Genomics via Optical Mapping III: Contiging Genomic DNA

Thomas Anantharaman, Bud Mishra and David Schwartz
Courant Institute, New York University, 251 Mercer St. NYC, NY-10012

Abstract

In this paper, we describe our algorithmic approach to constructing an alignment of (*contiging*) a set of restriction maps created from the images of individual genomic (uncloned) DNA molecules digested by restriction enzymes. Generally, these DNA segments are sized in the range of 1–4Mb. 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 to create complete maps of various microorganisms (E. coli, P. falciparum and D. radiodurans). Experimental results match known sequence data.

The optical mapping approach (Cai et al. 1998; Anantharaman, Mishra and Schwartz 1997; Samad et al. 1995; Schwartz et al. 1993) can be used to determine an approximate restriction map (with ordering of fragments) from fluorescent microscopy images of individual DNA molecules. When the DNA molecules are derived from clones, an accurate restriction map can be obtained by combining the approximate restriction maps from a small number (50-200) of DNA molecules (Anantharaman and Mishra 1998a). We have previously described a Bayesian/Maximum-likelihood algorithm capable of automatically producing accurate maps for moderate size clones (e.g., BAC, Bacterial Artificial Chromosome) (Anantharaman, Mishra and Schwartz 1997; Cai et al. 1998; Anantharaman and Mishra 1998b).

In this paper, we extend our approach to explore if accurate restriction maps can be constructed using only genomic (uncloned) DNA. Each DNA molecule will be a random fragment of the genome. The approximate restriction maps of these DNA molecules generated by optical mapping need to be combined into a larger restriction map by computing the contig (alignment between them) from the overlaps despite the errors in the maps. At the same time errors in the single molecule restriction maps need to be eliminated by combining the information from multiple overlapped restriction maps.

In particular, we can use optical mapping to generate single molecule DNA restriction maps for random DNA *segments* say, between 1Mb and 4Mb. The individual restriction map produced from the images may have false negatives (up to 30% restriction cut sites missing as a result of partial digestion), false positives (up to 20% false optical cut sites produced by the image processing algorithm or DNA breakage), sizing errors (variations in the estimated distance between the actual restriction sites, ranging from 5-30% with an average value of 10-15%), the inability to tell the orientation of the DNA segment, and the loss of some fraction of the small restriction fragments etc. With the current technology developed in our laboratory, it is possible to create such "imperfect maps" for a large number of DNA segments with high throughput (Jing et al. 1999). For instance, we were able to map about 100 segments of length 700Kb to 1.4Mb from *Deinococcus radiodurans* and the resulting maps had a digestion rate exceeding 70%, a relative sizing error $\approx$ 15% and under 5% of all cuts observed were false positive.

A key to solving this shotgun optical mapping problem is a set of efficient algorithms for contiging individual maps with significant errors. The algorithms have been implemented in a program called Gentig and tested on real and simulated data.

The paper is organized as follows: In the next section (Section 2), we present the algorithms used in Gentig, based on a Bayesian/Maximum-likelihood formulation, to contig restriction 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 heuristic algorithms in order to derive an efficient implementation. In section 3 we present experimental results using Gentig. The final section discusses the worst-case complexity of the problem (it is NP-hard), a statistical analysis of the data under various error models (Mishra 1999), applications of our algorithms, and related open problems.

An overview of our restriction map generation process is illustrated in Figure (1).

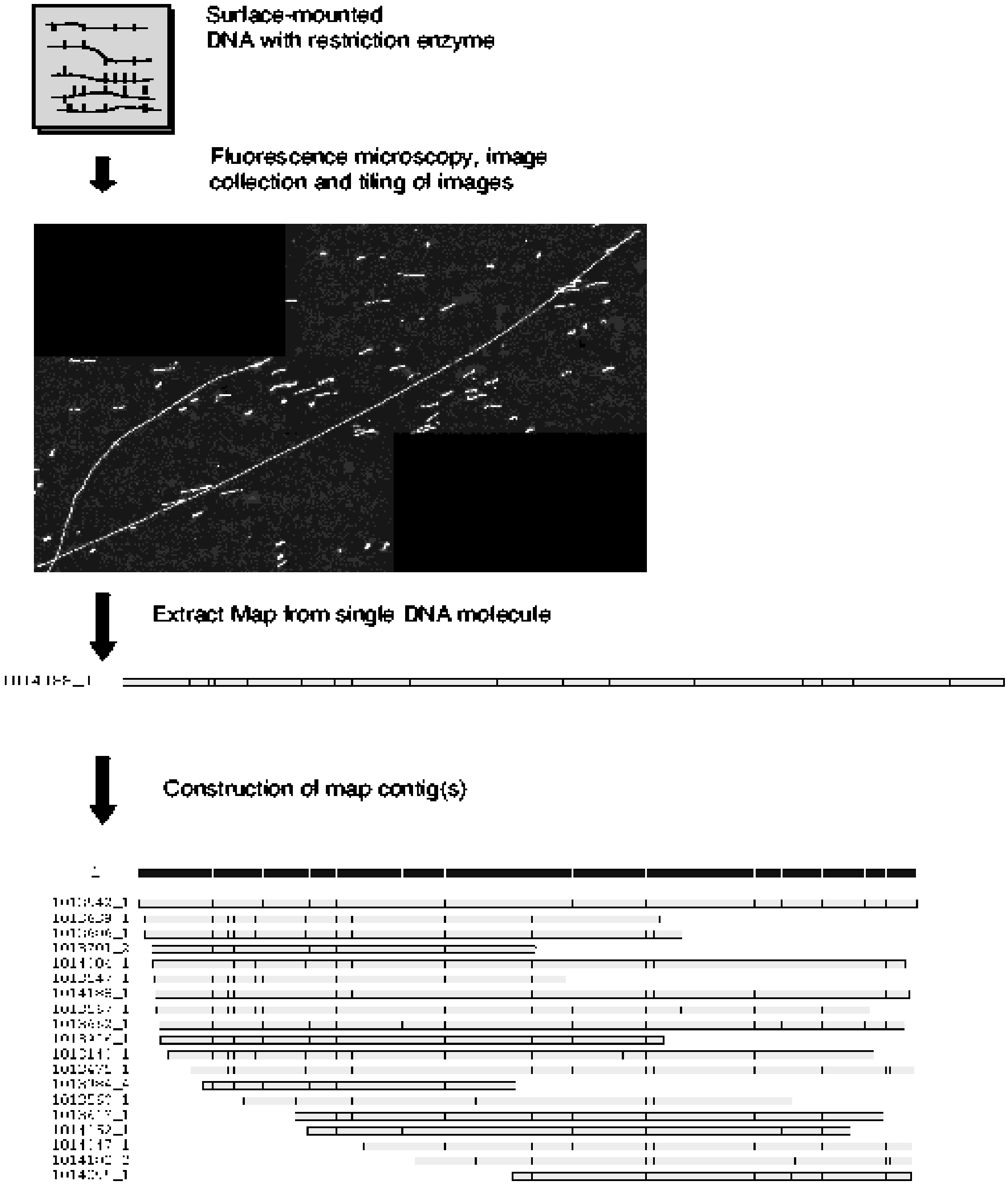

Figure 1: Shotgun Optical mapping of genomic DNA

Algorithms

Scoring Functions

We begin the description of Gentig with a probabilistic scoring function to compare different possible placements of possible restriction map overlaps and a heuristic algorithm for finding the placement with the best score. The input to our algorithm is the set of (approximate) restriction maps to be contiged and a parameter denoting a maximum acceptable *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 plan is to maximize the Bayesian probability density subject to never creating contigs whose false positive overlap probability exceeds the threshold specified by the user. We describe a number of different algorithms all of which have in common that they repeatedly combine two islands that increase the probability density while having an acceptable false positive overlap probability and continue until no further progress is possible.

Note that the primary purpose of the false overlap threshold is to allow the search of the Bayesian probability density space to proceed in two phases. In the first phase the maps and resulting contigs may have many local errors, the only thing we are sure about is the approximate location of the overlaps. For this to work either the quality of the data must be high or the size of overlapping DNA must be large (say 0.5 Mb or more of overlap). Enough data must be collected to make up for the fraction (typically 50%) of data that cannot be contiged in this phase. In the second phase the contig restriction map can be improved locally by using gradient methods to optimize the contig hypothesis using methods similar to those described in (Anantharaman, Mishra and Schwartz 1997). This would eliminate most of the errors in the input maps. The input maps discarded in the first phase could be placed probabilistically on the contig to further reduce errors, particularly sizing errors, without affecting the overall contig.

In this paper we describe the first phase of the process which is sufficient to allow an approximate restriction map to be computed with good error characteristics as shown in the results section.

The algorithms use somewhat different strategies for selecting the pairs of islands to combine next, and some allow for backtracking when no further progress is possible to see if an alternate selection strategy would produce a different/better solution.

Since speed is critical for this application, some approximations are used in the computation of the Bayesian probability which are unlikely to affect the contigs but may result in slightly suboptimal contig fragment size estimates. Once the final contig has been determined better contig fragment size estimates could be obtained by performing a more exact Bayesian probability maximization with the contig fragment sizes as parameters.

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*. The Bayesian probability density estimate is computed in two steps:

1. The most likely alignment (orientation and placement) for the two component maps is computed.
2. The best contig hypothesis corresponding to the most likely alignment is computed.

The Bayesian probability density computation here is similar to that in (Anantharaman, Mishra and Schwartz 1997) with only two maps to combine, except that we approximate the computation to some extent, as explained earlier. In particular, we approximate the following:

1. The best contig hypothesis is computed by a simple averaging of the contiged fragment sizes (using linear interpolation for missing cuts) from the best possible alignment, rather than a true Bayesian probability density maximization with fragment sizes as parameters. This avoids the expensive global and local optimization over possible contigs as performed in (Anantharaman, Mishra and Schwartz 1997).
2. Good estimates of the global error model parameters are assumed to be known a priori, but will be further improved by using a limited number (2-3) of iterations of a Bayesian probability density maximization step over all contigs.

When the input data consists of genomic optical maps computed from single instances of DNA fragments this approximation may sometimes 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 correct 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 (Anantharaman, Mishra and Schwartz 1997):

$$\begin{aligned} f(\mathcal{H}|D_1 \ldots D_M) &= f(\mathcal{H}) \prod_{j=1}^{M} f(D_j|\mathcal{H}) / \prod_{j=1}^{M} f(D_j) \\ &\propto f(\mathcal{H}) \prod_{j=1}^{M} f(D_j|\mathcal{H}). \end{aligned}$$

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 is 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 could adjust K until the computed total length is approximately of this length. We have found that the most convenient and quickets way of setting K is to estimate the value of K that would exactly balance the expected decrease in probability density from correct overlaps (given the error parameters) and setting K to some small multiple (typically 1.5-2.0) of this value to allow for some outliers in errors.

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. 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.
4. 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 contig (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:

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

Here m_j is the number of cuts in input map D_j. The prior component of Λ for each contig is simply $2KX\sigma^2$, where X is the length of the island, and K a constant as discussed earlier. The other components of Λ can be computed for each island and then summed up.

For fragment sizing errors, consider each fragment of the proposed contig, and let the contig fragment be composed of overlaps from several map fragments 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 $(r_p - \mu)^2/2$ for the largest such partial map fragment for each contig fragment.

Now consider the presence of missing cuts (restriction sites) with $p_c < 1$. 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.

The False Positive Match Probability

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

The false positive match score S_{FP} for a pair of maps to be contiged is approximated as the product of the above probability ratio for each fragment in the contig, which is estimated as follows:

Let the average fragment size be ℓ and the fragment size distribution be modeled by the exponential distribution $f(x) = \exp(-x/\ell)/\ell$. 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 maps/contigs 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.

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:

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

$\mathbb{E}[p_c]$ is a local estimate of p_c, obtained by counting the number of misaligned cuts (of each 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 derived similarly.

If false cuts are present ($p_f > 0$), an upper bound of S_{FP} can be obtained by treating all false cuts as real cuts with the corresponding matching cuts all missing and using $\ell/(p_c + \ell p_f)$ as the (apparent) average distance between two consecutive cuts of the same enzyme in the previous expressions.

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 (typically just two) iterations suffice for the global search, given good initial values for σ, p_c and p_f.

One property of our global search heuristic is that it is greedy in combining pairs of maps/contigs and therefore may be suboptimal if the data quality is poor. However, the false positive threshold will automatically reduce the amount of contiging if the data quality is poor, hence the use of a greedy search is sufficient if all data errors were modeled.

We have since discovered that there is a very small chance that an input map is actually a combination of two pieces of DNA that just happened to lie end-to-end in the image (Optical Chimerism). In such cases a greedy search strategy can get stuck in a sub-optimal local maxima. To handle such un-modeled errors, the final contigs are analyzed for statistically unlikely distributions in the contig depth : Any contiging errors are likely to result in very shallow contigs at the error location, if these errors are low probability events. Contigs can then be broken up at these shallow points and the input maps that formed the bridges at these points discarded and then further contiging attempted. This provides a simple form of back-tracking that is sufficient to recover from low probability missing terms in the error model.

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 for finding the best alignment between a single molecule map and a map hypothesis. The problem here is different from finding the best alignment of a data map against a hypothesis map, as described in (Anantharaman, Mishra and Schwartz 1997). Here 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.

Consider two input data maps or contigs D_i and D_j. Let the locations of the restriction site cuts on the input data be $X_{i1} \dots X_{im_i}$ and $X_{j1} \dots X_{jm_j}$. Denote the length of the maps by L_i and L_j respectively. For the Dynamic Programming formulation we define three tables P_{IJ}, L_{IJ} and R_{IJ} of size $I = 1 \dots m_i$ by $J = 1 \dots m_j$. Their values are defined to be:

1. P_{IJ} : For all alignments of D_i and D_j such that cuts X_{iI} and X_{jJ} are aligned, the least possible penalty to the right of X_{iI} and X_{jJ} in the best such alignment.
2. L_{IJ} : For all alignments of D_i and D_j such that cuts X_{iI} and X_{jJ} are the left most aligned cuts, the least possible penalty to the left of X_{iI} and X_{jJ}.
3. R_{IJ} : For all alignments of D_i and D_j such that cuts X_{iI} and X_{jJ} are the right most aligned cuts, the least possible penalty to the right of X_{iI} and X_{jJ}.

We first compute L_{IJ} and R_{IJ} for all possible I and J by computing the penalty for the remaining (misaligned) cuts in the overlap end region specified. For each cut choose from amongst three alternate explanations the one with the least penalty : The cut is a false penalty, (Λ_f) or the cut is missing from the other map ($\Lambda_c + \Lambda_n$) or the cut matches the end of the map ($(r_p - \mu)^2/2$). For input maps (not contigs) this can be written as :

$$
\begin{aligned}
L_{IJ} \quad = \quad & \sum_{n=I+1}^{m_i} (X_{iI} - X_{in} \geq X_{jJ} - X_{J0})?0 : \\
& \min(\Lambda_f, \Lambda_c + \Lambda_n, (X_{iI} - X_{in} - X_{jJ} + X_{j0})^2/2) \\
& + \sum_{t=J+1}^{m_j} (X_{jJ} - X_{jt} \geq X_{iI} - X_{i0})?0 : \\
& \min(\Lambda_f, \Lambda_c + \Lambda_n, (X_{jJ} - X_{jt} - X_{iI} + X_{i0})^2/2) \\
R_{IJ} \quad = \quad & \sum_{n=1}^{I-1} (L_i - X_{in} \geq L_j - X_{jJ})?0 : \\
& \min(\Lambda_f, \Lambda_c + \Lambda_n, (L_i - X_{in} - L_j + X_{jJ})^2/2) \\
& + \sum_{t=1}^{J-1} (L_j - X_{jt} \geq L_i - X_{iI})?0 : \\
& \min(\Lambda_f, \Lambda_c + \Lambda_n, (L_j - X_{jt} - L_i X_{jJ} + X_{iI})^2/2)
\end{aligned}
$$

The table P_{IJ} is computed using dynamic programming : For each pair of aligned cuts I and J we consider all possible combinations of the next pair of aligned cuts n, t to the right of I, J :

$$
\begin{aligned}
P_{IJ} = \min\Bigg(& R_{IJ}, \\
& \min_{I+1 \leq n \leq m_i} \ \min_{J+1 \leq t \leq m_j} \big[(X_{in} - X_{iI} - X_{jt} + X_{jJ})^2/2 \\
& + (n - I - 1 + t - J - 1) \min(\Lambda_f, \Lambda_c + \Lambda_N) \big] \Bigg)
\end{aligned}
$$

We also remember the best next alignment n and t for each I,J (or NIL if R_{IJ} was the least penalty term)

To find the best alignment, we just need to locate its leftmost alignment I, J given by the smallest possible sum $P_{IJ} + L_{IJ}$ over all I, J. The next alignment is then given by the remembered best values of n and t which can be iterated to determine the entire alignment.

When combining contigs of maps rather than input maps, the Dynamic programming structure is the same, except that the exact penalty values are slightly different and computed as the increase in penalty of the new contig over the penalty of the two shallower contigs being combined. The expressions for computing the penalty of deep contigs has been described previously.

The resulting algorithm has a time complexity of $O(m_i^2 m_j^2)$ in the worst case. However one can incorporate heuristics to bring down the average case complexity to $O(m_i + m_j)$:

1. The probability that two pairs of aligned cuts are separated by very many unaligned cuts is very small, which allows the number of n and t values that need to considered in computing P_{IJ} to be limited to some small constant number. In practice one could limit the total number of misaligned cuts between pairs of aligned cuts to no more than about 7, resulting in $8 \times 9/2 = 36$ pairs of n and t values. This reduces the complexity for large m_i and m_j to $O(m_i m_j)$.
2. All aligned cuts in the best alignment will correspond to similar offsets between the maps/contigs. This means that all the good (low) penalty scores in the

P_{IJ} table will lie along some diagonal (not necessarily the center diagonal). It is possible to locate the approximate offset value that this diagonal corresponds to by filling in only a small region of the table P_{IJ} corresponding to large I or large J values, then only compute those remaining elements of P_{IJ} with similar offset values. This reduces the complexity to $O(m_i + m_j)$.

This dynamic programming needs to be combined with a global search that tries all possible pairs of the M input maps for possible overlaps. This could add a worst case multiplicative complexity of $O(M^3)$: checking $O(M^2)$ pairs of maps to find the next pair to combine, and doing this $O(M)$ times. Moreover, the overall complexity then is not $O(nM^3)$, where n is the average number of cuts per input map, but $O(n^2M^3)$, if the contigs are grown from one side adding one new map at a time. This can be improved to $O(nM^{1+\epsilon})$ by the following heuristics :

1. The next pair of maps to combine can be determined by using a hashing scheme suggested by Dennis Shasha, which will be detailed in a subsequent publication. It allows all pairs of matching maps to be determined in $O(nM^{1+\epsilon})$ time, where ϵ ranges from 0 to 1 depending on the severity of the error process ($\epsilon = 0.25$ is typical for the errors we encounter).

2. By combining maps, so that all maps/contigs remain of comparable size, the total time to compute all $O(M)$ alignments becomes $O(n{*}Mlog(M))$, since the combination effectively involves $O(log(M))$ phases in which all maps are combined in $O(nM)$ time (their total size). Combining this with the complexity for hashing, we get the previously stated total complexity.

Experimental Results

We, first, present some results of experiments with Gentig using data from Chromosome 2 of *Plasmodium falciparum*. The results have been verified from available sequence data.

The data from Chromosome 2 of *Plasmodium falciparum* had previously been collected by Junping Jing for the BamHI restriction enzyme (Jing et al. 1999), and used to manually assemble a contig over a period of several weeks. The same data was also run through Gentig and produced essentially the same contigs in a matter of a few seconds. The contig also agrees with the sequence data that has since become available. All three contigs are shown in the table below, with fragment sizes in kilobases.

Method	
Manual Assembly	77.1 19.9 7.5 26.1 9.9 41.0 12.4 3.7 34.8 21.1 62.2 49.7 43.5 74.6 47.2 80.8 2.0 8.9 18.6 80.8 19.9 31.1 17.4 28.6 52.2 2.0 24.9 6.0 34.8
Contig Program	79.8 19.6 6.9 26.1 10.9 45.3 13.7 4.0 31.9 19.5 58.2 49.9 34.0 54.3 51.6 82.2 2.1 10.7 18.3 81.0 19.8 28.9 17.7 27.8 49.6 2.1 25.2 5.6 38.9
Sequence Data	77.5 21.0 6.8 27.1 9.8 43.3 13.6 3.7 36.0 20.2 61.8 55.2 40.8 70.3 46.9 87.3 1.8 11.6 18.0 84.0 20.7 30.4 18.0 30.8 50.0 1.8 24.8 5.3 28.1

Note that a major difference in the maps is in the size of the right end fragment. The sequence data shows a smaller right end fragment. However it is known that the telomeric regions near the chromosome ends are full of repeats and therefore probably compressed by the sequence assembler. Conversely Gentig appears to produce end fragment size estimates that are slightly larger than the Manual Assembly. The reason is that Gentig did not model chromosome ends and actually returns the largest value in the sample for each end fragment, on the assumption that the contig might continue on either end. In contrast the manual assembler recognized the end fragments and averaged their size.

Complete restriction maps of all 14 chromosomes of *Plasmodium falciparum* for the BamHI restriction enzyme, were computed from a single set of genomic DNA molecules: 12 of the chromosomes were automatically identified as separate contigs by Gentig, and the remaining 2 chromosomes appeared as 2 contigs each, which could be identified based on the known approximate sizes of the chromosomes. Similar maps of all 14 chromosomes have also been computed for the NheI enzyme. These maps are published in their entirety elsewhere, but are not shown here since only chromosome 2 can be verified from sequence data.

To test the effectiveness of the software on larger genomes, 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 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			
Coverage, c	**4**	**8**	**16**
Contigs	**4**	**2**	**1**
Longest contig	**12,248** Kb	**15,599** Kb	**19,849** Kb
Computed			
Contigs	4	2	2
Maps Rejected	0	0	1
Longest contig	12,251 Kb	15,601 Kb	19,963 Kb
Est. SD, σ	2.99 Kb	2.96 Kb	3.00 Kb
Est. p_c	0.815	0.808	0.805
Est. p_f	0.6×10^{-6}	0.9×10^{-6}	0.9×10^{-6}

Conclusion

The Worst-case Complexity

Note that the Bayesian approach inherent to the Gentig algorithm relies heavily on the knowledge of a good prior distribution of the errors: partial digestion, false cuts and sizing errors. In the absence of this information, the worst case behavior of any algorithm to solve the genomic contiging problem is likely to be exponential. In particular, there are pathological examples (consisting of artificially constructed input maps) for which this problem can be shown to be NP-hard by a transformation from the Hamiltonian path problem for a cubic graph. The proof of this theorem is rather technical and is given in the appendix.

Statistical Analysis of Feasible Errors

We want to understand how the feasibility of our genomic contiging approach varies with parameters in the error models. In order to bound the range of error parameter values for which unique solutions to the contiging problem exist, it is instructive to examine a rather simple overlap rule. We shall consider two important parameters: the relative sizing error, β and partial digestion rate p_c. For the sake of simplicity, we assume that the number of true restriction fragments in each input restriction map is n and number of detected restriction fragments, $m = np_c$. The "overlap threshold ratio" (the minimum fraction of one input map, in base pairs, that has to be overlapped by the other input map to recognize the overlap) is denoted by θ. A low θ value is desirable to achieve a contig with low coverage.

The overlap rule that we examine here takes the partial digestion into account. The rule is fairly simple and as follows: Consider an alignment, where input map D_i is aligned with respect to another input map D_j with an overlap ratio bigger than θ (in 4 possible relative orientations); at least k of the restriction fragments of the input maps match positionally and the numbers of unmatched fragments in the prefixes are bounded by r.

It is easily seen that in order for the false negative probability to be small, it is required that

$$k \le \frac{np_c^4\theta}{2} \quad \text{and} \quad r \ge \frac{k_1}{p_c^4}, \quad k_1 \approx 2.$$

Thus, if in fact input maps D_i and D_j overlap by θ, then we will detect it with probability bigger than

$$(1 - e^{-k_1})(1 - e^{-np_c^4\theta/8}).$$

Now consider the false positive probability and consider an arbitrary alignment (not necessarily satisfying the constraints on the unmatched prefixes). Let the random variable W denote the number of fragments in input map D_i that positionally match with the fragments of input map D_j. Note that

$$\begin{aligned} \mathbb{E}\left[\binom{W}{i}\right] &= \binom{m}{i}(\beta/2)^i \\ &= \frac{1}{i!}\left(\frac{np_c\beta}{2}\right)^i. \end{aligned}$$

By Brun's sieve, we see that

$$\text{Prob}[W = i] = \frac{1}{i!}(\beta np_c/2)^i e^{-\beta np_c/2}.$$

Thus the false positive probability is

$$4r\sum_{i=k}^{\infty}\frac{1}{i!}(\beta np_c/2)^i e^{-\beta np_c/2}.$$

and we need to make r as small as possible and k as large as possible, in order to guarantee that the false positive probability remains small.

Hence, we need to satisfy the following constraints:

$$\frac{3\beta np_c}{4} \le k \le \frac{np_c^4\theta}{2} \quad \text{and} \quad \frac{k_1}{p_c^4} \le r \le \frac{1}{\beta}.$$

Thus

$$1 \ge \theta \ge \frac{3\beta}{2p_c^3} \quad \text{and} \quad \beta \le \frac{2p_c^3}{3}.$$

Thus our approach of contiging genomic optical maps works well when the partial digestion probability is rather high (close to 1), i.e.,

$$p_c \ge (3\beta/2)^{\frac{1}{3}}.$$

or the relative sizing error is quite low.

Applications

The algorithms described in this paper together with the optical mapping procedure can be used to construct genome wide restriction maps in a rapid and scalable manner without using any cloning and requiring very little DNA sample. This has a number of interesting applications.

First the genome wide restriction maps can be used for sequence verification. Since the maps have long range fidelity, they will not reflect errors due to repeats (unless the repeats are larger than about 1-4 Megabases).

Second, the genome wide restriction maps can be used in sequence assembly. Since any reasonably large sequence contig (say 10 kb or longer) can typically be placed uniquely on a multi-enzyme genome wide restriction map, this allows the gaps in the sequence assembly to be pin-pointed accurately.

Finally, the ability to rapidly construct genome wide restriction maps opens the possibility to making meaningful population genomic studies without doing any sequencing. For example large scale DNA insertions and deletions such as occur in cancer cell lines, compared to non-cancerous cell lines, could be detected just by comparing their genome wide restriction maps.

Open Problems

Among many remaining related unsolved problems, the most interesting one involves the situation where optical maps to be contiged are those coming from K populations. If the populations of DNA have restriction maps that are sufficiently different from each other this would be no different than the current problem of computing the restriction map of an entire genome with multiple chromosomes. However if some of the members in the population are very similar, they may be combined into the same contig or even into two contigs involving erroneous crossovers. A common example is to be able to distinguish the diploid chromosomal pairs. The same problem has been studied for optical maps from clones in (Mishra and Parida 1998).

Other open problems include the ability to model missing fragments, systematic sizing error (an entire DNA molecule has all its fragments too large or too small), the ability to use additional side information from the optical mapping process (such as when the size of a particular fragment is likely to be less reliable).

Acknowledgements

Our thanks go to C. Aston, J. Jing, J. Lin, Z. Lai and R. Qi of Keck Laboratory, Chemistry Department for the genomic maps and the contigs for *Plasmodium falciparum* (chromosome 2), S. Paxia of Courant Institute and Z. Lai for the help with the map display and to E. Huff of Chemistry Department for the help with the synthetic data used in testing the software. Our thanks also go to D.J. Carucci and S.L. Hoffman of Naval medical Research Institute and to M. Gardner, H. Tettelin, L.M. Cummings and J.C. Venter of The Institute for Genomic Research for providing the sequences used in the experimental verification of the Gentig algorithm.

The research presented here was partly supported by an NSF Career grant: IRI-9702071, an NIH Grant: NIH R01 HG0025-07 and a grant from the **Chiron Corporation**.

Appendix: Worst Case 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

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

Theorem 0.1 *The problem GC is NP-hard.*

Proof. We give a simple transformation from Hamiltonian path problem restricted to a cubic graph (Garey and Johnson 1979). 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

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

The cut locations are then

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

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. $\square$

References

T.S. Anantharaman and B. Mishra, "Genomics via Optical Mapping I: Probabilistic Analysis of Optical Mapping Models," *NYU technical report 1998-770*, August 1998.

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.

T.S. Anantharaman and B. Mishra, "Genomics via Optical Mapping II(A): Restriction Maps from Partial Molecules and Variations," *NYU technical report 1998-759*, March 1998.

W. Cai, et al., "High-Resolution Restriction Maps of Bacterial Artificial Chromosomes Constructed by Optical Mapping," *Proc. Natl. Acad. Sci. USA*,**95**:3390–3395, 1998.

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.

J. Jing, et al., "Optical Mapping of *Plasmodium falciparum* Chromosome 2," *Genome Research*,**9**:175–181, 1999.

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

B. Mishra, "Algorithmic Biology," *Courant Lecture Notes Series*, 1999 (Tent.).

B.Mishra and L. Parida, "Partitioning K-Clones: Inapproximability Results and a Practical Solution to the K-Population Problem," *Recomb 98*,192–201, 1998

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

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

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