J. Appl. Prob. **38,** 335–356 (2001)
Printed in Israel
© *Applied Probability Trust* 2001

LOCAL MATCHING OF RANDOM RESTRICTION MAPS

MENGXIANG TANG [*] [**] AND
MICHAEL S. WATERMAN,[*] *University of Southern California*

Abstract

Optical mapping is a new technique to generate restriction maps of DNA easily and quickly. DNA restriction maps can be aligned by comparing corresponding restriction fragment lengths. To relate, organize, and analyse these maps it is necessary to rapidly compare maps. The issue of the statistical significance of approximately matching maps then becomes central, as in BLAST with sequence scoring. In this paper, we study the approximation to the distribution of counts of matched regions of specified length when comparing two DNA restriction maps. Distributional results are given to enable us to compute p-values and hence to determine whether or not the two restriction maps are related. The key tool used is the Chen–Stein method of Poisson approximation. Certain open problems are described.

Keywords: Restriction maps; matching; Chen–Stein method

AMS 2000 Subject Classification: Primary 60C05
Secondary 60F20

1. Introduction

Before a genome is sequenced, restriction maps of the genome are often obtained. Restriction maps are examples of physical maps. The restriction map of the DNA sequence of an organism is a very useful tool for DNA sequencing and other genome studies. Restriction maps are aligned for many reasons including: constructing longer restriction maps by overlapping short maps, finding evolutionary relationships and, given genome restriction maps, locating smaller pieces of DNA in that genome. The basis of restriction map alignment is that, if two DNA sequences are identical, their restriction maps using the same set of enzymes are expected to be almost identical, which means they should have the same number of restriction sites and each ordered pair of restriction fragments should have almost the same length ('almost' is due to the experiment error in measuring the lengths of restriction fragments, detailed explanation for different kinds of error sources are described in Tang (2000a).

In 1987, Kohara *et al.* constructed an eight-enzyme restriction map of the entire genome of *Escherichia coli*. One key step in their strategy of construction is that they searched for overlapping pairs of clones by matching the clones' eight-enzyme restriction maps. Overlapping clones are detected when they match at more than 5 consecutive fragments. Lander and Waterman (1988) presented a mathematical analysis for physical mapping by fingerprinting random clones, and also evaluated different types of fingerprinting schemes. Their paper contains simplified calculations for the probability that random clones will be declared an overlap.

Received 30 May 2000; revision received 11 December 2000.

[*] Postal address: Mathematics Department, University of Southern California, Los Angeles, CA 90089-1113, USA.
[**] Email address: metang@hto.usc.edu

Waterman *et al.* (1984) first modelled the map alignment problem to study the evolution between homologous genes from closely related organisms. The goal is to align the maps, which is the so-called global alignment problem. An algorithm was provided to compute an optimal alignment of two maps. The algorithm allows for site indels as well as differences in distance between sites. Later, Waterman and Raymond (1987), in a study of geological strata matching, made a modification to the algorithm that handles one type of the restriction map data errors from experiments: closely spaced sites of the same enzyme are merged as a single site. Huang and Waterman (1992) extended the previous map comparison model to handle another type of data error: closely spaced sites for different enzymes are ordered incorrectly, and that gave a dynamic algorithm program. Rudd *et al.* (1990) used restriction map alignment software to align DNA sequence, genetic, and physical maps of *E. coli* to form an integrated genomic map, after which the integrated genomic map of *E. coli* was used to refine the genetic map and to locate newly sequenced or mapped genes. The restriction map alignment software considers restriction maps as strings analogous to DNA or protein sequences, except that two values, enzyme name and DNA base address, are associated with each position on the string.

A new technique to generate restriction maps of DNA, optical mapping, has been developed by David Schwartz *et al.* (1993). A single DNA molecule is stretched out and fixed to a surface. A restriction enzyme is added to cut the DNA sequence, and the cut sites are visualized as gaps and the lengths of fragments are estimated by an optical method (the DNA molecules are coated with a fluorochrome and imaged). Optical mapping can generate restriction maps easily and quickly. Schwartz *et al.* (1993) applied optical mapping to *Saccharomyces cerevisiae*, and Lin *et al.* (1999) obtained the whole-genome restriction map of *Deinococcus radiodurans*. Potentially, this method can rapidly produce restriction maps of megabases of DNA, possibly of genomes. To relate, organize, and analyse these maps it will be necessary to rapidly compare maps. Anantharaman *et al.* (1997) provided the first detailed model of the data produced by the optical mapping process. They formulated a statistical algorithm for the problem of producing optical maps and implemented the algorithm. Statistical issues of the significance of approximately matched maps then become central as they do in BLAST with sequence scoring.

The global matching of two random restriction maps has been studied in Tang (2000b). The probability that two random restriction maps match is very small. Looking at two restriction maps, there might be two segments of consecutive restriction fragments on the two restriction maps, respectively, that do match. A pair of segments of matched fragments is referred to as a matching region in this paper. The number of fragments in the matching region, which is called length of the matching region, varies from 1 to some large value. Obviously, the larger the length, the less chance such a matching region occurs at random. If there exists a matching region of a specified length, there might be a significant similarity between the two restriction maps. But how large should the length be to detect the significant similarity? Karlin *et al.* (1983) outlined such a study for nucleic acid and protein sequences. Later on, this statistical significance problem has been studied widely for DNA sequences (see Arratia *et al.* (1990), Karlin *et al.* (1990), Karlin and Altschul (1990), Waterman and Vingron (1994), and C. Neuhauser (1994)). Here we will study this problem for restriction maps. There is much similarity with DNA or protein sequence matching, but enough differences to challenge us. As with sequence matching, maps present their own set of difficulties.

We consider single-enzyme DNA restriction maps with the model as follows. The occurrences of cut sites along a DNA sequence is a Poisson process with rate λ, where λ is determined by the specific pattern of the cut site. Hence, the fragment lengths are i.i.d., each fragment length following an exponential distribution with density function $\lambda e^{-\lambda x}, \ x \geq 0$.

We consider two fragments equal if their lengths differ by no more than a constant σ (which is usually small compared to λ^{-1}) and the probability of this event is denoted as p, which is later computed from the exponential distribution. We will study two types of matching regions here. One is a matching region with all the consecutive pairs matched in the region. We are interested in the longest matching region between two random restriction maps. Since base mutations and deletions occur in the evolution of DNA sequences, to enable us to reveal the relationship between DNA sequences even after these changes, we will study another model of matching regions by allowing a few mismatches in a matching region. The main goal is to find an approximation to the distribution of counts of matching regions of specified length. From those distributional results, we can compute the tail probability for a matching region of specified length or greater, and hence test whether or not the two restriction maps are related.

The outline of the paper is as follows. In Section 2, we give an introduction to the key tool used in the paper: the Chen–Stein method. We approximate the distribution of the counts of matching regions of specified length in Section 3, where only matched fragment pairs are allowed in matching regions. Also, we test the results using simulations. In Section 4, we extend the approximation for matching regions to allow a few mismatches. Finally, we discuss an unsolved problem in Section 5, which we term the merge-matching problem, similar to the indel-problem of sequence alignment.

2. The Chen–Stein Method of Poisson approximation

The basic method employed in this paper is the Chen–Stein method, which is a method used to approximate the distribution of occurrences of dependent events by the Poisson distribution. A brief introduction is given in this section. The method is based on the work by Stein (1972) and was developed by Chen (1975). It was generalized to a multivariate context by Arratia *et al.* (1989), and below, we will state a version of the Chen–Stein method following their paper.

Let I be an arbitrary index set, and for $i \in I$, X_i is an indicator function to indicate whether or not some event occurs. The total number of occurrences of events is

$$W = \sum_{i \in I} X_i.$$

The set of events $\{X_i\}_{i \in I}$ could be dependent. The Chen–Stein method is a general approach to approximate the distribution of W by a Poisson distribution Z via bounding the total variation distance between W and Z. Let $h : \mathbb{Z}^+ \to \mathbb{R}$, where $\mathbb{Z}^+ = \{0, 1, 2, \dots\}$, and write $\|h\| = \sup_{k \geq 0} |h(k)|$. The total variation distance between W and Z is denoted by

$$\|W - Z\| = \sup_{\|h\|=1} |\mathrm{E}h(W) - \mathrm{E}h(Z)| = 2 \sup_{A \subset \mathbb{Z}^+} |\mathrm{P}(W \in A) - \mathrm{P}(Z \in A)|.$$

Before we state the theorem, we present more notation used in the approximation. Let J_i be an index set such that $j \notin J_i$ if X_j is independent of X_i. The approximation is related to the first and second moments of $\{X_i\}_{i \in I}$, b_1 and b_2, which are defined as

$$b_1 = \sum_{i \in I} \sum_{j \in J_i} \mathrm{E}(X_i)\mathrm{E}(X_j),$$

and

$$b_2 = \sum_{i \in I} \sum_{i \neq j \in J_i} \mathrm{E}(X_i X_j).$$

Theorem 2.1. *Let W be the number of occurrences of dependent events, and let Z be a Poisson random variable with* $\mathrm{E}(Z) = \mathrm{E}(W) = \lambda$. *Then*

$$\|W - Z\| \leq 2(b_1 + b_2)\frac{1 - \mathrm{e}^{-\lambda}}{\lambda} \leq 2(b_1 + b_2),$$

and in particular

$$|\mathrm{P}(W = 0) - \mathrm{e}^{-\lambda}| \leq (b_1 + b_2)\frac{1 - \mathrm{e}^{-\lambda}}{\lambda}.$$

This theorem is a process version of the Poisson approximation which is useful when we have to use the entire process of indicators $\{X_i\}_{i \in I}$. If b_1 and b_2 are small, then W will be approximately Poisson distributed with rate $\mathrm{E}(W)$. Thus, to establish the Poisson approximation, we should check that the quantities b_1 and b_2 are small.

3. Matching region with matched fragment pairs

As in the global matching problem, a restriction map is represented as a string of capital letters with indices, such as $A_1A_2 \ldots A_n$. Each A_i denotes the length of the ith fragment in the restriction map in a fixed orientation and n is the total number of fragments in the restriction map (on maps with several enzymes, the cut site can also carry the identity of the enzyme as well). A pair of fragments A_i and A_j are matched if their lengths differ by no more than a small constant σ; matching is denoted by $A_i =_\triangledown A_j$ in this paper. A matching region between two restriction maps consists of two series of contiguous restriction fragments from each of the two restriction maps that have the same number of fragments and in which each corresponding fragment pair matches (see Figure 1). The number of fragment pairs in the matching region is defined to be the length of the matching region. When two restriction maps are aligned locally, we are interested in the matching region with the maximum length observed between the two restriction maps. If a matching region of length greater than or equal to some test value t is observed, can we conclude that there is a high similarity or relation between the two sequences? To answer the question, we wish to know the p-value of such an observation.

As is often used for the studies of restriction maps, the occurrences of cut sites along a DNA sequence is assumed to be a Poisson process of rate λ, which is used to denote the cutting rate along DNA sequence. Therefore, the length of a restriction fragment has an exponential distribution with mean $1/\lambda$. The value of λ depends on the cut site pattern and the distribution of nucleotides in the DNA sequences.

3.1. Main results

Let $\boldsymbol{A} = A_1A_2 \ldots A_n$ and $\boldsymbol{B} = B_1B_2 \ldots B_m$ denote two restriction maps from the same enzyme of length n and m respectively. We wish to find the local similarities between $\boldsymbol{A}$ and $\boldsymbol{B}$; that is, we are interested in finding the matching regions of specified length between $\boldsymbol{A}$ and $\boldsymbol{B}$.

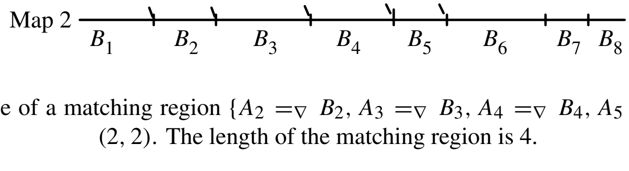

Figure 1: Example of a matching region $\{A_2 =_\triangledown B_2, A_3 =_\triangledown B_3, A_4 =_\triangledown B_4, A_5 =_\triangledown B_5\}$ starting at $(2, 2)$. The length of the matching region is 4.

A matching region between $\boldsymbol{A}$ and $\boldsymbol{B}$ can start at any index pair (i, j) for $1 \leq i \leq n$ and $1 \leq j \leq m$. For a test value t, we define the index set I to be $\{(i, j) : 1 \leq i \leq n - t + 1,\ 1 \leq j \leq m - t + 1\}$. For any index pair $\nu = (i, j) \in I$, it is possible to observe a matching region of length t starting at ν, which is $\{A_{i+k} =_\nabla B_{j+k} : k = 0, 1, \ldots, t - 1\}$.

To state our results, we need a few more definitions. We define Y_ν to be an indicator function to denote the occurrence of the matching region of length t starting at $\nu = (i, j)$, that is,

$$Y_\nu = \mathbf{1}_{\{A_{i+k} =_\nabla B_{j+k} : k=0,1,\ldots,t-1\}}.$$

Thus, $Y_\nu = 1$ denotes the occurrence of a matching region of length t starting at ν. The probability of observing such an occurrence can be easily calculated by the following argument. The probability that $A_{i+k} =_\nabla B_{j+k}$, for $k = 0, 1, \ldots, t - 1$, is

$$\mathrm{P}(Y_\nu = 1) = \prod_{k=0}^{t-1} \mathrm{P}(A_{i+k} =_\nabla B_{j+k}) = p^t,$$

since $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$ have i.i.d. exponential distributions. Here, $p = 1 - \mathrm{e}^{-\lambda\sigma}$ is the probability of two random restriction fragments matching, and is computed later in Lemma 3.2. When the fragment pair immediately before the matching region matches, matching regions occur in clumps. To de-clump, we define X_ν to be the indicator function for the start of a new clump of matching region. Hence,

$$X_\nu = Y_\nu, \quad \text{if } i = 1 \text{ or } j = 1,$$

otherwise

$$X_\nu = \mathbf{1}_{\{A_{i-1} \neq B_{j-1},\, A_{i+k} =_\nabla B_{j+k} : k=0,1,\ldots,t-1\}}.$$

Therefore, it is easy to show that

$$\mathrm{P}(X_\nu = 1) = \begin{cases} p^t, & i = 1 \text{ or } j = 1, \\ (1 - p)p^t, & \text{otherwise.} \end{cases}$$

Let $W(t) = \sum_{\nu \in I} X_\nu$ denote the number of occurrences of de-clumped matching regions and $S_{n,m}$ denote the length of the longest matching region between $\boldsymbol{A}$ and $\boldsymbol{B}$. We will prove the following theorem, where we set

$$\lambda_{n,m}(t) = [(n + m - 2t + 1) + (n - t)(m - t)(1 - p)]p^t$$

and define

$$n \vee m = \max\{n, m\}.$$

Theorem 3.1. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two random restriction maps of length n and m respectively, as defined above, and $W(t)$ be the number of de-clumped matching regions of length t between $\boldsymbol{A}$ and $\boldsymbol{B}$. Then*

$$\mathrm{E}(W(t)) = \lambda_{n,m}(t).$$

Let b_1 and b_2 be as in Section 2. Then

$$\begin{aligned} |\mathrm{P}(W(t) = 0) - \mathrm{e}^{-\lambda_{n,m}(t)}| &\leq b_1 + b_2 \\ &\leq 2\lambda_{n,m}(t)(2t - 1)(n \vee m)p^t \\ &\quad + nm(2(2t - 1)(n \vee m)p^{(3/2+3c)t} + (2t - 1)^2 p^{(1+2c)t}), \end{aligned}$$

since

$$b_1 \le 2\lambda_{n,m}(t)(2t-1)(n \vee m)p^t$$

and

$$b_2 \le nm(2(2t-1)(n \vee m)p^{(3/2+3c)t} + (2t-1)^2 p^{(1+2c)t}).$$

The expected value of $W(t)$ can easily be derived as follows. Since

$$\mathrm{E}(W(t)) = \sum_{\nu \in I} \mathrm{E}(X_\nu) = \sum_{\nu \in I} \mathrm{P}(X_\nu = 1),$$

and there are $(n-t+1)+(m-t+1)-1$ distinct starting indices in I with $i=1$ or $j=1$ and $(n-t)(m-t)$ distinct starting indices in I with $i>1$ and $j>1$. The proof for the second result in Theorem 3.1 is the main goal in the next subsection.

From this theorem, we can derive a corollary to enable us to compute the tail probability of $S_{n,m}$ approximately. Before we start our study of approximating the tail probability of $S_{n,m}$, we present an asymptotic result. Arratia and Waterman (1985) proved for sequence matching that

$$\lim_{n,m\to\infty} \frac{S_{n,m}}{\log_{1/p}((1-p)nm)} = 1, \quad \text{with probability 1}$$

when $n=m$. By the same argument, they obtain the same result even if $n \neq m$, when the growth rate of n and m satisfies certain conditions.

The second result of Theorem 3.1 can be formulated under some conditions about the relative growth rate of n, m and t. Suppose that the growth rate of n and m follows

$$\frac{\log(n)}{\log(nm)} \to \rho > 0.$$

Then t can be scaled appropriately with n and m so that $\lambda_{n,m}(t)$ stays bounded away from 0 and ∞. Actually, from the asymptotic property of $S_{n,m}$, setting $t = \log_{1/p}((1-p)nm)+s$ will keep $\lambda_{n,m}(t)$ between 0 and ∞. We will be more specific on how to choose the relative growth rate at the end of Subsection 3.2. Using this growth rate, we will approximate the distribution of $W(t)$ by a Poisson distribution that has the same expected value as $W(t)$ and then derive the probability of $W(t)=0$ when $n, m, t \to \infty$. Thus, we obtain the following corollary of Theorem 3.1.

Corollary 3.1. *Under the conditions on the relative growth rate of n, m and t described above, there exists constants $C, \gamma > 0$, such that*

$$|\mathrm{P}(S_{n,m} < t) - \mathrm{e}^{-\lambda_{n,m}(t)}| = |\mathrm{P}(W(t)=0) - \mathrm{e}^{-\lambda_{n,m}(t)}| \le C(\log nm)^{-\gamma}.$$

3.2. Approximate distribution for $W(t)$

To establish the Poisson estimate for $W(t)$, we use the Chen–Stein method as introduced in Section 2. The first and second moments of $W(t)$ should be well behaved to achieve our goal. Therefore, we need find bounds for b_1 and b_2 in the method and show their convergence to 0 as $n, m \to \infty$. As in Theorem 2.1, for a given X_ν, where $\nu \in I$, let J_ν denote the set of potential dependence, i.e.

$$J_\nu = \{\mu = (i', j') \in I : |i - i'| \le t \text{ or } |j - j'| \le t\}.$$

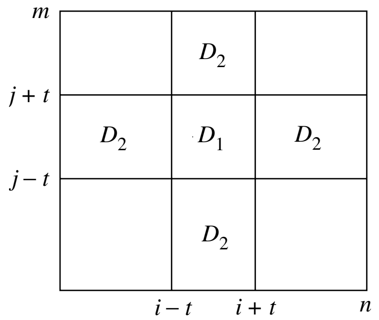

FIGURE 2: The four arms are D_2, the central part is D_1.

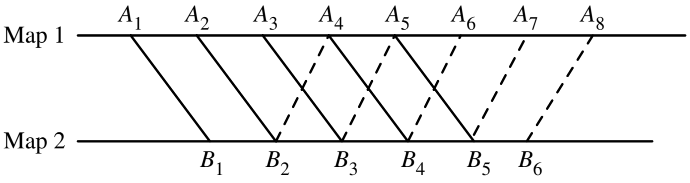

FIGURE 3: Solid lines and dashed lines represent two dependent matching regions of length 5, $Y_{(1,1)}$ and $Y_{(4,2)}$, respectively. The two matching regions separate all the involved fragments to two 2-components and two 5-components. They are $\{A_1 =_\nabla B_1\}$, $\{A_8 =_\nabla B_6\}$, $\{A_2 =_\nabla B_2, B_2 =_\nabla A_4, A_4 =_\nabla B_4, B_4 =_\nabla A_6\}$ and $\{A_3 =_\nabla B_3, B_3 =_\nabla A_5, A_5 =_\nabla B_5, B_5 =_\nabla A_7\}$.

If $\mu \in J_\nu$, the two matching regions starting at μ and ν share common restriction fragments, and so X_ν and X_μ can be dependent; otherwise, if $\mu \notin J_\nu$, the two matching regions share no common fragments, and so X_ν and X_μ are independent.

The estimation of b_1 is easily found to be

$$\begin{aligned} b_1 &= \sum_{\nu\in I}\sum_{\mu\in J_\nu} \mathrm{E}(X_\nu)\mathrm{E}(X_\mu) \le \sum_{\nu\in I}\mathrm{E}(X_\nu)(2t+1)(n+m)p^t \\ &\le \sum_{\nu\in I}\mathrm{E}(X_\nu)2(2t+1)(n\vee m)p^t = 2\lambda_{n,m}(t)(2t+1)(n\vee m)p^t. \end{aligned}$$

The estimation of the upper bound of b_2 is not as straightforward as the estimation of b_1. We notice that J_ν consists of a horizontal and a vertical strip. We divide J_ν into two disjoint subsets (see Figure 2). Let

$$D_1 = \{\gamma = (i', j') \in I : |i - i'| \le t \text{ and } |j - j'| \le t\}$$

be the intersection of the two strips and $D_2 = J_\nu - D_1$.

We estimate b_2 by computing upper bounds for $\mathrm{E}(X_\nu X_\mu)$ for $\mu \in D_i$, $i = 1, 2$, separately. Since $X_\nu \le Y_\nu$, an upper bound for $\mathrm{E}(Y_\nu Y_\mu)$ is also an upper bound for $\mathrm{E}(X_\nu X_\mu)$. Here we only study the bound for $\mathrm{E}(X_\nu X_\mu)$ when $i_\nu - i_\mu \ne j_\nu - j_\mu$, since $X_\nu = X_\mu = 0$ when $i_\nu - i_\mu = j_\nu - j_\mu$ by our definition of de-clumping. In the following figures, we represent a fragment as a node and a solid or dashed line connecting two matched fragments. All the fragments involved in the two matching regions are separated into independent connected components (see Figure 3). We define the size of a component as the number of fragments

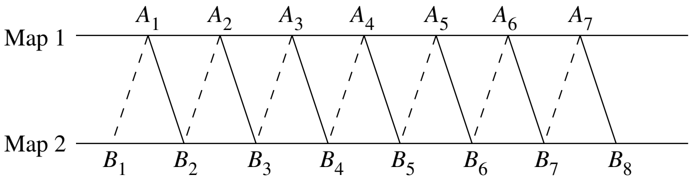

FIGURE 4: Full connected matching regions.

(nodes) involved in the component; an r-component is a component of size r. The distinct components do not share any common fragments, so they are independent. The probability that $Y_\nu = Y_\mu = 1$ can be written as the product of the probabilities of the occurrence of those independent components.

The following lemma about the components defined by two matching regions will be useful in estimating the upper bound for b_2. It is the same result as in sequence matching, but we include the proof for completeness.

Lemma 3.1. *Let Y_ν, Y_μ be variables defined for matching regions as above, that is, $\mu \in J_\nu$, and $i_\nu - i_\mu \neq j_\nu - j_\mu$. All the fragments involved in the two matching regions are separated into independent connected components. Let x_r, $r = 2, \ldots, 2t + 1$, denote the number of r-components, then*

$$\sum_{r=2}^{2t+1} x_r(2(r-1)) = 4t.$$

Proof. It is obvious that the size of a component is at least 2; the largest component formed from two matching regions of length t is that where all the fragments in the two matching regions are connected to one component (see Figure 4) and the size is $2t + 1$.

We count the number of fragments involved as follows. If a fragment appears in one matching region, it is counted once; if it appears in both matching regions, it is counted twice. Since there are $2t$ fragments in each matching region, there is a total of $4t$ fragments. For each r-component, the middle $r - 2$ fragments appear in both matching regions and are counted twice; the two end fragments appear in only one matching region and are counted once. So an r-component contains $2 + 2(r - 2) = 2(r - 1)$ counted fragments. Summing over all the components, we obtain

$$\sum_{r=2}^{2t-1} x_r(2(r-1)) = 4t.$$

Let p_r denote the following probability,

$$p_r = \mathrm{P}(A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3 \text{ and } \ldots\ A_{r-1} =_\nabla A_r),$$

where A_i, $i = 1, \ldots, r$, are i.i.d. and exponentially distributed with density function $\lambda e^{-\lambda x}$. We use p_r in the calculation of $\mathrm{E}(Y_\nu Y_\mu)$, but it is difficult to compute. Since we only need to estimate the upper bound of $\mathrm{E}(Y_\nu Y_\mu)$, we give the following two lemmas for a similar purpose to that of Lemma 11.5 in Waterman (1995). The proof of the two lemmas are given in the appendices.

Lemma 3.2. *Let p_3 denote the probability of observing a 3-component, then $p_3 = p^{3/2+3c}$, for a constant $0 < c < \frac{1}{6}$.*

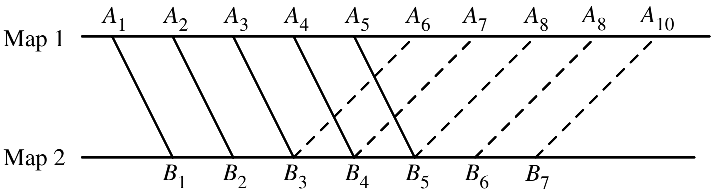

FIGURE 5: There are $y = 4$ 2-components and $x = 3$ 3-components. The length of matching region is 5, and $4x + 2y = 20 = 4t$.

Using Lemma 3.2, p_r can be bounded through p for any $r \geq 2$.

Lemma 3.3. *For any $r \geq 2$, there is*

$$p_r \leq p^{(1/2+c)(r-1)},$$

where c is as in Lemma 3.2.

Now, we consider the upper bound for b_2. First, we will consider $\mu \in D_2$ (see Figure 2). In this case, there are only 2-components and 3-components (see Figure 5), since no two matched fragments in the matching region starting at ν are involved in the matching region starting at μ and vice versa.

Let x denote the number of 3-components and y denote the number of 2-components; then $4x + 2y = 4t$. Since $3c < \frac{1}{2}$, we have $3c - \frac{1}{2} < 0$ and $(3c - \frac{1}{2})x > (3c - \frac{1}{2})t$ for $x \leq t$. Thus,

$$\begin{aligned} \mathrm{E}(Y_\nu Y_\mu) &= p^y p_3^x = p^{2t-2x}(p^{3/2+3c})^x = p^{2t+(3c-1/2)x} \\ &\leq p^{2t+(3c-1/2)t} = p^{(3/2+3c)t}, \quad \text{for } \mu \in D_2. \end{aligned}$$

Next, we consider $\mu \in D_1$. In this case, the size of component varies from 2 to $2t - 1$. If x_r denotes the number of r-components formed from the two matching regions, then $\sum_{r=2}^{2t-1} x_r(2(r-1)) = 4t$. We write

$$\begin{aligned} \mathrm{E}(Y_\nu Y_\mu) &= p^{x_2} p_3^{x_3} \cdots p_{2t-1}^{x_{2t-1}} \leq \prod_{r=2}^{2t-1} (p^{(1/2+c)(r-1)})^{x_r} \\ &= p^{(1/2+c)\sum_{r=2}^{2t-1}(r-1)x_r} = p^{(1/2+c)2t} = p^{(1+2c)t}, \quad \text{for } \mu \in D_1. \end{aligned}$$

The inequality is due to Lemma 3.2, $p_r \leq p^{(1/2+c)(r-1)}$, for $r \geq 2$.

We have bounded $\mathrm{E}(Y_\nu Y_\mu)$ for μ in D_1 and D_2 separately. From the definition of D_1, we know that $|D_1| = (2t + 1)^2$ and $|D_2| = (2t + 1)(n + m)$. To obtain the bound for b_2 we combine the above results:

$$\begin{aligned} b_2 &= \sum_\nu \sum_{\nu \neq \mu \in J_\nu} \mathrm{E}(X_\nu X_\mu) \leq \sum_\nu \sum_{\nu \neq \mu \in J_\nu} \mathrm{E}(Y_\nu Y_\mu) \\ &= \sum_\nu \left(\sum_{D_2} + \sum_{D_1} \right) \mathrm{E}(Y_\nu Y_\mu) \\ &\leq nm((2t+1)(n+m)p^{(3/2+3c)t} + (2t+1)^2 p^{(1+2c)t}) \\ &\leq nm(2(2t+1)(n \vee m)p^{(3/2+3c)t} + (2t+1)^2 p^{(1+2c)t}). \end{aligned}$$

Combining with the bound for b_1, we obtain

$$b_1 + b_2 \le 2\lambda(2t+1)(n \vee m)p^t + nm(2(2t+1)(n \vee m)p^{(3/2+3c)t} + (2t+1)^2 p^{(1+2c)t}).$$

We have proved Theorem 3.1. To derive Corollary 3.1, we need to show the bound for $b_1 + b_2$ goes to zero as $n, m, t \to \infty$ under the relative growth rate of n and m

$$\frac{\log(n)}{\log(nm)} \to \rho > 0 \quad \text{and} \quad \frac{\log(m)}{\log(nm)} \to 1 - \rho > 0,$$

and make a choice of t such that $\lambda_{n,m}(t) \in (0, \infty)$. The expected value of $W(t)$ has already been calculated and is denoted by $\lambda_{n,m}(t)$. Obviously,

$$\lambda_{n,m}(t) \approx_\infty nmp^t,$$

where $\approx_\infty$ means the asymptotic equality of the logarithms of the two quantities, that is,

$$Q_1 \approx_\infty Q_2$$

implies that

$$\frac{\log(Q_1)}{\log(Q_2)} \to 1, \quad \text{as } n, m \to \infty,$$

where Q_1 and Q_2 are two quantities depending on n and m. From the asymptotic equality between Q_1 and Q_2, it is easy to derive that $Q_1 \to 0$ as $Q_2 \to 0$ for $n, m \to \infty$ and vice versa.

If $\rho \ge \frac{1}{2}$, then $(n \vee m) \approx_\infty n \approx_\infty (nm)^\rho$, and thus

$$\begin{aligned} b_1 &\approx_\infty 2\lambda_{n,m}(t)(2t+1)(n \vee m)p^t \\ &\approx_\infty \lambda_{n,m}(t)(2t+1)(nm)^\rho p^t \\ &\approx_\infty \lambda^2_{n,m}(t)(2t+1)(nm)^{\rho-1}, \end{aligned}$$

and

$$\begin{aligned} b_2 &\approx_\infty (2t+1)(nm)^{1+\rho} p^{(3/2+3c)t} + (2t+1)^2 (nm) p^{(1+2c)t} \\ &\approx_\infty \lambda_{n,m}^{(3/2+3c)}(t)(2t+1)(nm)^{\rho-(1/2)-3c} + \lambda_{n,m}^{(1+2c)}(t)(2t+1)^2(nm)^{-2c}. \end{aligned}$$

To ensure that these bounds go to 0, the following conditions should be satisfied:

$$\rho - 1 < 0, \quad \rho - \tfrac{1}{2} - 3c < 0, \quad \text{and} \quad -2c < 0.$$

The first and third inequalities are satisfied automatically, so ρ should assume values to satisfy the second inequality, and we obtain $\rho < \frac{1}{2} + 3c$. When $\rho \le \frac{1}{2}$, $(n \vee m) \approx_\infty m \approx_\infty (nm)^{1-\rho}$, and repeat the above argument, to obtain $\rho > \frac{1}{2} - 3c$. From the analysis, if we let $\rho \in (\frac{1}{2} - 3c, \frac{1}{2} + 3c)$, and t be chosen so that $\lambda_{n,m}(t)$ is bounded away from 0 and ∞, then $b_1 + b_2 \to 0$ when $n, m, t \to \infty$. We conclude that $W(t)$ is approximately distributed as a Poisson random variable with mean $\lambda_{n,m}(t)$. Since

$$|\mathrm{P}(W(t) = 0) - \mathrm{e}^{-\lambda_{n,m}(t)}| \le (b_1 + b_2)\frac{1 - \mathrm{e}^{-\lambda_{n,m}(t)}}{\lambda_{n,m}}(t) \le b_1 + b_2,$$

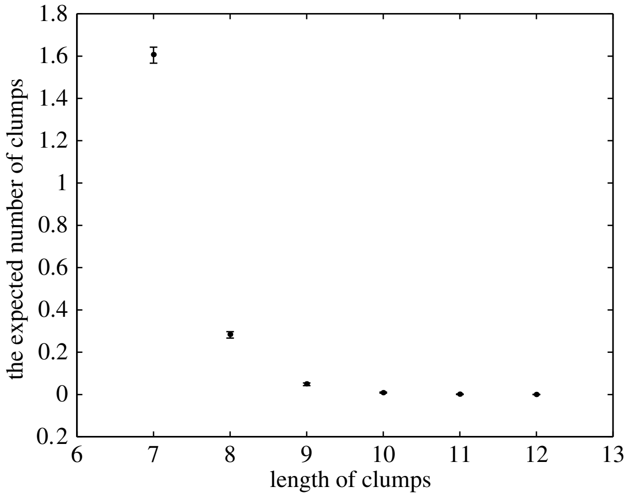

FIGURE 6: Comparison of $\lambda_{600,600}(t)$, the expected number of $W(t)$, of our calculated formula and simulation results with 2 standard deviation.

and letting $t = \log_{1/p}(nm) + s,\ s > 0$, to keep $\lambda_{n,m} \in (0, \infty)$, there exists some constants $C > 0$ and $\gamma > 0$ such that

$$b_1 + b_2 \le C(\log nm)^{-\gamma},$$

under the relative growth rate of n, m and t as discussed in the above asymptotic study. Therefore,

$$|\mathrm{P}(W(t) = 0) - \mathrm{e}^{-\lambda_{n,m}(t)}| \le C(\log nm)^{-\gamma}.$$

Finally, we wish to find the p-value of $S_{n,m}$ given a test value t. If $W(t) = 0$, then there is no matching region of length t between $\boldsymbol{A}$ and $\boldsymbol{B}$, which implies that there is no matching region of length greater than or equal to t between the two restriction maps, that is, $S_{n,m} < t$; on the other hand, $S_{n,m} < t$ implies that there is no matching region of length t between $\boldsymbol{A}$ and $\boldsymbol{B}$, and thus $W(t) = 0$. We conclude that the p-value of $S_{n,m}$ is the same as $1 - \mathrm{P}(W(t) = 0)$, so

$$|\mathrm{P}(S_{n,m} < t) - \mathrm{e}^{-\lambda_{n,m}(t)}| =\le C(\log nm)^{-\gamma}.$$

3.3. Testing the model

In the previous subsection, we showed that $W(t)$ is approximately Poisson distributed. We will do tests simulating restriction map matching to show how well the distribution is approximated. Our simulation tests $\lambda_{n,m}(t)$ and the distribution of $W(t)$ compared with the Poisson distribution.

We use $n = m = 600$, $\sigma = 200$ and $\lambda = 1/1024$ (corresponding to a 5-letter cutter). In our simulation, 5000 pairs of random restriction maps are compared. For each comparison, we count the number of de-clumped matching regions of some specified length t, where t assumes values 8, 9, 10, 11 and 12 separately. The average number of de-clumped matching regions of length t for the 5000 comparisons is computed and compared with the values from our formula for $\lambda_{n,m}(t)$. From Figure 6, we see that $\lambda_{n,m}(t)$ agrees with the simulation results very well, which is due to the fact that $\lambda_{n,m}(t)$ is the exact analytical expected value of $W(t)$ under our assumptions.

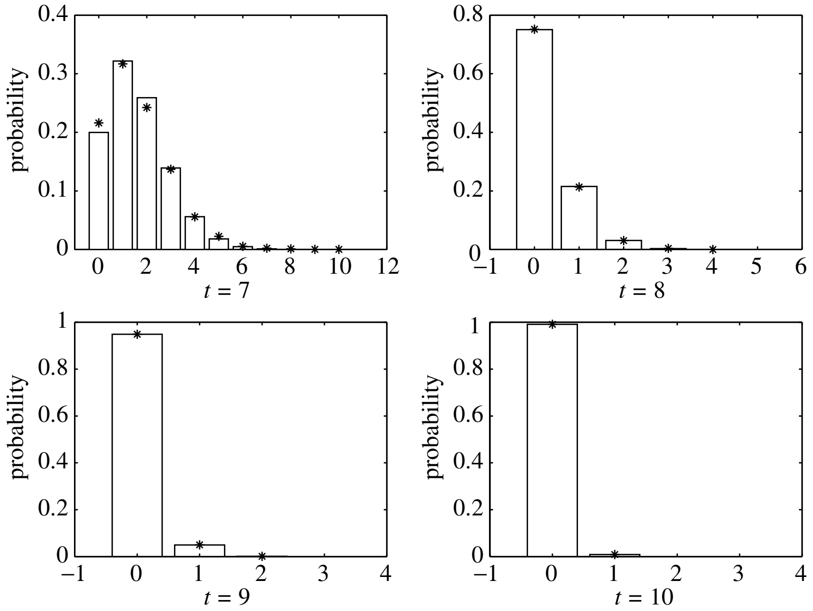

FIGURE 7: Poisson approximation for the distribution of $W(t)$, for $t = 7, 8, 9, 10$. Here, * denotes the empirical distribution of $W(t)$ from our simulation. The vertical bar denotes the Poisson distribution with mean value $\lambda_{600,600}(t)$.

To test the distribution of $W(t)$, we collect the lengths of matching regions for many (5000) matchings of restriction maps of length $n = m = 600$. Then, for a given t, we count the number of compared restriction map pairs with maximum matching regions of length over t. In general, for larger t the Poisson approximation is better. To show how well $W(t)$ is approximated by the Poisson distribution, we compare the empirical distribution of $W(t)$ with the Poisson distribution Z with the same expected value $\lambda_{n,m}(t)$. From Figure 7, we see that the larger t is, the better the empirical distribution is approximated by Poisson.

4. Matching regions with a few mismatches

In the evolution of DNA sequences, when deletions occur (here we refer to large segment deletions, which reduce the length of a restriction fragment), the restriction map of the DNA sequence with deletions is quite different from the restriction map of the original DNA sequence. If the deletion occurs within a restriction fragment, i.e. between two adjacent cut sites, then the lengths of the two restriction fragments in the two DNA sequences do not match (see Figure 8). In this case, even if the two DNA sequences are highly related, it might not be reported since the deletion shortens the length of matching region between the two restriction maps.

4.1. The expected counts of matching regions

To be able to reveal highly related restriction maps even if there are large segment deletions within single restriction fragments, we allow a few mismatches in a matching region. Instead of the model of exact matching, in which all fragment pairs in a matching region are matched, we study the model of *imperfect match*, in which we consider a long run of matches of length

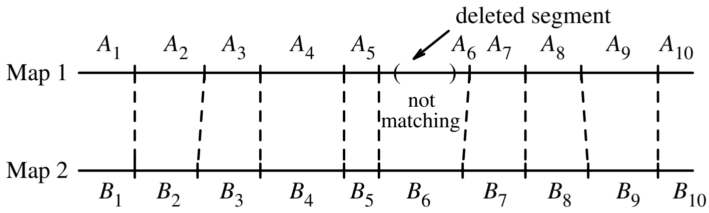
deleted segment
A_1 A_2 A_3 A_4 A_5 A_6 A_7 A_8 A_9 A_{10}
Map 1
not matching
Map 2
B_1 B_2 B_3 B_4 B_5 B_6 B_7 B_8 B_9 B_{10}

FIGURE 8: The occurrence of a segment deletion in map 1 causes A_6 not to match B_6.

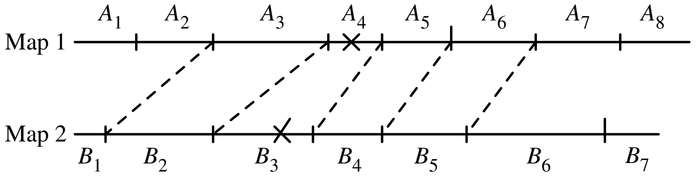
A_1 A_2 A_3 A_4 A_5 A_6 A_7 A_8
Map 1
Map 2
B_1 B_2 B_3 B_4 B_5 B_6 B_7

FIGURE 9: Example of a matching region of length 4 with 1 mismatch.

t except for k mismatches, i.e.

$$\{A_{i+r} =_{\nabla} B_{j+r} : r = 0, \dots, t-1, \text{ except for } k \text{ mismatches}\},$$

as a matching region of length t (see Figure 9), and it is called a matching region of type $(t; k)$, where $k < t$. Also, we call the t consecutive fragment pairs a window of size t starting at (i, j). For a fixed k, we wish to find the longest window with no more than k mismatches. Let $S^k_{n,m}$ denote the maximum length of a window between $\boldsymbol{A}$ and $\boldsymbol{B}$ including at most k mismatches. We wish to find the distribution of $S^k_{n,m}$ in order to estimate the p-value. We study the problem for fixed t first.

When looking at a window of size t, the more matching pairs we observe the higher the similarity is. We begin by finding the probability of observing k mismatches in a window of size t. As usual, $I = \{(i, j) : 1 \le i \le n - t + 1,\ 1 \le j \le m - t + 1\}$ denotes the index set. To begin with, we consider fixed t and k, and refer to matching regions without mentioning type $(t; k)$. We give some definitions below that are used in the proof. Let Y_ν be the indicator function for indicating the occurrence of a matching region starting at $\nu \in I$. (Remember that Y_ν actually depends on k in this section. For simplicity, we ignore k in all the notations.) Since we allow k mismatches in a matching region, there are many possible matchings starting at ν of the same type. Let $Y^\nu_1, Y^\nu_2, \dots, Y^\nu_{K_\nu}$ be indicator functions of all possible matchings starting at $\nu \in I$; then

$$\begin{aligned} Y_\nu &= \mathbf{1}_{\{Y^\nu_1=1 \text{ or } Y^\nu_2=1 \text{ or ... or } Y^\nu_{K_\nu}=1\}} \\ &= \mathbf{1}_{\{A_{i+r}=_\nabla B_{j+r}:r=0,\dots,t-1 \text{ except for } k \text{ mismatches}\}}, \end{aligned}$$

where K_ν is the number of all possible distinct matchings starting at ν. Clearly,

$$K_\nu = \binom{t}{k} \le t^k,$$

which is a constant for all $\nu \in I$.

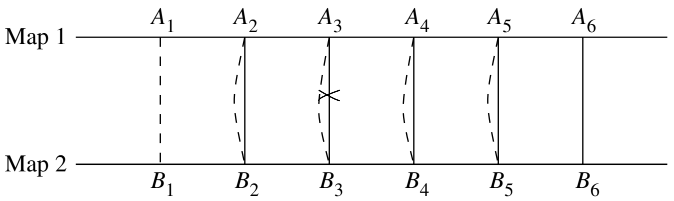

FIGURE 10: Two highly correlated matching regions.

In each possible matching region, there are $t-k$ matching pairs and k mismatches. It is easy to calculate the probability $\mathrm{P}(Y_i^\nu = 1) = p^{t-k}(1-p)^k$. Let $p_{t,k}$ denote the probability that there is at least one matching region starting at position $\nu = (i_\nu, j_\nu) \in I$ with k mismatches, i.e. $p_{t,k} = \mathrm{P}(Y_\nu = 1)$. For $1 \le i \ne j \le K_\nu$, $Y_i^\nu = Y_j^\nu = 1$ implies there is an $0 \le l < t$ with $A_{i_\nu+l} =_\nabla B_{j_\nu+l}$ in one matching region and $A_{i_\nu+l} \ne B_{j_\nu+l}$ in the other one, which is a contradiction. Therefore, the events $\{Y_i^\nu\}_{i=1}^{K_\nu}$ are mutually exclusive, and we obtain

$$\begin{aligned} p_{t,k} &= \mathrm{P}(\{A_{i+r} =_\nabla B_{j+r} : r = 0, 1, \dots, t-1, \text{ except for } k \text{ mismatches}\}) \\ &= \sum_{l=0}^{K_\nu} \mathrm{P}(Y_l^\nu = 1) = \binom{t}{k} (1-p)^k p^{t-k}, \end{aligned}$$

which is also $\mathrm{E}(Y_\nu)$. The two matching regions starting at $\nu = (i_\nu, j_\nu) \in I$ and $\mu = (i_\mu, j_\mu) \in I$ and $\mu \ne \nu$, if $i_\nu - i_\mu = j_\nu - j_\mu$ and $|i_\nu - i_\mu| < t$, are highly correlated (see Figure 10) since they share $t - |i_\nu - i_\mu|$ common compared fragment pairs in the two windows. Thus, the matching regions tend to occur in clumps.

To be able to obtain a Poisson approximation, we should count the clumps and define the indicator function that a clump begins at ν to be

$$X_\nu = Y_\nu(1 - Y_{\nu-1})(1 - Y_{\nu-2}) \cdots (1 - Y_{\nu-t+1}),$$

where $\nu - l$, for $l = 1, \dots, t-1$, is the index pair $(i_\nu - l, j_\nu - l)$. When $i_\nu - l < 1$ or $j_\nu - l < 1$, we always have $Y_{\nu-l} = 0$. To simplify the proof of the next lemma, we neglect the edge condition and set $I' = \{(i,j) : t-1 < i < n-t+1,\ t-1 < j < m-t+1\}$. We state a result for approximating the expected value of X_ν, $\nu \in I'$, by the expected value of Y_ν. The proof for this lemma is almost the same as the proof of a similar lemma in Arratia *et al.* (1990) and is therefore omitted.

Lemma 4.1. *Let $p < a = (t-k)/t \le 1$, for $\nu \in I'$*

$$a - p \le \frac{\mathrm{E}(X_\nu)}{\mathrm{E}(Y_\nu)} \le (a-p) + C(k)t^k p^t,$$

where $C(k)$ is a positive constant depending on k and p.

For fixed k, $C(k)t^k p^t \to 0$ as $t \to \infty$, so $\mathrm{E}(X_\nu) \to (a-p)\mathrm{E}(Y_\nu)$ when t goes to infinity. For large t we can use $(a-p)\mathrm{E}(Y_\nu)$ to approximate $\mathrm{E}(X_\nu)$.

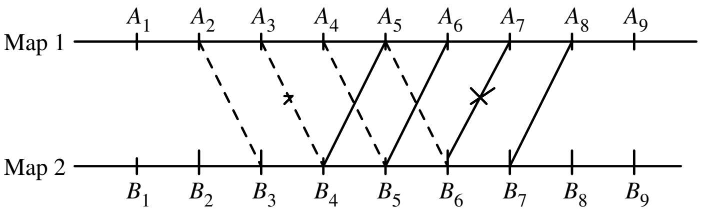

FIGURE 11: Solid lines and dashed lines each represent two matching regions of length $t = 4$ and $k = 1$. There are $x_2 = 2$ 2-components, $x_3 = 2$ 3-components and 2 mismatches in total. Therefore, $2 \times 2 + 2(2(3-1)) + 4 \times 1 = 16 = 4t$.

Theorem 4.1. *Let $W(t,k) = \sum_{\nu \in I} X_\nu$ be the number of de-clumped matching regions of type $(t;k)$. We have*

$$1 \le \frac{\mathrm{E}(W(t,k))}{e_{t,k}} \le 1 + \frac{C(k)t^k p^t}{a-p} + \frac{(n+m-2t+2)(t-1)}{(a-p)(n-2t+2)(m-2t+2)},$$

where $e_{t,k} = \sum_{\nu \in I'} (a-p)\mathrm{E}(Y_\nu) = (a-p)(n-2t+2)(m-2t+2)\mathrm{E}(Y_\nu)$.

The proof for this theorem is quite straightforward (the extra term in the upper bound is from end effects). Thus, we can approximate $\mathrm{E}(W(t,k))$ by neglecting the edge conditions. With this approximation, we state the following distribution approximation for $W(t,k)$.

Theorem 4.2. *Let $W(t,k) = \sum_{\nu \in I} X_\nu$ be the number of de-clumped matching regions of type $(t;k)$ and $\lambda_{t,k} = \mathrm{E}(W(t,k))$. Then b_1 and b_2 for $\{X_\nu\}_{\nu \in I}$ are bounded as*

$$\begin{aligned}|\mathrm{P}(W(t,k)=0) - \mathrm{e}^{-\lambda_{t,k}}| &\le b_1 + b_2 \\ &\le 2(2t+1)\lambda_{t,k}(n \vee m)p_{t,k} \\ &\quad + nm(2t+1)t^{2k}(C_k(n+m)p^{(3/2+3c)t} + D_k(2t+1)p^{(1+2c)t}).\end{aligned}$$

We can approximate $\lambda_{t,k}$ by $e_{t,k}$ in Theorem 4.1 when t is large. The distribution of $W(t,k)$ will be approximated by the Poisson distribution using the Chen–Stein method, and then we derive the estimation of the probability that $W(t,k) = 0$ from the distributional result. To approximate the distribution of $W(t,k)$, we need to show that the two quantities b_1 and b_2 are small when n, m are big. The proof for this theorem is similar to the proof of the exact match model. Only the combinatorial result of the components is a little different, since we do not count the mismatches in a component. This difference does not change the proof much. When k, the number of mismatches, is fixed, the terms related to k can always be absorbed as a constant factor in the approximation. The following is the combinatorial lemma.

Lemma 4.2. *Let Y_i^ν and Y_j^μ be the two dependent variables defined in this section and suppose that $i_\nu - i_\mu \ne j_\nu - j_\mu$. All the fragments in the two matching regions are separated into connected components. Let x_r, $r = 2, \dots, 2t+1$, denote the number of r-components. Then*

$$\sum_{r=2}^{2t+1} x_r(2(r-1)) = 4t - 4k.$$

Figure 11 shows an example of the resulting components from two imperfect matching regions.

One difference from the exact match model in the imperfect match model is the selection of t, the asymptotic centring constant, to keep the $\lambda_{t,k}$ away from 0 and ∞. In analogy with the discussion of Arratia *et al.* (1986), we let

$$t = \log_{1/p}(nm) + k \log_{1/p} \log_{1/p}(nm) + s, \qquad s > 0.$$

Thus, $p^t = (nm)^{-1}(\log_{1/p}(nm))^{-k} p^s$. Therefore,

$$\begin{aligned}\lambda_{t,k} &\approx (a-p)(n-2t+2)(m-2t+2)\binom{t}{k}(1-p)^k p^{t-k}\\ &\approx_\infty nmp^t t^k (1-p)^k p^{-k}\\ &= (nm)(nm)^{-1}\frac{t^k}{(\log_{1/p}(n))^k} p^{s-k}(1-p)^k\\ &= \frac{t^k}{(\log_{1/p}(nm))^k} p^{s-k}(1-p)^k\end{aligned}$$

is bounded away from 0 and ∞, since $t/\log_{1/p}(nm) \to 1$ as $n, m \to \infty$. Now, we can derive a corollary from Theorem 4.2 as we did in the exact match model.

Corollary 4.1. *Under the same conditions on the relative growth rate of n, m in the exact match model and letting $t = \log_{1/p}(nm) + k \log_{1/p} \log_{1/p}(nm) + s$, $s > 0$, there exist C, $\gamma > 0$ such that*

$$|\mathrm{P}(W(t,k) = 0) - \mathrm{e}^{-\lambda_{t,k}}| \le C(\log nm)^{-\gamma}.$$

4.2. Approximate distribution of $S^k_{n,m}$

In the previous section, we studied the approximate distribution of observing a window of size t including k mismatches. Fixing the number of mismatches, k, allowed in a window, the distribution of $S^k_{n,m}$ is studied in this section. We wish the approximate distribution of $W(t,k)$ to derive the tail probability of $S^k_{n,m}$ and have the following theorem.

Theorem 4.3. *Let $S^k_{n,m}$ denote the length of maximum matching regions between $\boldsymbol{A}$ and $\boldsymbol{B}$ of length n, m with at most k mismatches. Under the conditions on the relative growth rate of n, m, k and t described in the last section, there exist C_1, $\gamma_1 > 0$ such that*

$$|\mathrm{P}(S^k_{n,m} < t) - \mathrm{e}^{-\lambda_{t,k}}| < C_1(\log nm)^{-\gamma_1}.$$

This result does not follow exactly as in the exact matching case. To derive this result from the previous results about $W(t,k)$, we fix the window size t and find the distribution of the maximum number of matching pairs in all the windows of size t. Let M_t denote the maximum number of matching pairs in a window of size t. It is obvious that $\{M_t < t-k\} \subset \{W(t,k) = 0\}$. The inequality is due to the existence of windows with more than $t-k$ matching pairs.

Lemma 4.3. *If there exists $\nu \in I$ such that $Y_\nu = 1$, then there exists $\nu' \in I$ such that $X_{\nu'} = 1$, which implies that $W(t,k) > 0$.*

The lemma states that the existence of a matching region of type $(t;k)$ implies the existence of such a clump. Before we start the proof, for $\nu = (i_\nu, j_\nu)$, $\mu = (i_\mu, j_\mu) \in I$, we define a partial order relation, $\nu < \mu$ if $i_\mu - i_\nu = j_\mu - j_\nu > 0$.

Proof. If $Y_\nu = 1$ for $\nu \in I$, then $X_\nu = 1$ or there exists $\nu > \nu_1 \in I$ such that $Y_{\nu_1} = 1$ from the definition of X_ν. If $X_\nu = 1$, then we are done; otherwise, $Y_{\nu_1} = 1$ implies that $X_{\nu_1} = 1$ or $Y_{\nu_2} = 1$ for some $\nu_1 > \nu_2 \in I$. Repeating this discussion, we obtain a strictly decreasing index series $\nu, \nu_1, \nu_2, \cdots \in I$ such that $Y_{\nu_i} = 1$. Since the index set has lower bound $(1, 1)$, the strictly decreasing index series stops at some $\nu_h \in I$, such that $Y_{\nu_h - l} \neq 1$ for $1 \leq l \leq t-1$, which implies that $X_{\nu_h} = 1$.

The next theorem shows the difference of the probability between observing no window of size t with k mismatches and observing no window of size t with k mismatches but observing a window of size t with less than k mismatches.

Theorem 4.4. *Let M_t and $W(t,k)$ be defined as above, then*

$$0 \leq \mathrm{P}(W(t,k)=0) - \mathrm{P}(M_t < t-k) \leq C_2(\log nm)^{-\gamma_2},$$

where C_2, $\gamma_2 > 0$ are two constants.

Proof. The first inequality is obvious since $\{M_t < t-k\} \subset \{W(t,k)=0\}$. When $M_t \geq t-k$, there exists at least one window with no more than k mismatches. If this window has exactly k mismatches, from Lemma 4.3 we know that $W(t,k) > 0$. To have $W(t,k) = 0$ and $M_t \geq t-k$, each window should have more or less than k mismatches and at least one window with less than k mismatches, say the window at $\nu = (i, j)$. Let $N_t(\nu)$ be the number of matching pairs in a window of size t starting at ν, which is then greater than $t-k$. The window starting at $\nu - 1$ has at most one less matching pair than the window starting at ν, so $N_t(\nu - 1) \geq t-k$. In the case of equality, we can derive $W(t,k) > 0$ from Lemma 4.3, hence strict inequality holds. We keep moving the window left by one pair until we reach the smallest possible index, and obtain

$$\begin{cases} N_t((1,1)) > t-k, & \text{if } i = j, \\ N_t((i-j+1, 1)) > t-k, & \text{if } i > j, \\ N_t((1, j-i+1)) > t-k, & \text{if } i < j. \end{cases}$$

So, if $W(t,k) = 0$ and $M_t \geq t-k$, at least one of the windows with at least one 1 in the starting point, as shown above, has more than $t-k$ matching pairs. Thus,

$$\begin{aligned} \mathrm{P}(W(t,k)=0 \text{ and } M_t \geq t-k) &\leq (1+n-t+m-t)\mathrm{P}(N_t((1,1)) > t-k) \\ &\leq 2(n \vee m)\sum_{l=0}^{k-1} p_{t,l} \leq 2C(k)(n \vee m)t^k p^t, \end{aligned}$$

since

$$\sum_{l=0}^{k-1} p_{t,l} = \sum_{l=0}^{k-1} \binom{t}{l} p^{t-l}(1-p)^l \leq \sum_{l=0}^{k-1} t^k p^t \left(\frac{1-p}{p}\right)^l \leq C(k)t^k p^t,$$

where $C(k)$ is a constant depending on k and p. Under the relative growth rate of n, m and $\lambda_{t,k}$, $\lambda_{t,k} \approx_\infty nmt^k p^t$ is bounded away from 0 and ∞ as in Section 3.2,

$$C(k)(n \vee m)t^k p^t \approx_\infty \begin{cases} (nm)^\rho (nm)^{-1} = (nm)^{\rho - 1}, & \rho \geq \frac{1}{2}, \\ (nm)^{1-\rho}(nm)^{-1} = (nm)^{-\rho}, & \rho < \frac{1}{2}, \end{cases}$$

with $0 < p < 1$, which goes to zero when $n, m \to \infty$. So there exist C_2, $\gamma_2 > 0$ such that

$$\mathrm{P}(W(t,k)=0) - \mathrm{P}(M_t < t-k) \leq C_2(\log nm)^{-\gamma_2}.$$

Combined with Theorem 4.1, the tail probability of M_t can be approximated by the following theorem.

Theorem 4.5. *Under the same conditions on the growth rate of n, m and t as in Corollary 4.1, there exist* $C_1, \gamma_1 > 0$ *such that*

$$|\mathrm{P}(M_t < t - k) - \mathrm{e}^{-\lambda_{t,k}}| \le C_1(\log nm)^{-\gamma_1}$$

under certain conditions for the relative growth rate of n, m and t as described in Theorem 4.3.

This can be proved easily, since

$$\begin{aligned}|\mathrm{P}(M_t < t - k) - \mathrm{e}^{-\lambda_{t,k}}| &\le |\mathrm{P}(M_t < t - k) - \mathrm{P}(W(t,k) = 0)| + |\mathrm{P}(W(t,k) = 0) - \mathrm{e}^{-\lambda_{t,k}}| \\ &\le C(\log nm)^{-\gamma} + C_2(\log nm)^{-\gamma_2} \\ &\le C_1(\log nm)^{-\gamma_1},\end{aligned}$$

for some $C_1, \gamma_1 > 0$.

Now we return to the distribution of $S^k_{n,m}$. For any t, we have the following equivalences:

$S^k_{n,m} < t$ $\iff$ there is no window of size no smaller than t with k or fewer mismatches
$\iff$ there is no window of size t with k or fewer mismatches
$\iff$ $M_t < t - k$.

Therefore,

$$\mathrm{P}(S^k_{n,m} < t) = \mathrm{P}(M_t < t - k),$$

and Theorem 4.3 is proved through Theorem 4.5.

5. Open problem

Since mutations occur in DNA sequences, it is possible that a mutation might create a new cut site or make an existing cut site disappear. When a cut site is mutated into a non-cut-site, the two fragments from the cut site in the restriction map of the DNA sequence before the mutation merge into one big fragment in the restriction map of the DNA sequence after the mutation. A mutation causing a new cut site divides one fragment including the new cut site into two smaller fragments. To make our results more powerful and useful, we consider this kind of mutation in restriction map matching. This matching is considered in a general form in the algorithms of Huang and Waterman (1992).

The mutation of a cut site might result in the merge of two restriction fragments to one fragment or vice versa, so we allow a few merge-matches in a matching region as depicted in Figure 12. A *merge-match* is defined to occur when the sum of lengths of two adjacent

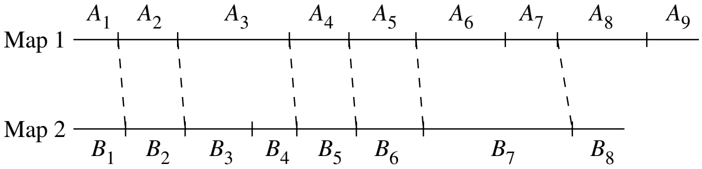

FIGURE 12: An example of a matching region of length 5 including 2 merge-matches, $A_3 =_\nabla B_3 + B_4$ and $A_6 + A_7 =_\nabla B_7$.

fragments differs from the length of the corresponding fragment by no more than σ. Three fragments are involved in a merge-match. A mutation in a cut site might induce a merge-match.

For fixed t and k, a matching region of length t is defined as a long run of t matching fragment pairs except for at most k merge-matches between two restriction maps. The authors undertook a study of this type of matching regions, wishing to obtain similar results to those obtained for exact matchings and imperfect matchings. Because the merge-match brings greater complexity to the combinatorial analysis in the study, no good results have yet been obtained. The statistical results for this more complicated matching will be more powerful in detecting significant similarity between maps.

For practical application, we recommend simulation studies as in Waterman and Vingron (1994). At the basis of that work was a decomposition of the scoring parameter space into linear and logarithmic growth. The corresponding generalization of Arratia and Waterman (1989) could be established for map matching.

Appendix A. Proof of Lemma 3.2

To show the relation between p and p_3, we compute them from the exponential distribution first:

$$\begin{aligned}
p &= \mathrm{P}(A_1 =_\nabla A_2) = \mathrm{P}(|A_1 - A_2| < \sigma)\\
&= \int_0^\sigma \int_0^{x+\sigma} \lambda \mathrm{e}^{-\lambda x} \lambda \mathrm{e}^{-\lambda y}\, \mathrm{d}y\mathrm{d}x + \int_\sigma^\infty \int_{x-\sigma}^{x+\sigma} \lambda \mathrm{e}^{-\lambda x} \lambda \mathrm{e}^{-\lambda y}\, \mathrm{d}y\mathrm{d}x\\
&= \int_0^\sigma \lambda \mathrm{e}^{-\lambda x}(1 - \mathrm{e}^{-\lambda(\sigma+x)})\, \mathrm{d}x + \int_\sigma^\infty \lambda \mathrm{e}^{-\lambda x}(\mathrm{e}^{-\lambda(x-\sigma)} - \mathrm{e}^{-\lambda(x+\sigma)})\, \mathrm{d}x\\
&= (1 - \mathrm{e}^{-\lambda\sigma}) + \tfrac{1}{2}\mathrm{e}^{-\lambda\sigma}(\mathrm{e}^{-2\lambda\sigma} - 1) + \tfrac{1}{2}(\mathrm{e}^{\lambda\sigma} - \mathrm{e}^{-\lambda\sigma})\mathrm{e}^{-2\lambda\sigma}\\
&= 1 - \mathrm{e}^{-\lambda\sigma},
\end{aligned}$$

and

$$\begin{aligned}
p_3 &= \mathrm{P}(A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3)\\
&= \mathrm{P}(|A_1 - A_2| < \sigma \text{ and } |A_2 - A_3| < \sigma)\\
&= \int_0^\sigma \int_0^{x+\sigma} \int_0^{x+\sigma} \lambda \mathrm{e}^{-\lambda x} \lambda \mathrm{e}^{-\lambda y} \lambda \mathrm{e}^{-\lambda z}\, \mathrm{d}z\mathrm{d}y\mathrm{d}x\\
&\quad + \int_\sigma^\infty \int_{x-\sigma}^{x+\sigma} \int_{x-\sigma}^{x+\sigma} \lambda \mathrm{e}^{-\lambda x} \lambda \mathrm{e}^{-\lambda y} \lambda \mathrm{e}^{-\lambda z}\, \mathrm{d}z\mathrm{d}y\mathrm{d}x\\
&= \int_0^\sigma \lambda^{-\lambda x}(1 - \mathrm{e}^{-\lambda(x+\sigma)})^2\, \mathrm{d}x + \int_\sigma^\infty \lambda \mathrm{e}^{-\lambda x}(\mathrm{e}^{-\lambda(x-\sigma)} - \mathrm{e}^{-\lambda(x+\sigma)})^2\, \mathrm{d}x\\
&= \mathrm{e}^{\lambda\sigma} \int_0^\sigma (1 - \mathrm{e}^{-\lambda\sigma-\lambda x})^2\, \mathrm{d}(1 - \mathrm{e}^{-\lambda\sigma-\lambda x})\\
&\quad + (\mathrm{e}^{\lambda\sigma} - \mathrm{e}^{-\lambda\sigma})^2 \int_\sigma^\infty \lambda \mathrm{e}^{-\lambda x}\mathrm{e}^{-2\lambda x}\, \mathrm{d}x\\
&= \tfrac{1}{3}\mathrm{e}^{\lambda\sigma}((1 - \mathrm{e}^{-2\lambda\sigma})^3 - (1 - \mathrm{e}^{-\lambda\sigma})^3) + \tfrac{1}{3}(\mathrm{e}^{\lambda\sigma} - \mathrm{e}^{-\lambda\sigma})^2\mathrm{e}^{-3\lambda\sigma}\\
&= \tfrac{1}{3}\mathrm{e}^{\lambda\sigma}((1 - \mathrm{e}^{-2\lambda\sigma})^3 - (1 - \mathrm{e}^{-\lambda\sigma})^3) + \tfrac{1}{3}\mathrm{e}^{-\lambda\sigma}(1 - \mathrm{e}^{-2\lambda\sigma})^2.
\end{aligned}$$

Next, p_3 is written in terms of p. Since $p = 1 - \mathrm{e}^{-\lambda\sigma}$, $\mathrm{e}^{-\lambda\sigma} = 1 - p$, and then

$$\begin{aligned} p_3 &= \frac{1}{3(1-p)}((1-(1-p)^2)^3 - p^3) + \tfrac{1}{3}(1-p)(1-(1-p)^2)^2 \\ &= \frac{(1-(1-p)^2)^2}{3}\left(\frac{1-(1-p)^2}{(1-p)} + (1-p)\right) - \frac{p^3}{3(1-p)} \\ &= \frac{(p^2-2p)^2}{3}\frac{1}{(1-p)} - \frac{p^3}{3(1-p)} \\ &= \frac{p^2}{3(1-p)}((p-2)^2 - p) = \frac{p^2}{3(1-p)}((p-1)^2 + 3(1-p)) \\ &= p^2(\tfrac{1}{3}(1-p)+1) = p^{3/2}\sqrt{p}(\tfrac{1}{3}(1-p)+1) = p^{3/2}h(p), \end{aligned}$$

where

$$h(p) = \sqrt{p}(\tfrac{1}{3}(1-p)+1) = \sqrt{p}(\tfrac{4}{3} - \tfrac{1}{3}p).$$

To show that $p_3 < p^{3/2}$, we only need show that $h(p) < 1$ for $0 < p < 1$. Consider the derivative of $h(p)$:

$$h'(p) = \tfrac{4}{3}\tfrac{1}{2}p^{-1/2} - \tfrac{1}{3}\tfrac{3}{2}\sqrt{p} = \frac{2}{3\sqrt{p}}(1 - \tfrac{3}{4}p) > 0, \quad \text{when } 0 < p < 1,$$

so $h(p)$ is increasing in (0, 1), and $h(p) < h(1) = 1$ when $0 < p < 1$. Therefore, $p_3 < p^{3/2}$ and there exists a constant $c > 0$ such that $p_3 = p^{3/2+3c}$. From the above, we have

$$p_3 = p^2(\tfrac{1}{3}(1-p)+1) > p^2,$$

which implies that

$$\tfrac{3}{2} + 3c < 2, \quad \text{that is} \quad c < \tfrac{1}{6}.$$

Now, we can conclude that

$$p_3 = p^{3/2+3c}, \quad \text{where } 0 < c < \tfrac{1}{6}.$$

Appendix B. Proof of Lemma 3.3

It is obvious that the event $\{A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3 \text{ and } \ldots\ A_{r-1} =_\nabla A_r\}$ is included in the event $\{A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3,\ A_4 =_\nabla A_5 \text{ and } A_5 =_\nabla A_6,\ \ldots\}$. To show the inequality for p_r, the upper bound for the latter is found,

$$\begin{aligned} p_r &= \mathrm{P}(A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3 \text{ and } \ldots\ A_{r-1} =_\nabla A_r) \\ &\le \mathrm{P}(A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3,\ A_4 =_\nabla A_5 \text{ and } A_5 =_\nabla A_6, \ldots) \\ &= \mathrm{P}(A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3)\mathrm{P}(A_4 =_\nabla A_5 \text{ and } A_5 =_\nabla A_6) \cdots . \end{aligned}$$

Above, the r fragments are divided into small distinct groups with three fragments in each group, except that the last group might have fewer than three fragments. We will discuss three cases. Let $r = a$, mod 3, and $a = 0$, 1 or 2. We discuss the three cases separately.

Case 1. For $a = 0$,

$$p_r \le \mathrm{P}(A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3) \cdots \mathrm{P}(A_{r-2} =_\nabla A_{r-1} \text{ and } A_{r-1} =_\nabla A_r)$$
$$= p_3^{r/3} \le p_3^{r/3} p^{-(1/2+c)}.$$

Case 2. For $a = 1$,

$$p_r \le \mathrm{P}(A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3) \cdots \mathrm{P}(A_{r-3} =_\nabla A_{r-2} \text{ and } A_{r-2} =_\nabla A_{r-1})$$
$$= p_3^k = p_3^{r/3} p_3^{-1/3} = p_3^{r/3} p^{-(1/3)(3/2+3c)} = p_3^{r/3} p^{-(1/2+c)}$$

Case 3. For $a = 2$,

$$p_r \le \mathrm{P}(A_1 =_\nabla A_2 \text{ and } A_2 =_\nabla A_3) \cdots \mathrm{P}(A_{r-4} =_\nabla A_{r-3} \text{ and } A_{r-3} =_\nabla A_{r-2})$$
$$\times \mathrm{P}(A_{r-1} =_\nabla A_r)$$
$$= p_3^k p = p_3^{r/3} p_3^{-2/3} p = p_3^{r/3} p^{-(2/3)(3/2+3c)} p = p_3^{r/3} p^{-(1+2c)+1}$$
$$= p_3^{r/3} p^{-2c} \le p_3^{r/3} p^{-(1/2+c)},$$

since $c < \frac{1}{6}$, then $2c < \frac{1}{3} < \frac{1}{2} + c$.

Combining the above results, we obtain

$$p_r \le p_3^{r/3} p^{-(1/2+c)} < (p^{3/2+3c})^{r/3} p^{-(1/2+c)} = p^{(1/2+c)(r-1)}.$$

Acknowledgements

Once again we are indebted to Davis Schwartz for pioneering another area of scientific inquiry that is rich in problems for computational biology. This work was supported by grants from National Institutes of Health (GM36230) and the National Science Foundation (BIR 95-04393). The first author also expresses gratitude to the University of Southern California for his C. W. Trigg Fellowship and Dissertation Fellowship.

References

Anantharaman, T. S., Mishra, B. and Schwartz, D. C. (1997). Genomics via optical mapping II: ordered restriction maps. *J. Comput. Biol.* **4,** 91–118.

Arratia, R. and Waterman, M. S. (1985). An Erdős–Rényi law with shifts. *Adv. Math.* **55,** 13–23.

Arratia, R. and Waterman, M. S. (1989). The Erdős–Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Prob.* **17,** 1152–1169.

Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen–Stein method. *Ann. Prob.* **17,** 9–25.

Arratia, R., Gordon, L. and Waterman, M. S. (1986). An extreme value theory for sequence matching. *Ann. Statist.* **14,** 971–993.

Arratia, R., Gordon, L. and Waterman, M. S. (1990). The Erdős–Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.* **18,** 539–570.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Prob.* **3,** 534–545.

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

Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci. USA* **87,** 2264–2268.

Karlin, S., Dembo, A. and Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Prob.* **18,** 571–581.

Karlin, S. *et al.* (1983). New approaches for computer analysis of nucleic acid sequences. *Proc. Nat. Acad. Sci. USA* **80,** 5660–5664.

Kohara, Y., Akiyama, K. and Isono, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of large genomic libraries. *Cell* **50,** 495–508.

Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2,** 231–239.

Lin, J. *et al.* (1999). Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* **285,** 1558–1562.

Neuhauser, C. (1994). A Poisson approximation for sequence comparisons with insertions and deletions. *Ann. Statist.* **22,** 1603–1629.

Rudd, K. E. *et al.* (1990). Alignment of *Escherichia Coli* k12 DNA sequences to a genomic restriction map. *Nucleic Acids Res.* **18,** 313–321.

Schwartz, D. C. *et al.* (1993). Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262,** 110–114.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Prob.* Vol. 11, *Probability Theory*, eds L. M. Le Cam *et al.* University of California Press, Berkeley, pp. 583–602.

Tang, M. (2000a). Topics in computational genome analysis: (I) matching restriction maps and (II) evolution of gene families by duplication. Doctoral Thesis, University of Southern California.

Tang, M. (2000b). Global matching of random restriction maps. *Methodol. Comput. Appl. Prob.* **2,** 183–201.

Waterman, M. S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman and Hall, New York.

Waterman, M. S. and Raymond, R. (1987). The match game: new stratigraphic correlation algorithm. *Math. Geol.* **19,** 109–127.

Waterman, M. S. and Vingron, M. (1994). Sequence comparison significance and poisson approximation. *Statist. Sci.* **9,** 367–381.

Waterman, M. S., Smith, T. F. and Katcher, H. L. (1984). Algorithm for restriction map comparisons. *Nucleic Acids Res.* **12,** 237–242.