

A REVIEW OF OPTICAL MAPPING AS A METHOD OF WHOLE GENOME ANALYSIS

MAY 1, 2009

AUSTIN J. RAMME

INSTRUCTOR: ISABEL DARCY, PHD
COURSE: 22M:151

Key words: optical mapping, restriction mapping, map assembly, genome shotgun, graph theory

Mathematics used: Graphs, Hidden Markov Model, error reduction, statistical analysis

Mathematical Difficulty: Complex

Area of Application: Whole Genome Analysis

Application Area Difficulty: Very Complex

ABSTRACT

Optical mapping is a method of whole genome analysis that was originally introduced in 1995¹. It involves the generation of ordered restriction maps for entire genomes using classical molecular biology and biological microelectromechanical systems (BioMEMS) microfluidic techniques adopted from the semiconductor industry. The goal of this paper is to review the pertinent literature on optical mapping. The literature review will be broken into three sections: an overview of the optical mapping process, an in-depth investigation of the algorithms involved in the data analysis, and finally a look at relevance and applications of this technology. We aim to offer the reader a comprehensive review of this technology and an explanation of its future potential.

INTRODUCTION

HISTORY AND GOAL OF OPTICAL MAPPING:

In recent years, medical research has largely focused on identifying genetic causes for disease in hopes of advancing both diagnosis and treatment. Researchers have found diseases associated with a single genetic cause; however, the number of identified polygenetic diseases is ever increasing. This has shifted the focus of genetics research from individual genes to analysis of entire genomes. In recent years, genetics research has greatly progressed due to advances in techniques and the sequencing of the human genome. Individual clinical genomic analysis is still a future goal, and researchers have made giant strides towards this objective. In the past, the mapping of the human genome primarily depended on restriction enzyme mapping. Restriction enzymes are specialized proteins able to cut phosphodiester bonds of DNA sequences in repeatable, consistent, and specific patterns depending on a given DNA base pair sequence. Traditionally, restriction map construction has been dependent on gel electrophoretic analysis as shown in Figure 1¹. Restriction maps provide precise genomic distances that are useful in providing spatial information for specific genetic loci. This technique has successfully mapped a number of bacteria including *E. coli*, *S. cerevisiae*, and *C. elegans*, but has not been able to map high order organisms due to limitations intrinsic to the methods¹. For a long time, the lack of commercial software for gel electrophoretic analysis inhibited its advancement as a method of genome analysis. The low throughput of the process and extensive manual labor required were two additional roadblocks to using restriction mapping as a method of whole genome analysis.

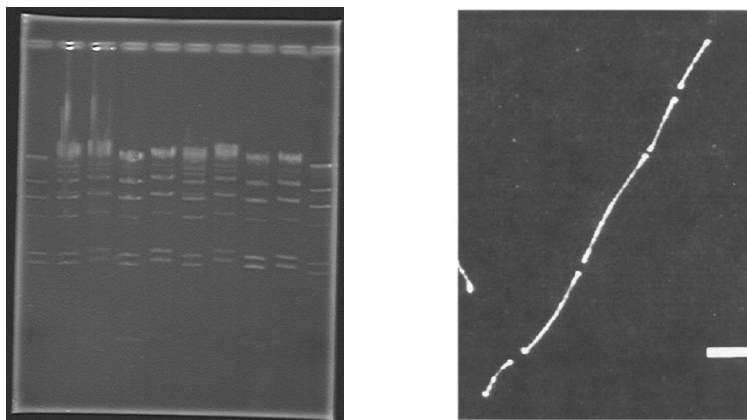


Figure 1: Left- An image of a 1D electrophoresis gel displaying successful restriction enzyme digest of λ viral DNA². The different bands represent different lengths of DNA. Each individual band represents a

collection of DNA molecules of equivalent length. Right- A strand of genomic DNA elongated on a charged glass surface and digested with a restriction enzyme to produce visible gaps between different sized segments of a single DNA molecule, also called an optical map¹.

Optical mapping is an automated nonelectrophoretic method of ordered restriction map generation with a goal of whole genome analysis¹. This method significantly differs from traditional gel electrophoresis in both its throughput and also its methods. A primary difference from gel electrophoresis is that in optical mapping restriction fragments from individual DNA molecules are used for analysis instead of large collections of fragments from multiple DNA molecules. Optical mapping technology offers a fully automated system for restriction map construction based on computer guided image acquisition and analysis systems¹. This technology is capable of generating high resolution genomic restriction maps without prior sequence knowledge. The generated restriction maps have a variety of uses including identifying genetic insertions, repeats, inversions, and deletions³. It also offers a means to establish genotype-phenotype correlations for clinical medicine.

THE ORIGINAL OPTICAL MAPPING PROCESS AND ITS LIMITATIONS:

In the original optical mapping method, fluorescently labeled DNA molecules were elongated in a flow of molten agar between a coverslip and microscope slide^{1,3}. The agar gelling process captured the DNA molecules in an elongated state. Restriction enzymes previously added to the agar, were activated by adding magnesium ions to the solution to initiate the digestion reaction. The DNA cut sites could be visualized as gaps between the DNA fragments as the DNA fragments recoiled. Time lapse fluorescence microscopy was used to image the single molecules; a sample image can be seen in Figure 1. A major limitation to this method was the random locations of the DNA molecules in the agar which made imaging difficult. The difficulties were primarily due to a lack of an organized structure of DNA molecules and different focal planes needed to image the molecules located in three-space.

The second generation of the optical mapping process eliminated the need for the agar suspension and instead used a polylysine treated glass surface^{1,3}. The positively charged glass surface allowed for physical interaction with negatively charged DNA molecules. Optimization of reaction conditions allowed for sufficient access of restriction enzymes to the DNA molecules for the reaction to occur. Surface mounting removed the necessity of time lapse imaging as the DNA molecules were fixed on the surface and easily remained in focus. This generation of optical mapping was based on genomic restriction map construction derived from a large number of small, overlapping DNA fragments³. Image analysis started with determining the correct number of DNA fragments and then developing histograms based on the sizes of the given fragments. Restriction maps were generated based on average restriction fragment size from histogram analysis, which was easily constructed since the order of the restriction fragments is maintained by optical mapping. This iteration was a step towards automated genomic analysis, but many key issues regarding data acquisition, organization, and analysis remained as clear roadblocks.

MODERN OPTICAL MAPPING PROCESS: *SECTION OMITTED FOR ONLINE PUBLISHING

DATA ANALYSIS IN OPTICAL MAPPING

IMAGE COLLECTION TO DATA ANALYSIS:

The modern optical mapping process is a drastic improvement over previous methods and has the potential to be fully automated from DNA purification to image collection and analysis. However, data collection and imaging only mark the beginning of the optical mapping process. Analyzing the extremely large quantities of data and correcting for confounds inherent to the optical mapping process are very challenging and complex.

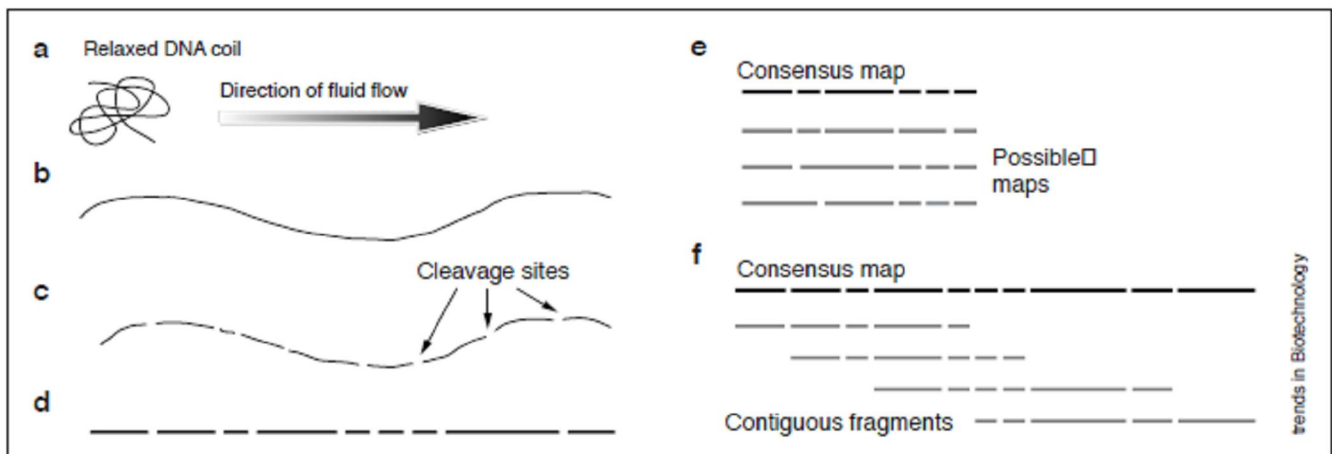


Figure 5: This figure shows the steps of optical mapping, starting with DNA elongation and surface deposition (a,b)⁵. The process continues with restriction enzyme digestion forming visible gaps between DNA segments (c). The DNA is imaged using fluorescence microscopy and converted into a series of “bar code” like bands (d). Using multiple optical maps, a graphing scheme is used to generate a consensus map based on multiple optical maps (e). Finally, the individual consensus map fragments are aligned to form a consensus map for the entire genome (f).

Figure 5 shows an overview of the full optical mapping process. Recall that the imaging process results in sheared DNA molecules with restriction fragments being separated by gaps within the fluorescently marked molecules⁴. The relative size of the DNA fragments is determined from an internal standard (λ DNA) of known length that is added to the plate prior to imaging as well as the intensity distribution from the fluorescent dye used to visualize the DNA. Essentially, the process yields an image collection containing a set of genomic restriction fragments of known length deposited in a reproducible pattern. Genomic DNA is very fragile and it is very difficult to obtain an entirely intact genomic DNA strand prior to restriction reaction. Thus optical mapping is based on imaging a collection of randomly sheared molecules (resulting from simple laboratory procedures, such as pipetting) representing the entire genome; this has been coined “Shotgun Optical Mapping⁴.” A huge number of randomly sheared genomic fragments are present in each optical mapping imaging study in such a way that the human genome is redundantly represented. The analysis of the imaged data and generation of high quality restriction maps is not a trivial process. Along with this vast quantity of data, we encounter a variety of challenges inherent to this process since we are looking at individual DNA molecules. Error reduction after image acquisition must be implemented to correct for confounds inherent to the optical mapping process. These errors include^{4,5}:

- Spurious, or false restriction sites
- Partial digestion (efficiency is typically 70-90%)
- Small (< 2kb) fragments that are underrepresented in the dataset
- Fragment sizing error
- Chimeric maps (artifactual overlapped molecules hampering analysis)

Accurate map generation is accomplished using statistical analysis of a number of imperfect maps. Experience has shown that combining the results from multiple optical mapping will often give more accurate final restriction maps. Recent advancements in algorithms have allowed for improved restriction map accuracy and efficiency using graphing techniques.

OVERVIEW ORDERED RESTRICTION MAP GENERATION

Shotgun optical mapping involves the shearing of genomic DNA molecules from random locations within the DNA molecule due to the inherent stresses of laboratory procedures such as pipetting⁴. Thus the optical maps represent random parts of the genome, not identical DNA molecules. This system aims to achieve 10-50x coverage of a given genetic locus. In the past, several groups have investigated methods to reconstruct restriction maps from small molecules. The problem has been considered NP-hard by some, but others have found polynomial time algorithms for low order eukaryotes (bacteria). However, these algorithms could not be applied to shotgun optical mapping, which has traditionally been the most useful form.

A new algorithm has been published that allows for optical map assembly from shotgun maps using a modified overlap-layout-consensus computational framework, which is commonly used in DNA sequence assemblers⁴. The overlap-layout-consensus method is a three step process corresponding to the three elements of its name. The overlap step of this process is responsible for establishing the overlapping regions optical maps. The layout step of this process is responsible for establishing the local and global connectivity of the overlapping maps. The consensus step computes the finished restriction map. To accomplish these steps, graphing techniques are employed to generate a connectivity graph to represent significant overlaps of optical maps. A distance based error reduction technique is also employed based on the graph structure. From the connectivity graph, a draft genomic restriction map is created from the composite of many optical maps. A map refinement process is finally employed to correct inaccuracies and to produce a final genomic restriction map.

The input to the whole-genome optical map assembly process is a set of optical maps (many fragmented, digested ordered DNA molecules) from an optical mapping experiment. The modified overlap-layout-consensus process actually consists of seven steps that are detailed in this section including: calculation of overlaps, overlap graph construction, graph correction, island identification, contig construction, draft consensus map construction, and consensus map refinement.

I. Calculation of Overlaps

The first step involves computing all alignments (overlaps) of optical maps within the input dataset⁴. A scoring system is used to identify accurate overlaps for the next step of the process, overlap graph construction. This scoring system accounts for missing cuts, false cuts, and sizing errors^{5,6}. Each map segment [i,k] consists of sites i through k with a matching pair between two maps defined as $(i,j;k,l)$ ⁴. Valouev, et al. describes the global alignment Π between two maps A

(with m cut sites) and B (with n cut sites) as a sequence of ordered matching pairs $(i_1, j_1; k_1, l_1)$ $(i_2, j_2; k_2, l_2)$... $(i_d, j_d; k_d, l_d)$, where $k_t < i_{t+1}$ and $l_t < j_{t+1}$ for each $t < d$. The locations of the cut sites are represented as q_x for a given site x on map A and as r_y for a given site y on map B . To calculate the score of the global overlap, we set $\lambda \geq 0$ and $\nu \geq 0$ and represent the score as⁶:

$$\begin{aligned} score(\Pi) = & \sum_{t=1}^d \sigma(i_t, j_t; k_t, l_t) + l(q_{i_1}, r_{i_1}) + \sum_{t=2}^d l((q_{i_t} - q_{k_{t-1}}), (r_{j_t} - r_{l_{t-1}})) \\ & + l((q_{i_m} - q_{k_d}), (r_{j_n} - r_{l_d})) - \lambda \left[m + n - \sum_{t=1}^d (k_t - i_t + 1) - \sum_{t=1}^d (l_t - j_t + 1) \right], \end{aligned}$$

where

$$\begin{aligned} \sigma(i_t, j_t; k_t, l_t) = & \nu \cdot (\# \text{ of matching sites pair in segment } [i_t, j_t; k_t, l_t]) \\ & + l((q_{k_t} - q_{i_t}), (r_{l_t} - r_{j_t})) - \lambda((k_t - i_t) + (l_t - j_t)). \end{aligned}$$

This scoring system rewards each pair of matching cut sites between two maps by ν in $\sigma(i_t, j_t; k_t, l_t)$. It also penalizes each cut site that is not matched between the two maps by λ ⁶. Distance discrepancy between the matching pairs is accounted for using a scoring function of length similarity, $l(a, b)$, for the two overlapping maps, one with length a and the other with length b .

By comparing, identifying, and rewarding matching cut sites and penalizing false or missing cut sites, this alignment scoring function is used for overlap graph construction. Since the number of optical maps is typically very large (for a human it is over 500,000), this is a very computationally expensive step. However, modern parallel processors allow this to be completed in a reasonable amount of time. The overlap for two fragments m and n is computed in $O(mn)$ time. This is equivalent to < 0.01 seconds per overlap on an average computer.

II. Overlap Graph Construction

The goal of the overlap graph is to represent overlaps between individual optical maps to allow for generation of a consensus map for the entire genome. Technically, the overlap graph is a digraph which would be described in terms of arcs; however, the literature describes calls it a graph and describes it in terms of edges. We will maintain the convention adopted in the literature. In our directed graph $G(V, E)$, the set of nodes V represent individual optical maps and the edges E represent high quality overlaps between pairs of maps⁴. A sample of this method of graph construction is seen in Figure 6. This graph structure conveniently allows all procedures to be performed using depth first searching, breadth first searching, and heaviest path searching.

The scoring system in the first step is used as an initial method of identifying accurate overlaps. Additionally, a quality score or “q-score” looks at the degree of site match for the overlaps. The final set of q-scores is used during graph construction such that only overlaps with q-scores exceeding a threshold value are included as edges in the graph. To ensure that the most confident edges are added first, the overlap q-score values are sorted prior to graph construction and the edges are added in the order of decreasing significance. False edges typically have low q-scores. By inspecting false edges late in graph construction and incorporating error reduction measures, an accurate graph can be made and inclusion of false edges can be minimized.

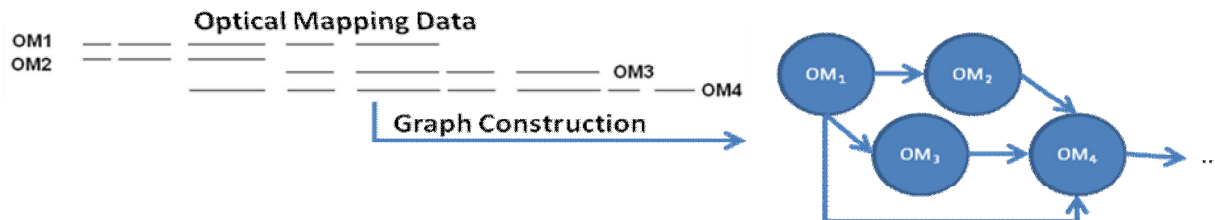


Figure 6: This is an example of the graph structure generated from four optical maps using the system described. Each node corresponds to an individual optical map and edges connecting nodes correspond to regions of overlap between the individual optical maps. The direction and weighting of the edges is determined using a calculation of map orientation and amount of overlap. Technically, this is a digraph, but the literature describes it as a graph and we will maintain that description.

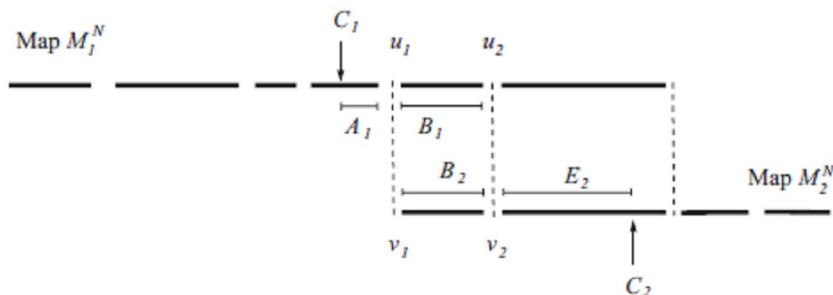


Figure 7: The distance between optical maps (M_1, M_2) is given by the distance between the centers (C_1, C_2). In this case the distance is calculated as $A_1 + (B_1 + B_2)/2 + E_2$ where u_1 represents the closest to C_1 matching site of the largest alignment block ($u_1, u_2; v_1, v_2$) that does not contain center points C_1 and C_2 ⁴. The distance calculation is used to weight and orient the edges.

Establishing the orientation of the edges that are added is challenging due to properties of the optical mapping process⁴. The sheared DNA molecules are not necessarily attached to the surface in the same orientation; thus molecules can either be considered to be normal or reversed based on their comparison to how the optical map is stored in the system. Orientation and edge weight determination are tied to a calculation of genomic distance between overlap regions as demonstrated in Figure 7. The edge weights are calculated as Map M_1 with respect to Map M_2 . The weight corresponds to a genomic distance respect to the midpoints or centers (C_1, C_2) and is given by $A_1 + (B_1 + B_2)/2 + E_2$ with the variables defined as given in Figure 7. The region that is measured is defined as the largest block of overlap that does not contain the midpoints. The sign of the distance is determined by the orientation of the individual maps. To orient the edge in graph construction, the edge directions are chosen such that the edge weight is positive. After all nodes, edges, weights, and orientations have been assigned to the graph, error reduction is the next step in the process.

III. Graph Correction Procedure

Errors in the overlap graph can lead to false connectivity of the genomic map if left unaddressed⁴. False edges correspond to falsely identified overlaps and false nodes correspond to chimeric maps (two molecules that physically overlap during imaging). A graph correction technique looks to account for involving spurious edges, orientation consistent false overlaps, orientation inconsistent false overlaps, and chimeric maps. Examples of these edges can be seen in Figure 8.

The elimination of false edges can be split into three categories: spurious, orientation inconsistent, and orientation consistent. Spurious edges connect two nodes in such a way that a cycle is present within the graph structure. This is an impossible configuration for linear genomes and thus these edges are removed. Orientation inconsistent false edges create an orientation conflict within the graph and management of the inclusion of these edges is accomplished at the graph construction stage. During graph construction, edges added to a graph must agree with the orientation of components that have already been added to the graph. If an edge does not agree with the already defined graph structure, the edge will simply be skipped and not added.

Orientation consistent false edges are edges that connect two unrelated portions of the genome. To deal with this issue, a depth-first search of specified depth is first performed for each node N_i and collects all nodes N_j that have multiple independent paths through the path connecting N_i to N_j . For each of the paths, the distance D_a that maximizes the size of the cluster of paths between the two nodes is calculated and compared to a normal distribution. Based on this calculation, if multiple paths are found with normally distributed distances, then the edge is mark confirmed since it has been shown to be connected with a reasonable distance. Edges that are not confirmed based on their connectivity and distance measurements are removed. If this removal results in isolated nodes, then the nodes are also removed.

Chimeric maps have a distinctive appearance in that they consist of two groups of nodes only connected via a single node. The region is not locally connected to any other region of the graph. Identification of chimeric maps is accomplished by using a breadth first search identifying nodes that locally disconnect the graph when removed. The chimeric node and all edges connected to it are removed. Isolated nodes or node groupings are also removed.

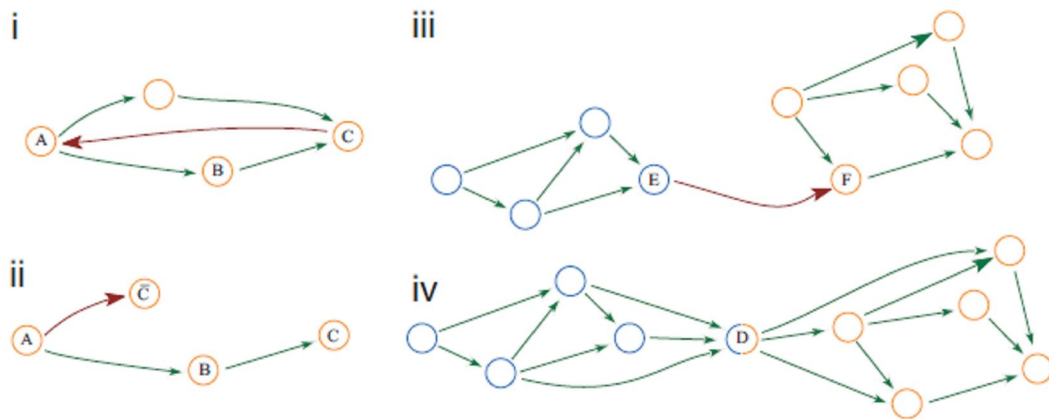


Figure 8: This figure demonstrates the four types of error corrected for in the overlap graph. i. This digraph demonstrates the presence of a spurious edge (red) causing a cycle, which is typically not allowed in linear genomes. ii. This digraph demonstrates an orientation inconsistent false edge (red) that is consistent with the orientation of the rest of the graph. iii. This digraph demonstrates an orientation consistent false edge (red) that connects two unrelated components of the genome. iv. This digraph demonstrates a chimera that results in the connection of two unrelated components of the genome⁴.

IV. Identification of Islands

After graph correction, the process next breaks the overlap graph into multiple components. These components are termed islands and are representations of the genomic regions spanned by overlapping optical maps. For each of the islands, we need to extract the genomic region corresponding to the island, a contig.

V. Contig Construction

Within each of the islands, contigs can be defined as paths from sources to sinks within the island subgraph of the overlap graph. To produce the most comprehensive representation of the genomic region of interest, the heaviest cycle-free path is chosen with the weight of the paths corresponding to the genomic distances assigned to the edges. This is accomplished by first identifying the sources and then performing a depth-first search from each source. The longest path based on the edge weights is assigned to each discovered node⁴. The heaviest path is found by finding the node with largest weight and the path ending at that node with the largest weight. This path is used to define a contig for the given island.

VI. Construction of Draft Consensus Map

Using the determined path, the nodes and edges are used to merge together the individual optical maps corresponding to each island. Each of these individual composite optical maps is stored for further analysis. The result is a draft consensus maps corresponding to each island as shown in Figure 9.

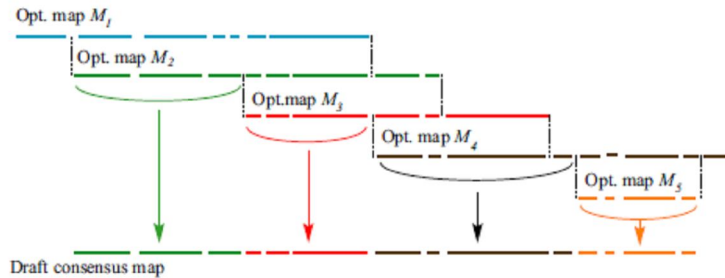


Figure 9: A draft consensus map is constructed by combining the regions of overlap from optical maps corresponding to each island. The final result is a composite of multiple optical maps representing a given island⁴.

VII. Consensus Map Refinement

Since the draft map is formed from individual optical maps, it may contain errors like missing cuts, false cuts, and inaccurate fragment sizing⁴. Large numbers of optical maps can be used to correct the discrepancies exhibited by a single draft map. This is accomplished by means of hypothesis testing, where optical maps are aligned with a draft consensus map to identify possible positions of additions or deletions in the graph structure. Fragment size re-estimation is accomplished by taking averages of distances from the draft consensus map and comparing them to average distances from other optical maps. These are iterative procedures that proceed until no further corrections can be made, typically requiring 13-15 iterations. The corrected consensus maps corresponding to each island can be pieced back together to establish the entire restriction map for the genome. We will now briefly look at a method of graph refinement that has been applied to the optical mapping algorithm.

The three major errors addressed in the refinement process are sizing errors, missing cuts, and false cuts⁷. The sizing errors are addressed using fluorescence intensity information and are represented with a normal distribution. A recorded DNA fragment of size Y is estimated to be size X for correction using a σ value of 0.6 giving normal distribution of $X \sim N(Y, \sigma^2 Y)$. The missing cuts result from a restriction enzyme inefficiency and are modeled using a Bernoulli event with a probability of success set to $p = 0.8$. The false cut errors are generally caused by DNA breakage not corresponding to restriction digestion. This process is assumed to be uniform and to follow a Poisson distribution with the rate $\zeta = 0.005 \times \text{Kb}^{-1}$.

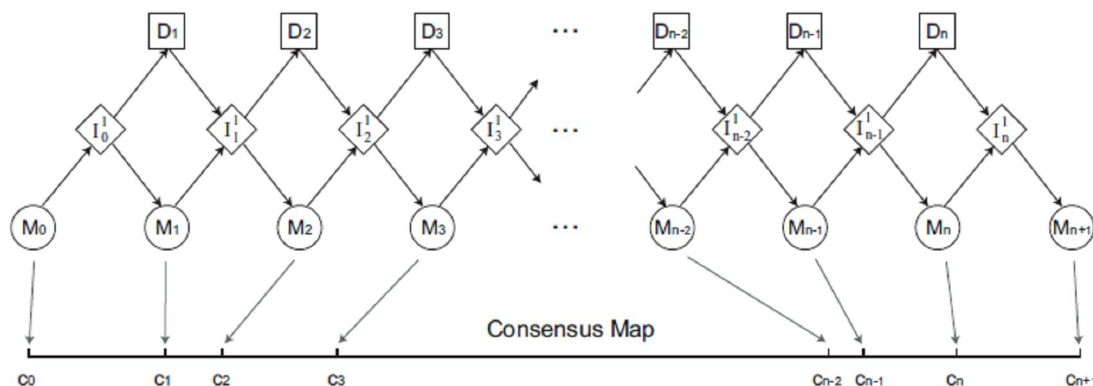


Figure 10: This is the Hidden Markov Model used in the mesh refinement process. The cut sites are represented by c_i on the consensus map. Each cut site has two corresponding states: a delete (D) and match (M) state. These states correspond to a missing cut and a matching cut, respectively. Between each cut site an insertion (I) state is represented, which represents the presence of an additional cut site. When an optical map is compared to the consensus map, this model is used to remove errors from the consensus map⁷.

A Hidden Markov Model (HMM) is used to address these two types of errors, where the HMM represents the consensus map that has been previously constructed⁷. By comparing optical maps to the HMM, we are able to refine our optical map. Valouev, et al describes the HMM as having n potential cut sites on the consensus map s_c each labeled as c_i ⁷. The model includes three different states: match, insertion, and deletion. The match state (M_i) has $n + 2$ components corresponding to each restriction site and the beginning and end of the consensus map. If the optical map under investigation contains a cut site corresponding to the c_i on the consensus map, then it passes through M_i at that position of the HMM. The delete state (D_i) has n components corresponding to each potential restriction site that can be missed on the consensus map. If the optical map under investigation is missing cut site c_i , then it passes through D_i at that position of the HMM. The insertion state (I_i^1) has $n+1$ components corresponding to potential insertion sites between known cut sites on the consensus map. If the optical map under investigation contains an extra cut site between c_i and c_{i+1} in the consensus map, then it passes through I_i^1 at location between c_i and c_{i+1} . Assuming that the optical maps being compared to the consensus map via the HMM are accurate, the match state represents restriction sites, the delete state represented a missed cut site, and the insert state represents an additional cut site. Figure 10 demonstrates the whole HMM and Figure 11 demonstrates a path through a HMM and its corresponding consensus map.

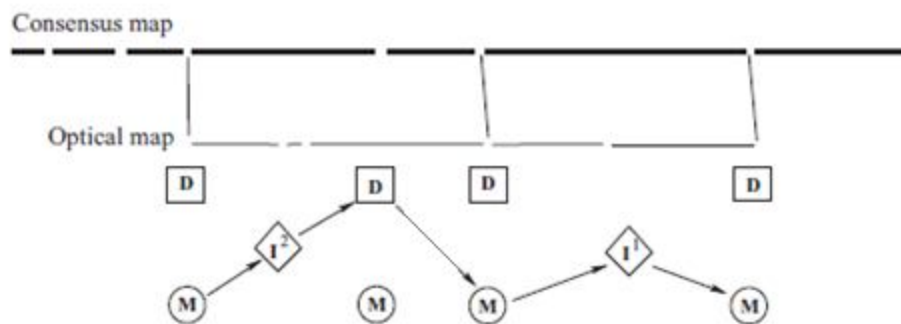


Figure 11: This figure shows the comparison between an optical map and a consensus map and the resulting path through the Hidden Markov Model where cut sites are deleted, added, and matched based on the comparison⁷.

For the HMM to be useful, we must align optical maps to the consensus map for the model to begin to correct the consensus map. This is performed by aligning the cut sites and fragment lengths between the two maps⁷. After alignment, the comparison between the two maps is represented as a path through the HMM with modifications being made along the way depending on whether or not deletions and insertions are present. A statistical analysis is involved in determining whether or not to include insertions and deletions in the consensus map, and is more appropriately discussed elsewhere. The re-estimation of fragment size is also based on a statistical analysis, and is also more appropriately discussed elsewhere. The reader is directed towards Valouev A, Zhang Y, Schwartz DC, Waterman MS (2006) for further description of the statistical analysis involved in this process⁷. Essentially, depending on the results of the statistical analysis for each potential modification, a given cut site will either be included or disregarded. Typically, it has been shown to take 13-15 iterations to reach an equilibrium state for the consensus map. The final result is a refined/corrected consensus map that represents a restriction map for the genome of interest.

APPLICATIONS OF OPTICAL MAPPING

The utility of generating whole genome restriction maps without prior sequence information is not limited to a single application. As aforementioned, restriction maps are useful in identifying genetic insertions, repeats, inversions, and deletions³. It also offers a means to establish genotype-phenotype correlations for clinical medicine and advance both the diagnosis and treatments of many human diseases. In addition to providing whole genome maps for humans, rice, and a variety of bacteria, optical mapping has already provided insight into disease mechanisms related to BRCA1/2, human chromosome 22, the Beckwith-Wiedman locus, and the human mitochondrial genome^{1,3}. The huge number of potential human diseases creates a great opportunity for a seemingly limitless number of potential applications. As the technology continues to improve, miniaturization will lead to greater accuracy, reduced costs, and increased throughput.

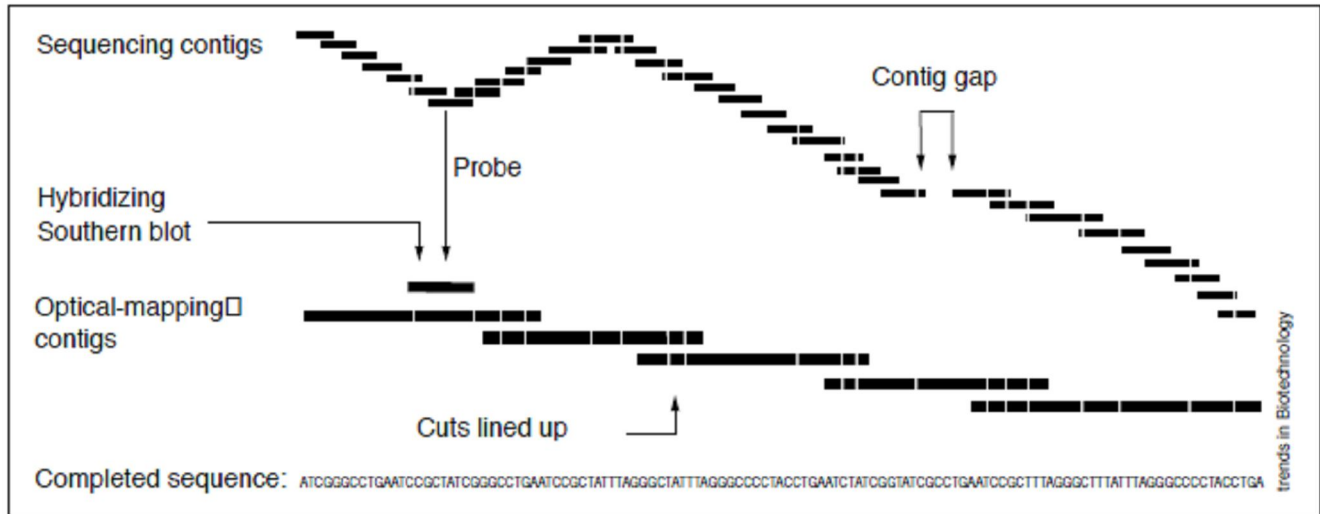


Figure 12: This figure demonstrates the ability of an optical map to be used as a scaffold for sequencing contigs⁵. Notice how the optical map covers much larger regions of DNA and offers a means to align the much smaller sequence contigs. Also notice how an optical map can be used to identify gaps and align sequencing contigs on either side of the gap. The figure also shows a hybridizing southern blot which is simply a validation tool. The complete sequence shown on the bottom simply emphasizes that in this application, a finalized sequence is the ultimate goal and optical mapping can serve as a guide to align the much smaller sequence contigs thereby improving the speed and accuracy of the sequencing effort.

Aside from genetic disease analysis, optical mapping also offers a potential improvement to modern sequencing efforts⁵. The cost of sequencing in recent years has reduced with advancements in technology, which has brought us closer to whole genome analysis on a patient-specific basis. The science of sequencing is better addressed elsewhere, but limitations in the process are very real in its current form. PCR based sequencing can often contain gaps between overlapping contigs as demonstrated in Figure 12. These gaps make it difficult to align the contigs for sequence construction without prior knowledge of the gap location or sequence³. Ordered restriction maps generated from optical mapping without prior sequence knowledge can be useful in aligning sequencing contigs to ensure sequence integrity throughout the sequencing process⁵. Thus optical mapping offers a scaffold for DNA sequencing that could improve contig alignment efficiency and accuracy; thereby, further reducing the cost and time required for PCR based sequencing approaches.

CONCLUSIONS

We have presented a comprehensive review of the published literature related to optical mapping technology and its data analysis algorithms. Optical mapping is a method of restriction map generation for whole genome analysis using information attained from individual DNA molecules. This multidisciplinary technology incorporates components from biochemistry, molecular biology, genetics, BioMEMS, discrete mathematics, statistics, and computer science. The process spans from preparation of an optical mapping sample to imaging to data analysis. In the analysis of the data, graphing is used as a method of restriction map construction and error correction. A Hidden Markov Model is used for further error reduction and restriction map refinement. The final product is a highly accurate restriction map that can be used for a variety

of applications ranging from clinical phenotype-genotype correlations to identification of polymorphisms in a variety of diseases including cancer. In the future, optical mapping technology will help to realize the goal of patient-specific whole genomic analysis.

REFERENCES

1. Samad A, Huff EF, Cai W, Schwartz DC. Optical mapping: A novel, single-molecule approach to genomic analysis. *Genome Res.* 1995;5:1-4.
2. Ramme AJ. Personal image collection. .
3. Schwartz DC, Samad A. Optical mapping approaches to molecular genomics. *Curr Opin Biotechnol.* 1997;8:70-74.
4. Valouev A, Schwartz DC, Zhou S, Waterman MS. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc Natl Acad Sci U S A.* 2006;103:15770-15775.
5. Aston C, Mishra B, Schwartz DC. Optical mapping and its potential for large-scale sequencing projects. *Trends Biotechnol.* 1999;17:297-302.
6. Valouev A, Li L, Liu YC, et al. Alignment of optical maps. *J Comput Biol.* 2006;13:442-462.
7. Valouev A, Zhang Y, Schwartz DC, Waterman MS. Refinement of optical map assemblies. *Bioinformatics.* 2006;22:1217-1224.