False Positives in Genomic Map Assembly and Sequence Validation [1]

Thomas Anantharaman
Dept. of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI.
Bud Mishra
Courant Institute, New York, NY.

1 Introduction

In the recent years, genome-wide shot-gun restriction mapping of several microorganisms using optical mapping [Lai99, Lin99] have led to high-resolution restriction maps that directly facilitated sequence assembly avoiding gaps and compressions or validated shotgun sequence assembly [CMS99]. The simplicity and scalability of shot-gun optical mapping suggests obvious extensions to bigger and more complex genomes, and in fact, its applications to human and rice are underway. Furthermore, a good-quality human map is likely to play a critical role in validating several currently available but unverified sequences.

The key computational component of this process involves the assembly of large numbers of partial restriction maps with errors into an accurate restriction map of the complete genome. The general solution has been shown to be NP-complete, but a polynomial time solution is possible if a small fraction of false negatives (wasted data) is permitted. The critical component of this algorithm is an accurate bound for the false positive probability that two maps that appear to match are in fact unrelated.

The map assembly and alignment problems are related to the much more widely studied sequence assembly and alignment problems. The primary difference in the problem domains is that the sequence alignment problem involves discrete data only in which errors can be modeled as discrete probabilities, whereas map alignment involves fragment sizing errors and hence requires continuous error models. However, even in the case of sequence alignment, statistical significance tests play a key role in eliminating false positive matches and are included in many sequence alignment tools such as BLAST (see for example chapter 2 in [DEKM98]).

A simple bound using Brun's sieve can be easily derived [AMS99], but such a bound often fails to exploit the full power of optical mapping. Here, we derive a much tighter but more complex bound that characterizes the sharp transition from infeasible experiments (requiring exponential computation time) to feasible experiments (polynomial computation time) much more accurately. Based on these bounds, a newer implementation of the Gentig algorithm for assembling genome-wide shot-gun maps [AMS99] has improved its performance in practice.

A close examination shows that the false positive probability bound exhibits a computational phase-transition: that is, for poor choice of experimental parameters the probability of obtaining a solution map is close to zero, but improves suddenly to probability one as the experimental parameters are improved continuously. Thus careful optimized choice of the experimental parameters analytically has strong implication to experiment design in solving the problem accurately without incurring unnecessary laboratory or computational cost. In this paper, we explicitly delineate the interdependencies among these parameters and explore the trade-offs in parameter space: e.g., sizing error vs. digestion rate vs. total coverage. There are many direct applications of these bounds apart

[1] The research presented here was partly supported by NSF Career Grant IRI-9702071, DOE Grant 25-74100-F1799, NYU Research Challenge Grant, NYU Curriculum Challenge Grant.

from the alignment and assembly of maps in Gentig: Comparing two related maps (e.g. chromosomal aberrations), Validating a sequence (e.g. shot-gun assembly-sequence) or a map (e.g., a clone map) against a map, etc. Specific usage of our bounds in these applications will appear elsewhere [AMAP00].

1.1 A Sub-quadratic Time Map Assembly Algorithm: Gentig

For the sake of completeness we give a brief but general description of the basic Gentig (GENomic conTIG) map assembly algorithm previously described elsewhere in details [AMS99]. Roughly, Gentig can be thought of as a greedy algorithm that in any step considers two islands (individual maps or map contigs) and postulates the best possible way these two maps can be aligned. Next, it examines the overlapped region between these two islands and weighs the evidence in favor of the hypothesis that "these two islands are unrelated and the overlap is simply a chance occurrence." If enough evidence favors this "false positive" hypothesis, Gentig rejects the postulated overlap. In the absence of such evidence, the overlap is accepted and the islands are fused into a bigger/deeper island. What complicates these simple ideas is that one needs a very quantitative approach to calculate the probabilities, the most likely alignment and the criteria for rejecting a false positive overlap—all of these steps depending on the models of the error processes governing the observations of individual single molecule maps. Ultimately, the Gentig algorithm can be seen to be solving a constrained optimization problem with a Bayesian inference algorithm to find the most likely overlaps among the maps subject to the constraints imposed by the acceptable false positive probability. False Positive constraints limit the search space, thus obviating full-scale back-tracking and avoiding an exponential time complexity. As a result, the Gentig algorithm is able to achieve a sub-quadratic time complexity.

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 that placement while allowing for various modeled data errors: *sizing errors*, *missing restriction cut sites*, and *false optical cuts sites*.

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 $(1 \le j \le M)$ 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]:

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

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. It is also possible to model the standard deviation as some polynomial of the true fragment size which will be described in a future paper. (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.

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 $x_1, \ldots, x_N$. 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:

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

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 computed by a Dynamic Programming algorithm determines which cuts are missing.

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. Finally, consider the presence of false optical cuts when $p_f > 0$. For each false cut, we add a penalty $\Lambda_f = 2\sigma^2 \log[1/(p_f \sqrt{2\pi}\sigma)]$ in order to model a "scaled" multiplicative penalty of p_f. A modified penalty term is required for the end fragments of each map which might be partial fragments, as described in [AMS99]. 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 resulting alignment algorithm has a time complexity of $O(m_i^2 m_j^2)$ in the worst case, but an average case complexity of $O(m_i + m_j)$, achieved with several simple heuristics. The basic dynamic programming is combined with a global search that tries all possible pairs of the M input maps for possible overlaps. A sophisticated implementation in Gentig achieves an average case time complexity of $O([mM]^{1+\epsilon})$ ($\epsilon = 0.40$ is typical for the errors we encounter), where m is the average value of m_j. It relies on several heuristics based on "geometric hashing" while avoiding any backtracking.

1.2 Summary of the New Results

Before proceeding further with the technical details of our probabilistic analysis, we summarize the two main formulae that can be used directly in estimating the false positive probability for a particular map alignment, or in designing a bio-chemical experiment with the goal of bounding the false positive probability below some acceptable small value (typically $\leq 10^{-3}$).

1.2.1 The formula for false positive probability

Consider a population of M ordered restriction maps with errors of the kind described earlier. Assume that the best matching pair of maps (under a Bayesian formulation) has n aligned cuts and r misaligned cuts, and R is some average of the relative sizing error of aligned fragments in the overlap. Then FPT_r denotes the probability that the two maps are unrelated and the detected overlap is purely by chance.

$$FPT_r \leq 4 \binom{M}{2} \binom{2n + r + 2}{r} P_n e^{\frac{rR}{\sqrt{2\pi}}}$$

$$\text{where } P_n = \frac{(R\sqrt{\frac{\pi e}{8}})^n}{\sqrt{n\pi}}.$$

Note that if $r = 0$ (implying that the best match has all the cuts aligned and the only error source is sizing error), then $FPT_0 = 4\binom{M}{2}P_n$. If $R \ll 1$ then as n gets larger FPT_0 exhibits an exponential decay to 0, and this property remains true for non-zero values of r.

1.2.2 The formula for feasible genome-wide shotgun optical mapping

Consider an optical mapping experiment for genome-wide shotgun mapping for a genome of size G and involving M molecules each of length L_d. Thus the coverage is ML_d/G. Let the a fragment of true size X have a measured size $\sim \mathcal{N}(X, \sigma^2 X)$. Let the average true fragment size be L, and the digestion rate of the restriction enzyme be P_d. Thus the average relative sizing error $R = \sigma\sqrt{P_d/L}$ and the average size of aligned fragments will be $L/{P_d}^2$. As usual, let θ represent the minimum "overlap threshold." Hence the expected number of aligned fragments in a valid overlap is at least $n = \theta L_d {P_d}^2/L$. Let $d = 1/P_d$, the inverse of the digest rate. Feasible experimental parameters are those that result in an acceptable (e.g. $\leq 10^{-3}$) False Positive rate FPT:

$$FPT \quad \approx \quad 2M^2 \binom{\lceil 2nd+2 \rceil}{\lfloor 2n(d-1) \rfloor} \frac{(R\sqrt{\frac{\pi e}{8}})^n}{\sqrt{n\pi}} e^{\frac{2(d-1)nR}{\sqrt{2\pi}}}$$

To achieve acceptable false positive rate, one needs to choose an acceptable value for the experimental parameters: P_d, σ, L_d and coverage. FPT exhibits a sharp phase transition in the space of experimental parameters. Thus the success of a mapping project depends extremely critically on a prudent combination of experimental errors (digestion rate, sizing), sample size (molecule length and number of molecules) and problem size (genome length). Relative sizing error can be lowered simply by increasing L with a choice of rarer-cutting enzyme and digestion rate can be improved by better chemistry [Reed98].

As an example, for a human genome of size $G = 3,300Mb$ and a desired coverage of $6\times$, consider the following experiment. Assume a typical value of molecule length $L_d = 2Mb$. If the enzyme of choice is PAC I, the average true fragment length is about $25Kb$. Assume a minimum overlap[2] of $\theta = 30\%$. Assume that the sizing error for a fragment of $30kb$ is about $3.0kb$, and hence $\sigma^2 = 0.3kb$. With a digest rate of $P_d = 82\%$ we get an unacceptable $FPT \approx 0.0362$. However just increasing P_d to 86% results in an acceptable $FPT \approx 0.0009$. Alternately, reducing average sizing error from $3.0kb$ to $2.4kb$ while keeping $P_d = 82\%$ also produces an acceptable $FPT \approx 0.0007$.

Obviously one should allow some margin in choosing experimental parameters so that the actual experimental parameters will be a reasonable distance from the phase transition boundary. This is needed both to allow some slippage in experimental errors as well as the possibility that there may be additional small errors not modeled by the error model.

2 A Technical Probabilistic Lemma

The key to understanding the false positive bound is the following technical lemma that forms the basis of further computation. Let $\mathbf{X} = \langle x_1, \ldots, x_n \rangle$ and $\mathbf{Y} = \langle y_1, \ldots, y_n \rangle$ be a pair of sequences of positive real numbers, each sequence representing sizes of an ordered sequence of restriction fragments. We rely on a "matching rule" to decide whether $\mathbf{X}$ and $\mathbf{Y}$ represent the same restriction fragments in a genome, by comparing the individual component fragments. We proceed by computing a "weighted

[2]This value should be selected to minimize FPT

squared relative sizing error" that is then compared to a specific threshold Θ. The "weighted squared relative sizing error" is simply

$$\sum_{i=1}^{n} w_i \left(\frac{X_i - Y_i}{X_i + Y_i}\right)^2,$$

where w_i's are chosen to match the error model. For example, if the sizing error variance for a fragment with true size X is $\sigma^2 X^p$, where $p = 0..2$, we can use $w_i \approx \frac{x_i + y_i{}^{2-p}}{\sigma^2}$.

Lemma 2.1 *Let* $\mathbf{X} = \langle X_1, \ldots, X_n\rangle$ *and* $\mathbf{Y} = \langle Y_1, \ldots, Y_n\rangle$ *be a pair of sequences of IID random variables* X_i*'s and* Y_i*'s with exponential distributions and pdf's* $f(x) = \frac{1}{L}e^{-x/L}$. *Then*

1. $\mathbb{P}\mathrm{r}(|X_i - Y_i|/(X_i + Y_i) \le \Theta) \le \Theta$, *for all* $0 \le \Theta$ *and with equality holding, if* $\Theta \le 1$.

2. $\mathbb{P}\mathrm{r}(\sum_{i=1}^n w_i(\frac{X_i - Y_i}{X_i + Y_i})^2 \le \Theta) \le \frac{(\frac{\pi}{4}\Theta)^{n/2}}{(\frac{n}{2})! \prod_{i=1}^n \sqrt{w_i}}$, *for all* $0 \le \Theta$ *and with equality holding, if* $\Theta \le \min_{1\le i \le n} w_i$.

Proof.

The first identity can be shown by integrating the relevant portion of the joint distribution of X_i and Y_i:

$$\begin{aligned} P_1 &\equiv \mathbb{P}\mathrm{r}(|X_i - Y_i|/(X_i + Y_i) \le \Theta) \\ &= \int_{X_i=0}^{\infty} \int_{Y_i = X_i \frac{1-\Theta}{1+\Theta}}^{X_i \frac{1+\Theta}{1-\Theta}} \frac{1}{L^2} e^{-\frac{X_i + Y_i}{L}}\, dY_i\, dX_i = \Theta. \end{aligned}$$

Note that this means that for each pair of random fragment sizes X_i, Y_i the statistic $U_i \equiv |X_i - Y_i|/(X_i + Y_i)$ is uniformly distributed between 0 and 1.

We can now compute the overall probability P_n for all n fragment pairs:

$$\begin{aligned} P_n &\equiv \mathbb{P}\mathrm{r}(\sum_{i=1}^n w_i(\frac{X_i - Y_i}{X_i + Y_i})^2 \le \Theta) \\ &= \mathbb{P}\mathrm{r}(\sum_{i=1}^n w_i U_i{}^2 \le \Theta) \end{aligned}$$

Note that $U_1, ..., U_n$ are IID uniform distributions over [0..1], hence this probability is just that part of the volume of the n-dimensional unit cube that satisfies the condition $\sum_1^n w_i U_i{}^2 \le \Theta$. For small sizing errors such that $\Theta \le \min(w_1, ..., w_n)$, this region is one orthant of an n dimensional ellipsoid with radius values of $\sqrt{\Theta/w_i}$ in the ith dimension. In general this volume is an upper bound and hence:

$$P_n \le \frac{(\frac{\pi}{4}\Theta)^{n/2}}{(\frac{n}{2})! \prod_{i=1}^n \sqrt{w_i}}$$

Here $n!$ is defined in terms of the Gamma function for fractional n: $n! \equiv \Gamma(n+1)$. **QED**

Lemma 2.2 *Let* $\mathbf{X} = \langle X_1, \ldots, X_n\rangle$ *and* $\mathbf{Y} = \langle Y_1, \ldots, Y_n\rangle$ *be a pair of sequences such that variables* X_i*'s and* Y_i*'s are given in terms of IID random variables* Z_j*'s with exponential distributions and pdf's*

$f(z) = \frac{1}{L}e^{-z/L}$. In particular, for $i = 1, \ldots, n$, if we can express X_i and Y_i in terms of exponential IID random variables $Z_1, \ldots, Z_{r_i}, Z_{r_i+1}, \ldots, Z_{r_i+s_i}$ as follows:

$$\begin{aligned}
\min(X_i, Y_i) &= \frac{1}{2}\sum_{k=1}^{s_i} Z_{r_i+k} \\
\max(X_i, Y_i) &= \sum_{k=1}^{r_i} Z_k + \frac{1}{2}\sum_{k=1}^{s_i} Z_{r_i+k}.
\end{aligned}$$

Then

1. $\mathbb{P}\mathrm{r}(|X_i - Y_i|/(X_i + Y_i) \le \Theta) \le \binom{s_i+r_i-1}{r_i}\Theta^{r_i}$, *for all* $\Theta \ge 0$.

2. $\mathbb{P}\mathrm{r}(\sum_{i=1}^n w_i(\frac{X_i-Y_i}{X_i+Y_i})^2 \le \Theta) \le \frac{\prod_{i=1}^n (r_i/2)!\binom{s_i+r_i-1}{r_i}(\Theta/w_i)^{r_i/2}}{(\sum_{i=1}^n r_i/2)!}$, *for all* $\Theta \ge 0$.

Proof.

Similar to the previous lemma. **QED**

3 Model of Random maps

Our model of random maps is that cut sites are randomly and uniformly distributed, so that the distance between cut sites is a random variable X with an exponential distribution and probability density $f(x) = \frac{1}{L}e^{-x/L}$, where L is the average distance between cut sites. Here we assume that all cut sites are indistinguishable from each other.

First we consider the case with no misaligned cuts, so that the only errors in the proposed overlap region are sizing errors. Thus our alignment data consists of two maps with fragment sizes $x_1, \ldots, x_n$ on one map that align with fragment sizes $y_1, \ldots, y_n$ on the other map, where n is the number of fragments in the overlap region. Here the quality of the alignment will be measured by a weighted squared relative sizing error, $E = \sum_1^n w_i(x_i - y_i)^2/(x_i + y_i)^2$, where w_i's are chosen as explained earlier. We need to compute $P_n = \mathbb{P}\mathrm{r}(\sum_1^n w_i(\frac{X_i-Y_i}{X_i+Y_i})^2 \le E)$, where $E \equiv \sum_1^n w_i(\frac{x_i-y_i}{x_i+y_i})^2$. By an application of the previous lemma, we have:

$$P_n \le \frac{(\frac{\pi}{4}\sum_{i=1}^n w_i(\frac{x_i-y_i}{x_i+y_i})^2)^{n/2}}{(\frac{n}{2})!\prod_{i=1}^n \sqrt{w_i}}.$$

Here, $n! \equiv \Gamma(n+1)$. For current purposes it suffices to note that $(\frac{1}{2})! = \frac{\sqrt{\pi}}{2}$. For example, $(\frac{3}{2})! = \frac{3}{2}(\frac{1}{2})! = \frac{3\sqrt{\pi}}{4}$.

To see more clearly how this probability scales with the sizing errors, let us define the weighted RMS relative sizing error R_n, and the average weight A_n:

$$R_n \equiv \sqrt{\frac{\sum_{i=1}^n w_i(\frac{x_i-y_i}{x_i+y_i})^2}{\sum_{i=1}^n w_i}}. \tag{1}$$

$$A_n \equiv \frac{1}{n}\sum_{i=1}^n w_i. \tag{2}$$

Then we can rewrite P_n using Sterling's expression for factorials as:

$$P_n \le \frac{(R_n/\sqrt{2/e\pi})^n}{\sqrt{n\pi}}\prod_{i=1}^n\sqrt{\frac{A_n}{w_i}}. \tag{3}$$

This shows that asymptotically the n-fragment false positive probability P_n will decrease with the nth power of the RMS relative error R_n provided that $R_n \leq \sqrt{2/e\pi} = 0.4839$.

To complete our computation of the False Positive Likelihood FP for a particular pair of maps D_1 and D_2, we need to consider the multiple possible choices of overlaps of n or more fragments. Let the two molecules contain N_1 and N_2 fragments with $N_1 \leq N_2$. If $n < N_1$ there will be exactly 4 possible ways of forming an overlap of n fragments. Otherwise, if $n = N_1$ there will be $2(N_2 - N_1 + 1)$ ways of forming an n fragment overlap. Each such overlap has the same independent probability P_n. Thus with 4 possible overlaps, the probability FP_n of finding at least 1 overlap of n fragments between two random maps as good as the actual alignment is bounded by the probability $FP_n \leq 1 - (1 - P_n)^4 \leq 4P_n$.

We also need to consider random overlaps of more than n fragments that are as good as the actual overlap. Under a typical Bayesian error model such as described in [AMS99], each overlap of more than n fragments can have slightly larger sizing errors than the actual alignment with the same probability density, since the prior probability density must be biased towards larger overlaps. For an error model such as in [AMS99] one can show [AM00] that the permissible increase in relative sizing error R_{n+1} vs. R_n is given approximately by:

$$A_{n+1}{R_{n+1}}^2 \leq \frac{nA_n{R_n}^2 + K/2}{(n+1)},$$

where K is a prior bias parameter, typically in the range $1 \leq K \leq 1.4$.

Hence for $n + k < N_1$ we can write FP_{n+k} as:

$$FP_{n+k} \quad \leq \quad 4P_n\sqrt{\frac{n}{n+k}}{R_n}^k \left(\frac{\pi A_n e^{K/2A_n{R_n}^2}}{2G_n}\right)^{k/2}$$

Here G_n is the geometric mean of $w_i, i = 1..n$.

If $n + k > N_1$ then $FP_{n+k} = 0$ and if $n + k = N_1$ we just need to replace the factor 4 by $2(N_2 - N_1 + 1)$.

We can now compute FP by combining overlaps of all possible number of fragments $(n..N_1)$:

$$\begin{aligned} FP &\leq \sum_{k=0}^{N_1-n} FP_{n+k} \\ &= 2P_n\left(2/(1-Z) + \left(N_2 - N_1 - \frac{1+Z}{1-Z}\right) Z^{N_1-n}\right) \\ &\qquad \text{where,} Z = R_n\sqrt{\frac{\pi A_n e^{K/2A_n{R_n}^2}}{2G_n}} \end{aligned}$$

This result applies to the case of two maps. The generalization to a population of many maps is considered for the more general case of missing cuts in the next section.

4 False Positive Probability with Missing Cuts

When misaligned cuts are present in the actual alignment, the false positive probability becomes larger. Assuming the maps are random, we have many possible alignments for a given overlap region, greatly increasing the odds of coming up with a good alignment.

In this case our actual alignment data consists of n pairs of fragment sizes $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$, as before, plus the total number of fragments m in the overlap region of the two maps, where $m = 2n + r$ and $r \geq 0$ is the number of misaligned cuts.

First consider the case where the number of misaligned cuts is fixed at $r = m - 2n$ and the number of aligned fragments is fixed at n in each map and we define the probability $P_{n,m}$ as the probability that two random maps with an overlap region of exactly m total fragments in both maps could produce an alignment of n fragments as good as the actual alignment.

The key to computing $P_{n,m}$ is a systematic way to enumerate possible alignments that can be applied to each sample of the two hypothesized random maps, then compute the probabilities that a particular enumerated alignment will have better sizing error than the actual alignment, and combine these probabilities over all enumerated alignments.

It simplifies matters if we consider random alignments between the left end of two random maps and compute the probability of finding an alignment involving the first m fragments from the left (on either of the two random maps) that is as good as the actual overlap.

We now claim that all possible alignments between the left end of two random maps involving m fragments can be enumerated, independent of the random map sample, as follows:

1. Pick any $n+1$ numbers $s_0, s_1, ..., s_n$ and another $n+1$ numbers $r_0, r_1, ..., r_n$ subject only to the constraints $s_i \geq 1$, $r_i \geq 1$, $\sum_{i=0}^{n}(s_i + r_i) \leq m + 2$

2. Align the two random map samples so that their left ends coincide. Then scan both maps from the left end and pick the s_0th cut site encountered on either map. Then scan further to the right on the other map until another r_0 cut sites have been encountered and align the r_0th one with the previous cut site. This defines the first aligned pair of cuts. The map is now re-aligned so that this pair of cuts coincide (rather than the left ends of the maps).

3. Repeat this step for $i = 1...n$: Starting from the previous aligned cut site scan to the right on both maps until s_i sites have been seen and mark the last one, which could be on either map. Then scan to the right from that cut site on the opposite map only until r_i cut sites have been seen and align the last (r_ith) one with the previously marked cut site. This defines the ith set of aligned fragments. Realign the maps so that this pair of cut sites coincide.

4. After aligning the last ($(n+1)$th) cut site, scan right on both maps until a total of $m+2$ cut sites have been seen (including all cut sites seen in previous steps). Mark the boundary of the aligned region anywhere between the last seen cut site and the next one.

For each enumerated alignment defined by a particular choice of $s_0, ..., s_n, r_0, ..., r_n$ we can compute the probability $PA_{n,m,s,r}$ that for two random maps that particular enumerated alignment will have relative sizing errors better than the actual map alignment. We can then compute an upper bound for $P_{n,m}$ as the sum of $PA_{n,m,s,r}$ over all enumerated alignments.

First we compute E_{r_0,s_0} the probability of no overhang as a function of r_0 and s_0. An overhang occurs if the sum of r_0 random fragments add up to more than the leftmost fragment in the same molecule. Using IID random variables $\delta_1, ..., \delta_r$ each drawn from $\frac{1}{L}e^{-X/L}$ to represent the r intervals, and $Z_1, ..., Z_s$, also drawn from $\frac{1}{L}e^{-X/L}$, to represent twice the s intervals used to select the first aligned cut, we can obtain by suitable integration:

$$E_{r,s} = \frac{1}{3^{r-1}}\left(\frac{1}{3^{s-1}}\binom{r+s-2}{r-1} + \sum_{k=1}^{s-1}\frac{1}{3^k}\binom{r+k-2}{k-1}\right)$$

Using the earlier lemma 2.2, We can write $PA_{n,m,s,r}$ as follows:

$$PA_{n,m,s,r} = E_{r_0,s_0}\mathbb{P}\mathrm{r}\left(\sum_{i=1}^{n} w_i\left(\frac{X_i - Y_i}{X_i + Y_i}\right)^2 \leq E\right)$$

$$\text{where } E \equiv \sum_{i=1}^{n} w_i\left(\frac{x_i - y_i}{x_i + y_i}\right)^2 = nA_n{R_n}^2,$$

and simplify it to

$$PA_{n,m,s,r} \quad \leq \quad E_{r_0,s_0} P_n (2eR_n{}^2 A_n)^{S/2} \left(\frac{n}{n+S}\right)^{(n+S+1)/2} \prod_{i=1}^{n} \frac{(\frac{r_i}{2})!}{(\frac{1}{2})!} \frac{\binom{s_i+r_i-1}{r_i}}{w_i{}^{\frac{r_i-1}{2}}}$$

$$\text{where } N \equiv \max_{1\leq i\leq n} r_i, \text{ and } S \equiv \sum_{i=1}^{n}(r_i - 1).$$

Here P_n is the result without misaligned cuts.

We will now sum up $PA_{n,m,s,r}$ over all possible choices of $s_0, ..., s_n$ while keeping $r_0, ..., r_n$ fixed subject to the constraint $\sum_{i=0}^{n}(s_i + r_i) \leq m + 2$ to produce:

$$\begin{aligned} PA_{n,m,r} &\equiv \sum_s PA_{n,m,s,r} \\ &\leq \frac{1}{2^{r_0}} \left(2\binom{m+1-r_0}{2n+S} + \binom{m+1-r_0}{2n+1+S}\right) \\ &\quad \times P_n (2eR_n{}^2 A_n)^{S/2} \left(\frac{n}{n+S}\right)^{(n+S+1)/2} \prod_{i=1}^{n} \frac{(\frac{r_i}{2})!}{(\frac{1}{2})! w_i{}^{(r_i-1)/2}} \end{aligned}$$

Next, we will add up $PA_{n,m,r}$ over all possible choices of r_0 then $r_1, \ldots, r_n$, where r_0 is constrained by $1 \leq r_0 \leq m - 2n - S + 1$ and $r_1, \ldots, r_n$ by $S \equiv \sum_{i=1}^{n}(r_i - 1) \leq m - 2n$. Approximating w_i by its geometric mean G_n where needed we get:

$$\begin{aligned} P_{n,m} &= \sum_{r_1..r_n} \sum_{r_0} PA_{n,m,r} \\ &\leq \sum_{r_1..r_n} P_n (2eR_n{}^2 A_n)^{S/2} \left(\frac{n}{n+S}\right)^{(n+S+1)/2} \binom{m+1}{2n+1+S} \prod_{i=1}^{n} \frac{(\frac{r_i}{2})!}{(\frac{1}{2})! w_i{}^{(r_i-1)/2}} \\ &\leq P_n \binom{m+1}{2n+1} \left(1 + \sum_{j=2}^{r+1} \left(\frac{(\frac{j}{2})!}{(\frac{1}{2})!}\right) \left(R_n \sqrt{\frac{2A_n}{G_n}}\right)^{j-1} \left(\frac{\binom{m+1}{2n+j}}{\binom{m+1}{2n+1}}\right)\right)^n . \end{aligned}$$

The resulting expression diverges for large values of r but the bound is quite tight for realistic values of m, n, R_n.

As a final step in computing the False Positive probability we need to combine the $P_{n,m}$ just computed over random alignments involving fewer misaligned cuts (smaller values of r) or more aligned fragments (larger n), as well as consider the possible ways the ends of two random maps could be aligned with each other. Using the same approach as for the case without misaligned cuts to model the permissible change in sizing error we can show that the result is:

$$\begin{aligned} FP_r &\leq 4P_n \binom{2n+r+2}{r} \left(1 + \sum_{j=2}^{r+1} \left(\frac{(\frac{j}{2})!}{(\frac{1}{2})!}\right) \left(R_n \sqrt{\frac{2A_n}{G_n}}\right)^{j-1} \left(\frac{\binom{2n+r+1}{r+1-j}}{\binom{2n+r+1}{r}}\right)\right)^n \\ &\quad \left(\frac{1}{1-Z} + \frac{1}{2}\left(N_2 - N_1 - \frac{1+Z}{1-Z}\right) Z^{N_1-n-r/2}\right) \\ &\quad \text{where,} \\ Z &= R_n \sqrt{\frac{\pi A_n e^{K/2A_n R_n{}^2}}{2G_n}} \left(\frac{2n+r+3}{2n+3}\right)^2 \end{aligned}$$

5 False Positive Probability : Population of Maps

Finally if there are multiple maps to choose from, we need to reject the possibility that the proposed map pair is merely the best matching amongst all possible map pairs. We need not consider maps with less than $n+r/2$ fragments. Let the number of maps with at least $n+r/2$ fragments be M, with number of fragments $N_i, i = 1..M$ arranged in ascending order. For each of the possible $M(M-1)/2$ map pairs we can compute the probability FP_r just described, but with N_1, N_2 suitably adjusted. The resulting probability FPT_r is given by:

$$\begin{aligned}
FPT_r &\leq \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} FP_r(N_i, N_j) \\
&\leq P_n \binom{2n+r+2}{r} \left(1 + \sum_{j=2}^{r+1} \left(\frac{(\frac{j}{2})!}{(\frac{1}{2})!} \right) \left(R_n \sqrt{\frac{2A_n}{G_n}} \right)^{j-1} \left(\frac{\binom{2n+r+1}{r+1-j}}{\binom{2n+r+1}{r}} \right) \right)^n \\
&\quad \left(\frac{2M(M-1)}{1-Z} + 4 \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \left(N_j + N_i - \frac{1+Z}{1-Z} \right) Z^{N_i - n - r/2} \right) \qquad (4) \\
&\quad \text{where,} \\
Z &= R_n \sqrt{\frac{\pi A_n e^{K/2A_n R_n{}^2}}{2G_n}} \left(\frac{2n+r+3}{2n+3} \right)^2 \qquad (5)
\end{aligned}$$

Which is to be used with the previous equations for P_n, R_n,A_n,G_n and the error model parameter K (and implicitly σ).

6 Experiment Design

In designing a shot-gun genome wide mapping experiment, one needs to ensure that the data allows correct map overlaps to be clearly distinguished from random map overlaps. If this is done using a False Positive threshold such as the FPT we have derived in this paper, the goal is to ensure that the expected FPT for correct map overlaps does not exceed some acceptable threshold (e.g. 10^{-3}). In this section we will estimate the expected value of FPT for a valid overlap based on the experimental error parameters.

In principle we just need to estimate the values of n, r, R_n, M for a correct overlap based on the experimental errors. However given the extreme sensitivity of FPT on n, the number of aligned fragments, we will compute FPT for correct map overlaps of a certain minimum size. By selecting a suitable value of θ, the minimum "overlap value", we can control the expected minimum value of n, at the cost of some reduction in effective coverage by the factor $1-\theta$ [Waterman95].

In addition to θ assume the following experimental parameters: G = Expected Genome size. L_d = Length of each map. C = Desired coverage (before adjustment for θ). L = average distance between restriction site in Genome. $\sigma\sqrt{X}$ = sizing error (standard deviation) for fragment of size X. P_d = The digestions rate of the restriction enzyme used.

Assuming $R \ll 1$ and $A_n \approx G_n$ we can then write FPT in terms of the experimental parameters as:

$$FPT \approx 2M^2 \binom{\lceil 2nd+2 \rceil}{\lfloor 2n(d-1) \rfloor} \frac{(R\sqrt{\frac{\pi e}{8}})^n}{\sqrt{n\pi}} \left(1 + (d-1)R\sqrt{\frac{2}{\pi}} \right)^n \qquad (6)$$

Where $d = \frac{1}{P_d}$,$n = \frac{L_d \theta}{L P_d{}^2}$, $R = \frac{\sigma}{\sqrt{L/P_d}}$, and $M = \frac{CG}{L_d}$.

7 Conclusion

In this paper we derived a tight False Positive Probability bound for overlapping two maps. This can be used in the assembly of genome wide maps to reduce the search space from exponential time to sub-quadratic time with only a small increase in false negatives. The False Positive Probability bound also can be used to determine if a sequence derived map has a statistically significant match with a map.

We also showed how the False Positive Probability bound can be used to select experimental parameters for whole-genome shot-gun mapping that will allow the genome wide map to be assembled rapidly and reliably and showed that the boundary between feasible and infeasible experimental parameters is quite narrow, exhibiting a form of computational phase transition.

References

[AMAP00] M. Antoniotti, B. Mishra, T. Anantharaman, and T. Paxia, "Genomics Via Optical Mapping IV: Sequence Validation via optical Map Matching", Submitted to *ISMB2001*, Feb 2001.

[AM00] T. Anantharaman and B. Mishra, "Contiging Restriction Maps in Whole Genome Optical Mapping", Technical Report "`http://www.cs.wisc.edu/~tsa/gentig.ps`," 2000.

[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.

[AMS99] T.S. Anantharaman, B. Mishra and D.C. Schwartz, "Genomics via Optical Mapping III: Contiging Genomic DNA," *Proceedings 7th Intl. Conf. on Intelligent Systems for Molecular Biology: ISMB '99*, **7**:18–27, AAAI Press, 1999.

[ASE92] N. Alon, J.H. Spencer and P. Erdös, *The Probabilistic Method*, Wiley Interscience, John Wiley & Sons, New York, 1992.

[Cai+98] W. Cai, et al., "High Resolution Restriction Maps of Bacterial Artificial Chromosomes Constructed by Optical Mapping," *Proc. National Academy of Science*, **95**:3390–3395,1998.

[CMS99] C. Aston, B. Mishra and D.C. Schwartz, "Optical Mapping and Its Potential for Large-Scale Sequencing Projects", *Trends in Biotechnology*, **17**:297–302, 1999

[DEKM98] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis*, Cambridge University Press, 1998.

[GR80] I.S. Gradshteyn and I.M. Ryzhik, *Tables of Integrals, Series and Product*, Academic Press, New York, 1980.

[Lai99] Z. Lai et al., "A Shotgun Sequence-Ready Optical Map of the Whole *Plasmodium falciparum* Genome," *Nature Genetics*, **23,3**:309–313, Nov 1999.

[Lin99] L. Lin et al., "Whole-Genome Shotgun Optical Mapping of *Deinococcus radiodurans* ," *Science*, **285**:1558–1562, Sept 1999.

[Reed98] J. Reed et al., "A Quantitative Study of Optical Mapping Surfaces by Atomic Force Microscopy and Restriction Endonuclease Digestion Assays", *Analytical Biochemistry*, **259**:80–88, 1998

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

[Waterman95] M.S. Waterman, *Introduction to Computational Biology*, Chapman and Hall, 1995.