Genomics via Optical Mapping III: Contiging Genomic DNA and Variations

(Extended Abstract)

Thomas Anantharaman, Bud Mishra and David Schwartz[1]

Abstract

In this paper, we describe our algorithmic approach to constructing an alignment of (*contiging*) a set of optical maps created from the images of individual genomic DNA molecules digested by restriction enzymes. Generally, these DNA segments are sized in the range of 1–4Mb. The problem of assembling clone contig maps is a simpler special case of this contig problem and is handled by our algorithms. The goal is to devise contiging algorithms capable of producing high-quality composite maps rapidly and in a scalable manner. The resulting software is a key component of our physical mapping automation tools and has been used routinely to create composite maps of various microorganisms (E. coli, P. falciparum and D. radioduran). The experimental results appear highly promising.

1 Introduction

Single molecule approaches provide a new direction for characterizing structural and functional properties of individual DNA molecules. The use of single molecule substrates are now regarded as a powerful tool for facilitating the ongoing human genome mapping/sequencing effort. Some of these approaches (notably, optical mapping [AMS97, Ana+97b, Cai+95, Men+95, Sam+95, Sch+93, WHS95]) are emerging as formidable competitors against electrophoretic methods in terms of the accuracy and resolution, while providing accurate sizing between positions of markers. Such maps, provide ideal scaffolds for the assembly and verification of sequencing efforts, in addition to establishing deep insight into genome organization. For example, the use of restriction enzymes with CG-rich recognition motifs will mark the 5' position of approximately half of all genes. The ideal map will combine high resolution restriction enzyme maps with co-placement of traditional markers, such as STS's.

The key to enhancing the accuracy of these physical maps rely on combining many imperfect maps obtained from the copies of a single clone and an appeal to the statistics through the law of large numbers. For instance, in optical mapping approach, through a judicious choice between the control of the error processes (through biochemical manipulations, and surface conditions) and the model of the error processes (both physically and statistically), it has been possible to devise a Bayesian algorithm capable of automatically producing accurate maps of moderate size clones (e.g., BAC, Bacterial Artificial Chromosome) [AMS97].

In this paper, we pose the following question: Assuming that optical mapping provides only a "single physical map" (thus, possibly imperfect, and may not have been improved via reliance on a sample population), can such maps be used in constructing an alignment of a set of single molecules (*the contig problem*)? Note that the problem of contiging clone based maps is a simpler special case of this contig problem.

[1] *Authors' Current Address: Courant Institute, New York University, 251 Mercer St, NYC, NY-10012.* *The research presented here was partly supported by an NSF Career grant: IRI-9702071, an NIH Grant: NIH R01 HG0025-07 and grant from the* **Chiron Corporation**.

In particular, let us assume that we are given a single large DNA *segment* (say, between $1Mb$ and $4Mb$), for which we can create an "imperfect" ordered restriction map by the single molecule approach, e.g., optical mapping. Thus, the map we produce may have false negatives (missing restriction cut sites as a result of partial digestion), false positives (false optical cut sites produced by the image processing algorithm), sizing error (variations in the distance between the actual physical markers), error in assigning an orientation to the DNA segment and error due to loss of few small restriction fragments, etc. With the current technology available in our laboratory, it is possible to create such "imperfect maps" (with the error processes suitably tamed) for a large number of DNA segments with high throughput. For instance, it was possible to "map" more than 100 segments of length $700Kb$ to $1.4Mb$ from *Deinococcus radioduran* in slightly more than a month. The resulting maps had a digestion rate exceeding 0.7, a relative sizing error $\approx 15\%$ and negligible false positive rate. Thus, it is important to examine whether it is practical to align these imperfect maps to create contigs and to quantify the expected quality of contigs produced by this approach.

Clearly, the contig problem as formulated here is rather interesting from both statistical and computational complexity viewpoints, as it suggests practical trade-offs between statistical error processes and the inevitable computational complexity. Thus, the approach proposed here has the potential for creating a "very rough physical map" of a genome significantly quickly. But a much stronger motivation comes from the biological side, as in some cases the standard laboratory protocol for creating maps from cloned molecules may prove to be impossible, difficult or prohibitively expensive.

Some examples of the unavoidable necessity for direct single molecular manipulation and analysis are as follows:

- The ultimate goal is to analyze large populations of genomes, at high resolution and in their entirety. The requirement of constructing and verifying libraries for large numbers of individuals is impractical, given constraints of sheer sample bulk and labor considerations.

  PCR based analysis techniques can be high throughput and accurate, but can only sample a fraction of each genome. Generally, one can only assay known loci having somewhat well-defined variances. Such analysis may not necessarily provide sufficient information for understanding the complex functionality of the genome.

- The difficulty of cloning long segments of AT-rich chromosomes, as in the case of the unicellular malaria parasite *Plasmodium falciparum*.

- Creation of libraries and their analysis always involve some degree of error. Lingering questions of clone fidelity often plague many mapping and sequencing projects. Indeed, much effort goes into verification through laborious *in situ* hybridization techniques and PCR-based analysis. Direct, high resolution mapping of genomic DNA eliminates many of these concerns, and characterizes many types of clone artifacts (notably rearrangements) and deletions.

The paper is organized as follows: In the next section (Section 2), we formulate the "genomic contig problem" and study its computational complexity. We show that the problem is NP-hard by a transformation from the Hamiltonian path problem for a cubic graph. In Section 3, we present a new overlap rule (Type D), to determine if two genomic DNA segments overlap and study the probability of introducing false positive overlap. An example in that section shows how our probabilistic analysis may be used in designing "map making" experiments. In Section 4, we present an algorithm, based on a Bayesian formulation, to contig optical maps of genomic DNA segments subject to the constraint that the false positive overlap probability does not exceed some prespecified value. We also discuss a set of heuristics in order to derive an efficient implementation. A concluding section discusses the significance of the results of the paper and indicates future research direction. An appendix summarizes experimental results for several sets of artificial data. Our experiments with two other data sets obtained from chromosome 2 of *Plasmodium falciparum* and *Deinococcus radioduran* will be described elsewhere.

2 Problem Formulation and Complexity

We are assumed to be given M intervals (genomic DNA segments)

$$D_1, D_2, \ldots, D_M,$$

each of length L and each containing n cut sites (either true restriction cut sites or false optical cut sites; sizing error is ignored in the discussion of the complexity). For instance, the cut sites on the j^{th} interval D_j are given as

$$0 < c_{j1} < c_{j2} < \cdots < c_{jn} < L.$$

In the following complexity analysis, we assume that we are given an external parameter, $p_c \in [0, 1]$ that represents the digestion rate.

Our goal is to place these M intervals on the real line by fixing the alignment (the orientation and the position of the left end) of each interval. By $\overline{D}_j$, we denote the interval D_j after it has been placed on the real line; and by Interval($\overline{D}_j$), the interval spanned by D_j after it has been placed. For any such placement of the intervals, every connected subinterval of the union of the placed intervals (i.e., $\bigcup$Interval($\overline{D}_j$)) is an *island*; any island that is not a singleton interval is a *contig*. A placement is *admissible* if the union of the placed interval is connected (i.e., there is only one contig).

For any placement we define a composite map

$$0 < m_1 < m_2 < \cdots < m_K,$$

such that there is a cut at position m_i in the composite map iff the fraction of the placed intervals straddling m_i ($m_i \in$ Interval($\overline{D}_j$)) that has a cut at m_i ($m_i \in \overline{D}_j$) exceeds the parameter p_c.

$$\frac{|\{m_i \in \overline{D}_j\}|}{|\{m_i \in \text{Interval}(\overline{D}_j)\}|} > p_c.$$

Notice that every admissible placement induces a permutation of the intervals $\overline{D}_{\pi(1)}$, $\overline{D}_{\pi(2)}$, ..., $\overline{D}_{\pi(M)}$ determined by the placement of the left ends of the intervals (with any reasonable rule for tie-breaking). Define a metric of goodness for an admissible placement by

$$\chi(\overline{D}_1, \overline{D}_2, \ldots, \overline{D}_M) = \min_{1 \le j < M} |\{m_i \in \overline{D}_{\pi(j)} \cap \overline{D}_{\pi(j+1)}\}|.$$

We are interested in exploring the following decision problem:

GENOMIC CONTIG (GC) PROBLEM:

Given: M intervals D_1, D_2, ..., D_M, each of length L and each containing n cut sites; a rational number $p_c \in [0, 1]$; and a desired goodness given by a natural number $k > 3$.

Determine: If the intervals allow an admissible placement such that

$$\chi(\overline{D}_1, \overline{D}_2, \ldots, \overline{D}_M) \geq k.$$

Theorem 2.1 *The problem GC is NP-hard.*

Proof. We give a simple transformation from Hamiltonian path problem restricted to a cubic graph [GJ79]. Given a cubic graph $G = (V, E)$, with $|V| = M$ and $|E| = 3M/2$, we create M intervals, one for each vertex as follows: Corresponding to vertex v_j, we create an interval $D_j = [0, 18]$ that has exactly $k(> 3)$ cut sites in each of the subintervals $[3, 4]$, $[4, 5]$, $[5, 6]$, $[12, 13]$, $[13, 14]$ and $[14, 15]$. Let the three edges incident at v_j be

$$e_{j_1} = v_j v_1^j, \quad e_{j_2} = v_j v_2^j \quad \text{and} \quad e_{j_3} = v_j v_3^j.$$

Let

$$x_{j_1} = \frac{1}{k}\left(1 - \frac{1}{2M}\right)^{j_1}, \quad x_{j_2} = \frac{1}{k}\left(1 - \frac{1}{2M}\right)^{j_2}, \quad \text{and} \quad x_{j_3} = \frac{1}{k}\left(1 - \frac{1}{2M}\right)^{j_3}.$$

The cut locations are then

$$\begin{aligned} D_j \quad = \quad & \Big(3 + x_{j_1}, 3 + 2x_{j_1}, \ldots, 3 + kx_{j_1}, \\ & 4 + x_{j_2}, 4 + 2x_{j_2}, \ldots, 4 + kx_{j_2}, \\ & 5 + x_{j_3}, 5 + 2x_{j_3}, \ldots, 5 + kx_{j_3}, \\ & 12 + x_{j_1}, 12 + 2x_{j_1}, \ldots, 12 + kx_{j_1}, \\ & 13 + x_{j_2}, 13 + 2x_{j_2}, \ldots, 13 + kx_{j_2}, \\ & 14 + x_{j_3}, 14 + 2x_{j_3}, \ldots, 14 + kx_{j_3}\Big). \end{aligned}$$

We choose the desired goodness to be k and $p_c = \frac{3}{4}$.

Suppose G has a Hamiltonian path from v_1 to v_M which may be assumed to be (after suitable renumbering)

$$v_1, v_2, \ldots, v_M.$$

It is fairly straightforward to create an admissible placement of D_1 through D_M such that at any location at most two placed intervals Interval($\overline{D}_j$) and Interval($\overline{D}_{j+1}$) overlap. Furthermore, the composite map contains exactly the k cuts in $\overline{D}_j \cap \overline{D}_{j+1}$ and correspond to the edge $v_j v_{j+1}$ in the Hamiltonian path. Thus for this admissible placement

$$\chi(\overline{D}_1, \overline{D}_2, \ldots, \overline{D}_M) = k,$$

as desired.

Conversely, we need to show that if $D_1, \ldots, D_M$ allow an admissible placement $\overline{D}_1, \ldots, \overline{D}_M$ with a goodness of k or higher, then the resulting permutation π induced by the positions of the left ends of the placed intervals gives a Hamiltonian path

$$v_{\pi(1)}, \ldots, v_s = v_{\pi(j)}, v_t = v_{\pi(j+1)}, \ldots, v_{\pi(M)}.$$

Suppose that it is not a Hamiltonian path, i.e., for some j, $v_s = v_{\pi(j)}$ and $v_t = v_{\pi(j+1)}$ are nonadjacent in G. Then it is rather easy to see that $|\{\overline{D}_{\pi(j)} \cap \overline{D}_{\pi(j+1)}\}| \leq 2$; thus, contradicting the assumption that the initial placement has a goodness of k or more. □

3 Probabilistic Analysis

In spite of the pessimistic results of the preceding section, it may still be possible to contig the DNA segments, by exploiting the statistical structure of the (imperfect) ordered restriction maps and by making allowance for some false negatives in overlap (two maps that really overlap may not be in the same island in the computed placement). However, our goal will be to minimize false positive overlaps so that computed placement almost never wrongly overlaps two disjoint segments. We study the false positive overlap probability with respect to a somewhat relaxed overlap rule (taking into account some number of missing or false cuts in the ordered restriction maps).

The steps of the procedure are as follows:

1. Let G be the length of the genome. Let M denote the number of uniformly randomly chosen DNA segments each of length L that are further analyzed by optical mapping and produce the optical maps $D_1, \ldots, D_M$, as before.

2. Assume that each optical map is created with respect to one 6-cutter restriction endonuclease, say A (e.g., Ava I) or B (e.g, BamH I), etc. Since this optical map is created for a genomic DNA (without cloning), the map is subject to some missing cuts (with partial digestion rate $p_c < 1$) and some sizing error of the fragments (resulting in a relative error of β). We assume that the false cut rate is sufficiently small that its effect can be safely ignored.

3. Using these optical maps, we next contig the genomic DNA segments by detecting possible overlaps among the fragments. We use the following rule (called **type D**) to detect possible overlaps: Given two DNA segments, D_1 and D_2, we say they overlap if k or more of the restriction fragments align (subject to the sizing error) positionally. D_1 and D_2 are allowed to have other restriction fragments in the overlap region that may not align positionally. The allowed mismatches are accounted for by the partial digestion process.

4. While comparing two restriction fragments of lengths x and y, we say they match if

$$(1-\beta)x \leq y \leq (1+\beta)x.$$

The placement of the DNA segments are then determined by the overlap information determined according to the type D overlap rule.

5. Note that the expected number of restriction fragments per DNA segment is $n = Lp_c p_6$, where for a six-cutter enzyme $p_6 \approx 1/4000$. This value will vary enormously with the specific sequence, since most genomes have compositional biases and a granular structure often consisting of large families of repeated sequences. We define the overlap threshold ratio $\theta = T/L$, and only wish to detect those overlaps where the overlap region between two DNA segments exceed the length T. If D_1 and D_2 truly overlap, then on the average roughly $L\theta p_6 p_c^2 = k$ of the true restriction fragments should have their ends completely digested and thus k or more restriction fragments must align positionally. This is the parameter k chosen for the type D overlap rule.

Next we compute the probability that two randomly chosen DNA segments give rise to a false positive overlap (i.e., a declared overlap is a false positive) under the type D matching rule. Let D_1 and D_2 be two randomly chosen DNA segments, then

$$Pr[\text{There are exactly } k \text{ matches}] = 4/p_c^4 \binom{n}{k} (\beta/2)^k (1-\beta/2)^{n-k} \approx 4/p_c^4 e^{-\beta n/2} (\beta n/2)^k / k!.$$

Thus the probability of a false positive overlap is simply $4/p_c^4 e^{-\beta n/2} \sum_{i=k}^{\infty} (\beta n/2)^k / k!$.

Comparing this result, with similar results for fingerprint data (Type A/B) and for perfect ordered restriction map data (Type C) [LW88], we see Type D overlap rule provides significantly better overlap information than type A/B but somewhat worse than type C, for the same values of β, k, etc.

3.1 Example

In an experiment, we have studied the effect of this process on a chromosome of *Deinococcus radioduran*, with a chromosomal length of $G \approx 3Mb = 3 \times 10^6 b$. Starting with $N \approx 100$ copies of the chromosomes from a particular strain of the organism, we create a collection of DNA segments by breaking the chromosomes by mechanical shearing force and collecting those segments of length $\in [L, U] = [450Kb, 700Kb]$. For instance, if we assume that $p = 1/(5 \times 10^5)$ is the probability of breaking the DNA at a random location then we expect to get roughly $M = NGpf \approx 110$ segments, where the fraction f is given as

$$f = (pL+1)e^{-pL} - (pU+1)e^{-pU} = 0.181.$$

Note that the coverage $c = LM/G = 18$ and by Clarke-Carbon analysis [CC76], we see that the represented amount of genome in the collection of DNA segments is $1 - e^{-c} = 1 - 1.5 \times 10^{-8} \approx 100\%$.

Assume that the optical maps are created with the digestion rate $p_c = 0.5$ and a sizing error of $\beta \approx 10\%$. The expected number of restriction fragments in a DNA segment is $n = Lp_6 p_c \approx 60$ (with a six-cutter enzyme). Choosing a threshold ratio of $\theta = T/L = 0.5$, we see that we may choose a $k = L\theta p_6 \mu^2 = 15$ for the type D overlap rule. Thus the probability that a declared overlap is a false positive is given by $64 \sum_{i=k}^{\infty} e^{-3} 3^k / k! \approx 0.004\%$. Thus the contig created by our process can be accepted to be the "correct one" with very high confidence.

Note also that with an overlap threshold ratio of $\theta = 0.5$, we have now an effective coverage of $c\sigma = 9$, where $\sigma = 1 - \theta = 0.5$. Thus the expected number of contigs is $\lceil Me^{-c\sigma}[1 - e^{-c\sigma}] \rceil = 1$, the best one can hope for.

Further note that the expected number of DNA segments covering any restriction fragment of the genome is $R = (L - L_6)M/G = 18$, where $L_6 = 4000$ is the expected length of a true restriction fragment. As a result, we see that any particular restriction fragment of the genome occurs among the optical maps $Rp_c^2 \approx 4$ times (on the average) and thus the accuracy to which this particular restriction fragment can be sized in the complete genome-wide map improves to $\beta/\sqrt{Rp_c^2} = 5\%$.

We can repeat the procedure above, for another enzyme B. Now we have two genome-wide complete maps with respect to enzymes A and B—call them $\mathcal{M}_A$ and $\mathcal{M}_B$. In this step we would like to align $\mathcal{M}_A$ and $\mathcal{M}_B$ to create a multiple-digestion map $\mathcal{M}_{AB}$. Thus we need to orient (clock-wise, CW, or anti-clock-wise, ACW) $\mathcal{M}_A$ and $\mathcal{M}_B$ and then find a location where $\mathcal{M}_A$ and $\mathcal{M}_B$ must align. A simple way to achieve this would be to choose a random DNA segment L of length $500Kb$ and double-digest L with respect to A and B. Now we can find a position in $\mathcal{M}_A$ where L matches by using a simple variant of the type D matching rule. We can also do the same step for $\mathcal{M}_B$ and combine the maps. However, there is a small probability that this process may result in a false map alignment and can be computed as before. The result for exactly k of the fragments matching is given by $8/p_c^4 e^{-\beta n_{AB}/2}(\beta n_{AB}/2)^k/k!$, where the constant 8 accounts for all orientations (2 for $\mathcal{M}_A$ with respect to $\mathcal{M}_B$ and 2 for L) as well as false match with respect $\mathcal{M}_A$ or $\mathcal{M}_B$. Thus the false positive overlap probability can be computed and is negligibly small for any reasonable value of k (say 15). If necessary this last step can be repeated with more than one random DNA segments (L_1, L_2, ...) to reduce the error probability to as low a value as desired.

4 Algorithms

4.1 Scoring Functions

We begin with the description of a scoring function to compare different possible placements and a heuristic algorithm for finding the best scoring placement. The input to our algorithm is the set of maps (intervals) to be contiged and a parameter denoting a maximum allowable false positive overlap probability, specifying the worst-case probability that the final placement contains overlaps of maps whose DNA's do not in fact overlap.

The scoring function for a proposed contig has two components:

1. A Bayesian probability density estimate for the proposed placement, yielding a measure of goodness of fit.

2. An upper bound estimate of the false positive overlap probability that two unrelated pieces of DNA could have produced a Bayesian score as good as in the proposed placement.

The object is to maximize the Bayesian probability density subject to never creating contigs whose false positive overlap probability exceeds the threshold specified by the user. Our algorithm achieves this by repeatedly combining the two islands that produce the greatest increase in probability density, excluding any contigs whose false positive overlap probability is unacceptable.

4.2 The Bayesian Probability Density Estimate

The Bayesian probability density estimate for a proposed placement is an approximation of the probability density that the two distinct component maps could have been derived from the proposed placement as a result of various data errors. The data errors we model include *sizing errors*, *missing restriction cut sites*, and *false optical cuts sites*. First we hypothesize the most likely placement, then compute the probability density for the mismatch errors in the component maps given the hypothesized placement and our error model. The second step of our algorithm is similar to the Bayesian probability density computation in [AMS97], except that we approximate the computation to some extent, since speed is critical for this application. In particular, we approximate the following:

1. The hypothesized placement is computed by a simple averaging of the contiged fragment sizes, rather than a true Bayesian probability density maximization with fragment sizes as parameters.

2. Good estimates of the error model parameters are assumed to be known a priori, but further improved by a reestimation from the data using a limited number of iterations of a true Bayesian probability density maximization.

Where the input data consists of high quality optical maps based on clone data and estimated by the Bayesian procedure in [AMS97], these approximation will rarely have an effect on the computed contig. When the input data consists of genomic optical maps computed from single instances of DNA fragments the approximation may change the computed set of islands. But, by using a strict enough false positive overlap probability threshold, only the best data will be contiged together, and hence the resulting island(s) should give a good estimate of those parts of the actual contig supported by the best data.

The posterior conditional probability density for a hypothesized placement $\mathcal{H}$, given the maps, consists of the product of a prior probability density for the hypothesized placement and a conditional density of the errors in the component maps relative to the hypothesized placement. Let the M input maps to be contiged be denoted by data vectors D_j $(1 \leq j \leq M)$ specifying the restriction site locations and enzymes. Then the Bayesian probability density for $\mathcal{H}$, given the data can be written using Bayes rule as in [AMS97]:

$$f(\mathcal{H}|D_1 \dots D_M) = f(\mathcal{H}) \prod_{j=1}^{M} f(D_j|\mathcal{H}) / \prod_{j=1}^{M} f(D_j) \propto f(\mathcal{H}) \prod_{j=1}^{M} f(D_j|\mathcal{H}).$$

Note that for any reasonable error model, the probability density of the second term monotonically decreases as more and more maps are contiged, since the number of mismatches increases as more maps are contiged (overlapped). Thus the only way the probability density for a contig could be better than the probability density of its individual components is, if there is a strong prior bias in favor of producing more overlaps, reflected in the first component $f(\mathcal{H})$ (the prior density of $\mathcal{H}$).

We approximate the prior probability $f(\mathcal{H})$ as a decreasing function of the total contig length. In particular, we set the logarithm of $f(\mathcal{H})$ to be proportional to $-(KX)$, where X is the total length of contig hypothesis $\mathcal{H}$, and K is a constant. A larger K corresponds to a greater bias in favor of smaller total contig length (with greater overlaps). Note that, for good data the final contig should be stable over a fairly wide range of K values, since the errors due to excessive (and therefore incorrect) overlaps should be much greater than for correct overlaps. Thus, it is possible to find this region by increasing K gradually until the islands remain unchanged for several successive K values. On the other hand, with some knowledge of the total length of the islands, one can adjust K until the computed total length is approximately of this length.

The conditional probability density function $f(D_j|\mathcal{H})$ depends on the error model used. We model the following errors in the input data:

1. Each orientation is equally likely to be correct.

2. Each fragment size in data D_j is assumed to have an independent error distributed as a Gaussian with standard deviation σ.

3. Each input map D_j may have a proportionate scaling error (i.e., all fragments of D_j are either too large or too small by the same proportion), which is uniformly distributed over some range $[1 - V..1 + V]$ where $V < 1$ is a known maximum scaling error.

4. Missing restriction sites in input maps D_j are modeled by a probability p_c of an actual restriction site being present in the data.

5. False restriction sites in the input maps D_j are modeled by a rate parameter p_f, which specifies the expected false cut density in the input maps, and is assumed to be uniformly and randomly distributed over the input maps.

The Bayesian probability density components $f(\mathcal{H})$ and $f(D_j|\mathcal{H})$ are computed separately for each island of the proposed placement and the overall probability density is equal to their products. For computational convenience, we actually compute a *penalty function*, Λ, proportional to the logarithm of the probability density as follows:

$$f(\mathcal{H}) = \left(\prod_{j=1}^{M} \frac{1}{(\sqrt{2\pi}\sigma)^{n_j}} \right) \exp(-\Lambda/(2\sigma^2)).$$

Thus the prior component of Λ for each island is simply $2KX\sigma^2$, where X is the length of the island, and K a constant as discussed earlier. Thus, a typical value for K is simply the inverse of the average fragment length. The other components of Λ can be computed for each island and then summed up.

For fragment sizing errors, consider each map fragment of the proposed island, and let the map fragment be composed of overlaps from several maps of length $r_1, \ldots, r_N$. If $p_c = 1$ and $p_f = 0$ (the ideal situation), it is easy to show that the hypothesized fragment size μ and the penalty Λ are:

$$\mu = \frac{\sum_i^N r_i}{N}, \quad \text{and} \quad \Lambda = \sum_{i=1}^{N} (r_i - \mu)^2.$$

In addition if there is an overlap of this fragment by a partial map fragment of length $r_p > \mu$, we add an additional penalty to Λ equal to $(r_p - \mu)^2$ for the largest such partial map fragment for each contig fragment.

This ignores the extra degrees of freedom introduced if each map can be scaled proportionately over the range $[1-V..1+V]$, which tends to reduce Λ. A conservative way to compensate for the extra degree of freedom is to increase the additional penalty (increase in Λ) from each overlap by the ratio

$$\frac{\text{(number of overlapped fragments)}}{\text{(number of overlapped fragments} - 1)}.$$

In particular, if the number of overlapped fragments is one for any overlap in an island, the contig is disallowed, as in this case, any pair of fragment sizes can be made to match by suitable scaling.

Now consider the presence of missing cuts (restriction sites) when $p_c < 1$. To model the multiplicative error of p_c for each cut present in the contig we add a penalty $\Lambda_c = 2\sigma^2 \log[1/p_c]$ and to model the multiplicative error of $(1-p_c)$ for each missing cut in the contig we add a penalty $\Lambda_n = 2\sigma^2 \log[1/(1-p_c)]$. The alignment determines which cuts are missing, and the method for finding the best alignment is described later.

The computation of μ is modified in the case of missing cuts by assuming that the missing cuts are located in the same relative location (as a fraction of length) as in overlapping maps that do not have the corresponding cut missing. Also, the penalty for partial map fragments is modified in order not to exceed the Λ_n, since it is now possible that a real restriction cut site is missing in the partial map fragment.

Finally, consider the presence of false optical cuts when $p_f > 0$. For each false cut (as determined by the alignment in the proposed placement) we add a penalty $\Lambda_f = 2\sigma^2 \log[1/(p_f\sqrt{2\pi}\sigma)]$ in order to model a multiplicative penalty of p_f and the absence of one $1/(\sqrt{2\pi}\sigma)$ term normally present in the Gaussian error term.

Additionally, the penalty for the partial map fragments must now be bounded by the possibility that the partial map fragment is correct and the corresponding shorter internal fragments aligned against it all have a false cut.

4.3 The False Positive Match Probability

The score function corresponding to the false positive overlap is based on an estimate of the ratio of the probability that a random DNA would have matched at least as well or better than the actual DNA map (in terms of of the Bayesian penalty score) to the probability that random DNA would have matched at best as well or worse. If one is considering M maps to contig, and thus looking at $\binom{M}{2}$ pairs of maps to contig, the smallest score one would expect from random DNA is roughly $2/(M^2 - M)$. A conservative strategy to keep the false positive overlap probability for the best pair of maps to be contiged below the user specified level of FP is to require each contig's false positive match score S_{FP} to be below $2\text{FP}/(M^2 - M)$.

The false positive match score for a pair of maps to be contiged is estimated as follows:

Let the average fragment size be ℓ and the fragment size distribution be modeled by the exponential distribution $p(x) = \exp(-x/\ell)/\ell$. If the maps to be contiged have restriction sites for multiple enzymes, one would model each restriction enzyme by a separate value of ℓ to exploit the differences between rare cutting and frequent cutting enzymes.

First assume that $p_f = 0$ and $p_c = 1$ (the ideal situation). Consider two maps or contigs being considered for potential overlap, and let the fragment sizes in the overlap region be $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$, respectively. Also, let the two islands contain a total of N_x and N_y fragments, respectively. Then allowing for two orientations and at most $(N_x + N_y - 2N + 1)$ possible alignments with that many overlapping fragments, one can estimate an upper bound for the false positive match score S_{FP} by integrating over the ways that each pair of fragment sizes could be as close as they are by mere chance:

$$S_{\text{FP}} = 2(N_x + N_y - 2N + 1) \prod_{i=1}^{N} \frac{p(x_i, y_i)}{1 - p(x_i, y_i)},$$
$$\text{where } p(x_i, y_i) = \exp(-X_i/\ell) - \exp(-(X_i + 2D_i)/\ell),$$
$$\text{and } X_i = \min(x_i, y_i), \quad D_i = |x_i - y_i|.$$

For partial map fragments of size x_p overlapping an internal fragment of size μ, the value of S_{FP} can be further reduced, by allowing for the fact that the partial fragment has a true size of at least x_p. The detail expression for this case will be given in the full paper.

If we have missing cuts with $p_c < 1$, S_{FP} is modified as follows. Let n_x,n_y be the number of actual cuts in the overlap region of the two maps respectively, and let m be the number of those that are aligned:

$$S_{\text{FP}} = 2(N_x - n_x + N_y - n_y + 1) \prod_{i=1}^{m} \frac{p(x_i, y_i)}{1 - p(x_i, y_i)},$$
$$\text{where } p(x_i, y_i) = \textsc{Cor}(\mathbb{E}[p_c])(\exp(-p_c X_i/\ell) - \exp(-p_c(X_i + 2D_i)/\ell)),$$
$$D_i = \text{fragment size error relative to previous alignment,}$$
$$X_i = \text{smaller of distances to previous cut of same enzyme on either map, and}$$
$$\ell/p_c = \text{expected distance between cuts of same enzyme.}$$

$\mathbb{E}[p_c]$ is a local estimate of p_c, obtained by counting the number of misaligned cuts (of any enzyme type) between the current and previous cut alignment. $\textsc{Cor}(\mathbb{E}[p_c])$ is a pessimistic estimate of the number of times M consecutive identical alignments would increase the number of ways equally good alignments (with M total aligned fragments) could be achieved by chance with random DNA. This expression can be derived by considering M identical consecutive alignments such as the current one and counting the number of ways M or more restriction site alignments (other than the leftmost alignment) could be obtained by choosing M of the $M/\mathbb{E}[p_c]$ possible alignments sites of one of the molecules to align with random sites in the other molecule. Taking the M^{th} root of this number as $M \to \infty$ results in $\textsc{Cor}(\mathbb{E}[p_c])$. The expression for the partial end fragments is the similar.

If false cuts are present ($p_f > 0$), an upper bound of S_{FP} can be obtained by assuming that all false cuts are real cuts with corresponding matching cuts all missing, and increasing ℓ to $\ell/(1 + \ell p_f/p_c)$ (the new average distance between two consecutive cuts of the same enzyme).

4.4 Global Search

As mentioned earlier, the heuristic global search for the best placement is based on repeatedly combining those two islands that produce the greatest increase in Bayesian conditional posterior probability density and excluding any contig whose false positive overlap probability is unacceptable. The algorithm stops when there no longer is any pair of islands that can be combined to improve the probability density with acceptable false positive overlap probability. The final contig set is then used to estimate more accurate values for the parameter σ,p_f and p_c. If these significantly differ from the known input values, the entire global search is repeated. Our initial experiments show that only a few (limited to just two) iterations suffice for the global search, given initial values for σ, p_c and p_f.

One property of our global search heuristic is that it is greedy in combining the best island and therefore may be suboptimal if the data quality is poor. However, it should be pointed out that it is still superior to the more simple method of simply trying to contig each pair of the original maps and use the resulting directed connectivity graph to find a linear path connecting all maps. More details will appear in the full paper.

4.5 Overlapping Islands by Dynamic Programming

Finding the best offset and alignment between a pair of maps is potentially exponential in complexity since each cut site in one map could be aligned with almost any cut site in the other map. The solution is to use a dynamic programming algorithm similar to that described in [AMS97], for finding the best alignment between a single molecule map and a map hypothesis. The problem here is different since we have two maps and each possible alignment of the two maps generates a contig hypothesis which determines the Bayesian score of the two maps. The resulting algorithm has a time complexity of $O(n^4)$ in the worst case, but we have incorporated several heuristics in order to bring the average case complexity down to $O(n)$. Details of the dynamic programming algorithm is postponed to the full paper.

Combining the local and global search, we have an overall worst case complexity of $O(M^2n^4)$, where M is the total number of maps, and n is the maximum number of cut sites in any map. With various heuristics, the time complexity has been further improved in the average case to $O(M^2n)$.

5 Conclusion

In this paper, we make four contributions toward the construction of composite restriction maps from optical mapping data (the contig problem).

1. We formulate and analyze the worst-case complexity of the problem of constructing composite restriction map from individual optical map data. The underlying model as well as the complexity study has played an important role in the formulation of a Bayesian approach.

2. We analyze the probability of false positive overlap probability in the placement of optical maps, and provide simple rules in designing optical map experiments for making high-quality contigs.

3. We formulate a Bayesian algorithm for this problem that relies on several scoring functions derived from a carefully modeled prior distribution.

4. We have implemented the algorithm (in C, running on Sparc 20's) and experimented extensively. The experiments yield high-quality composite maps, consistent with the best result one can expect from the input data.

Among many remaining related unsolved problems, the most interesting one involves the situation where optical maps to be contiged are those coming from K (usually, $K = 2$, and the expected composite maps similar) populations. We believe that the analysis presented here will be helpful in providing efficient algorithms for this general problem.

Acknowledgment. Our thanks go to Chris Aston for his help and advice, Junping Jing, Jieyi Lin and Rong Qi for the genomic maps and the contigs for *Plasmodium falciparum* (chromosome 2) and *Deinococcus radioduran*, and to Ed Huff for the help with the synthetic data used in testing the software.

References

[AMS97] T.S. Anantharaman, B. Mishra and D.C. Schwartz, "Genomics via Optical Mapping II: Ordered Restriction Maps," *Journal of Computational Biology*,**4**(2):91–118, 1997.

[Ana+97b] T.S. Anantharaman et al., "Statistical Algorithms for Optical Mapping of the Human Genome," *1997 Genome Mapping and Sequencing Conference*, Cold Spring Harbor, New York, May, 1997.

[Cai+95] W. Cai et al., "Ordered Restriction Endonuclease Maps of Yeast Artificial Chromosomes Created by Optical Mapping on Surfaces," *Proc. Natl. Acad. Sci., USA*, **92**:5164–5168, 1995.

[CC76] L. Clarke and J. Carbon, "A Colony Bank Containing Synthetic ColE1 Hybrid Plasmids Representative of the Entire E. coli Gene," *Cell*, **9**:91–101, 1976.

[GJ79] M.R. Garey and D.S. Johnson, *Computer and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Co., San Francisco 1979.

[GGKS95] P.W. Goldberg, M.C. Golumbic, H. Kaplan and R. Shamir, "Four Strikes Against Physical Mapping of DNA," *Journal of Computational Biology*,**2**(1):139–152, 1995.

[LW88] E.S. Lander and M.S. Waterman, "Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis," *Genomics*, **2**:231–239, 1988.

[Men+95] X. Meng et al., "Optical Mapping of Lambda Bacteriophage Clones Using Restriction Endonuclease," *Nature Genetics*, **9**:432–438, 1995.

[MP97] S. Muthukrishnan and L. Parida, "Towards Constructing Physical Maps by Optical Mapping: An Effective Simple Combinatorial Approach." In *Proceedings First Annual Conference on Computational Molecular Biology*, (RECOMB97), ACM Press, 209–215, 1997.

[Sam+95] A. Samad et al., "Mapping the Genome One Molecule At a Time—Optical Mapping," *Nature*, **378**:516–517, 1995.

[Sch+93] D.C. Schwartz et al., "Ordered Restriction Maps of *Saccharomyces cerevisiae* Chromosomes Constructed by Optical Mapping," *Sciences*, **262**:110–114, 1993.

[WHS95] Y.K. Wang, E.J. Huff and D.C. Schwartz, "Optical Mapping of the Site-directed Cleavages on Single DNA Molecules by the RecA-assisted Restriction Endonuclease Technique," In *Proc. Natl. Acad. Sci. USA*, **92**:165–169, 1995.

[Wat95] M.S. Waterman, *An Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman Hall, 1995.

A Experimental Results

Simulated data was created for a genome of length $20Mb$ from which segments were sampled of random length in the range of 1—$3Mb$ located randomly along the genome. Restriction sites for the specific enzyme ACGTTGAC (8-cutter, non-palindromic restriction sequence, chosen at random) were located in each segment (subject to error). The restriction fragment sizes of average length $50Kb$ were further randomly scaled to produce a sizing error in the fragments with a standard deviation, $\sigma \approx 3Kb$ (equivalently, a 95% error of $\pm 12\%$). Restriction sites were randomly omitted (removed) for a simulated digestion rate of $p_c = 0.80$ and randomly located false cuts were introduced on all segments at a rate of 1 per Mb ($p_f = 10^{-6}$). The experiments were conducted on three sets of data according to the specifications described here. The number of segments in each set were varied to include 40, 80 and 160 segment maps corresponding to a coverage c of $4\times$, $8\times$ and $16\times$, respectively.

The contig program was run with a false positive overlap probability threshold of $0.01\%(= 10^{-4})$, and a K value (in the penalty function) of 1.7 per average restriction fragment overlapped. The actual number of contigs present and the computed number of contigs are shown in the following table. The largest contig was checked against the simulated genome to locate any errors, and *none was found*.

Theoretical Results			
Coverage, c	**4**	**8**	**16**
# Contigs (present)	**4**	**2**	**1**
Length of the longest contig	**12,248 Kb**	**15,599 Kb**	**19,849 Kb**
Simulated Results			
# Contigs (found)	4	2	2
# Uncontiged Maps	0	0	1
Length of the longest contig	12,251 Kb	15,601 Kb	19,963 Kb
Estimated Std. Dev., σ	2.99 Kb	2.96 Kb	3.00 Kb
Estimated p_c	0.815	0.808	0.805
Estimated p_f	0.6×10^{-6}	0.9×10^{-6}	0.9×10^{-6}