 1/10), again set from the knowledge of the error parameters in the experimental set up. We look upon these as sites where there are cuts at i and $\bar{i}$. Now we hypothesize that the remaining sites have the property that precisely one of i or its conjugate $\bar{i}$ will be a cut, and that reduces the problem to the

3% error scales to an error in placement of a restriction site of roughly 1000 base pairs (bps). This segment of ambiguity is well within the limits of current sequencing technologies, so if we needed more refined physical maps, we can do so by additional conventional sequencing guided by the output of our algorithm.

exclusive BFC problem above.

We remark that the description above is only conceptual and that implementation details differ. For instance, we do not explicitly set τ_p a priori. We consider the sites in order of decreasing "suitability" for the exclusive BFC problem and we discard trailing sites which has the effect we state above. Also, the precise values for τ's have to be carefully set to filter the two types of site. For instance, assume all the molecules are parallel and i is a cut while $\bar{i}$ is not. Then, i has several 1's as determined by the false negative errors and $\bar{i}$ has a few 1's due to false positive errors. But in reality the molecules are not all parallel and a substantial fraction of them are in a flip state. In that case, the 1's in their site i appear as 1's on $\bar{i}$ in the input. Thus a cut site might have number of 1's anywhere between the one we expect from false positive rates and that from false negative rates because of molecule flips. For this reason, we found that it was effective to set the thresholds in terms of not merely the number of 1's in the columns, but also in terms of the 1's common to a column and its conjugate.

Step 2. This is the technical crux. We make some observations about the structure in the exclusive BFC problem.

Theorem 1 *Given an assignment of flips to the molecules, we can determine the assignment of cuts at i or at $\bar{i}$ (but not both) for each i, such that the number of 1's in the consensus cuts is maximum in $O(nm)$ time in all. Similarly, given an assignment of exclusive consensus cuts amongst i and $\bar{i}$ for each i, $1 \leq i \leq n/2$, we can determine an assignment of flips to the molecules, so that the number of 1's in the consensus cuts is maximum, again in $O(nm)$ time.*

We omit the algorithms for both the parts of the theorem above; in both cases, simple greedy approach works. That theorem is useful since any solution we find for the exclusive BFC problem may be postprocessed by retaining one of the two sets of answers (namely the flip assignment or the cut assignment) and optimizing for the other on that basis and thereby hope to improve local non-optimal solutions.

In what follows, we present a theoretical approximation algorithm for the problem, and a simple heuristic that is highly effective in practice.

2.1 Exclusive BFC Problem: Theoretical Solution

Here we provide a polynomial time approximation scheme (PTAS) for the exclusive binary flip-cut problem. For simplicity, we consider only the case $n = O(m)$ and leave the general scenario for the final version.

Theorem 2 *For any fixed $\epsilon < 1$, there is a polynomial time algorithm that finds flips and cuts for the exclusive BFC problem with total weight at least $1 - \epsilon$ of the maximum weight.*

Proof. We only show the sketch. We formulate a quadratic optimization problem from the given flip-cut problem. Let Y_i be the indicator variable for site i, $1 \leq i \leq m/2$. Then $Y_i = 1$ if i is a cut and it is 0 otherwise (that is, $\bar{i}$ is a cut). Let X_i be the indicator variable for the molecule i. Then $X_i = 1$ implies it appears as-is, that is, without being flipped; $X_i = 0$ implies it is flipped from the input. Also, M_{ij} is the site j in molecule i and $\overline{M}_{ij} = M_{i(n-j)}$. The quadratic optimization problem is:

$$\max \sum_{i=1}^{i=n} \sum_{j=1}^{j=n/2} \begin{array}{l} Y_j(M_{ij}X_i + (1 - X_i)\overline{M}_{ij}) + \\ (1 - Y_j)(X_i\overline{M}_{ij} + (1 - X_i)M_{ij}) \end{array}$$

$$Y_i = 0, 1; \quad X_i = 0, 1$$

Rearranging terms, the objective function becomes:

$$\max \sum_{i=1}^{i=n} \sum_{j=1}^{j=n/2} \begin{array}{l} \left(Y_jX_i(M_{ij} - \overline{M}_{ij}) + Y_j\overline{M}_{ij}\right) \\ -Y_jX_i(\overline{M}_{ij} - M_{ij}) - Y_jM_{ij} \\ +X_i(\overline{M}_{ij} - M_{ij}) + M_{ij} \end{array}$$

Collecting terms, the objective function becomes:

$$\max \sum_{i=1}^{i=n} \sum_{j=1}^{j=n/2} \begin{array}{l} 2Y_jX_i(M_{ij} - \overline{M}_{ij}) \\ +Y_j(\overline{M}_{ij} - M_{ij}) \\ +X_i(\overline{M}_{ij} - M_{ij}) + M_{ij} \end{array}$$

Let W^* be the maximum solution for the above. We claim without providing the proof that we can now use recent techniques due to Arora et al [4] to conclude the following: for every $\epsilon < 1$, there is a polynomial time algorithm which solves the above and returns a solution W such that $W \geq W^* - \epsilon n^2$.

We also claim a lower bound on W^*, namely $W^* = \Omega(n^2)$. This is because, consider the flip which provides the optimal value W^*. For each i, clearly we can choose the column i or its conjugate whichever has more 1's than the other. That way $W^* \geq \frac{n}{2}\frac{n\tau_p}{2}$ where the first term on the right is the number of cuts and the second term is the lower bound on the average of the number of 1's in a conjugate pair of cuts. Since τ_p is a constant, it follows that $W^* = \Omega(n^2)$.

Combining both, we get a PTAS for the exclusive BFC problem. □

2.2 Exclusive BFC Problem: Practical Solution

In this section, we describe an extremely simple algorithm for the exclusive BFC problem. We do not guarantee any approximate or exact performance for this algorithm (except that it is a 1/2 approximation, but that is trivial). However as our experimental results show, this algorithm is remarkably accurate in predicting the consensus cuts on both synthetic data and on real data.

In what follows, we specify the main ingredients of the algorithm. Our algorithm resembles a greedy algorithm at the high level. However, there are two orthogonal manners to be greedy about (namely by fixing the consensus cuts or by fixing molecule flips). We try to attain as many 1's as possible in candidate

cut sites by greedily flipping the molecules appropriately to the extent we can. However, as we accumulate cut sites, existing molecule orders hinder procuring additional potential 1's. In that case, we reverse the flips of the molecules selectively before proceeding further. This may be thought of as limited backtracking on the choice of molecule flips. Although one can envisage situations where more levels of backtracking will improve the solution (we can construct such data sets easily), our experimental experience suggests that such a limited backtracking suffices. We also experimented with incorporating limited backtracking on the cut sites, and although there are cases where it improves the performance, our strong experimental intuition is that it is not crucial.

In what follows, we only provide the sketch of the algorithm. We experimented with a number of variations of this algorithm differing in details and the experimental results based on those variations and their comparisons will be presented in the full version of this paper. In this paper, we present experimental results based only on an implementation that is closely related to the description below.

Short Algorithm Sketch. We first calculate for each site, its potential for getting 1's. That is, for each site i, $1 \leq i \leq n/2$, we calculate C_i, the maximum number of 1's that can be made to align at that site by flipping the molecules. That is, $C_i = |\{j | M_{ij} = 1 \; OR \; \overline{M}_{ij} = 1\}|$. Then we consider sites in the decreasing order of their potential. For this, we sort the C_i's in decreasing order and process them in that order, for each deciding whether i or its conjugate $\bar{i}$ should be designated as a cut site. Say we have processed j of these. For each molecule i, we keep dif_i, which is the number of cut sites it has 1 in, minus the number of the conjugates of these cut sites it has a 1 in. For a subset of the molecules, we would have assigned flip directions (called *touched* molecules) and others are *untouched*. All molecules with flip directions i will satisfy $dif_i \geq 0$, and all untouched molecules have $dif = 0$.

Now we show how we add the $j+1$th of the sites in the decreasing sorted order. Let this be site k. We calculate the number of 1's that fall in site k from the touched molecules (or if it falls in site $\bar{k}$ for a touched molecule that has $dif_i = 0$ since we can flip that molecule without affecting the solution thus far), plus the number of 1's obtained by assigning flip directions to relevant untouched molecules to get as many 1's as possible in site k. This gives a count for k. Calculate the same quantity for its conjugate $\bar{k}$. Then, whichever has the higher count will be designated as the cut site. Note that this is a greedy decision for the sites subject to best possible flipping of flexible molecules. When the cut site is decided to k or $\bar{k}$, the dif values of the molecules are updated. For any molecule i for which $dif_i < 0$, its flip direction is reversed and its new dif value is computed. This is the limited backtracking on the molecules alluded to earlier.

That completes the description of the algorithm. It is easy to see that the whole algorithm works in time linear in the input size, that is $O(nm)$, and uses $O(nm)$ space in all.

We experimented extensively with this algorithm on simulated data and obtained extremely positive results on predicting the restriction sites. However, the real data is considerably different from the exclusive BFC case and therefore we do not present the experimental results with the simulated data in this writeup. Instead we focus on real data.

3 The WFC problem

Here we consider the problem of dealing with the real data. As mentioned earlier, the input to us is a binary $n \times m$ matrix. However a critical problem is that the input is "fuzzy", that is, it does not depict the location of restriction sites accurately because of the error inherent in measuring the lengths of fragments that remain after digestion by the restriction enzyme. Specifically, a 1 at some site in the molecule might in fact signal a restriction site in one of its neighbors. This fuzziness is the result of coarse resolution and discretization, other experimental errors, or errors in preprocessing the data prior to constructing physical maps such as in the image processing phase.

We tackle this problem as follows: we fix a window width w as a threshold parameter. If $M_{ij} = 0$, we locate the closest k to j where $\delta = |k - j| \leq w$, such that $M_{ik} = 1$ (if it exists), and set $M_{ij} = e^{-\frac{\delta^2}{(w/2)^2}}$. If such a k does not exist, M_{ij} is unaffected. (The underlying assumption we make is that the position of a cut is normally distributed around its true site with a standard deviation of $w/2$.) This results in entries that are not necessarily binary; they have values between 0 and 1 at each position. Next we sample the sites for potential cut sites. For this, we compute $s_j = \sum_i \max(M_{ij}, \overline{M}_{ij})$ and detect local maxima of $s(j)$, $1 \leq j \leq n/2$. These positions, and their conjugates, are selected as the potential cut sites and the remaining columns are discarded. That leaves us with the problem of finding the flip-cut as before except that now the entries are rationals between 0 and 1. We call this generalization of the BFC problem to the weighted case as the *Weighted Flip-Cut* (WFC) Problem.

In our programs, we solve the WFC problem. Just as in Section 2, we can perform eliminations to derive the exclusive WFC problem and the solutions there (both the practical one and the theoretical one) can be extended to the weighted case with some modifications. The practical algorithm now takes time $O(nm + n \log n)$. We omit all details.

4 Experiments

All the real data we used for our experiments were obtained from the W. M. Keck Laboratory for Biomolecular Imaging of the Department of Chemistry, New York University. The input to our problem from the laboratory is a set of molecules, each having unit length, with the positions of their restriction cuts given as lengths from the left end. We discretize this to 200 units[4]. All the DNA molecules in our experiment shown here are lambda vectors having 48,000 bps. Thus, each discretization unit here represents about 240 bps. The restriction enzymes used were *AvaI* and *EcoRI*. In all the experiments carried out, including the one with about 3000 molecules, the program takes less than a minute to compute the restriction map. Here we have shown a sample of four experimental results on real data.

For exposition, consider Figure 1.

1. The image on the left is the input data and the one on the right is the output of our algorithm. Each row of the image is a molecule with a white dot indicating a restriction site. The output image shows the molecules flipped as per the solution computed by the algorithm, with the computed cuts marked on the top of the image by tiny bars. In some cases, the output image is sparser than the input image since we do not display the molecules with $dif = 0$. (For explanation, see Section 2).

2. At the bottom of the two images, we display a smaller image which has two rows of bars: the top row shows the true position of cuts as provided by the laboratory and the second gives the positions computed by our algorithm.

3. The table below (2) shows the position of the true and the computed cuts in $1 - 200$ scale.

4. The table below (3) displays statistics on the distribution of the cuts in the input relative to the computed ones. The standard deviation of the displacement of a restriction site of a molecule from the computed cut site is shown, as is the expression of a cut site in the molecules (the number of molecules that have a 1 in that column) in percentage. □

We compare the map determined by our algorithm with the true map as follows. We define a one-to-one correspondence between the restriction sites in both maintaining a left-to-right order. The number of the restriction sites in both must match. The MAX ABS error is the maximum of the absolute distance of a cut site in the true map from its corresponding cut in the computed map. In our experiments, the MAX ABS percentage error is $0 - 3\%$, and we never missed a cut or found an extra cut. Also, there have been input instances (eg., one with a sample of 2910 molecules) with 0% error!

[4] A larger discretization does not give higher accuracy, as the precision of the imaging and the pre-processing stages is limited. This number of 200 was chosen after discussing with the researchers from the laboratory.

It is worth noting that Figures 2 and 3 represent two experiments for the same DNA-restriction enzyme pair. The result of one the experiments (Figure 3) is noisier than the other (Figure 2), yet our algorithm works robustly.

5 Concluding remarks

We have initiated the study of complexity of physical mapping with the emerging technology of OM. Our approach is combinatorial, and we identified two problems, namely, the exclusive BFC and WFC problems. For both, we presented polynomial time approximation schemes. However, our main contribution is a simple heuristic algorithm that rapidly and accurately (with in 3% error) constructs the physical map from data with immense experimental errors and imprecision (even with only 10% expression of a restriction site in the molecules).

Recently the exclusive BFC problem was proved to be NP-hard [6], thus our polynomial approximation scheme is of interest. Are other cost functions appropriate in BFC and WFC problems? We can design PTAS algorithms for these problems with many different cost functions using the technique in Theorem 1.

Our strong experimental results seem to indicate that although OM has immense experimental imprecision, the errors appear to be "local" and hence more easily manageable than the ones in other approaches such as gel electrophoresis where the errors appear "global". Also, our results seem to indicate that OM may be the appropriate technology for rapid construction of fairly accurate physical maps although they may not generate suitable data for very fine physical maps. Both these indications are extremely preliminary. Thorough study embracing all the different aspects of OM simultaneously is needed. We are aware of such studies underway [2], [3].

Finally, OM is an evolving technology. As chemical procedures change, new computational problems arise. For example, plans are underway to detect the flip of a molecule by a suitable chemical modification to the procedure in the lab. In the absence of data, the effectiveness of this change in general is unknown. Of further interest is the case when some of the fragments are missing; this might happen despite the careful construction of the experimental set up to hold the undigested pieces firmly on the slide. We have not considered this complication.

6 Acknowledgment

We are grateful to the team lead by Dr. David Schwartz at the W. M. Keck Laboratory for Biomolecular Imaging of the Department of Chemistry, NYU, for providing us the data. Thanks in particular to Joanne Edington and Junping Jing whose data we

have used in this paper. We would also like to thank Estarose Wolfson, Ernest Lee, Alex Shenker, Brett Porter and Ed Huff whose "machine-vision" component of the system was used to obtain the data in the form used by our algorithm. Finally, sincere thanks to Thomas Ananthraman, Davi Geiger and Bud Mishra for their insightful comments and feedback.

References

[1] F. Alizadeh, R.M. Karp, D.K. Weisser, G. Zweig, *Physical Mapping of Chromosomes Using Unique Probes*, J. Comp. Bio., 2(2):153-158, 1995.

[2] T. Ananthraman, B. Mishra B, D. Schwartz. *Optical Mapping II (Restriction Maps)*, under preparation.

[3] D. Geiger, L. Parida, *A Model and Solution to the DNA Flipping String Problem*, May, TR1996-720,1996.

[4] S. Arora, D. Karger and M. Karpinski. Polynomial time approximation schemes for dense instances of NP-Hard problems. *STOC*, 1996.

[5] N. G. Cooper(editor). *The Human Genome Project - Deciphering the Blueprint of Heredity*, University Science Books, Mill Valley, California, 1994.

[6] V. Dancik and S. Hannenhalli. NP-Hardness of Exclusive Binary Flip-Cut Problem. *Manuscript*, 1996.

[7] P.W. Goldberg, M.C. Golumbic, H. Kaplan, R. Shamir, *Four Strikes Against Physical Mapping of DNA*, J. Comp. Bio., 2(1):139-152, 1995.

[8] W. Huff, D. Schwartz. *Optical Mapping of site-directed cleavages on single DNA molecules by the RecA-assisted restriction endonuclease technique*, Proc. Nat. Acad. Sci., 92, pp 165-169, January 1995.

[9] R. Karp, *Mapping the Genome: Some Combinatorial Problems Arising in Molecular Biology*, 25th ACM STOC '93-5/93/CA, USA.

[10] M. Krawczak, *Algorithms for Restriction-Site Mapping of DNA Molecules*, Proc. Nat. Acad. Sci., 85, pp 7298-7301, 1988.

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

[12] P. Pevzner. *DNA Physical Mapping*, Computer Analysis of Genetic Texts, 154-158, 1990.

[13] P. Pevzner, A. Mironov. *An Efficient Method for Physical Mapping of DNA Molecules*, Molec. Bio.. 21:788-796, 1987.

[14] P. Pevzner, M. Waterman. *Open Combinatorial Problems in Computational Molecular Biology*, Proceedings of the Third Israel Symposium on Theory of Computing and Systems, Jan 4-6, 1995, Tel Aviv, Israel.

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

[16] M. Waterman, J. R. Griggs, *Interval Graphs and Maps of DNA*, Bull. of Math. Biol., 48:189-195, 1986.

Positions (1-200)								
True Sites	38	45	64	71	84	114	120	180
Computed Sites	37	44	63	71	83	109	115	180

Cuts	Standard Deviation of displacement from computed cut	Expression of cuts (%)
1	0.140	15.8
2	0.164	11.6
3	0.177	14.4
4	0.215	8.5
5	0.225	7.4
6	0.240	7.6
7	0.198	10.6
8	0.128	15.5

Figure 1: Clone: λ DNA, Enzyme: *AvaI*. Number of molecules = 800. The MAX ABS percentage error is 2.5%. (Data produced by Junping Jing.)

Positions (scale of 1-200)					
True Sites	87	107	130	160	183
Computed Sites	88	107	131	160	184

Cuts	Standard Deviation of displacement from computed cut	Cut expression (%)
1	0.606	2.1
2	0.490	3.0
3	0.219	9.0
4	0.335	4.2
5	0.286	6.3

Figure 2: Clone: λ DNA, Enzyme: *EcoRI*. Number of molecules = 333. The MAX ABS percentage error is 0.5%. (Data produced by Joanne Edington.)

Positions (scale of 1-200)					
True Sites	87	107	130	160	183
Computed Sites	89	108	130	160	184

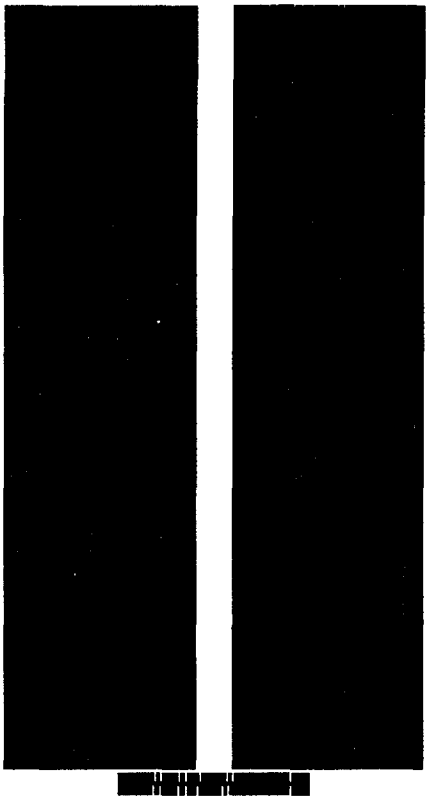
Cuts	Standard Deviation of displacement from computed cut	Cut expression (%)
1	0.436	2.7
2	0.309	4.5
3	0.243	8.4
4	0.243	7.4
5	0.193	11.9

Figure 3: Clone: λ DNA, Enzyme: *EcoRI*. Number of molecules = 403. The MAX ABS percentage error is 1.0%. Note that this input is noisier than the one shown in Figure 2, yet the algorithm is robust. (Data produced by Joanne Edington.)

Positions (scale of 1-200)					
True Sites	66	88	95	124	132
Computed Sites	65	86	94	125	132

Cuts	Standard Deviation of displacement from computed cut	Cut expression (%)
1	0.329	6.4
2	0.328	6.0
3	0.242	11.7
4	0.215	14.2
5	0.353	5.3

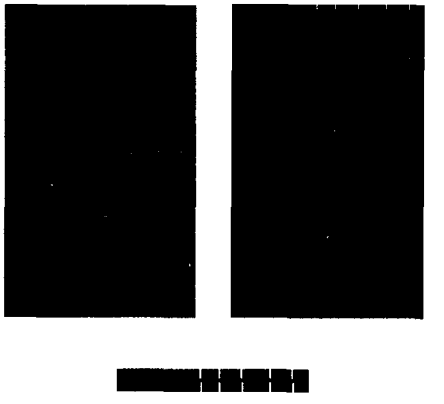
Figure 4: Clone: λ DNA, Enzyme : *ScaI*. Number of molecules = 281. The MAX ABS percentage error is 1.0%. (Data produced by Junping Jing.)



Positions (1-200)								
True Sites	38	45	64	71	84	114	120	180
Computed Sites	37	44	63	71	83	109	115	180

Cuts	Standard Deviation of displacement from computed cut	Expression of cuts (%)
1	0.140	15.8
2	0.164	11.6
3	0.177	14.4
4	0.215	8.5
5	0.225	7.4
6	0.240	7.6
7	0.198	10.6
8	0.128	15.5

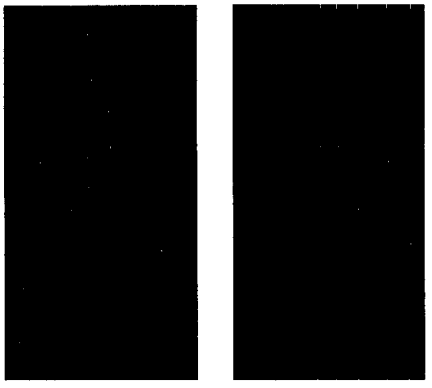
Figure 1: Clone: λ DNA, Enzyme: *AvaI*. Number of molecules = 800. The MAX ABS percentage error is 2.5%. (Data produced by Junping Jing.)



Positions (scale of 1-200)					
True Sites	87	107	130	160	183
Computed Sites	88	107	131	160	184

Cuts	Standard Deviation of displacement from computed cut	Expression of cuts (%)
1	0.606	2.1
2	0.490	3.0
3	0.219	9.0
4	0.335	4.2
5	0.286	6.3

Figure 2: Clone: λ DNA, Enzyme: *EcoRI*. Number of molecules = 333. The MAX ABS percentage error is 0.5%. (Data produced by Joanne Edington.)



Positions (scale of 1-200)					
True Sites	87	107	130	160	183
Computed Sites	89	108	130	160	184

Cuts	Standard Deviation of displacement from computed cut	Expression of cuts (%)
1	0.436	2.7
2	0.309	4.5
3	0.243	8.4
4	0.243	7.4
5	0.193	11.9

Figure 3: Clone: λ DNA, Enzyme: *EcoRI*. Number of molecules = 403. The MAX ABS percentage error is 1.0%. Note that this input is noisier than the one shown in Figure 2, yet the algorithm is robust. (Data produced by Joanne Edington.)

Positions (scale of 1-200)					
True Sites	66	88	95	124	132
Computed Sites	65	86	94	125	132

Cuts	Standard Deviation of displacement from computed cut	Expression of cuts (%)
1	0.329	6.4
2	0.328	6.0
3	0.242	11.7
4	0.215	14.2
5	0.353	5.3

Figure 4: Clone: λ DNA, Enzyme : *ScaI*. Number of molecules = 281. The MAX ABS percentage error is 1.0%. (Data produced by Junping Jing.)