JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 2, 2006
© Mary Ann Liebert, Inc.
Pp. 442–462

Alignment of Optical Maps

ANTON VALOUEV,[1] LEI LI,[1] YU-CHI LIU,[2] DAVID C. SCHWARTZ,[3] YI YANG,[1]
YU ZHANG,[4] and MICHAEL S. WATERMAN[2]

ABSTRACT

We introduce a new scoring method for calculation of alignments of optical maps. Missing cuts, false cuts, and sizing errors present in optical maps are addressed by our alignment score through calculation of corresponding likelihoods. The size error model is derived through the application of Central Limit Theorem and validated by residual plots collected from real data. Missing cuts and false cuts are modeled as Bernoulli and Poisson events, respectively, as suggested by previous studies. Likelihoods are used to derive an alignment score through calculation of likelihood ratios for a certain hypothesis test. This allows us to achieve maximal descriminative power for the alignment score. Our scoring method is naturally embedded within a well known DP framework for finding optimal alignments.

Key words: optical mapping, alignment score, restriction mapping, dynamic programming, likelihood ratio.

1. INTRODUCTION

Optical mapping is a powerful technology that allows construction of ordered restriction maps. Each optical map is a single DNA molecule digested by the restriction enzyme and imaged by the optical system. The map is comprised of estimates of fragment sizes in the order they appear on the imaged molecule. Hence, there are two types of information associated with each optical map: sizes of restriction fragments on the molecule and their relative order.

A broad spectrum of problems can be effectively addressed by means of optical mapping. Among the most important are analysis of genomic variation (deletions, insertions, and rearrangements on a scale of thousands of base pairs), construction of genomewide restriction maps without knowledge of the original DNA sequence, validation and finishing of sequence contigs, and genomic placement during sequencing projects. Despite the fact that optical mapping lacks single nucleotide resolution, it is capable of quickly

[1] Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113.
[2] Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089.
[3] Laboratory for Molecular and Computational Genomics, Departments of Genetics and Chemistry, University of Wisconsin-Madison, Madison, WI 53706.
[4] Department of Statistics, Harvard University, Cambridge, MA 02138.

assessing genomic differences as well as genotyping a human in a matter of a few hours, which is presently not achievable by means of sequencing. The domain of application of optical mapping is somewhat complementary to that of sequencing: quick genomewide analysis with resolution of several hundreds of base pairs is now routinely done for human-sized genomes.

To address these problems by means of optical mapping, tools for revealing homologies between optical maps need to be developed. More generally, such tools should include genomic placement of individual optical maps and potentially be capable of genomewide restriction mapping while remaining computationally feasible. Similar problems have been solved for DNA sequences. Development of sequence alignment dynamic programming (DP) algorithms (Waterman *et al.*, 1984; Huang and Miller, 1991) for accurate comparison of DNA and protein sequences provided tools for genomewide shotgun sequencing which was successfully implemented and applied to a variety of organisms (Myers, 1999; Huang and Madan, 1999).

It is therefore natural to explore the possibility of applying similar ideas to optical maps and study the feasibility of such methods. A significant amount of work has already been done in this direction. DP algorithms have been successfully used for restriction map alignments (Waterman *et al.*, 1984; Huang and Waterman, 1992; Myers and Huang, 1992). With the development of the Optical Mapping System (Dimalanta *et al.*, 2004; Zhou *et al.*, 2004; Ambrust *et al.*, 2004), much effort has been made to develop methods for genome wide restriction mapping. Ananthraman *et al.* (Ananthraman *et al.*, 1997, 1999, 2001; Ananthraman and Mishra, 2001) have designed an extensive Bayesian framework for optical mapping with potential for global map assembly. Its use, however, remained limited as laboratories now confront large genomes such as human or mouse.

In this paper, we present a complete statistical model that enables us to design a likelihood ratio based alignment score with the optimal discriminant power for distinguishing correct alignments from spurious ones. Statistical models corresponding to different error types in optical maps allow us to calculate the likelihoods of data observed in optical maps. Overall, the exact form of the alignment score follows from a probabilistic model associated with the way optical maps are generated. Appropriate conditioning allows a natural decomposition of scores into a sum of two components, first to account for sizing errors, and second to account for the presence of false cuts and missing cuts. The designed alignment score is implemented within a standard DP alignment framework similar to that used for sequences (Smith and Waterman, 1981) and restriction maps (Waterman *et al.*, 1984). Complexity of finding an optimal alignment for two maps with m and n fragments is $O(m^2n^2)$, but can be accurately approximated by a restricted $O(\delta^2mn)$ version (for all practical purposes $\delta \leq 5$).

2. MODELS

The errors associated with optical maps include missing cuts and false cuts, missing fragments, sizing errors, and chimeric reads. We will explore each in more detail.

Missing cuts and false cuts. The efficiency of DNA digestion by the restriction endonuclease is never perfect. As a result, some restriction sites on DNA remain uncut by endonuclease even after the DNA is subjected to digestion. After the molecule is imaged, corresponding restriction fragments remain concatenated in the output data, appearing as if they came from a single restriction fragment of the combined size. Digestion rate is monitored after the digestion has taken place. Many copies of a $\lambda-$phage of known size and number of restriction sites are comounted in the solution together with the target DNA, so that digestion rate in the solution can be screened. False cuts result from random DNA breaks or nonspecific action of endonuclease. Under our model assumptions, random breaks show no preference to particular regions of DNA and thus occur equally likely in all regions.

Missing fragments. After being extracted from cells, genomic DNA is deposited inside microfluidic channels and attached to the glass surface by means of capillary action. Digestion creates restriction fragments of different sizes, some of which are too small to firmly hold to the glass surface. These fragments are carried away by the flow of the buffer through the microfluidic channel and therefore are not captured by the imaging system. Many fragments shorter than 2 kb are often missing in the data. To simplify the calculations, we do not address missing fragments in our likelihoods. However, when actual alignments are calculated, short fragments are removed from both optical maps and reference maps to avoid unnecessary penalties.

Sizing errors. During the fluorescent marking of DNA, the dye is attached along the span of the molecule in a random fashion. The size of each restriction fragment within the optical map is detected by measuring the fluorescence intensity emitted by the fluors attached to the corresponding piece of DNA. To determine the size of the fragment, its intensity is compared to the measurement standard that corresponds to the amount of intensity associated with the DNA of a known size.

Consider a fragment y bps long. Suppose that W_i is the amount of intensity per i-th nucleotide of the fragment and $EW_i = \mu$, $Var(W_i) = \gamma^2$. The total amount of intensity accumulated along the length of the fragment is given by $W = \sum_{i=1}^{y} W_i$. The estimated fragment size is given by $X = \hat{y} = W/\mu$. Hence, after applying the central limit theorem, we conclude that $X \sim N(y, \sigma^2 y)$, where $\sigma^2 = (\gamma/\mu)^2$.

Further, suppose that ϵ, defined as $\epsilon = X - y$, is the measurement error. Of course, $\epsilon \to N(0, \sigma^2 y)$ in distribution as n increases. To show the validity of our size error model, we have collected pairs (X, y) from accurate alignments of optical maps of *Y. pestis* and reference genome, where X represents estimated fragment sizes on optical maps and y represents their underlying reference sizes. Since, according to our model, $X \sim N(y, \sigma^2 y)$, then for $e = \frac{X-y}{\sqrt{y}}$ we have $e \sim N(0, \sigma^2)$. Thus e should be homoscedastic across y and marginally normal. Figure 1 illustrates that the variance of e is homogenous across y and e is close to a normal variable.

The described error model agrees well with the data from relatively long fragments ($\geq$ 4 kb) observed in the optical maps (Fig. 1). However, for small fragments, the normality assumption fails due to lack of convergence to normal density. Other physical effects contribute as well. Balling of DNA at the ends of fragments reduces the accuracy of size estimation for short fragments. To take this effect into account, the sizing error for short fragments is modeled as $\epsilon \sim N(0, \eta^2)$, for an appropriate choice of η^2.

Chimeric reads. When two molecules cross, there is no way to confidently disambiguate the strands at the resolution of light microscopy. Consequently, some reads are chimeric, meaning that their portions represent multiple unrelated genomic regions. In this paper, chimeric maps are not addressed statistically. In other problems, when chimeras need to be identified, we use the alignment score to distinguish them from correct optical maps.

Restriction fragment size distribution. If we assume that DNA sequences are generated by independent sampling from a four-letter alphabet, then it can be shown that distribution of restriction fragment sizes can be approximated by an exponential curve. However, it is known that sequences are nonrandom and therefore some care is needed to address the issue. Churchill *et al.* (1989) studied the distribution of *E. coli* restriction fragment sizes and found that exponential density fits best for this distribution compared to several other densities they looked at. Our experience confirms that this assumption remains valid for other genomes. The average restriction fragment size needs to be estimated from the reference genome or optical maps themselves, since it depends on the nucleotide content as well as the restriction sequence. In our approach, we estimate this number for an organism and a restriction enzyme of interest.

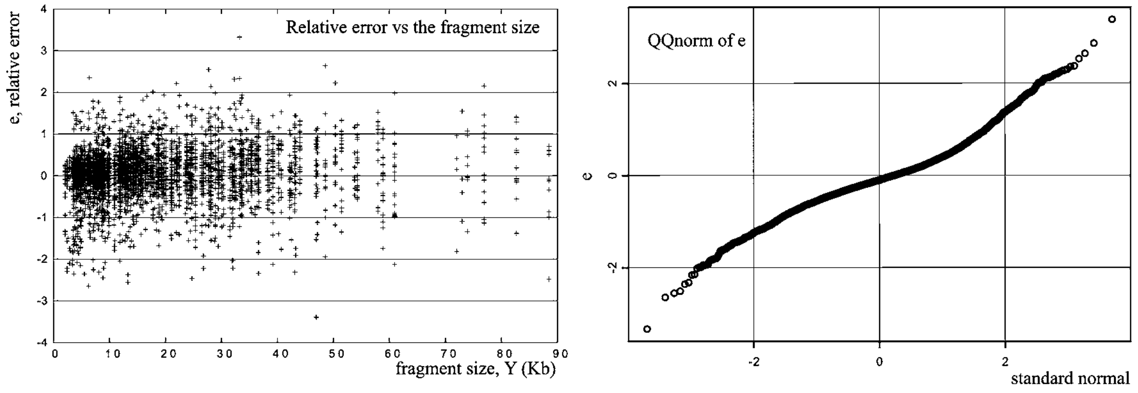

FIG. 1. Sizing error model. **Left** graph shows relative error $e = \frac{X-y}{\sqrt{y}}$ across different values of y, where X are sizes of fragments on optical maps and y are their true underlying sizes. It can be seen that e is homoscedastic. **Right** graph shows that the marginal density of e is close to normal.

Model assumptions and parameters. To summarize, we use the following set of assumptions hereafter in our analysis:

1. Sizes of genomic restriction fragments Y have exponential density with mean $EY = \lambda$, with λ depending on a particular endonuclease. Consequently, the number of restriction sites in s kb of linear DNA is a Poisson process with intensity s/λ.
2. Observation of a restriction site on the optical map is a Bernoulli event with probability p, which we also refer to as the digestion rate. Further, we assume that restriction sites are being digested independently, and hence observed restriction sites are explained by a thinned Poisson process with intensity p/λ. In most reported maps, digestion rate is high, i.e., $0.5 \leq p \leq 1$.
3. The number of false cuts per s kb of linear DNA is a Poisson process with intensity ζs.
4. The length discrepancy between the optical fragment of size X and its underlying true size Y, given by $\epsilon = X - Y$, has a normal density $N(0, \sigma^2 Y)$ for $X \geq \Delta$. Further, $\epsilon \sim N(0, \eta^2)$ for $0 < X < \Delta$. For all practical situations, $\Delta = 4$ kb is a reasonable threshold separating the two error models. Also, both X and Y always assume nonnegative values, because negative fragment sizes are never observed in maps.
5. For the purpose of approximations, we require $\lambda \gg \sigma^2$, which indeed holds in all practical situations.

3. LIKELIHOOD RATIOS FOR DIFFERENT MATCH TYPES

The purpose of designing a score for matching regions of the compared restriction maps is to utilize it in the alignment calculation. The general idea behind this method is to find an optimal alignment configuration between compared maps that will maximize the alignment score. In our approach, alignment score is designed as the log of ratios of likelihoods calculated under two hypothesis. The null hypothesis H_0 corresponds to the situation when compared maps are completely independent. Alternative hypothesis H_a corresponds to the situation when maps are related and represent the same genomic regions. As such, the optimal alignment will correspond to a configuration with maximum distance between compared hypothesis H_0 and H_a.

Our alignment will be given by the minimum partition of the alignment into blocks corresponding to the minimal matching regions between compared maps. Finding the minimal partition is accomplished through maximization of the alignment score, since mismatching sites are penalized by the alignment scores. Here, we define the alignment block as the two matching regions of maps that are flanked by matching restriction sites and do not contain any internal matching sites.

In this paper, we consider two situations. The first situation corresponds to comparison between the optical map and reference map. Reference map is usually derived from the known DNA sequence and hence can be thought to be free of inaccuracies. In this case, only one of compared maps contains randomness in the form of errors (missing cuts, false cuts, sizing inaccuracies). We refer to this situation as *reference match*. The second situation arises when two optical maps are compared to yield a common region. In this case, both maps will contain randomness in the form of errors. This situation is referred to as *optical match*.

We first present some probabilistic results that we use to derive the likelihoods under specified hypotheses. In the second part of this section, we will present theorems that give expressions for alignment scores. Our likelihood ratios are derived under certain conditioning of probabilities. The details of this are not important for understanding the idea behind the method; interested readers can find details of our derivations in the appendix.

3.1. *Likelihoods associated with matching regions of optical maps*

In this section, we present some preliminary results we use to simplify calculation of the alignment score. Most of the facts are trivial, and therefore proofs are placed in the appendix.

Lemma 1. Distribution of number of cut sites on optical maps. *Under the model assumptions, the number X of cut sites per s kb of linear DNA on an optical map is comprised of restriction cuts and random breaks. It has Poisson distribution with intensity s/τ, where $\tau = \left(\zeta + \frac{p}{\lambda}\right)^{-1}$.*

Lemma 2. Distribution of fragment sizes on optical maps. *Under the model assumptions, measured sizes X of fragments on optical maps have exponential density with the mean $\theta = \left[\frac{1}{\sigma}\sqrt{\frac{2}{\tau} + \frac{1}{\sigma^2}} - \frac{1}{\sigma^2}\right]^{-1}$ for $X \geq \Delta$.*

Proposition 1. Size distribution of matching regions between optical and reference maps. *Under the model assumptions, reference sizes of matching regions between optical and reference maps have exponential density with the mean $\upsilon = \lambda/p$.*

Proposition 2. Size distribution of matching regions between optical maps. *Under the model assumptions, reference sizes underlying matching regions between two optical maps have exponential density with the mean $\phi = \lambda/p^2$.*

Proposition 3. Size distribution for sum of multiple optical map fragments. *Consider $X = \sum_{i=1}^{m} X_i$ to be the total size of a region of an optical map comprised of m fragments. Then, under normality assumptions of the sizing error model, $X \sim N(Y, \sigma^2 Y)$, where Y is the underlying size of reference region from where $X_1, X_2, \ldots, X_m$ originate. This implies that the sizing error is independent of occurrences of false cuts and missing cuts.*

3.2. Likelihood ratios for optical-reference map match

In this section, we provide the formulas for likelihood ratios that we use for the calculation of the alignment score. The likelihood ratios are given for the minimal decomposition of the alignment into the blocks. In other words, since the decomposition is minimal, there are no matching sites contained within the blocks (i.e., between minimally decomposed regions of the compared maps). This implies that all sites within the block can be treated as false cuts and missing cuts, and hence double counting of matches inside the block is not necessary. The maximization of the alignment score implemented by the algorithm is given later in this paper. This algorithm accomplishes minimal decomposition into alignment blocks; hence, the use of the alignment scores in this form is appropriate.

Consider the situation of comparing the optical map to the reference map. Suppose that a region of an optical map of size X is comprised of m consecutive fragments $X_1, \ldots, X_m$ so that $X = \sum_{i=1}^{m} X_i$. Likewise, suppose Y is the size of the reference region comprised of n consecutive fragments on the reference map ($Y = \sum_{j=1}^{n} Y_j$). The likelihood ratio for comparing sizes of two regions is given by the following theorem.

Theorem 1. Reference size match LR. *The likelihood ratio for matching the region of size x comprised of m of optical map fragments to the region of size y comprised of n reference map fragments is given by*

$$LR(x|y,n,m) = \frac{\sqrt{2\pi y}\sigma x^{m-1}}{\Gamma(m)\theta^m} exp\left[\frac{(x-y)^2}{2\sigma^2 y} - \frac{x}{\theta}\right]$$

for $x > \Delta$. Furthermore, for $0 < x \leq \Delta$, the likelihood ratio has the form

$$LR(x|y,n,m) = \frac{\sqrt{2\pi}\eta x^{m-1}}{\Gamma(m)\theta^m} exp\left[\frac{(x-y)^2}{2\eta^2} - \frac{x}{\theta}\right].$$

Note that formulas for likelihood ratios do not involve the reference fragment number n because the sizing error is independent of n. The last theorem gives the expression for comparing sizes conditioned on the number of fragments within each region. The next theorem will present a comparison of marginal likelihoods for the number of fragments for matching regions of compared maps. This allows one to account for optical map errors in the form of missing cuts and false cuts that will appear in our alignments in the form of nonmatching sites.

Theorem 2. Reference site LR. *The likelihood ratio for the matching region comprised of m optical map fragments and the reference region comprised of n reference map fragments given the reference region* *of size y has the form*

$$LR(m|y,n) = \frac{e^{\zeta y}(m-1)!\, f_M(m)}{(1-p)^{n-1}(\zeta y)^{m-1}},$$

where $f_M(m)$ is the marginal density of m (the exact form of $f_M(m)$ is discussed in the proof).

3.3. Likelihood ratios for optical map match

Consider now a situation when two optical maps are compared to identify common genomic regions. Let x_1 and x_2 be the sizes of two regions of such maps. Suppose these regions are comprised of m_1 and m_2 consecutive optical map fragments, respectively. We want to design a test for identifying whether regions (x_1, m_1) and (x_2, m_2) of two maps represent the same genomic regions.

Theorem 3. Optical map size match LR. *The likelihood ratio for matching the region of size x_1 comprised of m_1 consecutive optical map fragments and the region of size x_2 comprised of m_2 consecutive optical map fragments from another map has the form*

$$LR(x_1, x_2|m_1, m_2) = \frac{\pi\phi\sigma^2 x_1^{m_1-1} x_2^{m_2-1}}{\Gamma(m_1)\Gamma(m_2)\theta^{m_1+m_2}} \times \frac{exp\left[-\left(\frac{1}{\theta}+\frac{1}{\sigma^2}\right)(x_1+x_2)\right]}{K_0\left(2\sqrt{\frac{1}{\phi}+\frac{1}{\sigma^2}}\sqrt{\frac{x_1^2+x_2^2}{2\sigma^2}}\right)},$$

where $K_\nu(z)$ is a modified Bessel function of the second kind.

The previous theorem gives the likelihood ratio conditioned on the number of fragments in each of the regions. Next, we give an expression for the likelihood ratio of marginal densities of fragment numbers within two regions of optical maps. This expression accounts for false cuts and missing cuts contained within such maps. Such cuts will not be matched when regions are compared.

Theorem 4. Optical map site LR. *Consider two regions of optical maps comprised of m_1 and m_2 fragments, respectively. The likelihood ratio for matching m_1 and m_2 has the form $LR(m_1, m_2) = \frac{C}{B}$, where $C = Pr(m_1, m_2)_{H_0} = f_{M_1,M_2}(m_1, m_2)$ and*

$$B = \frac{1}{\phi}\sum_{n=0}^{\infty}\frac{1}{\lambda^n n!}\sum_{k_1=0}^{m_1-1\wedge n}\ \sum_{k_2=0}^{m_2-1\wedge n-k_1} A(k_1,k_2|n) \times \frac{\zeta^{(m_1+m_2)-(k_1+k_2)-2}}{\left[\frac{1}{\phi}+\frac{1}{\lambda}+2\zeta\right]^{n+(m_1+m_2)-(k_1+k_2)-1}}$$
$$\times \frac{\Gamma(n+(m_1+m_2)-(k_1+k_2)-1)}{\Gamma(m_1-k_1)\Gamma(m_2-k_2)},$$

where

$$A(k_1,k_2|n) = \frac{\binom{n}{k_1\ k_2} p^{k_1+k_2}(1-p)^{2n-(k_1+k_2)}}{\sum_{i_1=0}^{n}\sum_{i_2=0}^{n-i_1}\binom{n}{i_1\ i_2} p^{i_1+i_2}(1-p)^{2n-(i_1+i_2)}}.$$

The exact form of $f_{M_1,M_2}(m_1, m_2)$ is discussed in the proof.

4. CALCULATION OF THE ALIGNMENTS

Two types of alignments are discussed in this paper: fit and overlap alignment. Other types of alignments are computed similarly with a slight modifications to initialization (see Waterman [1995]). Fit alignment is used to calculate a proper fit of the the optical map into the reference. Indeed, if the whole restriction map of the organism is known, any of its genomic optical maps should properly fit into the reference. Therefore, the fit alignment is used for the purpose of finding the best fit of the optical map against the reference to locate the genomic region represented by that optical map. Overlap alignment allows detection of shared genomic regions represented by two optical maps. This corresponds to the situation when two maps can be aligned only partially and one of the tails of each map may remain unaligned.

In this section, we outline an alignment algorithm that utilizes an alignment score that we have derived. For the fit alignment, suppose that maps R and O correspond to the reference restriction map with n sites and the optical map with m sites. As before, we count both the start and end positions as first and last sites. Additionally, suppose that numbers $0 = q_0 < q_1 < \cdots < q_m = s_O$ and $0 = r_0 < r_1 < \cdots < r_n = s_R$ mark the positions corresponding to the sites on optical and reference maps, respectively (here s_O and s_R are sizes of optical and reference maps, respectively). The fit alignment Π is defined as the ordered set of aligned sites $(i_0, j_0)\ (i_1, j_1)\ \ldots\ (i_d, j_d)$, where $0 \leq i_0 < i_1 < \cdots < i_d \leq m$ and $0 = j_0 < j_1 < \cdots < j_d = m$ corresponding to the indices of the restriction sites on the reference and optical maps, respectively.

For overlap alignment, consider optical maps O_1 and O_2 with m_1 and m_2 restriction sites, respectively. Also, suppose that numbers $0 = q_0 < q_1 < \cdots < q_{m_1} = s_{O_1}$ and $0 = r_0 < r_1 < \cdots < r_{m_2} = s_{O_2}$ mark the positions corresponding to the sites on both maps (here s_{O_1} and s_{O_2} are sizes of two optical maps). The overlap Π is defined as the ordered set of aligned sites $(i_0, j_0)\ (i_1, j_1)\ \ldots\ (i_d, j_d)$, where either $0 \leq i_0 < i_1 < \cdots < i_d = m_1$ and $0 = j_0 < j_1 < \cdots < j_d \leq m_2$, or $0 = i_0 < i_1 < \cdots < i_d \leq m_1$ and $0 \leq j_0 < j_1 < \cdots < j_d = m_2$ corresponding to the indices of matching sites on maps O_1 and O_2, respectively.

Alignment algorithm. Let $X(i, j)$ be the score of the largest scoring alignment with the right-most aligned pair of sites (i, j). The alignment score is calculated according to the Algorithm 1 of Huang and Waterman (1992).

Algorithm 1. Dynamic programming for calculation of alignments of optical maps.

```
Initialization:
Fit: $X(i,0) \leftarrow -\infty,\ i = 1, \ldots, m;\ X(0,j) \leftarrow 0,\ j = 0, \ldots, n$
Overlap: $X(i,0) \leftarrow 0,\ i = 1, \ldots, m;\ X(0,j) \leftarrow 0,\ j = 0, \ldots, n$

Recursion:
for $i \leftarrow 1$ to $m$ do
    for $j \leftarrow 1$ to $n$ do
        $y \leftarrow -\infty$;
        for $g \leftarrow max(0,\ i - \delta)$ to $i - 1$ do
            for $h \leftarrow max(0,\ j - \delta)$ to $j - 1$ do
                $y \leftarrow max\{y,\ X(g,h) + S\left(q_i - q_g, r_j - r_h, i - g, j - h\right)\}$;
            end
        end
        $X(i,j) \leftarrow y$;
    end
end
```

Score. The exact form of the score $S\left(q_i - q_g, r_j - r_h, i - g, j - h\right)$ is computed according to likelihood ratios presented above. For the fit alignment, it has the form

$$S\left(q_i - q_g, r_j - r_h, i - g, j - h\right) = -log(LR(q_i - q_g; r_j - r_h, i - g, j - h)) \\ - log(LR(i - g; r_j - r_h, j - h)),$$

where the first ratio is given by Theorem 1 and the second is given by Theorem 2. Similarly, for the overlap alignment, the score has the form

$$S\left(q_i - q_g, r_j - r_h, i - g, j - h\right) = -log(LR(q_i - q_g, r_j - r_h; i - g, j - h)) - log(LR(i - g, j - h)),$$

where the first ratio is given by Theorem 3 and the second is given by Theorem 4.

Complexity. Computational complexity of both fit and overlap alignments is $O(\delta^2 mn)$. In practice, δ is taken to be small ($\delta \leq 5$), and this both improves the alignments and reduces computation complexity to $O(mn)$, where m and n are numbers of fragments in the compared maps.

5. PRACTICAL ISSUES

The models in Section 2 assume a set of distribution parameters. In this section, we describe these parameters with respect to the specific experimental setup.

Sizing error σ^2. We can estimate σ^2 using several different methods. For the first method, a sample of prepared and imaged lambda DNA of known size s can be used to estimate the average intensity $s\hat{\mu}$ and intensity variance $s\hat{\gamma}^2$. Then σ^2 is estimated as $\hat{\sigma}^2 = \frac{s\hat{\gamma}^2}{s\hat{\mu}^2} = \frac{\hat{\gamma}^2}{\hat{\mu}^2}$. Alternatively, σ^2 can be estimated by empirical means, and then the fits of the optical maps into the reference are calculated. High scoring (or confident) alignments are chosen. From these alignments, we can infer the pairs (X, Y) where X is the size of the fragment on the optical map and Y is its true underlying size (matching fragment on the reference). From this data, we can estimate distribution parameters of $E = \frac{X-Y}{\sqrt{Y}}$ (mean, variance, based on the fact that $E \sim N(0, \sigma^2)$). For the *Y. Pestis* optical mapping project, the value of σ^2 was chosen to be 0.306. Generally speaking, although σ^2 may vary from experiment to experiment, this variation is small and σ^2 may be assumed the same for different mapping projects, once it has been confidently estimated.

Sizing error for small fragments. Another important parameter for the error model is η, which corresponds to the standard deviation of the error ϵ when the underlying size X is small ($X \leq \Delta$). The choice of η^2 depends on the amount of confidence assigned to matching of short fragments. For the *Y. Pestis* mapping project, we have set $\eta^2 = 5$.

Average size λ of restriction fragments. As has been discussed above, the distribution of sizes for reference restriction fragments is taken to be exponential with the mean λ. The value of λ depends on the frequency of cut sites as they appear along the DNA chains digested by restriction endonuclease. Endonuclease is site specific and recognizes a specific palindromic sequence of nucleotides. In the *Y. Pestis* optical mapping project, a 6-cutter $Xho - I$ was used to perform a digestion. This enzyme is specific to a sequence of six consecutive nucleotides $5' - CTCGAG - 3'$. It performs a cleavage after the second nucleotide leaving sticky ends. If the DNA sequence was to assume a uniform distribution of nucleotides, then the average size of the restriction fragment would be $4^6 \approx$ 4,000 bp. However, it is well-known that the sequences are nonrandom. From the reference sequence of *Y. Pestis*, obtained from the NCBI sequence database, the average size of the restriction fragment was estimated to be about 17 kb.

Digestion rate p. The digestion rate p can be estimated both from the alignments and λ-DNA. Here, λ-DNA is certainly more preferable since it has no missing fragments, which can affect the estimation of p. However, we should also mention that partially digested λ-DNA is much harder to identify during imaging, so that a bias in estimation of p may become a problem.

DNA breakage rate ζ. Another important parameter is the rate of random breakage ζ. Its estimate comes from the alignments. We have assumed $\zeta = 0.005$ for our calculations.

Size of maximum matching region δ. Finally, we should discuss the choice of δ. This parameter defines how far back we look in the alignment matrix for finding the optimal score for each entry. Thus, δ defines an upper bound on the number of fragments allowed within the matching region. For the regions of the reference map, this number rarely exceeds three since $P(\textit{more than } 3 \textit{ sites in a row are undigested}) \leq p^3 \approx 0.008$. Similarly, the probability of having more than three false cuts per average-sized matching region is less than 0.002. If more accurate alignments are needed, δ can be chosen appropriately. It is therefore reasonable to set $\delta = 5$ to capture most alignments. Restricting the value of δ to a small integer also significantly reduces the speed of the alignment computation.

Due to the nature of the likelihood ratio based score, rare matching fragments score higher than frequent ones. Consequently, longer fragments will score higher than short ones. As a result, unusually high scores can appear in regions containing large fragments. Such high scores may result in spurious matches of regions and ruin calculation of score based p-values. To address this issue, we use a score truncation where this becomes a problem. In other words, an upper score threshold is chosen, above which the returned score is that threshold. This helps to eliminate situations where many maps spuriously align to some genomic position due to an unusual restriction pattern at that genomic location.

Short fragments also present a challenge in calculation of alignment scores. A large number of short fragments is missing in optical map data. As a result, when optical maps are aligned to a reference region with an abundance of short fragments, alignments receive score penalties due to these missing fragments. To overcome this limitation, short fragments are removed from the reference, thus allowing for a better alignment of optical maps.

It is worth mentioning that a more accurate size error model may be utilized to describe measured fragment sizes as the ratio of two normal random variables. Although perhaps more accurate, this model will drastically increase the complexity of score computation with little improvement in terms of score quality.

6. RESULTS

We have tested our likelihood ratio based score on a variety of organisms and sets of optical maps. Our scoring method is highly capable of producing accurate alignments.[1] Figure 2 illustrates one such alignment along with the scoring pattern. Overall, spurious alignments have low scores, usually close to or less than zero (natural logs), while correct ones usually produce very high scores, usually in the neighborhood of 100 (natural logs) depending on the map size. Naturally, longer maps are easier to align correctly because they contain more information in the form of fragment sizes.

Next we demonstrate the advantage of our new likelihood based alignment score over the heuristic alignment score for restriction maps proposed by Waterman *et al.* (1984). Hereafter, "likelihood score" refers to the score described in this paper, and "heuristic score" refers to the heuristic alignment score due to Waterman *et al.* (1984). Table 1 demonstrates some comparative results between these two alignment scores. We have synthetically generated 1,000 maps from a region of human chromosome 13. For generation of maps, we have used our model assumptions for sizing errors, missing cuts, and false cuts. This made it possible to infer exact positions from where maps were generated, as well as the exact positions of all true cut sites, together with information concerning which cuts are spurious. We then aligned these maps to the reference genomic restriction map of the human chromosome 13. By comparing original and fit positions, we were able to infer whether the fits were correct. For correct fits, we have collected information about correctly and incorrectly aligned sites and summarized them in Table 1.

As is evident, both alignment methods produce similar results for synthetic maps with our new method, likelihood having a slight advantage. However, the most important feature of the score is its discriminative power for distinguishing true and spurious alignments. It turns out to be significantly different for the two scoring schemes (refer to Table 2). For each synthetic map that fits into the correct reference location (using both likelihood and heuristic scores), we have collected its 20 best alignment scores (10 in each orientation, including the score for the true fit) by declumping dependent alignments. The idea is that if the discriminative power of the score is high, no other independent alignments will have scores in the proximity of the score of the true alignment (corresponding to the correct fit). To measure this, we counted the number of other alignment scores (from the 19) within k standard deviations of the true alignment score. Standard deviation of that score was estimated based on fit scores from 10,000 random maps with the same number of fragments as the map of interest. The results for both methods are summarized in Table 2.

As now becomes apparent, the likelihood score possesses significantly more discriminative power since even within three standard deviations of the true fit score, on average 0.17 spurious fit scores are observed, while the same number for the heuristic score is 6.28 ($\approx$ 37 times more).

[1] The source code of the alignment program is available upon request.

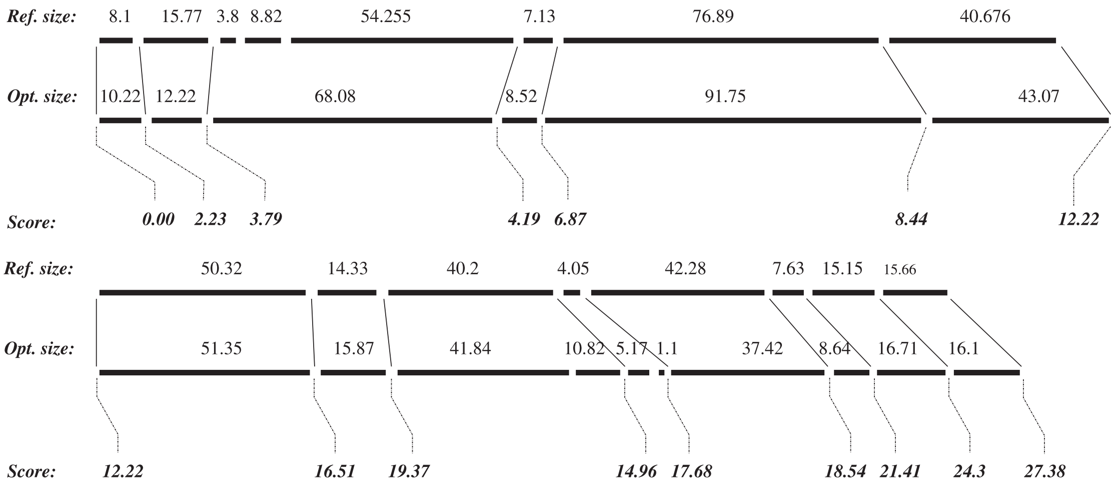

FIG. 2. Representation of alignments. Alignment of part of optical map from *Y. Pestis* against its reference genomic region. Two adjacent pieces of alignment are displayed along with the map fragment sizes and alignment scores. Gaps between fragments represent cut sites; thin lines connect matching sites. Scores are calculated directionally (from left to right) and are displayed at the matching sites. Score for each alignment block is obtained by taking the difference of the scores at the two flanking matching sites. Fragment sizes are displayed in kb.

Not surprisingly, our likelihood score, when compared to the heuristic score, shows even more discriminative power for real optical maps when compared to the same result based on synthetic maps. Hence, our alignment score produces significantly more confident alignments for real optical map data compared to the heuristic score.

We used a Kim strain of *Y. Pestis* commonly known as plague to study the performance of the alignment score on real data. The data set was comprised of 251 optical maps. The reference restriction map was inferred from the NCBI sequence database. This bacterial genome consists of 4.6 Mb and 267 restriction

TABLE 1. COMPARATIVE FITTING FOR TWO SCORING SCHEMES (BASED ON 1,000 SYNTHETICALLY GENERATED MAPS FROM A 40 Mb REGION OF HUMAN CHROMOSOME 13)

	% of maps fitted into correct reference locations	*% of false positive cut sites*	*% of false negative cut sites*
lik. sc.	95.1	7.9	10.1
heur. sc.	92.6	10.5	15.3

[a] Maps are being fitted into the reference. False positives and false negatives rates are calculated for the maps fitted into correct reference location using both methods. False positives represent the percentage of false cuts incorrectly matched to the reference, false negatives represent percentages of unmatched true restriction sites. The corresponding alignment parameters are chosen to be $\mu = 0.2$, $\nu = 1$, $\lambda = 2$ for the heuristic score and $\sigma^2 = 0.306$, $\lambda = 32$, $p = 0.8$, $\zeta = 0.005$, $\eta^2 = 5$ for likelihood score.

TABLE 2. OPTIMAL SCORE DISCRIMINATIVE POWER FOR TWO SCORING SCHEMES (BASED ON 1,000 ARTIFICIALLY GENERATED MAPS FROM A 40 Mb REGION OF HUMAN CHROMOSOME 13)

Stdev Num	*1 SD*	*2 SD*	*3 SD*	*4 SD*	*5 SD*	*6 SD*	*7 SD*	*8 SD*	*9 SD*
Av. number of lik. scores	0.006	0.036	0.17	0.494	1.413	3.206	5.983	8.89	11.82
Av. number of heur. scores	0.281	2.157	6.275	11.471	15.334	17.546	18.566	18.91	18.99

[a] Scoring parameters for likelihood and heuristic scores are taken to be the same as in Table 1. Of the 20 top scores of independent alignments (10 in each orientation), the table gives the average number of scores (other than optimal) within k standard deviations ($k = 1, 2, \ldots, 9$) of the optimal score. Hence, table entries never exceed 19.

Table 3. Comparative Fitting for Two Scoring Schemes
(Based on Optical Maps from a 4.6 Mb Genome of *Y. Pestis*)

Map id	*0499*	*1661*	*0387*	*0344*	*0517*	*0920*	*0944*	*1634*	*1659*
Fragments	70	34	77	27	85	37	71	23	79
Lik. sc.: optimal	80.56	45.24	62.32	31.79	113.61	51.11	47.54	33.60	90.24
Lik. sc.: optimal p-value	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0018	$< 10^{-4}$	$< 10^{-4}$	0.0003	0.00065	$< 10^{-4}$
Heur. sc.: optimal score	−31.76	−13.13	−35.59	−1.53	−39.50	−15.86	−51.22	−1.25	−34.68
Heur. sc.: optimal p-value	0.0005	0.083	0.002	0.0074	0.00055	0.181	0.1603	0.006	0.00135

[a] Likelihood score refers to likelihood ratio scoring scheme ($\sigma^2 = 0.306$, $\lambda = 17$, $p = 0.8$, $\zeta = 0.005$, $\eta^2 = 5$) and heuristic score refers to a score suggested by Waterman *et al.* (1984) ($\mu = 0.2$, $\lambda = 2$, $\nu = 1$). Optimal score corresponds to the top alignment score for the given map. Optimal p-value corresponds to the p-value of the optimal score. Total of 251 different maps were fitted. For likelihood, 174 passed 0.002 p-value threshold; for heuristic, 119 passed 0.002 p-value threshold.

sites specific to the XhoI endonuclease that was used for digestion of its DNA during collection of optical maps.

For real optical maps, we have calculated alignments using both likelihood and heuristic scores. Furthermore, a p-value threshold of 0.002 was used to select confident fits. As a result, the likelihood score allowed 174 maps to pass the p-value threshold, while only 119 maps passed the same threshold using the heuristic score. Notably, our alignment score allowed us to confidently align 47% more maps than did the heuristic score, which is significantly larger than the result for synthetic maps. Selected comparisons are shown in the Table 3.

P-values were computed based on simulation. For an optical map of n fragments, 10,000 random maps (each composed of n fragments) are generated by sampling from the fragment size distribution for optical maps in the current mapping project. Then, each random map is fitted into the reference, and its optimal score is stored. For the optimal alignment score of an optical map, the p-value is based on the number of optimal scores in simulation (out of 10,000) exceeding the given alignment score. It took 10,000 random maps in order to achieve the p-value accuracy of 10^{-4} in Table 3. Thus, we need to show results only for optical maps with one p-value larger than 10^{-4} and at least one significant score ($p - value < 0.002$). Nine maps satisfying these conditions were selected to illustrate the significance of the results.

7. CONCLUSION

In this paper, we formulated a novel statistical model for the measurements of fragment sizes that arise in optical maps. We also proposed a novel model to address false cuts in optical maps. These two models, along with previously known models, such as distribution of restriction sites and observation of cuts in optical maps, are combined to derive a likelihood ratio based alignment score that we report in this paper. We also show how this score can be used within an existing DP framework for finding optimal alignments. Our alignment score improves existing methods because it (1) does not require guessing of alignment parameters as is the case for heuristic scores (all distribution parameters are derived once from the mapping system) and (2) provides maximum discrimination power for scores of true alignments compared to spurious ones.

APPENDIX

A. Probabilistic framework

Alignments are calculated in order to find matching regions between maps. Calculation of optimal alignments depends profoundly on the alignment score. The utility of the alignment score is in finding the optimal configuration across all possible alignments.

Alignments of restriction maps are represented by pairs of matching sites between compared maps. Sites on maps represent loci of maps flanked by two adjacent fragments. According to our model, sites on optical maps can result either from a restriction enzyme cutting DNA at that position or from random DNA breakage.

Consider two maps D_1 and D_2 that are produced using a single restriction enzyme. In our notation, a pair of matching sites is denoted by $\langle i, j\rangle$; i.e., the site i located on map D_1 and site j located on map D_2 correspond to the same genomic locus. An alignment between D_1 and D_2 is represented by a set of pairs of matching sites $\langle i, j\rangle$ of maps D_1 and D_2. Matching regions are defined as regions of maps flanked by adjacent pairs of matching sites. Therefore, matching regions contain no matching sites other than the flanking ones. Naturally, alignment is represented by a set of ordered pairs of matching sites $\langle i_0, j_0\rangle$, $\langle i_1, j_1\rangle, \ldots, \langle i_d, j_d\rangle$, where order is given in the sense that $i_0 < i_1 < \cdots < i_d$ and $j_0 < j_1 < \cdots < j_d$. Consequently, regions $I_t = [i_{t-1}, i_t)$ and $J_t = [j_{t-1}, j_t)$ are matching regions since they are flanked by matching sites $\langle i_{t-1}, j_{t-1}\rangle$ and $\langle i_t, j_t\rangle$. Matching of regions is denoted by $\langle I_t, J_t\rangle$.

For simplicity, we assume that errors in different regions are independent of each other and hence their likelihoods can be written as product of likelihoods taken over matching regions. In other words, for $\langle I_m, J_m\rangle$ and $\langle I_n, J_n\rangle$ we have

$$Pr(\langle I_m, J_m\rangle, \langle I_n, J_n\rangle) = Pr(\langle I_m, J_m\rangle)Pr(\langle I_n, J_n\rangle),$$

if $m \neq n$. Here, $Pr(\langle I_m, J_m\rangle)$ is understood as the likelihood of the observed region size and fragment number under a specified hypothesis (H_0 or H_a). To define the alignment score, consider an alignment defined by matching sites $\langle i_0, j_0\rangle, \langle i_1, j_1\rangle, \ldots, \langle i_d, j_d\rangle$ of maps D_1 and D_2.

Here, we define two competing hypotheses: H_0 and H_a. Hypothesis H_0 corresponds to no dependence between compared regions of maps; i.e., the likelihoods can be calculated by taking a product of likelihoods of each region. As we specified before, under H_a, compared map regions represent the same genomic regions. The comparison score S is taken as a log of the likelihood ratio under H_0 and H_a. Hence, for two compared maps D_1 and D_2, we can define a score as follows:

$$S(D_1, D_2) = -log(LR(\langle D_1, D_2\rangle)),$$

where LR is the likelihood ratio calculated under competing hypotheses H_0 and H_a. Hence,

$$LR(\langle D_1, D_2\rangle) = \frac{f(D_1, D_2)_{H_0}}{f(D_1, D_2)_{H_a}} = \prod_{k=1}^{d} \frac{f(I_k, J_k)_{H_0}}{f(I_k, J_k)_{H_a}} = \prod_{k=1}^{d} LR(\langle I_k, J_k\rangle)$$

with the product taken over d matching regions. It therefore follows that for the total score

$$S(D_1, D_2) = \sum_{k=1}^{d} S(I_k, J_k) = \sum_{k=1}^{d} \left[-log\left(\frac{f(I_k, J_k)_{H_0}}{f(I_k, J_k)_{H_a}}\right)\right].$$

As we will see later, score $S(I_k, J_k)$ for the matching region $\langle I_k, J_k\rangle$ is calculated based on two pieces of information: total amount of DNA within I_k and J_k and the number of fragments contained in I_k and J_k, respectively.

Reference matching. To make things easier to follow, define $O := D_1$ and $R := D_2$ to represent optical (O) and reference (R) maps, respectively. Also, we define $O_k := I_k$ and $R_k := J_k$ to denote matching regions of optical and reference maps. Further, let $o_k := i_k$ and $r_k := j_k$ denote the matching sites of the optical map and reference map, respectively.

Denote s_k^O and s_k^R to be the total amount of DNA contained within matching regions O_k and R_k. Also, define f_k^O and f_k^R to be the number of fragments within O_k and R_k, respectively.

We can write the likelihood of the observed data as

$$\begin{aligned} f(O_k, R_k) &= f(O_k|R_k) \times f(R_k) = f(s_k^O, f_k^O|s_k^R, f_k^R) \times f(R_k) \\ &= f(s_k^O|s_k^R, f_k^R, f_k^O) \times f(f_k^O|s_k^R, f_k^R) \times f(R_k). \end{aligned}$$

Now the likelihood ratio can be written as

$$LR(\langle O_k, R_k\rangle) = \frac{f(O_k, R_k)_{H_0}}{f(O_k, R_k)_{H_a}} = \frac{[f(O_k|R_k) \times f(R_k)]_{H_0}}{[f(O_k|R_k) \times f(R_k)]_{H_a}}$$

$$= \frac{f(O_k|R_k)_{H_0} \times f(R_k)}{f(O_k|R_k)_{H_a} \times f(R_k)} = \frac{f(O_k|R_k)_{H_0}}{f(O_k|R_k)_{H_a}}.$$

Therefore, we can write

$$LR\langle O_k, R_k\rangle = \frac{f(s_k^O|s_k^R, f_k^R, f_k^O)_{H_0}}{f(s_k^O|s_k^R, f_k^R, f_k^O)_{H_a}} \times \frac{f(f_k^O|s_k^R, f_k^R)_{H_0}}{f(f_k^O|s_k^R, f_k^R)_{H_a}}$$

$$= LR(s_k^O|s_k^R, f_k^R, f_k^O) \times LR(f_k^O|s_k^R, f_k^R),$$

Thus, for the total score we have

$$S(O, R) = \sum_{k=1}^{d}[\alpha_k + \beta_k] = \sum_{k=1}^{d}\left[S(s_k^O|s_k^R, f_k^R, f_k^O) + S(f_k^O|s_k^R, f_k^R)\right],$$

with the sum taken over d matching regions.

Here,

$$\alpha_k = -log(LR(s_k^O|s_k^R, f_k^R, f_k^O))$$

is referred to as the size match score. It accounts for size differences between s_k^O and s_k^R. Similarly,

$$\beta_k = -log(LR(f_k^O|s_k^R, f_k^R))$$

is referred to as the site mismatch score. It accounts for missing and false cuts within a given matching region, which appear as unaligned sites.

Optical matching. Optical matching corresponds to the situation when two optical maps D_1 and D_2 are aligned to each other. Recall that I_k and J_k refer to matching regions between D_1 and D_2. As before, i_k, j_k define matching sites between aligned maps. Let s_k^I and s_k^J be the total amount of DNA contained within matching regions I_k and J_k, respectively. Also, define f_k^I and f_k^J to be the number of fragments within I_k and J_k. Both maps D_1 and D_2 define collections of random variables for which we can write the likelihood as

$$f(I_k, J_k) = f(s_k^I, s_k^J, m_k^I, m_k^J) = f(s_k^I, s_k^J|m_k^I, m_k^J) \times f(m_k^I, m_k^J),$$

and thus the likelihood ratio can be rewritten as

$$LR(\langle I_k, J_k\rangle) = \frac{f(s_k^I, s_k^J|m_k^I, m_k^J)_{H_0}}{f(s_k^I, s_k^J|m_k^I, m^J)_{H_a}} \times \frac{f(m_k^I, m_k^J)_{H_0}}{f(m_k^I, m_k^J)_{H_a}} = LR(s_k^I, s_k^J|m_k^I, m_k^J) \times LR(m_k^I, m_k^J).$$

Hence, we infer that for the score

$$S(I_k, J_k) = S(s_k^I, s_k^J|m_k^I, m_k^J) + S(m_k^I, m_k^J) = \alpha_k + \beta_k,$$

where

$$\alpha_k = -log(LR(s_k^I, s_k^J|m_k^I, m_k^J)).$$

The last expression accounts for the sizing differences between s_k^I and s_k^J. Also,

$$\beta_k = -log(LR(m_k^I, m_k^J))$$

accounts for the unaligned sites between optical maps within regions $\langle I_k, J_k\rangle$.

Therefore, the total alignment score can be rewritten as

$$S(D_1, D_2) = \sum_{k=1}^{d} (\alpha_k + \beta_k)\,.$$

B. Calculation of the scores

Proof (Lemma 1). Suppose W is the number of true cuts within the DNA of unit size that appears after digestion. Suppose also that V is the number of cuts that appear due to random breakage of DNA. Then, the total number of cuts per unit amount of DNA is $X = W + V$.

Number W results from a p-thinning of the Poisson process for site occurrences, where p is the digestion rate. Hence, W is Poisson with the rate p/λ.

According to model assumptions, the number of random breaks per unit of DNA has a Poisson distribution with the rate parameter ζ. Thus, it follows that X is the sum of two independent Poisson random variables and hence is Poisson with $EX = EW + EV$. It now follows that the total number of sites X is a Poisson with the mean $1/\tau = \zeta + \omega p = (\zeta + \frac{p}{\lambda})$. For s amount of DNA, the number of sites therefore follows a Poisson process with the parameter s/τ. More details on thinned processes may be found in Grimmett and Stirzaker (1982). ■

Proof (Lemma 2). Consider a fragment on an optical map of size $X > \Delta$. Size X is produced with a reference genomic region of size Y and a sizing error, i.e., $X = Y + \epsilon$. Lemma 1 asserts that Y has exponential density with the mean $\tau = (\zeta + p/\lambda)^{-1}$. Hence,

$$f_Y(y) = \frac{1}{\tau}e^{-\frac{y}{\tau}}.$$

Furthermore, by Proposition 3,

$$f_{X|Y}(x|y) = \frac{c_1(y)}{\sqrt{2\pi y}\sigma}e^{-\frac{(x-y)^2}{2\sigma^2 y}},$$

where $c_1(y)$ takes care of appropriate normalization

$$c_1(y) = \left[\int_0^\infty \frac{1}{\sqrt{2\pi y}\sigma}e^{-\frac{(x-y)^2}{2\sigma^2 y}}\,dx\right]^{-1}.$$

For $X \geq \Delta$, it is easy to show that $1 \geq c_1(Y) \geq 0.95$ with high probability and $c_1(y) \rightarrow 1$ as y increases. Thus, for all practical purposes, $c_1(y) \approx 1$ in our computations.

We conclude therefore that the joint density has the form

$$f_{X,Y}(x, y) = \frac{1}{\sqrt{2\pi y}\sigma\tau}exp\left[-\frac{(x-y)^2}{2\sigma^2 y} - \frac{y}{\tau}\right],$$

The desired marginal density is obtained by integrating:

$$f_X(x) = \int_0^\infty \frac{1}{\sqrt{2\pi y}\sigma\tau}exp\left[-\frac{(x-y)^2}{2\sigma^2 y} - \frac{y}{\tau}\right]dy.$$

The last integral is a special form of the Mellin integral transform for exponential functions. In fact, it is closely related to modified Bessel functions of the second kind for semi-integer order for which an exact analytic form can be written. Its solution can be found in Bateman and Erdelyi (1953). Here, we just use as a result that

$$\int_0^\infty \frac{1}{\sqrt{y}} exp\left[-\frac{(x-y)^2}{2\sigma^2 y} - \frac{y}{\tau}\right] dy = \frac{\sqrt{2\pi}}{\sqrt{\frac{1}{\sigma^2}+\frac{2}{\tau}}} exp\left[-x\left(\frac{1}{\sigma}\sqrt{\frac{1}{\sigma^2}+\frac{2}{\tau}} - \frac{1}{\sigma^2}\right)\right].$$

This immediately implies that

$$f_X(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

is the marginal density of X with

$$\theta = \left[\frac{1}{\sigma}\sqrt{\frac{2}{\tau}+\frac{1}{\sigma^2}} - \frac{1}{\sigma^2}\right]^{-1}$$

since

$$\sigma\tau\sqrt{\frac{1}{\sigma^2}+\frac{2}{\tau}} \approx \theta.$$

if $\lambda \gg \sigma$. ■

Proof (Proposition 1). In optical maps, not all sites are matched due to two factors: false cuts and missing cuts. False cuts do not affect sizes of matching regions, since false cuts should not be matched. Missing cuts, however, produce a thinned Poisson process from the original Poisson process of occurrence of restriction sites along the genome. The thinned Poisson process has, therefore, a rate $\frac{p}{\lambda}$, and thus the sizes of matching regions have an exponential distribution with mean $\upsilon = \lambda/p$. ■

Proof (Proposition 2). In a manner similar to the argument in the previous lemma, matching sites between two optical maps occur only when both restriction sites on both optical maps are being digested by endonuclease. This happens with probability p^2. Thinning of restriction sites imposed by this digestion of sites on both maps has, therefore, a Poisson distribution with rate p^2/λ. Hence, the sizes have exponential density with mean $\phi = \lambda/p^2$. ■

Proof (Proposition 3). By the normality assumption for error distribution, $\epsilon_i = X_i - Y_i \sim N(0, \sigma^2 Y_i)$. Thus, $X_i \sim N(Y_i, \sigma^2 Y_i)$ for $i = 1, 2, \ldots, m$. Since $X = \sum_{i=0}^m X_i$ is a linear combination of m independent normal variables, it also has a normal distribution. Finally, $EX = \sum_{i=0}^m EX_i = \sum_{i=0}^m Y_i = Y$ and $Var(X) = \sum_{i=0}^m Var(X_i) = \sigma^2 \sum_{i=0}^m Y_i = \sigma^2 Y$. ■

Lemma 3 (Bessel solution). *If $Re(\alpha^2 z) > 0$ and $Re(z) > 0$, then*

$$K_\nu(\alpha z) = \frac{\alpha^\nu}{2}\int_0^\infty t^{-\nu-1} exp\left[-\frac{z}{2}\left(t+\frac{\alpha^2}{t}\right)\right] dt,$$

where $K_\nu(z)$ is a modified Bessel function of the second kind.

Proof (Lemma 3). Consider the modified Bessel differential equation

$$z^2 y''(z) + zy'(z) - (z^2 + \nu^2)y(z) = 0.$$

The solution to this equation is given by

$$y(z) = a_1 I_\nu(z) + b_1 K_\nu(z),$$

where $I_\nu(z)$ is defined as the modified Bessel function of the first kind and $K_\nu(z)$ is defined as the modified Bessel function of the second kind.

Finally, the integral is described by Baterman and Erdelyi (1953). ■

Lemma 4 (Bessel solution). *If $a, b \in R$, then*

$$\int_0^\infty t^{-\nu-1} exp\left[-\left(a^2 t + \frac{b^2}{t}\right)\right] dt = 2\frac{a^\nu}{b^\nu} K_\nu(2ab).$$

Proof (Lemma 4). Follows from Lemma 3 after substituting

$$z = 2a^2, \qquad \alpha^2 = \frac{b^2}{a^2}.$$ ■

Proof (Theorem 1). Recall that in the calculation of the likelihood ratio for two competing hypotheses, H_0 corresponds to no match between x and y, and H_a corresponds to the match between x and y. The likelihood ratio is of the form

$$LR(x|y, n, m) = \frac{f_X(x|y, n, m)_{H_0}}{f_X(x|y, n, m)_{H_a}}.$$

Consider $X > \Delta$. Under H_0, X is the sum of independent exponential r.v. with mean θ given by Lemma 2. Thus, its likelihood is given by the marginal density of X:

$$f_X(x|y, n, m)_{H_0} = f_X(x|m) = \frac{1}{\Gamma(m)\theta^m} x^{m-1} e^{-\frac{x}{\theta}}.$$

Under H_a, X is the size measurement of the genomic region of size Y and thus under the error model

$$f_X(x|y, n, m)_{H_a} = \frac{c_1(y)}{\sqrt{2\pi y}\sigma} e^{-\frac{(x-y)^2}{2\sigma^2 y}}$$

where $c_1(y)$ gives appropriate normalization similar to the one discussed in Lemma 2:

$$c_1(y) = \left[\int_0^\infty \frac{1}{\sqrt{2\pi y}\sigma} e^{-\frac{(x-y)^2}{2\sigma^2 y}} dx\right]^{-1}.$$

As before, we use that $c_1(y) \approx 1$ for $X > \Delta$ to omit it from further derivations.

After substituting two likelihoods in the likelihood ratio, we obtain

$$LR(x|y, n, m) = \frac{\sqrt{2\pi y}\sigma x^{m-1}}{\Gamma(m)\theta^m} exp\left[\frac{(x-y)^2}{2\sigma^2 y} - \frac{x}{\theta}\right],$$

which gives the desired result for long fragments ($X > \Delta$).

The error model for short fragments $0 \leq X \leq \Delta$ implies that under the match

$$f_X(x|y, n, m)_{H_a} = \frac{c(y)}{\sqrt{2\pi}\eta} exp\left[-\frac{(x-y)^2}{2\eta^2}\right]$$

with appropriate normalization given by

$$c(y) = \left[\frac{1}{\sqrt{2\pi}\eta} e^{-\frac{(x-y)^2}{2\eta^2}}\right]^{-1}.$$

The marginal distribution of X for $0 \leq X \leq \Delta$ is not exactly exponential as is the case for $X > \Delta$. Due to significant sizing errors, however, for $0 \leq X \leq \Delta$ the likelihood ratio should not assume very small values to reflect the lack in test power. To accomplish this, we can choose any likelihood L_0 under H_0 such that it dominates L_a for small values of X. This is accomplished by appropriate choice of parameter η. Hence, we still can assume exponential marginal density for X as long as $\eta^2 \gg \sigma^2$. Also, $c(y)$ can be omitted.

Thus, under H_0, we use exponential density for likelihood of individual fragments, so that $X = \sum_{i=1}^{m}$. Therefore,

$$f_X(x|y,n,m)_{H_0} = f_X(x|m) = \frac{x^{m-1}}{\theta^m \Gamma(m)} e^{-\frac{x}{\theta}}.$$

Hence, we obtain the likelihood ratio in the form

$$LR(x|y,n,m) = \frac{\sqrt{2\pi}\eta x^{m-1}}{\Gamma(m)\theta^m} exp\left[\frac{(x-y)^2}{2\eta^2} - \frac{x}{\theta}\right]. \qquad \blacksquare$$

Proof (Theorem 2). As before, we have two competing hypotheses: H_0 corresponds to no match between two regions and H_a corresponds to the match. The likelihood ratio has the form

$$LR(m|y,n) = \frac{Pr(m|y,n)_{H_0}}{Pr(m|y,n)_{H_a}}.$$

Under H_0, there is no matching (m and n are independent), and hence

$$Pr(m|y,n)_{H_0} = Pr(m) = f_M(m).$$

(Exact calculation of f_m is discussed in the end of the proof.)

Under H_a, we have $m-1$ and $n-1$ sites observed on the optical and reference maps, respectively, within this matching region.

By definition, matching regions contain no internal matching sites. This implies that under H_a all $n-1$ sites are not digested and all $m-1$ sites on the optical map are random DNA breaks. Since p is the digestion rate and ζ is the frequency of random breakage per unit of DNA, we can rewrite this probability as

$$Pr(m|y,n)_{H_a} = (1-p)^{n-1} e^{-\zeta y} \frac{(\zeta y)^{m-1}}{(m-1)!}.$$

Therefore, the likelihood ratio has the form

$$LR(m|y,n) = \frac{e^{\zeta y}((m-1)! f_M(m)}{(1-p)^{n-1}(\zeta y)^{m-1}},$$

and this completes the proof.

Function $f_M(m)$ gives the marginal distribution of M for the choice of randomly selected regions of optical maps. In practice, regions longer than δ are never calculated. Thus, we assume that these regions occur with equal probability; hence, $f_M(m) = \frac{1}{\delta}$ for $1 \le m \le \delta$ or 0 otherwise. $\blacksquare$

Proof (Theorem 3). The likelihood ratio is of the following form:

$$LR(x_1,x_2|m_1,m_2) = \frac{f_{X_1,X_2}(x_1,x_2|m_1,m_2)_{H_0}}{f_{X_1,X_2}(x_1,x_2|m_1,m_2)_{H_a}}.$$

In this theorem, we assume only the first error model $\epsilon \sim N(0,\sigma^2 Y)$ for the sake of simplicity. The null hypothesis H_0 corresponds to no match between X_1 and X_2, meaning that they come from different genomic regions, and therefore X_1 and X_2 are independent. Hence, it follows that

$$f_{X_1,X_2}(x_1,x_2|m_1,m_2)_{H_0} = f_{X_1}(x_1|m_1) f_{X_2}(x_2|m_2),$$

where densities $f_{X_1}(x_1|m_1)$ and $f_{X_2}(x_2|m_2)$ are given by Lemma 2. Hence,

$$f_{X_1,X_2}(x_1,x_2|m_1,m_2)_{H_0} = \frac{x_1^{m_1-1}x_2^{m_2-1}}{\theta^{m_1+m_2}\Gamma(m_1)\Gamma(m_2)}e^{-\frac{x_1+x_2}{\theta}}.$$

Under the alternative hypothesis H_a, there is a match, and hence

$$f_{X_1,X_2}(x_1,x_2|m_1,m_2)_{H_a} = \int_0^\infty f_{X_1,X_2,Y}(x_1,x_2,y)dy,$$

where $Y = y$ is the true underlying size of the matching region.

Under H_a, X_1 and X_2 are the size measurements of Y and thus are conditionally independent:

$$f_{X_1,X_2,Y}(x_1,x_2,y) = f_{X_1|Y}(x_1|y)f_{X_2|Y}(x_2|y)f_Y(y).$$

Furthermore, according to Proposition 2, matching regions between two optical maps occur according to a Poisson process with rate p^2/λ; hence, their sizes occur according to exponential distribution with mean $\phi = \lambda/p^2$.

Hence, after combining, we deduce that

$$f_{X_1,X_2,Y}(x_1,x_2,y) = \frac{c_2(y)}{2\pi\sigma^2 y\phi} \times exp\left[-\frac{(x_1-y)^2}{2\sigma^2 y} - \frac{(x_2-y)^2}{2\sigma^2 y} - \frac{y}{\phi}\right],$$

where $c_2(y)$ guarantees appropriate normalization for the joint density to allow only positive values of X_1 and X_2:

$$\frac{1}{c_2(y)} = \int_0^\infty\int_0^\infty \frac{exp\left[-\frac{(x_1-y)^2}{2\sigma^2 y} - \frac{(x_2-y)^2}{2\sigma^2 y} - \frac{y}{\phi}\right]}{2\pi\sigma^2 y\phi}dx_1dx_2.$$

In the latter expression, $c_2(y)$ is close to 1 with high probability and hence can be replaced with 1 to make calculations easier.

To deduce the joint density of X_1 and X_2, first note that by Lemma 3

$$\int_0^\infty \frac{1}{y}exp\left[-\frac{(x_1-y)^2}{2\sigma^2 y} - \frac{(x_2-y)^2}{2\sigma^2 y} - \frac{y}{\phi}\right]dy = e^{\frac{x_1+x_2}{\sigma^2}}\int_0^\infty \frac{exp\left[-\left(\frac{1}{\phi}+\frac{1}{\sigma^2}\right)y - \left(\frac{x_1^2+x_2^2}{2\sigma^2}\right)\frac{1}{y}\right]}{y}dy$$

$$= 2 \times exp\left[\frac{x_1+x_2}{\sigma^2}\right]K_0\left[2\sqrt{\frac{1}{\phi}+\frac{1}{\sigma^2}}\sqrt{\frac{x_1^2+x_2^2}{2\sigma^2}}\right].$$

Hence, the joint density of X_1 and X_2 has the form

$$f_{X_1,X_2}(x_1,x_2|n)_{H_a} = \frac{exp\left[\frac{x_1+x_2}{\sigma^2}\right]}{\pi\sigma^2\phi} \times K_0\left[2\sqrt{\frac{1}{\phi}+\frac{1}{\sigma^2}}\sqrt{\frac{x_1^2+x_2^2}{2\sigma^2}}\right].$$

We can finally calculate the desired likelihood ratio:

$$LR(x_1,x_2|m_1,m_2) = \frac{\pi\phi\sigma^2 x_1^{m_1-1}x_2^{m_2-1}}{\Gamma(m_1)\Gamma(m_2)\theta^{m_1+m_2}} \times \frac{exp\left[-\left(\frac{1}{\theta}+\frac{1}{\sigma^2}\right)(x_1+x_2)\right]}{K_0\left[2\sqrt{\frac{1}{\phi}+\frac{1}{\sigma^2}}\sqrt{\frac{x_1^2+x_2^2}{2\sigma^2}}\right]}. \quad \blacksquare$$

Proof (Theorem 4). Our goal is to calculate the likelihood ratio for observing m_1 and m_2 within the matching regions of two optical maps. As before, the likelihood ratio has the form

$$LR(m_1, m_2|match) = \frac{Pr(m_1, m_2|match)_{H_0}}{Pr(m_1, m_2|match)_{H_a}}.$$

By definition of matching regions, they contain no internal matching sites. Hence,

$$\begin{aligned} Pr(m_1, m_2|match, s)_{H_a} &= \sum_{n=0}^{\infty} Pr(m_1, m_2, n|s, no\ other\ matches) \\ &= \sum_{n=0}^{\infty} Pr(m_1, m_2|s, n, no\ other\ matches) Pr(n|s), \end{aligned}$$

where n is the unobserved number of restriction sites per amount s of DNA corresponding to the reference region.

Finally, the joint density is calculated as

$$Pr(m_1, m_2|match) = \int_0^{\infty} Pr(m1, m_2|match, s) f_Y(s) ds.$$

Now, if k_1 and k_2 denote the numbers of correctly cut restriction sites in the matching region for each of two optical maps, then

$$\begin{aligned} Pr(m_1, m_2, k_1, k_2|s, n, no\ other\ matches) &= Pr(m_1, m_2, k_1, k_2|s, n, no\ sites\ in\ common) \\ &= Pr(k1,\ k2 \mid n, no\ sites\ in\ common) \\ &\quad \times e^{-\omega} \frac{\omega^{m_1-k_1-1}}{(m_1-k_1-1)!} \times e^{-\omega} \frac{\omega^{m_2-k_2-1}}{(m_2-k_2-1)!}, \end{aligned}$$

since it follows that $m_1 - k_1 - 1$ and $m_2 - k_2 - 1$ sites are due to random breakage. As before, $\omega = \zeta s$ corresponds to the Poisson parameter of the random DNA breakage process. Furthermore,

$$A(k_1, k_2|n) = Pr(k1,\ k2 \mid n, no\ sites\ in\ common) = \frac{Pr(k_1, k2, no\ sites\ in\ common|\ n)}{Pr(no\ sites\ in\ common|\ n)}$$

$$= \frac{\binom{n}{k_1\ k_2} p^{k_1+k_2}(1-p)^{2n-(k_1+k_2)}}{\sum_{i_1=0}^{n} \sum_{i_2=0}^{n-i_1} \binom{n}{i_1\ i_2} p^{i_1+i_2}(1-p)^{2n-(i_1+i_2)}}.$$

Hence, we can now calculate

$$\begin{aligned} Pr(m_1, m_2|s, n, no\ other\ matches) &= \sum_{k_1=0}^{(m1-1)\wedge n} \sum_{k_2=0}^{(m_2-1)\wedge(n-k_1)} A(k_1, k_2|n) \\ &\quad \times e^{-2\omega} \frac{\omega^{m_1+m_2-k_1-k_2-2}}{(m_1-k_1-1)!(m_2-k_2-1)!}. \end{aligned}$$

By Proposition 2, reference sizes underlying matching regions of two optical maps have exponential distribution with mean $\phi = \lambda/p^2$. Thus,

$$f_Y(s) = \frac{1}{\phi} e^{-\frac{s}{\phi}}.$$

Therefore,

$$
\begin{aligned}
B = Pr(m_1, m_2|match)_{H_a} &= \int_0^\infty \sum_{n=0}^{\infty} Pr(m_1, m_2, n|s, no\ other\ matches) f_Y(s) ds \\
&= \int_0^\infty \frac{1}{\phi} e^{-\frac{s}{\phi}} \sum_{n=0}^{\infty} e^{-\frac{s}{\lambda}} \frac{\left(\frac{s}{\lambda}\right)^n}{n!} \sum_{k_1=0}^{(m1-1)\wedge n} \sum_{k_2=0}^{(m_2-1)\wedge(n-k_1)} \left(A(k_1, k_2|n) \times e^{-2\omega} \frac{\omega^{m_1+m_2-k_1-k_2-2}}{(m_1-k_1-1)!(m_2-k_2-1)!} \right) \\
&= \frac{1}{\phi} \sum_{n=0}^{\infty} \frac{1}{\lambda^n n!} \sum_{k_1=0}^{(m_1-1)\wedge n} \sum_{k_2=0}^{(m_2-1)\wedge(n-k_1)} A(k_1, k_2|n) \frac{\zeta^{m_1+m_2-k_1-k_2-2}}{\Gamma(m_2-k_2)\Gamma(m1-k_1)} \\
&\quad \times \int_0^\infty e^{-s(\frac{1}{\phi}+\frac{1}{\lambda}+2\zeta)} s^{n+m_1+m_2-(k_1+k_2)-2} ds \\
&= \frac{1}{\phi} \sum_{n=0}^{\infty} \frac{1}{\lambda^n n!} \sum_{k_1=0}^{(m_1-1)\wedge n} \sum_{k_2=0}^{(m_2-1)\wedge(n-k_1)} A(k_1, k_2|n) \frac{\zeta^{(m_1+m_2)-(k_1+k_2)-2}}{\left[\frac{1}{\phi}+\frac{1}{\lambda}+2\zeta\right]^{n+(m_1+m_2)-(k_1+k_2)-1}} \\
&\quad \times \frac{\Gamma(n+(m_1+m_2)-(k_1+k_2)-1)}{\Gamma(m_1-k_1)\Gamma(m_2-k_2)}.
\end{aligned}
$$

Note that the last expression solely depends on m_1, m_2, and distribution parameters (λ, p, and ζ); hence, a table for these probabilities can be precomputed and used for the scoring.

Finally, under H_0, the regions are independent of each other, so that

$$C = Pr(m_1, m_2)_{H_0} = f_M(m_1) f_M(m_2) = Pr(m_1) \times Pr(m_2). \qquad \blacksquare$$

And thus finally the likelihood ratio has the desired form. Note that the marginal distribution $Pr(m)$ is defined by the choice of the null hypothesis H_0. Here we assume that the random regions of sizes from 1 to δ occur with equal probability and hence $Pr(m) = \frac{1}{\delta}$.

ACKNOWLEDGMENTS

We want to thank the anonymous reviewers who helped us to improve our paper and enhance the presentation of the material. This work is supported by NIH CEGS: Implications of Haplotype Structure in the Human Genome, grant No. P50 HG002790, and NIH, NHGRI, R01 HG00225-10.

REFERENCES

Ambrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kröger, N., Lau, W.W.Y., Lane, T.W., Larimer, F.W., Lippmeier, J.C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M.S., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P., and Rokhsar, D.S. 2004. The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science* 306(5693), 79–86.

Ananthraman, T., and Mishra, B. 2001. A probabilistic analysis of false positives in optical map alignment and validation. *Algorithms in Bioinformatics, Proc. 1st Int. Workshop (WABI 2001).*

Ananthraman, T., Mishra, B., and Schwartz, D. 1997. Genomics via optical mapping II: Ordered restriction maps. *J. Comp. Biol.* 4(2), 91–118.

Ananthraman, T., Schwartz, D., and Mishra, B. 1999. Genomics via optical mapping III: Contiging genomic DNA and variations. *Proc. 7th Int. Conf. on Intelligent Systems for Molecular Biology.*

Bateman, H., and Erdelyi, A. 1953. *Higher Transcendental Functions*, McGraw–Hill, New York.

Churchill, G.A., Daniels, D.L., and Waterman, M.S. 1989. The distribution of restriction enzyme sites *Escherichia coli*. *Nucl. Acids Res.* 18(3), 589–597.

Dimalanta, E.T., Lim, A., Runnheim, R., Lamers, C., Churas, C., Forrest, D.K., dePablo, J.J., Graham, M.D., Coppersmith, S.N., and Schwartz, D.C. 2004. A microfluidic system for large DNA molecule arrays. *Anal. Chem.* 76(18), 5293–5301.

Grimmett, G., and Stirzaker, D. 1982. *Probability and Random Processes*. Oxford University Press, Oxford, UK.

Huang, X., and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9, 868–877.

Huang, X., and Miller, W. 1991 A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.* 12, 337–357.

Huang, X., and Waterman, M.S. 1992. Dynamic programming algorithms for restriction map comparison. *Comput. Appl. Biosci.* 8(5), 511–520.

Myers, E.W. 1999. Whole genome DNA sequencing. *IEEE Computational Engineering and Science* 3(1), 33–43.

Myers, E.W., and Huang, X. 1992. An $O(N^2 logN)$ restriction map comparison and search algorithm. *Bull. Math. Biol.* 54(4) 599–618.

Smith, T.F., and Waterman, M.S. 1981. Comparison of biosequences. *Adv. Appl. Math.* 2, 482–489.

Waterman, M. 1995. *Introduction to Computational Biology*, Chapman and Hall, London.

Waterman, M.S., Smith, T.F., and Katcher, H. 1984. Algorithms for restriction map comparisons. *Nucl. Acids Res.* 12, 237–242.

Zhou, S., Kile, A., Bechner, M., Place, M., Kvikstad, E., Deng, W., Wei, J., Severin, J., Runnheim, R., Churas, C., Forrest, D., Dimalanta, E., Lamers, C., Burland, V., Blattner, F., and Schwartz, D. 2004. A single molecule approach to bacterial genomic comparisons via optical mapping. *J. Bacteriol.* 186(22), 7773–7782.

Address correspondence to:
Anton Valouev
Department of Mathematics
University of Southern California
MCB at 1050 Childs Way, 403M
Los Angeles, CA 90089-1113

E-mail: valouev@usc.edu