The Annals of Statistics
1994, Vol. 22, No. 3, 1603–1629

A POISSON APPROXIMATION FOR SEQUENCE COMPARISONS WITH INSERTIONS AND DELETIONS[1]

BY CLAUDIA NEUHAUSER

University of Southern California

We construct a statistical test for a sequence alignment problem which enables us to decide whether two given sequences are related. Such a test can be used in DNA and protein sequence comparisons. It is based on a comparison of two long sequences of i.i.d. letters taken from a finite alphabet. The test statistic typically employed is the length of the longest matching region between the two sequences in which a certain number of insertions and deletions but no mismatches are allowed. We give a distributional result which enables one to compute P-values, and hence to decide whether or not the two sequences are related. Its proof utilizes the Chen–Stein method for Poisson approximation. The test is based on a greedy algorithm that searches for the longest matching region. We show that this algorithm finds the longest matching region with probability approaching 1 as the lengths of the two sequences go to infinity.

1. Introduction. Sequence comparisons are important in a variety of fields, such as molecular biology, computer science, code and error control and in human speech research (see, e.g., [16]). Using the language of molecular biology, we illustrate the problem we wish to consider. In molecular biology, one compares DNA, RNA and protein sequences; here we focus our attention on DNA sequences. Such sequences can be represented as one-dimensional chains of letters taken from a finite alphabet consisting of the four letters A, C, G and T. These letters stand for the four bases adenine, cytosine, guanine and thymine, also referred to as the standard nucleotides. The nucleotides are linked together by a sugar–phosphate backbone, forming a one-dimensional chain. Their order determines the function of the sequence.

DNA sequences change over time. These changes, called mutations, are primarily caused by errors during chromosome replication (i.e., when DNA is copied). There are three main sources for mutations: substitutions, insertions and deletions. A substitution in a DNA sequence consists of replacing one nucleotide by another one. An insertion or deletion consists of inserting or deleting a nucleotide in the sequence; more than one letter can be inserted at a particular location in the sequence. There are several mechanisms that can cause these insertions and deletions: unequal crossing over between chromosomes, replication slipping during DNA replication, and DNA transposition (in which genetic material moves from one chromosomal location to another). Unequal

Received October 1992; revised March 1993.

[1] Supported in part by NSF Grants DMS-92-06139 and DMS-90-05833.

AMS 1991 *subject classifications.* Primary 62F05; secondary 92D20.

Key words and phrases. Chen–Stein method, sequence matching, Poisson approximation, DNA sequences, greedy algorithm.

crossing over and replication slipping typically cause only a few nucleotides to be either inserted or deleted, whereas a transposition can result in thousands of nucleotides being inserted or deleted at a particular location. (More on this can be found in, e.g., [12], [13] or [23].)

Currently, a large effort is under way in molecular biology to sequence DNA chains, that is, to find the order in which the letters are arranged. Once a sequence is known, one can investigate whether there are functional or evolutionary relationships with already known sequences. Closely related sequences typically show highly similar segments (contiguous subsequences) among them, that is, an alignment of these segments results in a high number of matching pairs. Such an alignment will typically also contain mismatches, insertions and deletions (insertions and deletions will be defined precisely later on). On the other hand, if two sequences possess unusually long highly similar segments compared to what one expects from completely unrelated sequences, then this might indicate a close relationship between the compared sequences. (Of course, the final decision on whether or not DNA sequences are related has to be made by a biologist since, due to chance, even completely unrelated sequences might once in a while give alignments of unusually long highly similar segments.) The decision on whether or not two sequences are similar can be formulated as a statistical test. In order to obtain P-values, one computes tail probabilities for the length of an alignment between two segments, one segment from each sequence. The null hypothesis will be that both sequences are unrelated, that is, independent.

Several aspects of this general problem have already been addressed (see, e.g., [3], [7], [8], [10] and [11]), with the most recent results appearing in [4]. (Reference [4] also includes a brief history on sequence comparison with references to earlier work.) The main result in [4] is a distributional Erdős–Rényi law for computing tail probabilities for the length of a matching region between two segments with a certain proportion of mismatches but no insertions and deletions. They posed the problem of finding a distributional result for segment comparisons when insertions and deletions are allowed. We address this problem in this article.

We study alignments between two i.i.d. sequences where insertions and deletions, but no mismatches, are allowed to occur. The main goal is to prove a distributional result that will allow us to compute tail probabilities for the length of such alignments. (We ultimately hope to prove a distributional result for an alignment between two sequences with mismatches *and* insertions and deletions.) We make the simplifying assumption that both sequences consist of independent letters. This is not true for DNA sequences; they can be described better by second-order Markov chains [20]. However, some justification for using i.i.d. sequences as an approximation when studying long matching regions can be found in [17]. In that paper, Monte Carlo simulations show that when comparing biological data and independent sequences with the same nucleotide frequencies, the distributions of the lengths of alignments are similar in both cases (see also [9]).

We consider the following model. Let μ and ν be distributions on a finite set E.

Consider the case where $\mathbf{A} = (A_1, A_2, \ldots, A_m)$, $A_i \in E$, and $\mathbf{B} = (B_1, B_2, \ldots, B_n)$, $B_i \in E$, are i.i.d. sequences which are independent of one another. The A_i's are distributed according to μ, that is, $P(A_i = l) = \mu_l$ for $l \in E$; the B_i's are distributed according to ν with $P(B_i = l) = \nu_l$ for $l \in E$. The probability of a *match* between a letter from $\mathbf{A}$ and a letter from $\mathbf{B}$ is denoted by p, that is, for all $1 \leq i \leq m$, $1 \leq j \leq n$,

$$p \equiv P(A_i = B_j) = \sum_{l \in E} P(A_i = B_j = l) = \sum_{l \in E} \mu_l \nu_l. \tag{1.1}$$

The last equality follows from the independence of the two sequences. Throughout the paper, we assume $0 < p < 1$. We wish to study local similarities between $\mathbf{A}$ and $\mathbf{B}$, that is, we are interested in finding *local alignments* between segments of $\mathbf{A}$ and $\mathbf{B}$, one from each sequence, with insertions or deletions (in short, *indels*) but no mismatches. We will call such local alignments of two sequences *matching regions*. A matching region therefore consists of contiguous subregions in which all pairs match and which are linked together by regions where insertions or deletions of one or more neighboring letters occur. These regions are called *links* or *indel regions*. (Note that links do not contain matching pairs.) For instance, the alignment

```
(1.2)   a g g t t c A C T T - -  G A A T C T T a T a C T T - - - C C A G
            a g A C T T a c  G A A T C T T - T - C T T a c t  C C A G  g t
```

consists of five matching subregions linked together by indels, the first link consisting of two indels, the next two links of one indel each and the last link of three indels. The total number of matching pairs is 19. (Letters in the matching region are capitalized.) We say that a matching region is of type $(t; k, l)$ if the total number of matching pairs is t, it consists of k indel regions (i.e., $k + 1$ matching subregions), and each link has at most l indels. Example (1.2) is therefore of type $(19; 4, l)$ for any $l \geq 3$. (When computing P-values, one chooses the smallest possible value for l.)

To visualize matching regions, one can represent matching pairs in a *dot matrix*. This is an $m \times n$ matrix where the entry $c_{ij} = 1$ if $A_i = B_j$ and equals 0 otherwise. (The 1's are the "dots" in the matrix.) A contiguous subregion in which all pairs match corresponds to a run of 1's along one of the diagonal lines $H_d = \{c_{ij}: j - i = d\}$ for some $d \in \{-m + 1, -m + 2, \ldots, n - 1\}$. Insertions or deletions of letters correspond to switching diagonal lines. Matching regions are then obtained by piecing together contiguous subregions.

In Section 4 we will give a detailed description of an algorithm that will search for matching regions; our results are based on this algorithm. An important feature of this algorithm is that it will only search for a particular type of matching region: since a matching region consists of contiguous subregions which are pieced together, there may be more than one way of piecing together two such subregions at a particular indel region. The algorithm will only search for those for which each of the subregions (except for the last one) is *maximal* in the sense that if (i, j) is the last site of a subregion (when represented in a dot matrix), then $(i+1, j+1)$ is a mismatch, that is, the algorithm switches diagonal lines only when it cannot continue along the same diagonal line. (A biological justification of this algorithm can be found at the end of Section 4.) In the following, we will restrict our attention to those matching regions that are found by the algorithm. We will see in Section 4 that, for m and n both large, most of the possible alignments the algorithm misses are not important, that is, the algorithm will find the longest alignment(s) with probability approaching 1 as $m, n \to \infty$. (For m and n both large, most of the missing alignments can be obtained by rearranging insertions and deletions in the indel regions. These missing alignments therefore do not provide us with really new alignments.) We wish to mention that in [15] an algorithm is described that searches for the longest common subsequence under the constraint that the number of insertions and deletions in an indel region is bounded by some fixed number. That algorithm is based on the Needleman and Wunsch algorithm and is different from our algorithm.

To state our results, we need a few definitions. Set $J = \{1, \dots, m\} \times \{1, \dots, n\}$. For $\alpha \in J$, let $\{\widehat{U}_1^\alpha, \widehat{U}_2^\alpha, \dots, \widehat{U}_{K(\alpha)}^\alpha\}$, $\widehat{U}_j^\alpha \subset J$, be the set of all possible matching regions of type $(t; k, l)$ starting at α. We wish to point out that we are only counting a subset of the matching regions, namely, those that can be found by our algorithm. Therefore, $\widehat{U}_j^\alpha$ is a matching region found by the algorithm if $c_\beta = 1$ for all $\beta \in \widehat{U}_j^\alpha$ (i.e., $|\widehat{U}_j^\alpha| = t$), the first k subregions are maximal in the sense described in the previous paragraph, and each indel region has at most l indels. Denote the event that $\widehat{U}_j^\alpha$ is a matching region found by the algorithm by U_j^α. (We suppress the dependence on t, k and l.) Note that if $\alpha = (i, j)$ is in the *interior* of J, that is, if $i \le m - (t + kl)$ and $j \le n - (t + kl)$, then $K(\alpha)$, the number of possible matching regions of type $(t; k, l)$, does not depend on α. In this case we denote $K(\alpha)$ simply by K. If α is not in the interior of J, then $K(\alpha)$ might be smaller than K since it might not be possible to realize all candidates for matching regions. (We might simply run out of sites in J before having obtained t matchings.) If for two matching regions $\widehat{U}_i^\alpha$ and $\widehat{U}_j^\beta$, $\widehat{U}_i^\alpha \cap \widehat{U}_j^\beta \ne \emptyset$, then the two matching regions are said to *share* matching pairs. We will now define an equivalence relation which partitions the set of matching regions found by the algorithm into equivalence classes: $\widehat{U}_i^\alpha$ and $\widehat{U}_j^\beta$ belong to the same *cluster* if we can find a sequence $\widehat{U}_{i_0}^{\alpha_0} = \widehat{U}_i^\alpha, \widehat{U}_{i_1}^{\alpha_1}, \dots, \widehat{U}_{i_s}^{\alpha_s} = \widehat{U}_j^\beta$ of matching regions of type $(t; k, l)$ found by the algorithm which satisfies $\widehat{U}_{i_0}^{\alpha_0} \cap \widehat{U}_{i_1}^{\alpha_1} \ne \emptyset, \dots, \widehat{U}_{i_{s-1}}^{\alpha_{s-1}} \cap \widehat{U}_{i_s}^{\alpha_s} \ne \emptyset$. (The definition of "cluster" used here is the same as the one used in [22].)

We now define the *starting point* of such a cluster. Let I_A denote the indicator function of the set A and $Y_\alpha = I_{\{U_1^\alpha \cup U_2^\alpha \cup \dots \cup U_{K(\alpha)}^\alpha\}}$; that is, $Y_\alpha = 1$ if there is at least one matching region of type $(t; k, l)$ starting at α that can be found by the algorithm. (Unless we say otherwise, from now on we simply say "matching region" when we really mean "matching region found by the algorithm.") For $1 \le i \le m$, $1 \le j \le n$, let $Z_{ij} = I_{\{A_i = B_j\}}$ and $V_{ij} = I_{\{Y_{ij} = 1 \text{ and } Z_{i-1, j-1} = 0\}}$. Let $\mathcal{A} = \{\alpha \in J\colon V_\alpha = 1\}$ (i.e., there is a matching region starting at α preceded by a mismatch). Partition $\mathcal{A}$ into disjoint sets $\mathcal{A}_1, \mathcal{A}_2, \dots$ so that $\alpha, \beta \in \mathcal{A}_k$ for some $k \geq 1$ if there is a matching region starting at α and a matching region starting at β which belong to the same cluster. We order the elements in $\mathcal{A}_k$, $k \geq 1$, as follows. Write $\|(i,j)\| = i + j$, $\pi_1(i,j) = i$ and $\pi_2(i,j) = j$. We say $\alpha \prec \beta$ if $\|\alpha\| < \|\beta\|$ or if $\|\alpha\| = \|\beta\|$ and $\pi_2(\alpha) < \pi_2(\beta)$. The smallest index in $\mathcal{A}_k$ according to the ordering "$\prec$" is designated as the starting point of the corresponding cluster. Also, let X_α be the indicator function of the event that α is the starting point of a cluster and let $W = \sum_{\alpha \in I} X_\alpha$ denote the number of clusters in J.

The following heuristics will describe the behavior we can expect in most situations that occur in practice. If t, the number of matching pairs in a matching region of type $(t; k, l)$ for fixed k and l, is sufficiently large (compared to m and n, the lengths of the two sequences), then there will be very few clusters in the dot matrix consisting of matching regions of that type. The clusters are of course not independent, but under the additional assumption (1.3) (see below) on the distributions μ and ν, we will be able to show that the dependency structure of the clusters is weak enough so that the total number of these sparse clusters will be approximately distributed according to a Poisson law. In this paper, a family of random variables $W(\iota)$, $\iota \in I$ (I some index set), will be called *approximately Poisson distributed with parameter* $\lambda(\iota)$ if for $Z(\iota)$ a Poisson random variable with mean $\lambda(\iota)$, $W(\iota) - Z(\iota) \to 0$ in the variation norm as $\iota \to \infty$. (For many more examples using Poisson approximation based on these heuristics, see [1].)

In order to formulate our results, we need the distributions μ and ν to satisfy the condition

$$\frac{\log c_h}{\log p} + \frac{\log c_v}{\log p} > 1, \tag{1.3}$$

where $c_h = P(B_2 = A_1 \mid A_1 = B_1)$ and $c_v = P(A_2 = B_1 \mid A_1 = B_1)$. Note that (1.3) automatically holds when $\mu = \nu$: in this case, $c_h = c_v = (1/p)\sum_{i \in E} \mu_i^3$ and (1.3) reduces to

$$\sum_{i \in E} \mu_i^3 < \left(\sum_{i \in E} \mu_i^2 \right)^{3/2}. \tag{1.4}$$

(This follows from the simple algebraic fact that, for $a_1, a_2, \dots, a_n > 0$, $q > 1$, then $(a_1 + a_2 + \cdots + a_n)^q > a_1^q + a_2^q + \cdots + a_n^q$.) Before we formulate the first theorem, we wish to explain (1.3). Given a matching pair at $\alpha = (i, j)$, the ratio $\log c_h / \log p$ (respectively, $\log c_v / \log p$) is a measure on how difficult it is to obtain a second matching by choosing a fresh letter for only B_2 (respectively, A_2). This second matching pair will then be on a horizontal (respectively, vertical) line through α in the dot matrix. The effect of condition (1.3) is to prevent the second moment of W from blowing up. As we will see in the next section, this is a crucial ingredient in the proof of the first theorem. We do not believe that (1.3) is necessary. (A reason for our belief is that the same condition is used when utilizing the Chen–Stein method to establish Poisson approximation for the case of matching two sequences without allowing mismatches or insertions and deletions. The author has shown in [14] for that case that Poisson approximation can be done outside of the regime in (1.3). This of course requires different techniques.)

The first theorem establishes the Poisson approximation for W. To do this, we have to scale m, n and t appropriately. We will be more specific on how to choose relative growth rates for m, n and t at the end of Section 3. Here, we only wish to mention that (1.3) allows us to choose relative growth rates for m and n so that $(\log n)/(\log nm) \to \rho$ as $m, n \to \infty$ for some

$$\rho \in \left(1 - \frac{\log c_v}{\log p}, \frac{\log c_h}{\log p}\right).$$

Then t can be scaled appropriately with m, n so that λ stays bounded away from 0 and ∞ [t is of order $\log(mn)$] and so that W is Poisson distributed in the limit as $m, n, t \to \infty$. (The rate of convergence of W in the total variation norm is faster than some negative power of mn.)

THEOREM 1. *Let* **A** *and* **B** *be two i.i.d. sequences, independent of each other, with distributions μ and ν, respectively. Let W denote the number of clusters of matching regions of type $(t; k, l)$ found by the algorithm and let $\lambda(m, n; t) = EW$. Set*

$$G(m, n; t) = mn(1 - p)\binom{t-1}{k} l^k p^{t-k} \left[\sum_{i \in E} \mu_i \nu_i (2 - \mu_i - \nu_i)\right]^k.$$

If (1.3) *holds, then, for fixed k and l, relative growth rates for m, n and t can be chosen so that W is approximately Poisson distributed with parameter $\lambda(m, n; t)$, where*

$$\frac{\lambda(m, n; t)}{G(m, n; t)} \to 1 \tag{1.5}$$

and $\lambda(m, n; t)$ stays bounded away from 0 *and* ∞ *as $m, n, t \to \infty$. The rate of convergence in* (1.5) *is faster than some negative power of t.*

The basic method employed in the proof of Theorem 1 is developed in [6]. This so-called Chen–Stein method enables us to establish the Poisson approximation in Theorem 1 since the first and second moments of W are well behaved. Bounds for these moments are somewhat routine and are proved along the lines of the corresponding proof in [4]. One also needs to estimate the parameter $\lambda(m, n; t)$. The upper bound for $\lambda(m, n; t)$ follows from a combinatorial argument and the first Bonferroni inequality. The lower bound for $\lambda(m, n; t)$ is considerably more complicated and is obtained by using the second Bonferroni inequality; this requires a detailed analysis of the geometry of the clusters. [A similar amount of work would be needed just to show that $G(m, n; t)$ gives the correct order of magnitude for $\lambda(m, n; t)$.]

Theorem 1 together with estimates stemming from the analysis of the algorithm and an elementary boundary estimate enables us to compute the P-values for the length of a local alignment between two sequences. This is the content of the next theorem. For this, let $S \equiv S(m,n;k,l)$ be the largest number of matching pairs in *any* matching region with k indel regions and at most l indels per indel region. We wish to emphasize that for S we consider *all* matching regions, not only the ones that can be found by our algorithm, whereas for W we consider only the ones found by the algorithm.

THEOREM 2. *Let* **A** *and* **B** *be two i.i.d. sequences, independent of each other, with distributions* μ *and* ν, *respectively. Let* S *be defined as above, and let* $\lambda = \lambda(m, n; t) = EW$ *be defined as in Theorem* 1. *If* (1.3) *holds, then there exist relative growth rates for* m, n *and* t, *and constants* $C, \gamma > 0$ *so that*

$$\left|P(S < t) - e^{-\lambda}\right| \le C\big(\log(mn)\big)^{-\gamma}.$$

Of course, the relative growth rates here will be chosen in the same way as in Theorem 1. The reader will see that the error estimate in Theorem 2 comes mainly from the error introduced by the algorithm, that is, it is basically the probability that the algorithm will fail to find the longest matching region. The proof of Theorem 2 consists of two parts. We will use the observation that if the length of the longest matching region found by the algorithm with k indel regions and at most l indels per indel region is less than t, then there are no clusters of matching regions of type $(t;k,l)$ (i.e., $W = 0$). This together with Theorem 1 and a boundary estimate constitutes one half of the proof. The analysis of the algorithm will provide us with estimates needed for the second half of the proof. These estimates will take care of the error introduced by comparing the length of the longest matching region found by the algorithm and the length of the longest matching region when we consider all matching regions.

REMARK. In both theorems we considered matching regions for which k, the number of indel regions, and l, the maximum number of letters allowed to be inserted, are fixed. If we are merely interested in showing that W can be approximated by a Poisson-distributed random variable, then, when l is fixed (a biologically reasonable assumption), we can let $k = o(t)$. However, in this case our estimates on $\lambda(m,n;t)$ are no longer good enough. If we insist on knowing the asymptotic behavior of $\lambda(m,n;t)$, then we can let k grow like $\log t$ (for fixed l) and (1.5) still holds. We will not show, this last fact since a growth rate of $\log t$ is too small to be useful.

The paper is organized as follows. The proofs of Theorems 1 and 2 are based on the Chen–Stein method, which is reviewed in Section 2(a). To employ this method, one establishes bounds on the first two moments of W. This is done in Section 2(b). Section 3 is devoted to the proof of Theorem 1 and to some of the estimates needed in the proof of Theorem 2. The main work here involves the study of the geometric structure of the clusters. The algorithm on which our results are based is described in detail in Section 4. In that section we will also show that for m and n both large, the algorithm finds the longest alignment(s) between sequences with probability close to 1. This allows us to finish the proof of Theorem 2, which is also done in Section 4.

2. Basic estimates. The proof of Theorem 1 is based on the Chen–Stein method. This method allows us to establish the Poisson approximation of W if we can show that the first two moments of W are well behaved. In this section we will first briefly explain the method, then give bounds on the moments of W and, finally, show that we can choose relative growth rates for m, n, and t so that λ will stay bounded away from zero and infinity.

(a) *The Chen–Stein method for Poisson approximation.* In this subsection, we will briefly review the Chen–Stein method for establishing Poisson approximation for dependent events. The method was developed by Chen [6] and is based on earlier work by Stein [18]. It is reviewed in [19]. It was generalized to a multivariable context in [2]. We follow [2] and [4] in our presentation of the setup.

Let J be an arbitrary index set. For $\alpha \in J$, let X_α be a random variable taking values in $\{0,1\}$ with $0 < P(X_\alpha = 1) < 1$. (Think of the X_α's as indicator functions for a certain event.) We wish to establish a Poisson approximation for the number of times this event occurs. For this, let

$$W = \sum_{\alpha \in J} X_\alpha \quad \text{and} \quad \lambda = EW, \tag{2.1}$$

where we assume that $\lambda \in (0,\infty)$. For each $\alpha \in J$, we choose a neighborhood set C_α such that the X_β's, $\beta \notin C_\alpha$, do not depend too strongly on X_α. The following three quantities are the key to establishing the Poisson approximation via the Chen-Stein method:

$$\begin{aligned} b_1 &\equiv \sum_{\alpha \in J} \sum_{\beta \in C_\alpha} EX_\alpha EX_\beta, \\ b_2 &\equiv \sum_{\alpha \in J} \sum_{\alpha \neq \beta \in C_\alpha} E(X_\alpha X_\beta), \\ b_3 &\equiv \sum_{\alpha \in J} E\big|E\big[X_\alpha - EX_\alpha \,|\, \sigma(X_\beta\colon \beta \in J - C_\alpha)\big]\big|, \end{aligned} \tag{2.2}$$

where $\sigma(X_\beta\colon \beta \in J - C_\alpha)$ is the σ-algebra generated by the X_β's which are outside of C_α. If b_1, b_2 and b_3 are all small, then W will be approximately Poisson distributed with parameter $\lambda = EW$. The quantity b_1 measures the neighborhood size. The influence of X_α on the occurrences of the X_β's for $\beta \in C_\alpha$ is measured by b_2, and b_3 quantifies the dependence between occurrences outside C_α and the event X_α. In many applications (such as ours), C_α can be chosen so that X_α is independent of X_β for $\beta \notin C_\alpha$, which implies that $b_3 \equiv 0$. Note that $b_2 - b_1 = EW^2 - \lambda - \lambda^2$ with $\lambda = EW$. Therefore, in cases where $b_3 = 0$, checking that the quantities in (2.2) are small is the same as showing that the first two moments of W are well behaved.

Let Z be a Poisson random variable with $EZ = EW = \lambda$. The approximation is phrased in terms of the total variation distance between the distributions of W, $\mathcal{L}(W)$, and Z, $\mathcal{L}(Z)$, which we denote by

$$\|\mathcal{L}(W) - \mathcal{L}(Z)\| = 2 \sup_{A \in \mathbf{N}_0} |P(W \in A) - P(Z \in A)|, \tag{2.3}$$

where $\mathbf{N}_0 = \{0, 1, 2, \dots\}$. The estimates of the Chen–Stein method are contained in the following lemma, which was presented in [4].

LEMMA 2.4. *Let W be the number of occurrences of dependent events, and let Z be a Poisson random variable with $EZ = EW = \lambda$. Then*

$$\|\mathcal{L}(W) - \mathcal{L}(Z)\| \leq 2(b_1 + b_2 + b_3)$$

and

$$|P(W = 0) - e^{-\lambda}| < \big(1 \wedge \lambda^{-1}\big)(b_1 + b_2 + b_3).$$

Thus, to establish the Poisson approximation, one has to show that the quantities b_1, b_2 and b_3 are small.

(b) *Bounds on b_1, b_2 and b_3.* As explained in the previous subsection, in order to employ the Chen–Stein method successfully, one has to establish good bounds on the quantities b_1, b_2 and b_3, defined in (2.2). This is the subject of this subsection.

Using the same notation as in the previous subsection, the index set $J = \{1, 2, \dots, m\} \times \{1, 2, \dots, n\}$ and X_α is the indicator function of the event that α is the starting point of a cluster of matching regions of type $(t; k, l)$ as defined in Section 1. We will next define the neighborhood set C_α. For $(i, j) \in J$, let

$$C_{(i,j)} = \big\{(i', j') \in J \colon |i - i'| < (t + kl) \text{ or } |j - j'| < (t + kl)\big\}; \tag{2.5}$$

that is, C_α consists of a horizontal and a vertical strip, each of width $2(t + kl) - 1$. We denote the horizontal strip by C^h_α and the vertical strip by C^v_α. Both are centered at α. Denote the intersection of the two strips by D_α, that is,

$$\begin{aligned} D_{(i,j)} &= C^h_{(i,j)} \cap C^v_{(i,j)} \\ &= \big\{(i', j') \in J \colon |i - i'| < (t + kl) \text{ and } |j - j'| < (t + kl)\big\}. \end{aligned} \tag{2.6}$$

Since matching regions of type $(t; k, l)$ which start outside of C_α will not have any letters of the sequences $\mathbf{A}$ or $\mathbf{B}$ in common with a matching region of type $(t; k, l)$ starting in α, the choice of C_α implies that X_α will be independent of X_β, for $\beta \notin C_\alpha$, and hence $b_3 \equiv 0$.

We start with estimating b_1. Clearly,

$$EX_\beta \leq \max_{\alpha \in J} EX_\alpha \quad \text{and} \quad |C_\alpha| \leq 2(t + kl)(m + n).$$

This together with $\lambda \equiv \sum_{\alpha \in J} EX_\alpha$ implies

$$b_1 = \sum_{\alpha \in J} \sum_{\beta \in C_\alpha} EX_\alpha EX_\beta \leq \Big(\max_{\alpha \in J} EX_\alpha \Big) |C_\alpha| \sum_{\alpha \in J} EX_\alpha \tag{2.7}$$
$$= 2\lambda(t + kl)(m + n)\Big(\max_{\alpha \in J} EX_\alpha \Big).$$

The final estimate on $\max_{\alpha \in J} EX_\alpha$ will have to wait until the end of Section 3.

The following lemma will be used to find an estimate for b_2. Recall that $p = P(A_i = B_j)$, $c_v = P(A_2 = B_1 | A_1 = B_1)$, $c_h = P(B_2 = A_1 | A_1 = B_1)$, and $K = \max_{\alpha \in J} K(\alpha)$.

LEMMA 2.8. *Let $\alpha \in J$. Then there exist $\theta_v \equiv (\log c_v)/(\log p) > 0$, $\theta_h \equiv (\log c_h)/(\log p) > 0$ and $\theta > 0$, such that, for t sufficiently large,*

$$E(X_\alpha X_\beta) \leq K^2 p^{t(1+\theta_v) - 8k}, \tag{2.9a}$$

for all $\beta \in C_\alpha^v - D_\alpha$, and

$$E(X_\alpha X_\beta) \leq K^2 p^{t(1+\theta_h) - 8k}, \tag{2.9b}$$

for all $\beta \in C_\alpha^h - D_\alpha$, and

$$E(X_\alpha X_\beta) \leq K^2 p^{t(1+\theta)}, \tag{2.9c}$$

for all $\beta \in D_\alpha$.

PROOF. First note that the event $\{X_\alpha = 1; X_\beta = 1\}$ is contained in the event $\{$there exists a pair (i, j) such that $\widehat{U}_i^\alpha \cap \widehat{U}_j^\beta = \emptyset$ and $U_i^\alpha \cap U_j^\beta$ occurs$\}$. (If $\widehat{U}_i^\alpha \cap \widehat{U}_j^\beta \neq \emptyset$, then the two matching regions would belong to the same cluster and therefore $X_\alpha X_\beta = 0$.) Therefore,

$$E(X_\alpha X_\beta) = P(X_\alpha = 1; X_\beta = 1)$$
$$\leq \sum_{\substack{i, j \\ \hat{U}_i^\alpha \cap \hat{U}_j^\beta = \emptyset}} P(U_i^\alpha \cap U_j^\beta). \tag{2.10}$$

To compute $P(U_i^\alpha \cap U_j^\beta)$, we will treat the two cases where $\beta \in D_\alpha$ and $\beta \in C_\alpha - D_\alpha$ separately. Suppose first that $\beta \in C_\alpha - D_\alpha$. (In this case, $\widehat{U}_i^\alpha \cap \widehat{U}_j^\beta = \emptyset$ holds automatically.) If β is in the vertical (respectively, horizontal) strip, we count the number of pairs of sites in the two matching regions which share letters from the **B** (respectively, **A**) sequence. Since we only need an upper bound on $P(U_i^\alpha \cap U_j^\beta)$, we can exclude the boundary sites of indel regions from our considerations, that is, we can exclude the first site after the last matching pair of each matching subregion (on the same diagonal) and the first site of each matching subregion. (We can include the first site of the first matching subregion.) There are $2k$ such sites. They may or may not share letters with sites from the other matching region. Hence there are at most $8k$ sites which we exclude. Suppose that, among the sites we do not exclude, there are $t - s$ matching pairs in each matching region which share letters. Then there are a total of at least $2(s - 4k)$ matching pairs in both matching regions which do not share any letters with any other matching pairs. The $2(s - 4k)$ matching pairs and the $t - s$ pairs of matching pairs are independent. With $c_v = P(A_2 = B_1 \mid A_1 = B_1)$ and $c_h = P(A_1 = B_2 \mid A_1 = B_1)$, it follows that $P(A_1 = A_2 = B_1) = c_v p$ and $P(A_1 = B_1 = B_2) = c_h p$. Since $\widehat{U}_i^\alpha$ and $\widehat{U}_j^\beta$ are two fixed candidates for a matching, it is easy to evaluate $P(U_i^\alpha \cap U_j^\beta)$: each of the $t - s$ pairs which share a letter, match with probability $c_v p$ (respectively, $c_h p$); each of the remaining pairs which we did not exclude match with probability p, irrespective of whether β is in the vertical (respectively, horizontal) strip. Therefore, if $\beta \in C_\alpha^v - D_\alpha$,

$$P(U_i^\alpha \cap U_j^\beta) \le p^{2(s-4k)}(c_v p)^{t-s}. \tag{2.11}$$

(If $\beta \in C_\alpha^h - D_\alpha$, replace c_v by c_h. We will only show the calculation for $\beta \in C_\alpha^v - D_\alpha$.) To estimate (2.11), note that

$$-\frac{1}{t}\log P(U_i^\alpha \cap U_j^\beta) \ge \log\frac{1}{c_v p} + \frac{2s}{t}\log\frac{1}{p} - \frac{s}{t}\log\frac{1}{c_v p} - \frac{8k}{t}\log\frac{1}{p}. \tag{2.12}$$

Note that $s/t \in [0, 1]$ and that

$$\log\frac{1}{c_v p} + \frac{s}{t}\left(2\log\frac{1}{p} - \log\frac{1}{c_v p}\right) \tag{2.13}$$

is linear in s/t, taking the value $\log(1/c_v p)$ for $s = 0$ and $2\log(1/p)$ for $s = t$. Since both values are positive, we can conclude that (2.13) is at least $\min(\log(1/c_v p), 2\log(1/p))$. Therefore, we can find a $\theta_v = ((\log c_v)/(\log p) \wedge 1) > 0$, which is independent of s so that

$$P(U_i^\alpha \cap U_j^\beta) \le \exp\Big\{\big(-t(1+\theta_v) + 8k\big)\log\big(1/p\big)\Big\} = p^{t(1+\theta_v)-8k}. \tag{2.14}$$

Since $c_v \ge p$, it follows that $(\log c_v)/(\log p) \le 1$ and hence $\theta_v \equiv (\log c_v)/(\log p)$. Substituting (2.14) back into (2.10) shows that

$$E(X_\alpha X_\beta) \le K(\alpha)K(\beta)p^{t(1+\theta_v)-8k} \le K^2 p^{t(1+\theta_v)-8k},$$

which is (2.9a) for $\beta \in C_\alpha^v - D_\alpha$ [(2.9b) follows from the same argument once c_v is replaced by c_h].

If $\beta \in D_\alpha$, we can use an argument which was developed in [5]. The same argument was also deployed in [4] for the corresponding estimate in their setting. (The situation here is somewhat more complicated than that in [4].) Let $\widehat{U}_i^\alpha$ and $\widehat{U}_j^\beta$ be the two candidates under consideration. When $\beta \in D_\alpha$, then the two matching regions might actually share sites, in which case they are contained in the same cluster which implies that $X_\alpha X_\beta = 0$. We can therefore assume that $\widehat{U}_i^\alpha \cap \widehat{U}_j^\beta \neq \emptyset$ as we did in (2.10). We will now describe the additional complication which did not arise in [4]: In addition to two matching pairs having just one letter (either from **A** or **B**) in common but otherwise being independent of the other sites in the two matching regions, there might be a sequence of sites $x_1, y_1, x_2, y_2, \dots$ contained in $\widehat{U}_i^\alpha \cup \widehat{U}_j^\beta$ (going back and forth between the two matching regions) in which x_1 and y_1 share a letter, y_1 and x_2 share a letter, x_2 and y_2 share a letter and so on. We can think of this sequence of sites as an *accordion*. (This notation was introduced in [4].) Whenever we refer to such sites, we will say that these sites are contained in an accordion. Furthermore, we require that accordions are *maximal* in the sense that an accordion always contains the maximal possible number of sites. The difference between our setting and [4] is that in their situation the accordions were independent, whereas here the accordions may be connected through sites from indel regions. Therefore, the accordions here are not necessarily independent. A little thought reveals that the only sites that can connect accordions are the mismatches that end matching subregions. Since each matching region of type $(t; k, l)$ has k indel regions, there are a total of $2k$ such sites that can connect accordions. We will call accordions that are connected in such a way *trees*. An accordion is said to be of length r if it consists of r sites (contained in the matching regions) going back and forth between the two matching regions. (An accordion of length 1 consists of just one matching pair which—because of our maximality condition—does not share any letters with any of the other sites contained in the two matching regions under consideration.) In [4], an energy argument was used to obtain the estimate. We will first illustrate the argument and show how to modify it in the following example. (We encourage the reader to draw a picture.) Let

$$\widehat{U}_i^\alpha = \{(1,1),(2,2),(3,3),(4,5),(5,6),(6,7),(7,8),(8,9),(9,10)\},$$

where the mismatch is at (4, 4), and let

$$\widehat{U}_j^\beta = \{(1,4),(2,5),(3,6),(4,7),(5,8),(9,9),(10,10),(11,11),(12,12)\},$$

where the mismatch is at (6, 9). Then both matching regions are of type (9; 1, 3). It contains two accordions of length 1 each, one accordion of length 5 and one tree. To reduce our situation to the one in [4], we use the following trick: by disregarding the two mismatches in the tree, the tree decomposes into three (now independent) accordions. Denote by $W(\alpha, \beta)$ the *energy* of the two matching regions, which is defined as

$$\begin{aligned} W(\alpha,\beta) = {} & \left(I_{\{A_{11}=B_{11}\}}\right) + \left(I_{\{A_{12}=B_{12}\}}\right) \\ & + \left(I_{\{B_3=A_3\}} + I_{\{A_3=B_6\}} + I_{\{B_6=A_5\}} + I_{\{A_5=B_8\}} + I_{\{B_8=A_7\}}\right) \\ & + \left(I_{\{B_1=A_1\}} + I_{\{A_1=B_4\}}\right) + \left(I_{\{B_4 \neq A_4\}}\right) \\ & + \left(I_{\{B_2=A_2\}} + I_{\{A_2=B_5\}} + I_{\{B_5=A_4\}} + I_{\{A_4=B_7\}} \right. \\ & \left. + I_{\{B_7=A_6\}}\right) + \left(I_{\{A_6 \neq B_9\}}\right) \\ & + \left(I_{\{A_8=B_9\}} + I_{\{B_9=A_9\}} + I_{\{A_9=B_{10}\}} + I_{\{B_{10}=A_{10}\}}\right). \end{aligned}$$

Terms enclosed in parentheses are referred to as *groups*. The first two groups correspond to the two accordions of length 1, respectively; the third group corresponds to the accordion of length 5; the fourth, sixth and eighth groups correspond to the three accordions the tree is made of; the fifth and seventh group are the two mismatches from the two indel regions. We denote the expression in the ith group by W_i and let $H(\alpha,\beta) = W(\alpha,\beta) - (I_{\{B_4 \neq A_4\}} + I_{\{A_6 \neq B_9\}})$. Then, for $\eta \in \mathbf{R}$, the *partition function* (or Laplace transform) of $W(\alpha,\beta)$ can be bounded by

$$\begin{aligned} E\exp\big[\eta W(\alpha,\beta)\big] &\le \exp(2\eta) E\exp\big[\eta H(\alpha,\beta)\big] \\ &\le \exp(2\eta) \prod_{\substack{i=1\\ i\neq 5,7}}^{8} E\exp(\eta W_i). \end{aligned} \tag{2.15}$$

Here we used that after "removing" the two mismatches the remaining accordions are independent. The right-hand side of (2.15) can now be estimated using Lemma 3 in [4]. Turning back to the general case where each matching region is of type $(t;k,l)$ and using the same notation as in the above example,

$$E\exp\big[\eta W(\alpha,\beta)\big] \le \exp(2\eta k) E\exp\big[\eta H(\alpha,\beta)\big].$$

We can now bound $P(U_i^\alpha \cap U_j^\beta)$, namely (as in [4]),

$$\exp(2t\eta) P\big(U_i^\alpha \cap U_j^\beta\big) \le E\exp\big[\eta W(\alpha,\beta)\big] \le \exp(2\eta k) E\exp\big[\eta H(\alpha,\beta)\big].$$

In Lemma 3 in [4] it was shown that, for some $\eta_0 \in \mathbf{R}$, there exists θ_0 with $2\theta_0 > \log(1/p)$ so that

$$\exp(-2t\eta) E\exp\big[\eta H(\alpha,\beta)\big] \le \exp(-2t\theta_0).$$

(Our θ_0 here is $I(1)$ in [4].) Therefore, there exists a $\theta > 0$ so that

$$P\big(U_i^\alpha \cap U_j^\beta\big) \le \exp\Big(-t\big[2\theta_0 - 2\eta_0 k/t\big]\Big) \le p^{t(1+\theta)},$$

for t sufficiently large, from which our claim (2.9c) follows. □

We can now use Lemma 2.8 to estimate b_2. With $K \equiv \max_{\alpha\in J} K(\alpha)$,

$$\begin{aligned} b_2 &\equiv \sum_{\alpha\in J} \sum_{\alpha\neq\beta\in C_\alpha} E(X_\alpha X_\beta) \\ &\le 2(t+kl)K^2 mn\big\{mp^{t(1+\theta_v)-8k} + np^{t(1+\theta_h)-8k} + 2(t+kl)p^{t(1+\theta)}\big\}. \end{aligned} \tag{2.16}$$

Therefore, combining (2.7) and (2.16), and recalling that $b_3 = 0$,

$$\begin{aligned} &b_1 + b_2 + b_3 \\ &\quad \le 2(t+kl)\Big\{K^2 mn\big[mp^{t(1+\theta_v)-8k} + np^{t(1+\theta_h)-8k} + 2(t+kl)p^{t(1+\theta)}\big] \\ &\qquad\qquad + \lambda(m+n)\Big(\max_{\alpha\in J} EX_\alpha\Big)\Big\}. \end{aligned} \tag{2.17}$$

3. Proof of Theorem 1. This section is devoted to proving Theorem 1. The main part is to establish upper and lower bounds on $\lambda \equiv EW = \sum_{\alpha \in J} EX_\alpha$, the expected value of the number of clusters of matching regions of type $(t; k, l)$ found by the algorithm when aligning two sequences of length m and n, respectively. This is the key part in establishing the Poisson approximation via the Chen–Stein method and requires a detailed analysis of the geometric structure of the clusters. The previous subsection showed that in order to employ the Chen–Stein method for Poisson approximation successfully, we need to find relative growth rates for m, n and t so that $\lambda \in (0, \infty)$ and at the same time the right-hand side of (2.17) goes to 0 as m, n and t (properly scaled) go to ∞. The following proposition establishes bounds on $\lambda = \lambda(m, n; t)$.

PROPOSITION 3.1. *There are constants $C_0, \gamma_0 > 0$ so that*

$$\frac{\big(m - 2(t + kl)\big)\big(n - 2(t + kl)\big)}{mn}(1 - C_0 t^{-\gamma_0}) \leq \frac{\lambda(m, n; t)}{G(m, n; t)} \leq 1,$$

where $G(m, n; t) = mn(1 - p)\binom{t-1}{k} l^k p^{t-k}[\sum_{i \in E} \mu_i \nu_i (2 - \mu_i - \nu_i)]^k$.

Recall that $Y_\alpha = I_{\{U_1^\alpha \cup \cdots \cup U_{K(\alpha)}^\alpha\}}$ and $V_{ij} = I_{\{Y_{ij} = 1 \text{ and } Z_{i-1, j-1} = 0\}}$. Let

$$J' = \big\{1 + (t + kl), \ldots, m - (t + kl)\big\} \times \big\{1 + (t + kl), \ldots, n - (t + kl)\big\}.$$

The proof of Proposition 3.1 is based on the following fact: if $\alpha \in J'$,

$$EX_\alpha = P(X_\alpha = 1) = \frac{P(X_\alpha = 1)}{P(V_\alpha = 1)} \frac{P(V_\alpha = 1)}{P(Y_\alpha = 1)} P(Y_\alpha = 1). \tag{3.2}$$

We will estimate the three factors on the right-hand side of (3.2) separately, starting with $P(Y_\alpha = 1)$. We will see that the asymptotic behavior of EX_α is mainly determined by the asymptotic behavior of EY_α. Since

$$EY_\alpha = P(Y_\alpha = 1) = P\big(U_1^\alpha \cup U_2^\alpha \cup \cdots \cup U_{K(\alpha)}^\alpha\big), \tag{3.3}$$

we can use the first two Bonferroni inequalities to estimate (3.3):

$$\sum_{j=1}^{K(\alpha)} P(U_j^\alpha) - \sum_{i<j} P(U_i^\alpha \cap U_j^\alpha) \leq EY_\alpha \leq \sum_{j=1}^{K(\alpha)} P(U_j^\alpha). \tag{3.4}$$

Note that, for $\alpha \in J'$, $P(U_j^\alpha)$ does not depend on α. Furthermore, for $\alpha \in J'$, $K \equiv K(\alpha) = \max_{\beta \in J} K(\beta)$, that is, $K(\alpha)$ takes on its maximum for $\alpha \in J'$ and is constant there. We will first estimate the upper bound in (3.4) for $\alpha \in J'$. This is the content of the next lemma.

LEMMA 3.5. *If $\alpha \in J'$, then*

$$\sum_{j=1}^{K} P(U_j^\alpha) = \binom{t-1}{k} l^k p^{t-k} \left[\sum_{i \in E} \mu_i \nu_i (2 - \nu_i - \mu_i)\right]^k.$$

PROOF. We wish to point out that we only count those matching regions that can be found by our algorithm. We begin with an exact computation of K. Since a matching region of type $(t; k, l)$ consists of k indel regions, there are $\binom{t-1}{k}$ ways of choosing sites which will be the starting points of the indel regions. For each indel region, we have to decide whether we insert letters into the **A** or **B** sequence. Suppose m of the indel regions insert letters into the **A** sequence. This can be done in $\binom{k}{m}$ ways. Once it is specified for each indel region where to insert letters, there are l sites (for each indel region) one can choose from which can serve as starting points of the next subregion. There are a total of t matching pairs, k of which are starting points of subregions other than the first subregion. These starting points have to be treated separately. Recall that our algorithm is such that an endpoint (i, j) has the property that $(i+1, j+1)$ is not a matching pair. This means that $(i+1, j+1)$ and the starting point of the subsequent matching subregion share a letter of one of the two sequences. Therefore,

$$\begin{aligned}\sum_{j=1}^{K} P(U_j^\alpha) &= \sum_{m=0}^{k} \binom{t-1}{k}\binom{k}{m} l^k p^{t-k}\left(\sum_{i\in E}\mu_i\nu_i(1-\nu_i)\right)^{k-m}\left(\sum_{i\in E}\mu_i\nu_i(1-\mu_i)\right)^{m}\\ &= \binom{t-1}{k} l^k p^{t-k}\left[\sum_{i\in E}\mu_i\nu_i(2-\nu_i-\mu_i)\right]^k. \qquad \square\end{aligned}$$

It follows from the proof of Lemma 3.5 that if $\alpha \notin J'$, then

$$\sum_{j=1}^{K(\alpha)} P(U_j^\alpha) \le \binom{t-1}{k} l^k p^{t-k}\left[\sum_{i\in E}\mu_i\nu_i(2-\nu_i-\mu_i)\right]^k, \tag{3.6}$$

since in this case we have fewer terms to sum over. (It may be that some or all of the possible candidates of matching regions cannot be realized since we simply run out of letters.)

For the lower bound in (3.4) we need an upper bound on $\sum_{i<j} P(U_i^\alpha \cap U_j^\alpha)$. This is more complicated and requires a detailed analysis of the geometric structure of matching regions that start at the same point. As we will see, we only need to do this for $\alpha \in J'$.

LEMMA 3.7. *If $\alpha \in J'$, then there are constants $C_1, \gamma_1 > 0$ so that*

$$\sum_{i<j} P(U_i^\alpha \cap U_j^\alpha) \le C_1 t^{-\gamma_1} \sum_{j=1}^{K} P(U_j^\alpha).$$

PROOF. Fix $\widehat{U}_i^\alpha$ and $\widehat{U}_j^\alpha$. Since both sets $\widehat{U}_i^\alpha$ and $\widehat{U}_j^\alpha$ have the same starting point but $i \ne j$, it follows that $\widehat{U}_i^\alpha \triangle \widehat{U}_j^\alpha \ne \emptyset$ where "$\triangle$" denotes the symmetric difference between the two sets. Denote by $\Delta_{ij} = \widehat{U}_i^\alpha \triangle \widehat{U}_j^\alpha$. Suppose $|\widehat{U}_i^\alpha \cap \Delta_{ij}|$

$= r_i$ and $|\widehat{U}_j^\alpha \cap \Delta_{ij}| = r_j$. Since $|\widehat{U}_i^\alpha \cap \Delta_{ij}^c| = t - r_i$, $|\widehat{U}_j^\alpha \cap \Delta_{ij}^c| = t - r_j$ and $|\widehat{U}_i^\alpha \cap \Delta_{ij}^c| = |\widehat{U}_j^\alpha \cap \Delta_{ij}^c|$, it follows that $r_i = r_j \equiv r$.

The following observations are important: $\widehat{U}_i^\alpha$ and $\widehat{U}_j^\alpha$ have the same starting point, therefore, they share at least the first matching subregion. The two matching regions can only get "separated" (i.e., stop sharing sites) at an indel region since we only consider matching regions that can be found by the algorithm. Likewise, once they are separated, they can only "recombine" (i.e., start sharing sites again) at locations at which at least one of the two matching regions has an indel region. Furthermore, since $i \neq j$, there is at least one subregion in each of the two matching regions which is not shared. Therefore, each of the matching regions must have at least one indel region whose distance to an adjacent indel region is at most r. We will split the estimate into two parts, one for $r > t^{\gamma_2}$, the other for $r \leq t^{\gamma_2}$ for some $0 < \gamma_2 < 1/2$.

(i) Let $r > t^{\gamma_2}$ and let $\theta > 0$ as in Lemma 2.8. Then the same argument as in the proof of (2.9c) yields

$$P(U_i^\alpha \cap U_j^\alpha) \leq C(k,l)\left(p^{1+\theta}\right)^{t^{\gamma_2}} p^{t - t^{\gamma_2}}, \tag{3.8}$$

where $C(k,l)$ depends only on k and l and takes care of the indel regions. Here, we do not have to be as careful as in the proof of Lemma 2.8. We can simply combine all the terms that take care of the indel regions into the quantity $C(k,l)$. The value of $C(k,l)$ may change from line to line. We denote by $\sum'$ the sum over all those pairs (i,j) where $r > t^{\gamma_2}$. From Lemma 3.5 we can conclude that $\sum_i P(U_i^\alpha) \geq C(k,l)Kp^t$. This together with (3.8) shows that

$$\begin{aligned} \frac{\sum_{i<j}' P(U_i^\alpha \cap U_j^\alpha)}{\sum_i P(U_i^\alpha)} &\leq C(k,l)\frac{K^2\left(p^{1+\theta}\right)^{t^{\gamma_2}} p^{t - t^{\gamma_2}}}{Kp^t} \\ &\leq C(k,l)t^k\left(\frac{p^{1+\theta}}{p}\right)^{t^{\gamma_2}} = C(k,l)t^k p^{\theta t^{\gamma_2}}. \end{aligned} \tag{3.9}$$

Here we used that $K \leq C(k,l)t^k$.

(ii) Let $r \leq t^{\gamma_2}$. Both matching regions have k indel regions. They share parts or all of some of the matching subregions as mentioned earlier. Since matching regions that are found by the algorithm can only get separated at indel regions, they have also some of the indel regions in common, that is, some of the mismatches are contained in both matching regions. Suppose they have $k - m$ of the indel regions in common, where $1 \leq m \leq k - 1$. We first choose the locations of those $k - m$ indel regions. We only have to do this for one of the matching regions since each indel region has a corresponding indel region at the same location in the other matching region. There are $\binom{t-1}{k-m}$ ways of choosing these locations. The locations of each of the remaining $2m$ indel regions (m for each matching region) have to be chosen within t^{γ_2} of one of the other indel regions (otherwise, $r > t^{\gamma_2}$), that is, each of them has to be within mt^{γ_2} of one of the previously chosen $k - m$ indel regions. Since $cp \leq p$, it suffices to compute the probabilities for matching pairs for only one of the two matching regions. We denote by $\sum''$ the sum over all those pairs (i,j) where $r \le t^{\gamma_2}$. Then

$$\begin{aligned}\sum_{i<j}{}'' P(U_i^\alpha \cap U_j^\alpha) &\le \sum_{m=1}^{k-1} C(k,l)\binom{t-1}{k-m}(k-m)^{2m}(2mt^{\gamma_2})^{2m}p^{t-k}\\ &\le C(k,l)p^{t-k}\sum_{m=1}^{k-1} k^{2m}t^{k-m}(2m)^{2m}t^{2\gamma_2 m}\\ &\le C(k,l)p^{t-k}t^k\sum_{m=1}^{k-1}\left(2k^2t^{\gamma_2-1/2}\right)^{2m}.\end{aligned}$$

If $0 < \gamma_2 < 1/2$, then, for t large, $0 < 2k^2t^{\gamma_2-1/2} < 1$, and the sum can be bounded by $4k^4t^{2\gamma_2-1}/(1-4k^4t^{2\gamma_2-1})$. Therefore,

$$\begin{aligned}\frac{\sum''_{i<j}P(U_i^\alpha\cap U_j^\alpha)}{\sum_j P(U_j^\alpha)} &\le C(k,l)\frac{p^{t-k}t^k4k^4t^{2\gamma_2-1}}{t^kp^{t-k}}\frac{1}{1-4k^4t^{2\gamma_2-1}}\\ &\le C(k,l)\frac{4k^4t^{2\gamma_2-1}}{1-4k^4t^{2\gamma_2-1}}.\end{aligned}\tag{3.10}$$

Combining (3.9) and (3.10), we see that for $\gamma_2 \in (0,1/2)$,

$$\begin{aligned}\sum_{i<j}P(U_i^\alpha\cap U_j^\alpha) &\le C(k,l)\left[t^kp^{\theta t^{\gamma_2}} + \frac{4k^4t^{2\gamma_2-1}}{1-4k^4t^{2\gamma_2-1}}\right]\sum_j P(U_j^\alpha)\\ &\le C_1t^{-\gamma_1}\sum_j P(U_j^\alpha),\end{aligned}$$

for some $0 < C_1, \gamma_1 < \infty$. □

Combining Lemmas 3.5 and 3.7 allows us to compute the bounds in (3.4) for $\alpha \in J'$ and therefore give bounds on the third factor on the right-hand side of (3.2).

We will now estimate the first and the second factor on the right-hand side of (3.2) in the following two lemmas.

LEMMA 3.11. *Let $\alpha \in J'$. Then*

$$\frac{P(V_\alpha = 1)}{P(Y_\alpha = 1)} = 1 - p.$$

PROOF. Let $\alpha = (i,j) \in J'$. Since $\{V_\alpha = 1\} \subset \{Y_\alpha = 1\}$,

$$\begin{aligned}\frac{P(V_\alpha = 1)}{P(Y_\alpha = 1)} &= P\big(V_\alpha = 1 \mid Y_\alpha = 1\big)\\ &= P\big(\text{mismatch at } (i-1, j-1)\big) = 1-p.\end{aligned}$$ □

LEMMA 3.12. *Let $\alpha \in J'$. Then there are constants $C_3, \gamma_3 > 0$, so that*

$$1 - C_3 t^{-\gamma_3} \le \frac{P(X_\alpha = 1)}{P(V_\alpha = 1)} \le 1.$$

PROOF. Note that since $\{X_\alpha = 1\} \subset \{V_\alpha = 1\}$, the upper bound is immediate. For the lower bound, we write

$$\frac{P(X_\alpha = 1)}{P(V_\alpha = 1)} = P\big(X_\alpha = 1 \,|\, V_\alpha = 1\big) = 1 - P\big(X_\alpha = 0 \,|\, V_\alpha = 1\big). \tag{3.13}$$

We will give an upper bound on $P(X_\alpha = 0 \,|\, V_\alpha = 1)$. On $\{V_\alpha = 1\}$, $\{X_\alpha = 0\}$ occurs if and only if there is a $\beta \prec \alpha$ with $Y_\beta = 1$ such that the matching regions starting in α and β are in the same cluster. (The relation "$\prec$" was defined in the Introduction.) We will first consider the case in which β and α are such that, for $\gamma_4 \in (0, 1/2)$, the first t^{γ_4} sites in the first matching subregion of the matching region starting at β do not share any letters with sites in the matching region starting at α. Denote this event by F_1. Denote the matching region starting at α by $\widehat{U}_i^\alpha$ and the one starting at β by $\widehat{U}_j^\beta$. Then

$$P\big(U_i^\alpha \cap U_j^\beta \cap F_1\big) \le C(k,l) p^{t + t^{\gamma_4}}. \tag{3.14}$$

The constant $C(k,l)$ takes care of the indel regions and boundary effects and may change from line to line. To compute $P(X_\alpha = 0;\, V_\alpha = 1)$, note that there are at most K^2 pairs of candidates ($\widehat{U}_i^\alpha$ and $\widehat{U}_j^\beta$). Therefore,

$$P(X_\alpha = 0;\, V_\alpha = 1) \le K^2 P\big(U_i^\alpha \cap U_j^\beta\big). \tag{3.15}$$

To estimate $P(V_\alpha = 1)$, we make use of Lemmas 3.5 and 3.7. Since by Lemma 3.11, $P(V_\alpha = 1) = (1 - p)P(Y_\alpha = 1)$, it follows that

$$P(V_\alpha = 1) \ge (1 - p)\big(1 - C_1 t^{-\gamma_1}\big) \sum_{j=1}^{K} P\big(U_j^\alpha\big). \tag{3.16}$$

From the proof of Lemma 3.5 we can conclude that

$$P\big(U_j^\alpha\big) \ge C(k,l)p^t \quad \text{and} \quad K \le C(k,l)t^k. \tag{3.17}$$

Therefore, combining (3.14)–(3.17),

$$\begin{aligned} \frac{P(X_\alpha = 0;\, V_\alpha = 1;\, F_1)}{P(V_\alpha = 1)} &\le C(k,l)\frac{K^2 p^{t + t^{\gamma_4}}}{(1 - p)(1 - C_1 t^{-\gamma_1})K p^t} \\ &= \frac{C(k,l)t^k}{(1 - p)(1 - C_1 t^{-\gamma_1})} p^{t^{\gamma_4}}. \end{aligned} \tag{3.18}$$

We will now investigate the case in which α and β are such that less than t^{γ_4} sites in the first matching subregion of the matching region starting at β

do not share any letters with sites in the matching region starting at α. This event is F_1^c. In this case, almost all sites in the two matching regions either will be contained in both matching regions or will share a letter with a site from the other matching region. We will use a similar argument as in the proof of Lemma 3.7. Let $F_2 = \{|\widehat{U}_i^\alpha \triangle \widehat{U}_j^\beta| - 2t^{\gamma_4} > t^{\gamma_5}\}$, for some $\gamma_5 \in (0, 1/2)$, where "$\triangle$" denotes the symmetric difference. That is, on $F_1^c \cap F_2$, there are at least t^{γ_5} sites in each matching region that are not in $\widehat{U}_i^\alpha \cap \widehat{U}_j^\beta$ but share letters. Then, as before in Lemma 2.8 or Lemma 3.7, there exists a $\theta > 0$ so that

$$\begin{aligned} P\big(U_i^\alpha \cap U_j^\beta \cap F_1^c \cap F_2\big) &\le C(k,l)p^{t - t^{\gamma_5}}\big(p^{1+\theta}\big)^{t^{\gamma_5}} \\ &= C(k,l)p^t p^{\theta t^{\gamma_5}}. \end{aligned} \tag{3.19}$$

Using (3.15)–(3.17) and (3.19) shows that

$$\begin{aligned} \frac{P\big(X_\alpha = 0;\ V_\alpha = 1;\ F_1^c \cap F_2\big)}{P(V_\alpha = 1)} &\le C(k,l)\frac{K^2 p^t p^{\theta t^{\gamma_5}}}{(1-p)(1 - C_1 t^{-\gamma_1})K p^t} \\ &= \frac{C(k,l)t^k}{(1-p)(1-C_1 t^{-\gamma_1})^l} p^{\theta t^{\gamma_5}}. \end{aligned} \tag{3.20}$$

On $F_1^c \cap F_2^c$, the two matching regions have to share some matching subregions. Suppose they share m such subregions. Then there are $\binom{t-1}{m}$ ways of choosing the locations of the corresponding indel regions. Since they do not necessarily share parts of the first matching subregion, there are $t^{2\gamma_5}$ ways of choosing the two locations of the first two indel regions. The remaining $k - m - 1$ indel regions are within distance (mt^{γ_5}) of the other indel regions. Then,

$$\begin{aligned} &P\big(X_\alpha = 0;\ V_\alpha = 1;\ F_1^c \cap F_2^c\big) \\ &\quad\le C(k,l)\sum_{m=1}^{k-1}\binom{t-1}{m} t^{2\gamma_5} k^{k-m-1}(2mt^{\gamma_5})^{2(k-m-1)} p^t \\ &\quad\le C(k,l)p^t t^{2k\gamma_5}\sum_{m=1}^{k-1}\big(t^{1-2\gamma_5}\big)^m \\ &\quad\le C(k,l)p^t t^{2k\gamma_5}\frac{t^{(1-2\gamma_5)k}}{t^{1-2\gamma_5} - 1} \\ &\quad= C(k,l)p^t \frac{t^k}{t^{1-2\gamma_5} - 1}. \end{aligned} \tag{3.21}$$

Here, we made use of the fact that $\gamma_5 \in (0, 1/2)$. Therefore, using (3.17) again,

$$\begin{aligned} &\frac{P\big(X_\alpha = 0;\ V_\alpha = 1;\ F_1^c \cap F_2^c\big)}{P(V_\alpha = 1)} \\ &\qquad\le C(k,l)\frac{p^t t^k}{(1-p)(1-C_1 t^{-\gamma_1})(t^{1-2\gamma_5} - 1)p^t t^k}. \end{aligned} \tag{3.22}$$

Combining (3.18), (3.20) and (3.22) shows that there are constants $C_3, \gamma_3 > 0$ so that

$$P\big(X_\alpha = 0 \mid V_\alpha = 1\big) \le C_3 t^{-\gamma_3}, \tag{3.23}$$

and Lemma 3.12 follows. □

We can now proceed with the proof of Proposition 3.1.

Proof of Proposition 3.1. It follows from (3.4) and Lemma 3.7 that if $\alpha \in J'$,

$$\big(1 - C_1 t^{-\gamma_1}\big) \sum_{j=1}^{K} P\big(U_j^\alpha\big) \le EY_\alpha \le \sum_{j=1}^{K} P\big(U_j^\alpha\big)$$

and hence, together with Lemmas 3.11 and 3.12, (3.2) can be bounded by

$$\begin{aligned} &(1-p)\big(1 - C_1 t^{-\gamma_1}\big)\big(1 - C_3 t^{-\gamma_3}\big) \sum_{j=1}^{K} P\big(U_j^\alpha\big) \\ &\quad \le EX_\alpha \le (1-p) \sum_{j=1}^{K} P\big(U_j^\alpha\big) \end{aligned} \tag{3.24}$$

as long as $\alpha \in J'$. If $\alpha \notin J'$, then $EX_\alpha \le (1-p)P(U_1^\alpha \cup \cdots \cup U_{K(\alpha)}^\alpha)$, which can be bounded by (3.6). Note that $|J| = mn$ and $|J'| = (m - 2(t + kl))(n - 2(t + kl))$. These two observations, together with (3.24) and $\Sigma_{\beta \in J'} EX_\beta \le \lambda \equiv \Sigma_{\beta \in J} EX_\beta$, show that

$$\begin{aligned} \lambda(m,n;t) &= \sum_{\beta \in J} EX_\beta \\ &\le (1-p)mn \binom{t-1}{k} l^k p^{t-k} \left[\sum_{i \in E} \mu_i \nu_i (2 - \mu_i - \nu_i)\right]^k \end{aligned} \tag{3.25}$$

and

$$\begin{aligned} \lambda(m,n;t) &\ge \sum_{\beta \in J'} EX_\beta \ge \sum_{\beta \in J'} (1-p)\big(1 - C_0 t^{-\gamma_0}\big) \sum_{j=1}^{K} P\big(U_j^\beta\big) \\ &= \big(m - 2(t+kl)\big)\big(n - 2(t+kl)\big)(1-p)\big(1 - C_0 t^{-\gamma_0}\big) \\ &\quad \times \binom{t-1}{k} l^k p^{t-k} \left[\sum_{i \in E} \mu_i \nu_i (2 - \mu_i - \nu_i)\right]^k, \end{aligned} \tag{3.26}$$

for arbitrary constants $C_0, \gamma_0 > 0$. Combining (3.25) and (3.26) proves Proposition 3.1. □

We are now ready to finish the proof of Theorem 1. First note that, by using Lemma 3.5, (3.6) and the observation following (3.24), we have

$$\max_{\alpha \in J} EX_\alpha \le (1-p)\binom{t-1}{k} l^k p^{t-k} \left[\sum_{i \in E} \mu_i \nu_i (2 - \mu_i - \nu_i)\right]^k .$$

Therefore, (2.17) can be bounded by

$$\begin{aligned} b_1 + b_2 + b_3 \le 2(t+kl)\Big\{ & K^2 mn \left[mp^{t(1+\theta_v)-8k} + np^{t(1+\theta_h)-8k} + 2(t+kl)p^{t(1+\theta)}\right] \\ & + \lambda(m+n)(1-p)\binom{t-1}{k} l^k p^{t-k} \\ & \times \left[\sum_{i \in E} \mu_i \nu_i (2 - \mu_i - \nu_i)\right]^k \Big\}. \end{aligned} \tag{3.27}$$

To finish the proof of Theorem 1, we still have to show that we can choose relative growth rates of m, n and t so that the right-hand side of (3.27) goes to 0 and at the same time λ stays bounded away from 0 and ∞. This is similar to what was done in [4]. We consider the case where m, $n \to \infty$ with

$$\frac{\log n}{\log(mn)} \to \rho > 0 \quad \text{and} \quad \frac{\log m}{\log(mn)} \to 1 - \rho > 0, \tag{3.28}$$

that is, $n \sim (mn)^\rho$ and $m \sim (mn)^{1-\rho}$. (The symbol $\sim$ means that the logarithms of the two quantities are asymptotic.) From Proposition 3.1 we can conclude that

$$\lambda \sim mnt^k p^t \tag{3.29}$$

and the estimate in (3.27) shows that

$$b_1 + b_2 + b_3 \sim mnt^{2k+1}\left[mp^{(1+\theta_v)t} + np^{(1+\theta_h)t}\right]. \tag{3.30}$$

If we choose m, n and t so that $\lambda \in (0, \infty)$, then

$$\begin{aligned} b_1 + b_2 + b_3 &\sim t^{1+k}\left[mp^{\theta_v t} + np^{\theta_h t}\right] \sim t^{1+k}\left[(mn)^{1-\rho} p^{\theta_v t} + (mn)^\rho p^{\theta_h t}\right] \\ &= t^{1+k}\left[(\lambda t^{-k})^{\theta_v}(mn)^{1-\rho-\theta_v} + (\lambda t^{-k})^{\theta_h}(mn)^{\rho-\theta_h}\right]. \end{aligned}$$

To ensure that this goes to 0 as m, $n \to \infty$, we therefore need that $1 - \rho - \theta_v < 0$ and $\rho - \theta_h < 0$. Therefore, if

$$\theta_h + \theta_v = \frac{\log c_h}{\log p} + \frac{\log c_v}{\log p} > 1, \tag{3.31}$$

that is, $\theta_h > 1 - \theta_v$, we can find $\rho \in (1 - \theta_v, \theta_h)$ and therefore relative growth rates for m and n so that $\lambda \in (0, \infty)$ and $b_1 + b_2 + b_3 \to 0$ as m, n (and thus t)

tend to 0 faster than some negative power of mn. This is exactly condition (1.3). (It is not surprising that this is the same condition as in [4].)

We will now explain why the claim in the remark at the end of Section 1 is true. Let

$$\begin{aligned}
\Lambda^r_{ij} &= \{i-r,\dots,i-1\}\times\{j-r,\dots,j-1\},\\
M^s_{ij} &= \big\{\text{the matching region starting at } (i,j) \text{ shares less than } s \text{ letters with}\\
&\qquad \text{any other matching region}\big\},\\
N^s_{ij} &= \big\{(i-1,j-1) \text{ is a mismatch; } \Lambda^r_{ij} \text{ contains no matches; } U^{ij}_1 \text{ occurs}\big\}.
\end{aligned}$$

We claim that when $k \ll s \ll r \ll t$, then $N^s_{ij} \cap M^r_{ij} \subset \{X_{ij} = 1\}$. Clearly, if $\{(i-1,j-1)$ is a mismatch; U^{ij}_1 occurs$\}$, then $\{V_{ij} = 1\}$. We choose k, s and r so that any matching region with index smaller than (i,j) that belongs to the same cluster as the matching region starting at (i,j) cannot intersect any matching region starting at (i,j). If there was a matching region with a smaller index than (i,j) that belonged to the same cluster as the matching region starting at (i,j), the matching region starting at (i,j) would have to share more than s letters with other matching regions. Therefore, for $k \ll s \ll r \ll t$ and l fixed, there are positive constants $C(l)$, c_1 and c_2 with $c_2 < 1$ so that

$$\begin{aligned}
P(X_{ij} = 1) &\geq P\big(N^r_{ij} \cap M^s_{ij}\big) \geq P\big(N^r_{ij}\big) - P\left((M^s_{ij})^c\right)\\
&\geq (1-p)\max_{a\in E}\left[\mu_a(1-\nu_a)\right]^r p^{t-k}\left[\sum_{a\in E}(2-\mu_a-\nu_a)\right]^k - Cp^{t-s}p^{(1-\theta)s}\\
&\geq C(l)p^{t+c_1t^{c_2}},
\end{aligned}$$

and hence

$$\lambda(m,n;t) \geq \big(m-2(t+kl)\big)\big(n-2(t+kl)\big)(1-p)C(l)p^{t+c_1t^{c_2}}.$$

Using this lower bound it is easy to see that under (1.3) for l fixed and $k = o(t)$, $b_1 + b_2 + b_3 \to 0$; m and n are scaled as before.

4. Proof of Theorem 2 and the algorithm. In this section we will prove a distributional result that will enable us to compute the P-values mentioned in the Introduction. Let $S_a \equiv S_a(m,n;k,l)$ be the largest number of matching pairs in any matching region between two sequences of lengths m and n found by the algorithm (hence the subscript "a"), where the matching region has k indel regions and each indel region has at most l indels. The subregions do not contain any mismatches. If W denotes the number of clusters of matching regions of type $(t;k,l)$, then $\{W \neq 0\} \subset \{S_a \geq t\}$. The two sets are not quite the same because of boundary effects. These boundary effects occur if $Y_\alpha = 1$ for some $\alpha = (i,j)$ with $i \leq t+kl$ or $j \leq t+kl$ but $V_\alpha = 0$ and hence $X_\alpha = 0$ and if there is no other $\beta \in J$ so that $X_\beta = 1$ and the matching region starting in α is contained in the cluster starting in β. Therefore,

$$\begin{aligned} |P(W=0) - P(S_a < t)| &= |P(W=0) - P(Y_\alpha = 0 \text{ for all } \alpha \in J)| \\ &\le \sum_{\alpha \in \partial J} P(Y_\alpha = 1), \end{aligned} \tag{4.1}$$

where $\partial J = \{(i,j)\colon i \le t + kl \text{ or } j \le t + kl\}$. This together with Lemma 2.4 can be used to show

$$\begin{aligned} |P(S_a < t) - e^{\lambda}| &\le |P(S_a < t) - P(W=0)| + |P(W=0) - e^{\lambda}| \\ &\le \sum_{\alpha \in \partial J} P(Y_\alpha = 1) + (b_1 + b_2 + b_3) \\ &\le (m+n)(t+kl)P(Y_\alpha = 1) + (b_1 + b_2 + b_3). \end{aligned} \tag{4.2}$$

Combining (3.3), (3.4) and Lemma 3.5, the right-hand side of (4.2) is less than or equal to

$$(m+n)(t+kl)\binom{t-1}{k} l^k p^{t-k} \left[\sum_{i \in E} \mu_i \nu_i (2 - \nu_i - \mu_i)\right]^k + (b_1 + b_2 + b_3). \tag{4.3}$$

It follows from Proposition 3.1 that the first term in (4.3) can be bounded by

$$\frac{\lambda(m+n)(t+kl)}{(1 - C_0 t^{-\gamma_0})(1-p)\big(m - 2(t+kl)\big)\big(n - 2(t+kl)\big)}. \tag{4.4}$$

This together with (3.27) and Proposition 3.1 shows that

$$\begin{aligned} &|P(S_a < t) - e^{-\lambda}| \\ &\quad\le \frac{\lambda(m+n)(t+kl)}{(1 - C_0 t^{-\gamma_0})(1-p)\big(m - 2(t+kl)\big)\big(n - 2(t+kl)\big)} \\ &\qquad + 2(t+kl)\Bigg\{ mnK^2\left[mp^{t(1+\theta_v) - 8k} + np^{t(1+\theta_h) - 8k} + 2(t+kl)p^{t(1+\theta)}\right] \\ &\qquad\qquad + \lambda(m+n)(1-p)\binom{t-1}{k} l^k p^{t-k}\left[\sum_{i \in E}(2 - \mu_i - \nu_i)\right]^k \Bigg\} \\ &\quad\le C_4\left[\lambda\frac{m+n}{mn} + mnt^{2k+1}\left(mp^{t(1+\theta_v)} + np^{t(1+\theta_h)}\right)\right] \end{aligned}$$

for large enough m, n and t and some arbitrary constant $C_4 > 0$. The choice of the relative growth rates for m, n and t at the end of Section 3 then implies that there are constants C, $\gamma > 0$ so that

$$|P(S_a < t) - e^{-\lambda}| \le C(mn)^{-\gamma}.$$

We will now give a detailed description of the algorithm on which our analysis is based, and we show that, for m and n large, it will provide us with an optimal alignment with probability close to 1. The algorithm can be called a *greedy algorithm*. It is not efficient, but it is easy to describe. A future paper will describe a more efficient algorithm that is easy to implement and uses techniques such as dynamic programming and hashing.

The algorithm which we now describe will not necessarily provide us with the best alignment but it will do so with probability approaching 1 as the lengths of the sequences go to infinity. The algorithm will check each site for whether or not a matching region of type $(t;k,l)$ starts at that particular site. Let $\alpha = (i,j)$ for some $1 \le i \le m$ and $1 \le j \le n$. If $Z_\alpha = 0$, stop and start with another site. Otherwise, find $r_1 \equiv \min_{r \ge 1} Z_{i+r,j+r} = 0$. The set of sites $\{(i,j), (i+1, j+1), \dots, (i+r_1-1, j+r_1-1)\}$ forms the first matching subregion. To connect this to another matching subregion, determine the set

$$F_1 = \{(i',j'): Z_{i',j'} = 1 \text{ where } (i',j') = (i+r_1+s, j+r_1) \text{ or } (i+r_1, j+r_1) \text{ for } 1 \le s \le l\}.$$

If $F_1 = \emptyset$, stop and check another site in the dot matrix. Otherwise, for each $(i',j') \in F_1$, find $r_{2,s} \equiv \min_{r \ge 1} Z_{i'+r,j'+r} = 0$ for $1 \le s \le |F_1|$; that is, the sets $\{(i',j'),(i'+1,j'+1),\dots,(i'+r_{2,s}-1, j'+r_{2,s}-1)\}$ for each $(i',j') \in F_1$ and $1 \le s \le |F_1|$, form the second matching subregions which are connected by indel regions with at most l indels. We can continue this until we find $k+1$ matching subregions linked together by indel regions, where each indel region has at most l indels. Now we have to check the resulting matching regions for their lengths. If there is one matching region of type $(t;k,l)$, we say that $Y_\alpha = 1$. We have to repeat this procedure for all sites $\alpha \in J$. After having determined the status of all Y_α in the dot matrix, we can check the status of the V_α's and subsequentially the status of all the X_α's. The algorithm is called greedy since it goes along diagonal lines until the first time it sees a mismatch and then switches to another diagonal line. This way we might miss some alignments, that is, there might be more than one way to connect two given matching subregions at a particular indel-link. For instance, if we wish to align AAGGGGCT and AAGGCT, the greedy algorithm would find

$$\begin{matrix} A\,A\,G\,G\,G\,G\,C\,T \\ A\,A\,G\,G\,-\,-\,C\,T \end{matrix} \tag{4.5}$$

whereas the following alignment, which the algorithm misses,

$$\begin{matrix} A\,A\,G\,G\,G\,G\,C\,T \\ A\,A\,G\,-\,G\,-\,C\,T \end{matrix} \tag{4.6}$$

would be possible as well. Alignment (4.6) can be obtained from (4.5) by simply rearranging letters in the indel region. It is therefore basically the same alignment as the one in (4.5) and should not be counted as a separate alignment. We will argue that when m and n are large, most alignments the algorithm misses are such that they can be obtained by simply rearranging letters in indel regions of the matching regions found by the algorithm. So it will be enough to count only those matching regions our algorithm can find.

We will show that when we fix k and l and let m, n and t (properly scaled) go to infinity, then, loosely speaking, a typical matching region will be of the following form: (i) adjacent indel regions are "far apart," and (ii) there is very little "overlap" where subregions are pieced together. We will make this rigorous in the following.

(i) For a matching region of type $(t;k,l)$, there are $\binom{t-1}{k}$ ways of choosing starting points for the k indel regions. We will show that in most of the resulting types, adjacent indel regions are at least t^{γ_6} units apart, where $\gamma_6 \in (0,\ 1)$. The number of ways of choosing starting points for the k indel regions such that all adjacent indel regions are at least t^{γ_6} units apart is at least

$$\frac{(t-1)(t-2-2t^{\gamma_6})(t-3-4t^{\gamma_6})\cdots\left(t-k-2(k-1)t^{\gamma_6}\right)}{k!}. \tag{4.7}$$

An upper bound on the number of ways where at least one pair of indel regions is at most t^{γ_6} units apart is therefore

$$\begin{aligned}\binom{t-1}{k} &- \frac{(t-1)(t-2-2t^{\gamma_6})(t-3-4t^{\gamma_6})\cdots\left(t-k-2(k-1)t^{\gamma_6}\right)}{k!}\\ &\leq \frac{(t-1)^k}{k!} - \frac{\left(t-k-2(k-1)t^{\gamma_6}\right)^k}{k!}\\ &= \frac{(t-1)^k}{k!}\left[1-\left(\frac{t-k-2(k-1)t^{\gamma_6}}{t-1}\right)^k\right]\\ &= \frac{(t-1)^k}{k!}\left[1-\left(1-\frac{k-1+2(k-1)t^{\gamma_6}}{t-1}\right)^k\right]\\ &\leq \frac{(t-1)^k}{k!}\frac{k(k-1)(1+2t^{\gamma_6})}{t-1},\end{aligned} \tag{4.8}$$

where we used $(1-x)^n \geq 1-nx$ in the last step. We can now bound the probability of finding a matching region of type $(t;k,l)$ somewhere, in which at least one pair of adjacent indel regions is at most t^{γ_6} units apart. Using the same argument as in the proof of Lemma 3.5, this is less than or equal to

$$mn\frac{(t-1)^k}{k!}\frac{k(k-1)(1+2t^{\gamma_6})}{t-1}l^k p^{t-k}\left[\sum_{i\in E}\mu_i\nu_i(2-\mu_i-\nu_i)\right]^k. \tag{4.9}$$

In Theorem 1 we showed that relative growth rates for m and n can be chosen so that

$$mn\binom{t-1}{k}l^k p^{t-k}\left[\sum_{i\in E}\mu_i\nu_i(2-\mu_i-\nu_i)\right]^k \tag{4.10}$$

converges to a constant which is bounded away from 0 and ∞. Therefore, since the additional term $[k(k-1)(1+2t^{\gamma_6})]/(t-1)$ in (4.9) tends to 0 as $t \to \infty$, the bound in (4.9) tends to 0 as $t \to \infty$; that is, with probability approaching 1 as $m, n, t \to \infty$, indel regions are at least t^{γ_6} units apart. The rate of convergence is of order $t^{\gamma_6 - 1}$ which is of order $(\log(mn))^{\gamma_6 - 1}$.

(ii) We will now show that there is not too much "overlap" between adjacent subregions. By this we mean the following: The greedy algorithm goes along a diagonal line until it hits a mismatch; it then switches to one of the neighboring diagonal lines to continue. The next subregion, however, might have started earlier, that is, the previous subregion and this next subregion may share letters in either of the two sequences. We will argue that this overlap is typically substantially smaller than t^{γ_7}, for some $\gamma_7 \in (0, 1)$, as m and n (and thus t,) tend to infinity. The probability that the overlap is bigger than t^{γ_7} in any matching region of type $(t; k, l)$ is less than or equal to

$$\begin{aligned} &C(k, l)mn\binom{t-1}{k} l^k p^{t-k-t^{\gamma_7}} \left(p^{1+\theta}\right)^{t^{\gamma_7}} \left[\sum_{i\in E} \mu_i \nu_i(2-\mu_i-\nu_i)\right]^k \\ (4.11) \qquad &= C(k, l)mn\binom{t-1}{k} l^k p^{t-k} \left[\sum_{i\in E} \mu_i \nu_i(2-\mu_i-\nu_i)\right]^k p^{\theta t^{\gamma_7}}, \end{aligned}$$

for some $\theta > 0$. Using again that (4.10) converges to a constant which is bounded away from 0 and infinity, and that $p^{\theta t^{\gamma_7}} \to 0$ as $t \to \infty$, this shows that the right-hand side of (4.11) tends to 0 as m and n (and thus t) tend to infinity. The rate of convergence is of order $O(p^{\theta t^{\gamma_7}})$, which is much faster than the rate of convergence obtained in (i).

If we choose $\gamma_7 < \gamma_6$, then, with probability close to 1, all indel regions will be isolated in the sense that none of the sites in different indel regions will share letters from either of the sequences. Therefore, a detailed knowledge of the structure of the indel regions is not necessary. Our greedy algorithm will find with probability at least $1 - C(\log(mn))^{-\gamma}$, for some γ, $C > 0$, local alignments of the type specified. It will produce those local alignments in which k is minimized for fixed l. Many algorithms that are currently used operate under a similar assumption, that is, they would rank (4.5) at least as high as (4.6). This is based on the (biological) assumption that it is at least as likely to have a deletion of, say, five nucleotides at a time at a particular location as to have five independent single deletions (see, e.g., [21]).

We can now finish the proof of Theorem 2. With the notation introduced earlier,

$$|P(S < t) - e^{-\lambda}| \leq |P(S < t) - P(S_a < t)| + |P(S_a < t) - e^{-\lambda}|.$$

Note that $|P(S < t) - P(S_a < t)|$ can be bounded by the probability that the algorithm fails to find the largest matching region. This probability, however, is at most $C(\log(mn))^{-\gamma}$ for some $\gamma, C > 0$.

Acknowledgments. The author wishes to thank M. Waterman and R. Arratia for suggesting the problem and for useful discussions. Furthermore, the author wishes to thank two referees for carefully reading the manuscript.

REFERENCES

[1] ALDOUS, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic*. Springer, New York.

[2] ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1989). Two moments suffice for Poisson approximations: The Chen–Stein method. *Ann. Probab.* **17** 9–25.

[3] ARRATIA, R., GORDON, L. and WATERMAN, M. S. (1986). An extreme value theory for sequence matching. *Ann. Statist.* **14** 971–993.

[4] 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.

[5] ARRATIA, R. and WATERMAN, M. (1989). The Erdös–Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.* **17** 1152–1169.

[6] CHEN, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab.* **3** 534–545.

[7] CHVÁTAL, V. and SANKOFF, D. (1975). Longest common subsequence of two random sequences. *J. Appl. Probab.* **12** 306–315.

[8] ERDÖS, P. and RÉNYI, A. (1970). On a new law of large numbers. *J. Analyse Math.* **23** 103–111.

[9] GOLDSTEIN, L. and WATERMAN, M. S. (1992). Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons. *Bull. Math. Biol.* **54** 785–812.

[10] KARLIN, S., GHANDOUR, G., OST, F., TAVARÉ, S. and KORN, L. J. (1983). New approaches for computer analysis of nucleic acid sequences. *Proc. Nat. Acad. Sci. U.S.A.* **80** 5660–5664.

[11] KARLIN, S. and OST, F. (1987). Counts of long aligned word matches among random letter sequences. *Adv. in Appl. Probab.* **19** 293–351.

[12] LEWIN, B. (1990). *Genes 4*. Oxford Univ. Press.

[13] LI, W.-H. and GRAUR, D. (1991). *Fundamentals of Molecular Evolution*. Sinauer, Sunderland, MA.

[14] NEUHAUSER, C. (1993). A phase transition for the distribution of matching blocks. Unpublished manuscript.

[15] SANKOFF, D. (1972). Matching sequences under deletion/insertion constraints. *Proc. Nat. Acad. Sci. U.S.A.* **69** 4–6.

[16] SANKOFF, D. and KRUSKAL, J. B. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, London.

[17] SMITH, T. F., WATERMAN, M. S. and BURKS, C. (1985). The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* **13** 645–656.

[18] STEIN, C. M. (1972). A bound for the error in the normal approximations to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **3** 583–602. Univ. California Press, Berkeley.

[19] STEIN, C. M. (1986). *Approximate Computation of Expectations*. IMS, Hayward, CA.

[20] TAVARÉ, S. and GIDDINGS, B. (1989). Some statistical aspects of the primary structure of nucleotide sequences. In *Mathematical Methods for DNA Sequences* (M. S. Waterman, ed.) 117–132. CRC Press, Boca Raton, FL.

[21] WATERMAN, M. S. (1984). Efficient sequence alignment algorithms. *J. Theoret. Biol.* **108** 333–337.

[22] WATERMAN, M. and EGGERT, M. (1987). A new algorithm for best subsequence alignments with application to tRNA–rRNA comparisons. *Journal of Molecular Biology* **197** 723–728.

[23] WATSON, J. D., HOPKINS, N. H., ROBERTS, J. W., STEITZ, J. A. and WEINER, A. M. (1987). *Molecular Biology of the Gene*, 4th ed., Benjamin/Cummings, Menlo Park, CA.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WISCONSIN
480 LINCOLN DRIVE
MADISON, WISCONSIN 53706