A Uniform Framework for Ordered Restriction Map Problems

Laxmi Parida

August 14, 1997

Abstract

Optical Mapping is an emerging technology for constructing ordered restriction maps of DNA molecules [1, 2, 3]. The underlying computational problems for this technology have been studied and several cost functions have been proposed in recent literature. Most of these propose combinatorial models; one of them also presents a probabilistic approach. However, it is not *a priori* clear as to how these cost functions relate to one another and to the underlying problem. We present a uniform framework for the restriction map problems where each of these various models is a specific instance of the basic framework. We achieve this by identifying the following approaches to the ordered restriction map problem: (1) using data consensus or agreement, and, (2) optimizing a characteristic function of the data. Our framework also opens up the possibility of exploring other cost functions. An additional feature is that we not only integrate the combinatorial models but also analyze the probabilistic model within the same framework. Finally, we indicate the open problems by including a survey of the best known complexity results for these problems.

1 Introduction

The Human Genome Project aims to determine the entire sequence of Human DNA and to extract genetic information from it. In this context an important step is to build *restriction maps* of portions of the DNA [4]. A restriction enzyme cleaves or cuts a DNA molecule at some fixed site called the *restriction site*. An ordered restriction map specifies the location of these identifiable markers or restriction sites along a DNA molecule. A microscope-based technique called Optical Mapping [1, 2, 3] is a very promising emerging technology for rapid production of ordered restriction maps.

We present a uniform framework for the ordered restriction map problems. In this process, we identify two approaches to the ordered restriction map problem: (1) using data consensus or agreement, and, (2) optimizing a characteristic function of the data. Various computational problems proposed in [5, 6, 7, 8, 9], turn out to be instances of problems in this uniform framework. Our framework also opens up the possibility of exploring other cost functions. Also, an interesting consequence has been the analysis of a general probabilistic approach [8] in the context of this combinatorial framework. While most of these problems turn out to be inapproximable as shown in the survey (Section 6), it is important to bear in mind that the hardness proofs are for the problems in their total generality. In real life, the data arise from a well-controlled (benevolent) process. For instance the Exclusive Binary Flip Cut has been shown to be MAX SNP-hard [9], but a dense instance[1], which the real data satisfies, has a polynomial time approximation scheme (PTAS) [6].

We conclude the paper with a brief survey of the known complexity results and open problems.

Organization of the paper. In Section 2, we give an informal introduction to the problem. Section 3 discusses the data consensus/agreement approach. Section 4 discusses the characteristic alignment approach. Section 5 presents the analysis of a general probabilistic approach. Section 6 presents a survey of related work.

2 The Ordered Restriction Map Problem

We will define the problem informally as follows. Let us view this as a game played by Ann and John. John has a string S, of length n, of 0's and 1's. He makes m copies of this string and, using some process, *alters* the m copies in some controlled manner.

John assures Ann that the number of these alterations is not very large. Now, this altered set of m strings, called the *data set*, is made available to Ann and she is required to guess the original string S John started with. Ann makes a (reasonable) guess by providing an S'. The problem that Ann solves is the Ordered Restriction Map problem.

Let us now look at the (reasonable) alterations John can make.

1. False Positives: John can change some 0's to 1's in the m copies. But he must assure Ann that the number of such changes is very small.

   In practice, these may be due to actual false cuts or due to errors in the pre-processing stage.

2. False Negatives: John can change some 1's to 0's in the m copies. But he must assure Ann that the number of such changes is no more than mc_j for each column j. Note that in the absence of this restraint on John (and with False Positives), Ann will have no way of guessing a reasonable S'.

[1] A problem is dense if the number of 1's or cuts in a column is $m\gamma$, where m is the number of molecules or rows and $0 \leq \gamma \leq 1$ is a fixed.

$(1 - c_j) = p_j$ is the *digestion rate* of the experiment or the minimum number of 1's required for a column j to be designated a *consensus site* [2].

3. **Spurious Molecules:** John can throw out some, say k, molecules from this data set and throw in k random strings of 0's and 1's in its place.

   In practice, some "bad" molecules get into the sample population; these need to be invalidated and not used in the map computation.

4. **Sizing Errors:** John moves the positions of some 1's in a *small* neighborhood, that is, for some $\delta > 0$, he can move the position of a 1 in the molecule at j to anywhere between $j - \delta$ and $j + \delta$.

   This corresponds to the possible sizing errors of the fragments. The input data 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. Thus 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.

5. **Orientation Uncertainties:** John flips some of the strings: if $s = x_1 x_2 \ldots x_{n-1} x_n$ is a string with $x_i = 0$ or $1, i = 1, 2, \ldots, n$, the flipped string is $x_n x_{n-1} \ldots x_2 x_1$.

   When the molecule is laid out on a surface, the left-to-right or right-to-left order is lost. However, the orientation information may be given in the data (using a more elaborate chemical protocol) with a vector arm on one fixed side of the molecule [1]. The model can view this as a consensus cut site at one end of the map. Notwithstanding this, there is a non-zero probability of the orientation of the molecule being still unknown.

6. **Missing Fragments:** John can remove some fragments of the string (the substring between two 1's).

   This corresponds to fragments that get washed away during the experiment, which is common for BAC DNA, although not for cosmids and λDNA [8].

The correspondence of the Ann and John game to the Ordered Restriction Map problem is as follows: a string is a molecule, the length of the string corresponds to the number of sites on each molecule, the 1's in the string refer to cuts and the 0's refer to no-cuts at that site. The string S is called the **map**, and the 1's on S are the **consensus cuts**. The changes that John makes correspond to the various experimental and/or pre-processing inaccuracies that creep in at various stages.

Ann is required to produce an S' that gives an *alignment* of the molecules that optimizes a cost function. **Alignment** of the rows/molecules refers to assignment of the following:

[2] It may be noted that if the number of false positives per column j is q_j, then Ann cannot make a reasonable guess, if the following holds: $p_j + q_j \approx 1$, for any j.

1) Labeling a molecule as *spurious* or not.
2) Labeling the orientation of the molecule as *flipped* or not.
3) Assigning a *left-flushed* or *right-flushed* or any other positioning of each molecule.

A *map* is an n-length string that designates each site as a *consensus cut* site or not.

Circular Ordered Restriction Map Problem. If John take the string S and glues the two free ends producing a "seamless" ring, the corresponding problem is the circular DNA problem. In this version John makes m altered *rings* (instead of linear molecules as in the previous case) available to Ann. The seamlessness refers to Ann not having any information about where John glued the ends. The problems in the linear version also appear in the circular configuration and we do not explicitly categorize any of these in the rest of the paper.

The Cost of an alignment is a function (measure) of the alignment which we optimize. This paper explores various forms of "reasonable" cost functions. Recall that a 1 in S' at location j implies that there are at least mp_j 1's in the aligned data set in column j. For the rest of the paper let $mp_j = c_j$, that is c_j is the minimum number of 1's required in column j for it to be a consensus cut column.

3 Consensus/Agreement with data

This approach uses the mutual agreement between the molecules to obtain an alignment of the molecules and a map. There are two views to this: one uses an explicit hypothesis and the other does not. We will discuss these views in the next two sections, and show that the latter is a particular case of the former.

3.1 Consensus/Agreement with a Hypothesis

Let hypothesis $\mathcal{H}$ have K restriction sites each at location $l_k, k = 1, 2, \ldots K$. As the location of a site is not exact, assume that it has a distribution $G_j()$ about the correct location l_j in $\mathcal{H}$, with standard deviation σ_j. Further assume that given $\mathcal{H}$ and a molecule i with some fixed alignment, we can designate every cut site in the molecule as *true* or *false*. A *true* site will correspond to l_j of $\mathcal{H}$, for some j, at a distance d_j from it, and, a *false* site will have no such correspondence. For an alignment of the rows/molecules define the following:

$$T_j = \sum_{i=1}^{m} (\text{number of } \textit{true} \text{ sites in molecule } i \text{ at } l), \tag{1}$$

$$F_j = \sum_{i=1}^{m} (\text{number of } \textit{false} \text{ sites in molecule } i \text{ at } l). \tag{2}$$

Let $F = \sum_j F_j$. Then $\mathcal{M}$, the match for an alignment with hypothesis $\mathcal{H}$ is defined as

$$\mathcal{M} = \sum_j \{f_j(T_j)G_j(d_j) + g_j(F_j)\}, \tag{3}$$

and the problem is to maximize $\mathcal{M}$. $f()$ and $g()$ are "suitable" functions on the number of *true* and *false* sites respectively at a location j, depending on whether j is a *consensus cut site* or not in the hypothesis $\mathcal{H}$. An alternate form of $\mathcal{M}$ can be obtained by defining function $\tilde{g}()$ on F instead of F_j in equation (3).

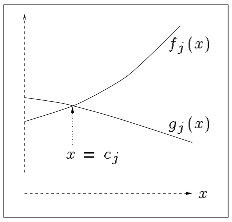
Agreement with a hypothesis uses the following optimization function:

$$\max_{(\text{over all hypotheses } \mathcal{H} \text{ \& alignments})} \left\{ \sum_{i=1}^{m} \mathcal{M}(\mathcal{H}, i) \right\}. \tag{4}$$

3.1.1 Properties of functions $f_j()$ and $g_j()$

Note that by their usage, $f_j, g, \tilde{g} : \{0, 1, 2, \ldots, m\} \to \mathbf{R}$. What must be the conditions on $f_j()$ and $g_j()$ (or $\tilde{g}_j()$) so that the "mutual agreement" of the molecules (or data) is not violated?

1. $f_j(x) < f_j(y), \quad \forall x < y, \text{ and } x, y \geq c_j.$
2. $g_j(x) \leq g_j(y), \quad \forall x > y, \text{ and } x, y \leq c_j.$
3. $g_j(x) > f_j(x), \quad \forall x < c_j,$
4. $g_j(x) < f_j(x), \quad \forall x > c_j.$



The first condition states that the "agreement" must steadily increase with increase in matches, else it violates it; the second condition has the same spirit. These conditions ensure that a consensus cut column at j has at least c_j 1's. Hence one can see that the cost functions can be "designed" using the constants c_j's (or probability p_j's) by defining appropriate $f_j()$ and $g_j()$ that satisfy the above conditions.

Let G_j have a very small $\sigma_j, \forall l$. This leades to the *idealized* version of the problem, where the location of a site is supposed to be exact; thus the molecules can be represented as a string of 0's and 1's. Thus the data can be represented in a $m \times n$ binary matrix $[M_{ij}]$ with each entry as either 0 or 1. Each row represents a molecule and each column refers to a site on the molecule: thus there are m molecules and n sites. A 1 at position (i, j) means that the j^{th} site (column) of the i^{th} molecule (row) is a cut. A 0 indicates the absence of a cut. A hypothesis is a map or a n-length vector with 1's representing a consensus cut site and 0 representing its absence. Let $S = \{p_1, p_2, \ldots, p_n\}$ be the digestion rates of the locations $j = 1, 2, \ldots, n$. The different problems that respect the "mutual agreement" criteria appear in the following sections.

3.1.2 Problem Instances

Binary Flip Cut (BFC) [5]: Given M_{ij} and the digestion rates S, find an alignment of the rows/molecules and a map that maximizes the number of 1's in the consensus cut columns (which is at least mp_j for the consensus cut column j). The alignment takes orientation uncertainties into account; incorporating the missing fragment errors gives rise to Binary Shift Cut (BSC) problem [8], [7], [9], and incorporating the good/spurious molecule error gives Binary Partition Cut (BPC) problem [8], [9].

Exclusive Binary Flip Cut (EBFC) [5, 6]: Given M_{ij} find an alignment that maximizes the number of 1's in consensus cut columns where only one of j or $\overline{j} = n - j + 1$ is a consensus cut satisfying the following condition: the column with the higher number of 1's (between j and $\overline{j}$) is the consensus cut.

The EBFC problem can alternately be defined in terms of the digestion rate, $p_j = c_j/m$. Note that this is an equivalent definition of the EBFC problem (see [9] for further explanation):

$$\begin{aligned} S_j &= |\{i | M_{ij} = 1 \text{ AND } M_{i\overline{j}} = 1\}| \\ \bar{S}_j &= |\{i | M_{ij} = 1 \text{ XOR } M_{i\overline{j}} = 1\}| \\ c_j &= S_j + \frac{\bar{S}_j}{2} \end{aligned} \tag{5}$$

Balanced BFC (BFC_B): Given M_{ij}, find an alignment of the rows/molecules and a map that maximizes the number of 1's in the consensus cut columns and the number of 0's in the columns that are not consensus cut columns. In a similar spirit as before we get the Balanced BSC (BSC_B) and the Balanced BPC (BPC_B) problems.

Conservative BFC (BFC_C): This gives a conservative evaluation of the cost per consensus cut column and is defined as follows. Given M_{ij}, find an alignment of the rows/molecules and a map that maximizes the number of 1's less the number of 0's in the consensus cut columns. In a similar spirit as before we get the Conservative BSC (BSC_C) and the Conservative BPC (BPC_C) problems.

Note that all these problems also appear in the circular ordered restriction map problem.

Lemma 1 *If $f_j(x)$ is a linear function and $f_j(x) = g_j(x)$, $0 \le x \le m$, $\mathcal{M}()$ is a constant function.*

Proof: Under these conditions, every column (irrespective of the alignments of the rows) can either be or not be a consensus cut column without affecting the cost. If $f_j(x) = \alpha x + \beta$, for some $\alpha > 0$, then the cost function is always $\alpha A + n\beta$ where A is the number of 1's in the input matrix and n is the number of columns. (Note that this is not true if the functions are not linear since $h(a+b) \neq h(a) + h(b)$ where $h()$ is not a linear function.) $\square$

Table 1 lists these problems and the forms of $f()$ and $g()$ and Figure 1 plots these functions for convenience. It is interesting to note that problems with linear $f()$ such as EBFC, BFC and others gives rise to inapproximable combinatorial problems [9]. However, attempting to simplify the cost function trivializes the problem as shown in Lemma 1.

3.2 Consensus/Agreement without (explicit) hypothesis

In this approach, we take d molecules, and solve the problem exhaustively: we take all possible d sets from the m molecules and get an alignment of all molecules. The approach is summarized as:

$$\max_{\text{(over all alignments)}} \left\{ \sum_{i_1=1}^{m} \sum_{i_2=i_1}^{m} \cdots \sum_{i_s=i_{d-1}}^{m} \mathcal{M}(i_1, i_2, \ldots, i_d) \right\}. \tag{6}$$

Informally, we are looking for d mutual agreements between molecules. In particular if $d = 2$, then the task is to maximize pairwise match. We will show that this formulation which does not have an explicit hypothesis is actually a special case of the one with a hypothesis.

Given s molecules, $i_1, i_2, \ldots, i_d$, they can be matched for an alignment (out of the 2^{d-1} possible configurations in the situation where orientation uncertainty is modeled). The total number of such alignments depend on the errors being modeled. For a $\delta > 0$, define

$$A^X(i_1, i_2, \ldots, i_d) = \text{ \# of cut sites that are within } \delta \text{ of each other, given } X,$$

and,

$$D^X(i_1, i_2, \ldots, i_d) = \text{ \# of cut sites that are not within } \delta \text{ of each other, given } X,$$

where X denotes one of the alignments and the match is made respecting this alignment. Let

$$\begin{aligned} \mathcal{M}_1^X(i_1, i_2, \ldots, i_d) &= A^X(i_1, i_2, \ldots, i_d) && (7)\\ \mathcal{M}_2^X(i_1, i_2, \ldots, i_d) &= A^X(i_1, i_2, \ldots, i_d) - D^X(i_1, i_2, \ldots, i_d). && (8) \end{aligned}$$

Notice that A^X and D^X incorporate all the errors being modeled.

Consensus/Agreement Criteria

	Problem	σ_j	Errors Modeled	$f(x)$	$g(x)$	c_j
				Linear $f()$		
A-1	Exclusive Binary Flip Cut (EBFC) [5], [6].		Orientation uncertainties			implicit (eqn 5)
A-2	Binary Flip Cut (BFC) [5], [6],[9]	small				
A-3	Binary Shift Cut (BSC) [8], [7]		Missing fragments			
A-4	Binary Partition Cut (BPC) [8], [9]		Good/spurious molecules	$x - c_j$	0	user defined
A-5	Weighted Flip Cut (WFC) [6]	fixed	Orientation uncertainties, sizing errors			
A-6	$\mathrm{EBFC}_{\max,\Sigma}$		Orientation uncertainties		$c_j - x$	implicit (eqn 5)
	Balanced problems					
A-7	BFC_B [7]		Orientation uncertainties			
A-8	BSC_B	small	Missing fragments	$2x - m$	$m - 2x$	implicit ($c_j = m/2$)
A-9	BPC_B		Good/spurious molecules			
	Conservative problems					
A-10	BFC_C [7]		Orientation uncertainties			
A-11	BSC_C	small	Missing fragments	x	$m - x$	implicit ($c_j = m/2$)
A-12	BPC_C		good/spurious molecules			
	Weighted Consistency Graph (WCG) problems			Quadratic $f()$		
A-13	Pairwise match under $\mathcal{M}_1$	small	All errors	$\binom{x}{2}$	0	No c_j
A-14	Pairwise match under $\mathcal{M}_2$			$2\binom{x}{2} - \binom{m}{2}$		implicit (eqn 10)
				degree-d $f()$		
A-15	d-wise match under $\mathcal{M}_1$	small	All errors	$\binom{x}{d}$	0	No c_j
A-16	d-wise match under $\mathcal{M}_2$			$2\binom{x}{d} - \binom{m}{2}$		implicit (eqn 9)

Table 1: Problems using the consensus/agreement criterion (identified primarily by monotonically increasing $f_j()$). All the problems also model false positive and false negative errors. Recall that $c_j = mp_j$ is the minimum number of 1's or cuts required in a column j for it to be a consensus cut column.

Theorem 1 *The matching with d-wise agreement is equivalent to the following optimization problems.*

$$\max_{\textit{(over all configurations)}} \sum_j \binom{T_j}{d}, \quad \textit{under } \mathcal{M}_1,$$

and,

$$\max_{\textit{(over all configurations)}} \sum_j \left(2\binom{T_j}{d} - \binom{m}{d} \right), \quad \textit{under } \mathcal{M}_2, \tag{9}$$

where T_j represents the number of cuts at the position l in that configuration.

Proof: Consider a configuration. The number of matches per column is $\binom{T_j}{d}$. We need to maximize this over all the columns, hence the result. Similarly the other cost function where the cost is

$$\binom{T_j}{d} - \left(\binom{m}{d} - \binom{T_j}{d} \right). \tag{10}$$

□

Corollary 1 *The pair-wise matching problem is equivalent to the following optimization problem:*

$$\max_{\textit{over all configurations}} \left(\sum_j T_j(T_j - 1) \right),$$

where T_j represents the number of cuts in the position l in that configuration.

Case $d = 2$: Construct a graph $\mathcal{G}$ corresponding to a given problem, with every vertex v_i corresponding to a molecule i. Let $X = S$ denote a configuration where both the molecules i and j are taken as is, and $X = O$ denote the configuration where one of them is flipped. Every edge $e_{ij} = v_i v_j$ is labeled and weighted as follows. Label $L(v_i v_j)$ is defined as:

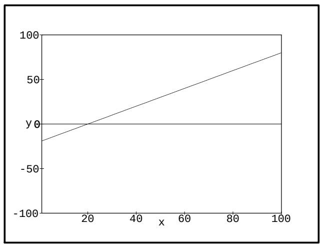
$$L(e_{ij}) = \begin{cases} 1 & \mathcal{M}^S(i,j) \geq \mathcal{M}^O(i,j), \\ -1 & \text{otherwise.} \end{cases}$$

Define $Flip(L(e_{ij})) = -L(e_{ij})$. Weight $Wt(e_{ij})$ is defined as (notice that all the weights are negative)[3]:

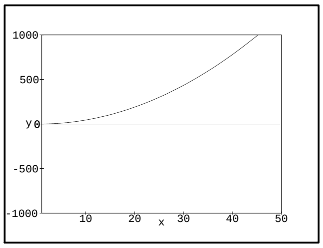
$$\begin{aligned} Wt(e_{ij}) &= \min(\mathcal{M}^S(i,j), \mathcal{M}^O(i,j)) - \max(\mathcal{M}^S(i,j), \mathcal{M}^O(i,j)) \\ &= -|\mathcal{M}^S(i,j) - \mathcal{M}^O(i,j)| \end{aligned} \tag{11}$$

[3] For readers uncomfortable with the use of negative weights, this problem may be viewed as a *maximization* problem with positive weights (by negating all the weights). However, we use the negative weight to retain the naturalness of the optimization problem.

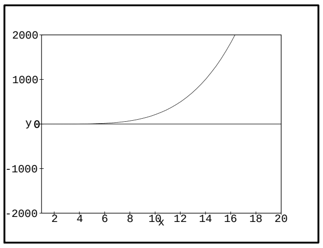
Linear | Quadratic | degree-d ($d = 4$)



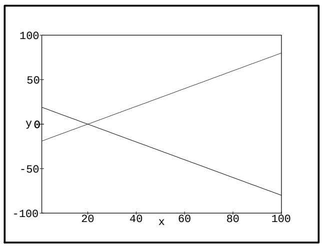
(a) Problems A-1 to A-5.



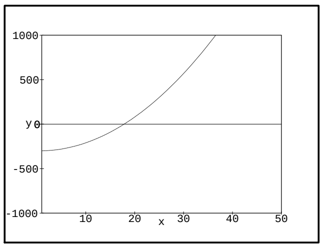
(c) Problem A-13.



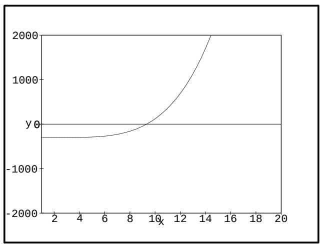
(e) Problem A-15.



(b) Problems A-6 to A-12.



(d) Problem A-14.



(f) Problem A-16.

Figure 1: The plots of the $f_j(x)$ and $g_j(x)$ functions for the different problems using the consensus/agreement criterion. $f_j(x)$ is the increasing function in all plots. $g_j(x)$ is the decreasing function in (b) and the horizontal line in the others. In each case, c_j is given by the zero of the polynomial $f_j(x) - g_j(x)$. (To create the plots certain values of m were assumed; since the purpose of these plots is to indicate the shape of these curves, we skip the other details that are not vital to the study of these function.)

Vertices v_i, v_j, v_k are *consistent* if $L(e_{ij})L(e_{jk})L(e_{ik}) = 1$. A labeled graph $\mathcal{G}$ is said to be *consistent* if every three vertices v_i, v_j, v_k are consistent.

Weighted Consistency Graph (WCG) Problem: Given the match problem with the corresponding graph $\mathcal{G}$, the problem is to obtain a set of edges $S = \{e^1, e^2, \ldots, e^L\}$ with new labels $\tilde{L}(e_{ij})$ defined as follows:

$$\tilde{L}(e_{ij}) = \begin{cases} -L(e_{ij}) & \text{if } e_{ij} \in S, \\ L(e_{ij}) & \text{otherwise} \end{cases}$$

such that

1. $\mathcal{G}$ is consistent under the labels $\tilde{L}()$, and,
2. $\sum_{l=1}^{L} Wt(e^l)$ is minimized.

If $\mathcal{G}$ is *consistent*, an alignment of all the molecules can be obtained by simply following any spanning tree, and computing the relative orientation of every pair of molecule.

An interesting starting point is to assume that $Wt(e)$ is constant for all edges and then find the minimum number of flips required to get a consistent graph. This is called the Consistency Graph Problem [9]. We make the following observations about a graph in this context.

Lemma 2 *For any graph G, there always exists a consistent labeling of the graph and an upper bound on the number of flips is the number of edges labeled "-1".*

Lemma 3 *For an inconsistent triplet of vertices (edges), exactly one edge can be flipped to make it consistent and that edge can be any one of the three.*

These observations may be useful in designing (approximate) algorithms to solve the problem.

4 Optimizing the characteristic of an alignment

Yet another way to formulate the problem is to optimize or bound a characteristic of an alignment.

4.1 Optimizing K, the number of consensus cuts

Minimizing number of consensus columns ($\mathsf{BFC}_{\min K}$): Given M_{ij} and the digestion rates S, find an alignment of the rows/molecules and a map that minimizes the number of consensus cut columns. In a similar spirit one can define $\mathsf{BSC}_{\min K}$ and $\mathsf{BPC}_{\min K}$.

Maximizing number of consensus columns ($\mathsf{BFC}_{\max K}$): Given M_{ij} and the digestion rates S, find an alignment of the rows/molecules and a map that maximizes the number of consensus cut columns. This cost function has been suggested in [8]. This has been defined as BFC_N in [9]. In a similar spirit one can define $\mathsf{BSC}_{\max K}$ and $\mathsf{BPC}_{\max K}$.

4.2 Optimizing discrepancy

Define a *discrepancy* d_j, between a column j and its conjugate $\overline{j}$ as the difference in the number of 1's between the column j and $\overline{j}$, given an alignment. Now, different cost functions can be defined based on d_j. Note that discrepancy is meaningful only in the context of EBFC problems.

1. $\mathsf{EBFC}_{\text{min,max}}$: Obtain an alignment and a map that minimizes the maximum of the individual (j = 1,2, ..., n/2) discrepancies. Note that this attempts to concentrate the 1's in the consensus cut columns.

2. $\mathsf{EBFC}_{\text{max,min}}$: Obtain an alignment and a map that maximizes the minimum of the individual (j = 1,2, ..., n/2) discrepancies. Note that this attempts to distribute the 1's as evenly as possible.

3. $\mathsf{EBFC}_{\text{max,max}}$, $\mathsf{EBFC}_{\text{min,min}}$: These can be defined in the same spirit but do not appear to be interesting values and the functions are very simple to compute.

4. $\mathsf{EBFC}_{\text{min},\Sigma}$: Obtain an alignment and a map that minimizes the sum of the discrepancies.

5. $\mathsf{EBFC}_{\text{max},\Sigma}$: Obtain an alignment and a map that maximizes the sum of the discrepancies. It is interesting to note that this satisfies the consensus/agreement criterion as the associated $f_j(x)$ and $g_j(x)$ functions satisfy the conditions discussed in Section 3.1.1. It can be verified that $\mathrm{EBFC}_{\max \Sigma}$ (A-6) attains its optima at the same alignment as EBFC (thus these two problems are identical).

4.3 Bounding Alignment Errors

Another class of problems is where we seek to obtain an alignment when the number of molecules that show the error is bounded by a fixed number, say γ. Let I_f denote the number of molecules that are flipped [4], let I_s denote the number of molecules that have missing fragments and let I_p denote the number of spurious molecules in an alignment. Thus $I_f < \gamma$ and so on for the optimal alignment. To retain consistency of notation we call these problems $\mathsf{BFC}_{<I}$, $\mathsf{BSC}_{<I}$ and $\mathsf{BPC}_{<I}$ when in the alignment I_f, I_s or I_p respectively are at most γ. One can also envisage problems

[4] In an alignment let j be the molecules flipped; then due to the symmetrical nature of this error $I_f = \min\{j, m-j\}$.

where multiple errors are being handled simultaneously. Note that all these problems also appear in the circular ordered restriction map problem.

The underlying chemical process that gives rise to the restriction map problem is fairly accurate and it is assumed that only a small percentage of error (orientation uncertainty or missing fragments or spurious molecules) creeps into the process. Hence these cost functions are meaningful.

Characteristic Optimization

	Problem	Errors Modeled	Cost Function		On c_j
			$\max\left\{\sum_j \{f_j(T_j)G_j(d_j) + g_j(F_j)\}\right\}$ (eqn 3)		
			$f_j(x)$	$g_j(x)$	
B-1	$\mathrm{EBFC}_{\min,\Sigma}$	Orientation uncertainties	$\begin{cases} 1 & x < c_j \\ 0 & \text{otherwise} \end{cases}$	0	implicit (eqn 5)
B-2	$\mathrm{BFC}_{\min K}$				user defined
B-3	$\mathrm{BSC}_{\min K}$	Missing fragments			
B-4	$\mathrm{BPC}_{\min K}$	Good/bad molecules			
B-5	$\mathrm{BFC}_{\max K}$	Orientation	$\begin{cases} 0 & x < c_j \\ 1 & \text{otherwise} \end{cases}$	0	user defined
B-6	$\mathrm{BSC}_{\max K}$	Missing fragments			
B-7	$\mathrm{BPC}_{\max K}$	Good/bad molecules			
B-8	Probabilistic Model [8]	potentially all	$x\ln(\frac{x}{m})+ (m-x)\ln(1-\frac{x}{m})$	$\tilde{g}(x) = x(\ln\frac{x}{m}-1)$	implicit
			Using discrepancy d_j		
B-9	$\mathrm{EBFC}_{\min,\max}$	Orientation uncertainties	$\min_{(\text{over all } \mathcal{H} \text{ \& aligns})}\{\max_j \ d_j\}$		implicit (eqn 5)
B-10	$\mathrm{EBFC}_{\max,\min}$		$\max_{(\text{over all } \mathcal{H} \text{ \& aligns})}\{\min_j \ d_j\}$		
B-11	$\mathrm{EBFC}_{\min,\Sigma}$		$\min_{(\text{over all } \mathcal{H} \text{ \& aligns})}\{\Sigma_j \ d_j\}$		
			Bounding alignment errors		
B-12	$\mathrm{BFC}_{<I}$	Orien uncert	$I_f < \gamma$ under cost function $\mathcal{C}$		depends on $\mathcal{C}$
B-13	$\mathrm{BSC}_{<I}$	Missing fragments	$I_s < \gamma$ under cost function $\mathcal{C}$		
B-14	$\mathrm{BPC}_{<I}$	Good/bad molecules	$I_p < \gamma$ under cost function $\mathcal{C}$		

Table 2: Classification of problems using the optimization of a characteristic of an alignment.

5 A Probabilistic Approach

A probabilistic model for the restriction map problem is present in [8]. We will analyze this model in the context of our framework.

We will not attempt to give a complete definition and description of the proposed model here. The reader is advised to look at [8] for notations and other details.

At the heart of the model is the following definition (equation (1) in [8], which we reproduce verbatim):

$$Pr[D_j^{(k)}|\mathcal{H}, good] = \left[\prod_{i=1}^{N}\left(p_{c_i}\frac{e^{-(s_{ijk}-h_i)^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i}\right)^{m_{ijk}}(1-p_{c_i})^{1-m_{ijk}}\right] \times \lambda_f^{F_{jk}}e^{-\lambda_f} \tag{12}$$

However, for the sake of ease of understanding, we use the example that the authors use to explain the above formula (which we again produce verbatim) shown below:

$$Pr_{jk} = Pr\left[D_j|\mathcal{H}, A_{jk}\right] = p_{c_1}\frac{e^{-(s_1-h_1)^2/2\sigma_1^2}}{\sqrt{2\pi}\sigma_1} \times (1-p_{c_2}) \times \lambda_f e^{-\lambda_f} \times \ldots \times p_{c_N}\frac{e^{-(s_N-h_N)^2/2\sigma_N^2}}{\sqrt{2\pi}\sigma_N} \tag{13}$$

This defines the underlying cost function for the probabilistic algorithm. $Pr\left[D_j|\mathcal{H}, A_{jk}\right]$ denotes the probability of the molecule j aligned (alignment index k) by A_{jk} with the hypothesis $\mathcal{H}$. $c_1, c_2, \ldots, c_N$ are the consensus cuts and $s_1, s_2, \ldots, s_N$ are their locations in the hypothesis $\mathcal{H}$. $h_1, h_2, \ldots, h_N$ denote the distances of each of these cuts, if they exist, in the molecule j from the exact locations under the alignment A_{jk}. M is the number of molecules and N is the number of consensus sites in the molecule. λ_f is the expected number of false cuts. F_{jk} is the number of false cuts in the data D_j for the alignment A_{jk} that do not match any hypothesis $\mathcal{H}$.

To understand this function better, consider the special case where every molecule is *good*. Then, the cost function that equation (13) suggests can be viewed as form (4). In [8] the log-likelihood, $\mathcal{L}$, which is to be maximized, is computed as (we again reproduce verbatim):

$$\mathcal{L} \equiv \sum_j \ln Pr[D_j|\mathcal{H}]. \tag{14}$$

Note that in the special case where every molecule is *good*, by the definition of $Pr[D_j|\mathcal{H}]$ in [8], the following holds:

$$Pr[D_j|\mathcal{H}] = \frac{1}{2}\sum_k Pr[D_j|\mathcal{H}, A_{jk}]. \tag{15}$$

Under the assumption that, when a molecule is matched against a hypothesis, the alignments that do not match give a very low probability, that is $P_{jk} \approx 0$, $\mathcal{L}$ corresponds to $\mathcal{M}$ of equation (4). Now, to understand how this function behaves, consider the special case where G_l has a constant

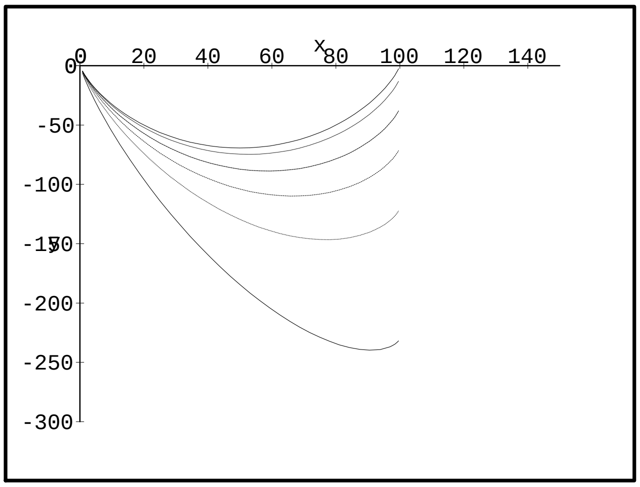


Figure 2: The plot of $f(x) = x\ln(\frac{a*x}{M}) + (M-x)\ln(1-\frac{x}{M})$ with $M = 100$ and $a = 0.1, 0.3, 0.5, 0.7, 0.9, 1.0$. The "perfect" cup shape is obtained when $a = 1.0$; the curve with the lowest dip corresponds to $a = 0.1$.

$\sigma_l, \forall l$; for different values of the constant we get plots of the function as shown in Figure 2 (using the equation (16) discussed below). For further discussion, we choose the curve, which corresponds to $G_l() = 1$ because of its appropriate shape [5]. Let $x = \sum_{j=1}^{M} m_{ijk}$ for a configuration. m_{ijk} is the indicator variable for molecule j and cut i of the hypothesis $\mathcal{H}$, under alignment A_k. Since p_{c_i} is the probability of cut i, $p_{c_i} = x/M$. Let $y = \sum_{j=1}^{M} F_{jk}$, where F_{jk} is the number of false cuts in data D_j. Thus $\lambda_f = y/M$, by definition of λ_f.

Notice that equation (12) uses indicator variables s_{ijk} and m_{ijk} and a count variable F_{jk} that gives the number of false cuts per molecule j for the alignment A_{jk} (although s_{ijk} does not appear explicitly in the equation it is nevertheless used, see [8] for details). The number (and the values) of the variables s_{ijk} and m_{ijk} depend on the number of cuts in the hypothesis $\mathcal{H}$ and the value of F_{jk} depends on decisions whether cuts in a molecule correspond to cuts in the hypothesis $\mathcal{H}$ or not. Hence we can effectively discretize the model depending on the number and values of these variables.

As $p_{c_j} = e^{\ln p_{c_j}}$, if there are x molecules showing a cut at location j (thus $M - x$ that do not) for

[5] Notice that if $G_l() = a = 0.1$ is used, the $\tilde{g}()$ curve in Figure 3 is such that $f(x) < \tilde{g}(x)$, for all $x > \delta > 0$, where δ is a very small number, which is not desired.

an alignment A_k, we have (using equation (14)),

$$\begin{aligned} f(x) &= \ln\left(\left(e^{\ln p_{c_j}}\right)^x \left(e^{\ln(1-p_{c_j})}\right)^{M-x}\right) \\ &= x\ln p_{c_j} + (M-x)\ln(1-p_{c_j}) \\ &= x\ln(\tfrac{x}{M}) + (M-x)\ln(1-\tfrac{x}{M}) \end{aligned} \tag{16}$$

assuming $G_j() = a = 1.0$ (see Figure 2). Similarly, if there are y false cuts, then,

$$\begin{aligned} \tilde{g}(y) &= M\left(\ln e^{-\lambda_f}\right) + y\ln\lambda_f \\ &= -M\lambda_f + y\ln\lambda_f \\ &= y\,(\ln\tfrac{y}{M} - 1) \end{aligned} \tag{17}$$

$f(x)$ is such that $f(a+b) < f(a) + f(b)$, for $a+b < M/2$ and $f(a+b) > f(a)+f(b)$ for $a+b > M/2$. Thus for $a+b < M/2$, the model prefers to put the 1's into one consensus site and for $a+b > M/2$ it tries to distribute the cuts between j and its conjugate $\bar{j}$.

Now, based on our framework, we suggest a possible modification that explicitly models the digestion rates p_j's.

1. Modeling the false cut at different locations separately, and,
2. Using a term $k_j \leq 1$ that ensures that the digestion rate is at least $p_j, 0 < p_j \leq 1$.

For clarity, we skip the subscript j of k_j, p_j in the rest of the discussion. Now equation (13) can be modified as:

$$Pr_{jk} = p_{c_1}\frac{e^{-(s_1-h_1)^2/2\sigma_1^2}}{\sqrt{2\pi}\sigma_1} \times (1-p_{c_2}) \times k\lambda_{f_l}e^{-\lambda_{f_l}} \times \ldots \times p_{c_N}\frac{e^{-(s_N-h_N)^2/2\sigma_N^2}}{\sqrt{2\pi}\sigma_N} \tag{18}$$

k can be viewed as a factor that "discourages" λ_{f_l} to be large and the extent of this "discouragement" is governed by equation (21) as shown below. Using equation (18), equation (3) modifies to:

$$\mathcal{M} = \sum_l \left\{f(T_l)G_l(d_l) + g(F_l)\right\}, \tag{19}$$

where

$$F_l = (\text{number of molecules that have a \textit{false} site at } l),$$

and,

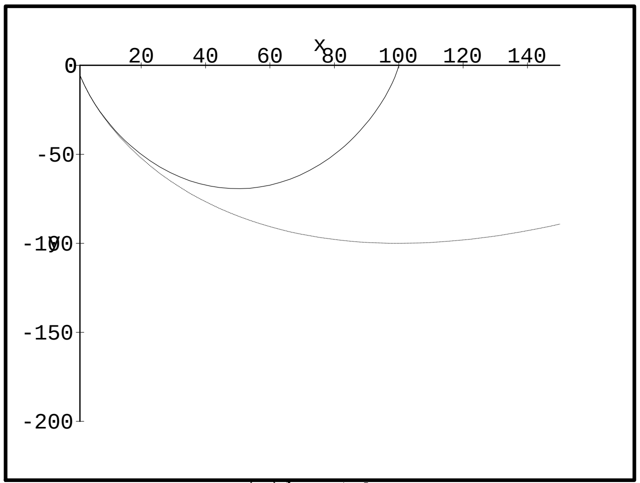
$$g(y) = y\left(\ln\frac{ky}{M} - 1\right). \tag{20}$$

k is such that

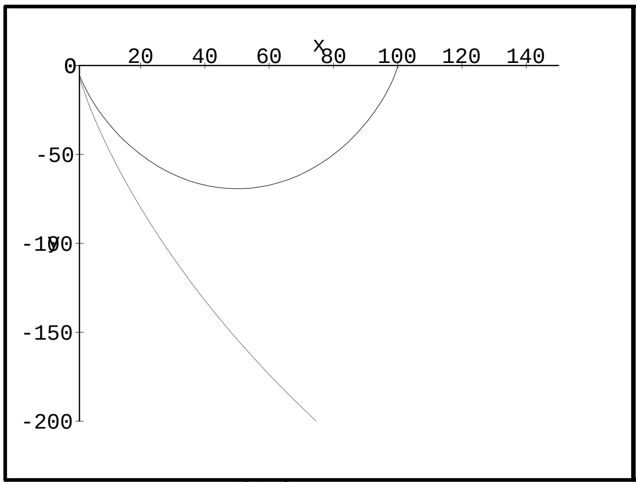
$$g(x) < f(x),\ \text{for } x > pM. \tag{21}$$

Figure 3 shows a plot of the functions $f()$ and $g()$ for different values of k.

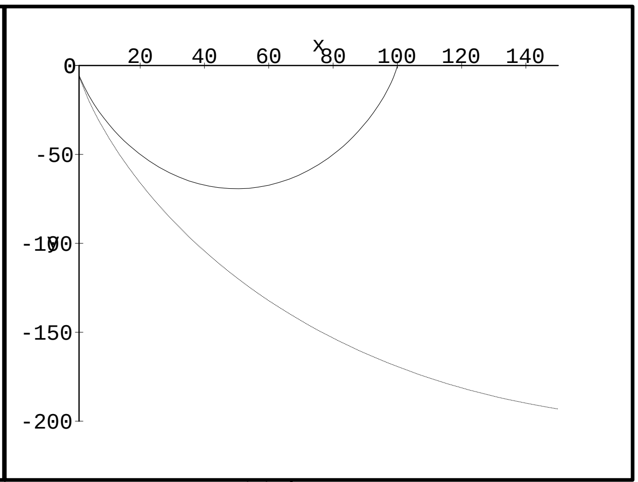
This concludes the analysis of the probabilistic model in our framework.



$(a) k = 1.0$



(b) $k = 0.25$



(c) $k = 0.5$

Figure 3: The plots of the $f(x)$ and $g(x)$ functions for the probabilistic model ($f(x)$ is the cup-shaped curve and the other is the $g(x)$ curve). Thus (a) shows the $f()$ and $\tilde{g}()$ functions for equation (12) and (b) and (c) show the $f(x)$ and $g(x)$ functions for equation (18).

6 Complexity Results

In this section we review the known complexity of the problems discussed in the previous sections.

The Exclusive Binary Flip Cut problem (A-1) was shown to be NP-Complete in [7]; this result was improved in [9] to show that the problem is MAX SNP-hard. MAX SNP hardness implies that obtaining an ϵ approximation for a problem, where $\epsilon > \epsilon_0$, for some fixed ϵ_0 (depending on the problem) is NP hard. Thus a MAX SNP hard problem can not admit a polynomial time approximation scheme (PTAS) [10]. In [6], it was shown that a dense instance of this problem admits a PTAS. An instance of a problem is dense if the number of 1's or cuts in a column is $m\gamma$, where m is the number of molecules or rows and $0 < \gamma < 1$ is a fixed constant. It can be verified that $\text{EBFC}_{\max\Sigma}$ (A-6) attains its optima at the same alignment as EBFC, hence $\text{EBFC}_{\max\Sigma}$ is also MAX SNP hard. Further, in [9], it is shown that achieving an approximation ratio $1 - \Upsilon/7$ for EBFC is NP-hard, where Υ denotes the upper bound on the polynomial time approximation factor of the well-known max cut problem.

The Binary Flip Cut (A-2) and Weighted Flip Cut (A-5) problems were shown to be MAX SNP hard in [9]. A different hue of this problem, $\text{BFC}_{\max K}$ (B-5) was shown in [8] to be NP-hard. This result was improved in [9] to show that the problem is MAX SNP hard. It is further shown in [9] that achieving an approximation ratio $1 - \Upsilon/7$ for BFC and WFC and an approximation ratio of $(1 - \Upsilon/7)p_{max}/p_{min}$ for $\text{BFC}_{\max K}$ is NP hard where $p_{min} = \min_j p_j$, and $p_{max} = \max_j p_j$ (the maximum and minimum of the digestion rates).

The Binary Shift Cut problem (A-3) was shown to be NP-Complete in [7], and another hue of this problem, $\text{BSC}_{\max K}$ (B-6), was shown to be NP hard in [8]. Both these results were improved in [9] where it was shown that the problems are MAX SNP hard. Further, achieving an approximation ratio of $1 - \Upsilon/6$ for BSC and a ratio of $(1 - \Upsilon/6)p_{max}/p_{min}$ for $\text{BSC}_{\max K}$ is NP-hard.

A version of Binary Partition Cut problem, under known number of bad molecules,was shown to be NP hard in [8]. It was shown that the general BPC problem actually has a polynomial time algorithm in [9].

Different hues of the Binary Flip Cut Problems, BFC_B (A-7) and BFC_C (A-10) were shown to be NP complete in [7]. It has been shown in [9] that the Weighted Consistency graph problems (A-13, A-14) and d-wise match problems (A-15, A-16) are MAX-SNP hard and achieving an approximation ratio of $1 - \Upsilon$ in each case is NP hard [9].

The complexity of the remaining problems is unknown at the time of writing this paper.

7 Conclusion

This paper presents a uniform framework to model problems arising from the emerging Optical Mapping technology in constructing ordered restriction maps. The attempt has been to integrate different approaches and problem formulations proposed in recent literature in this area, and to link the cost functions proposed in these approaches to one another and to the underlying problem.

We identify two main approaches to the ordered restriction map problem, one involving the use of a consensus or agreement method, and the other optimizing a characteristic function of the data. We use this to develop a framework where each of these models for the restriction map problem becomes a specific instance of the basic framework we propose. Interestingly, we have been able to encompass the combinatorial approaches with the probabilistic approach, all within the same framework. Finally, we have indicated the open problems by including a survey of the best known complexity results for these problems.

The availability of this framework opens up the possibility of exploring other cost functions, other variants of the problem and new approaches to solving these problems.

Acknowledgements

I would like to thank Bud Mishra and Thomas Ananthraman for their careful reading of the paper and their comments. I would also like to thank Saugata Basu and R Chandrasekar for helpful discussions.

References

[1] Y. Wang, E. 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.

[2] 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.

[3] D. Schwartz, X. Li, L. Hernandez, S. Ramnarain, E. Huff, Y. Wang. *Ordered Restriction Map of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping*, In Proc. Natl. Acad. Sci. USA, 92:165-169, 1995.

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

[5] D. Geiger, L. Parida, *A Model and Solution to the DNA Flipping String Problem*, Courant Inst. of Math. Sciences, TR1996-720, May, 1996.

[6] S. Muthukrishnan, L. Parida, *Towards Constructing Physical Maps by Optical Mapping: An Effective, Simple, Combinatorial Approach*, Procc of the International Conference on Computational Molecular Biology (RECOMB 97), ACM Press, Santa Fe, 1997.

[7] V. Dancik, S. Hannenhalli, S. Muthukrishnan, *Hardness of Flip-Cut Problems from Optical Mapping*, Journal of Computational Biology, 1997.

[8] T. Anantharaman, B. Mishra B, D. Schwartz, *Genomics via Optical Mapping II: Ordered Restriction Maps*, Journal of Computational Biology, vol 4, No 2, pp. 99-118, 1997.

[9] L. Parida, *Inapproximability of Flip-Cut and other related problems from Optical Mapping*, under preparation.

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

[11] M. Garey, D. Johnson, *Computers and Intractability: A Guide to the theory of NP-Completeness*, page 210, Freeman Press, 1979.