

# PATH PROPERTIES OF LCS-OPTIMAL ALIGNMENTS

J. Lember\*, H. Matzinger, A. Vollmer

October 4, 2007

Tartu University  
Institute of Mathematical Statistics  
Liivi 2-513 50409, Tartu, Estonia  
E-mail: jyril@ut.ee

University of Bielefeld  
Postfach 10 01 31, 33501 Bielefeld, Germany  
E-mail: matzing@mathematik.uni-bielefeld.de

Georgia Tech  
School of Mathematics  
Atlanta, Georgia 30332-0160, U.S.A.  
E-mail: matzing@math.gatech.edu

University of Bielefeld  
Postfach 10 01 31, 33501 Bielefeld, Germany  
E-mail: avollmer@mathematik.uni-bielefeld.de

**Keywords.** *Longest common subsequence, optimal alignments.*

## Abstract

We investigate the behavior of optimal alignment paths for related and non-related random sequences. An alignment between two finite sequences is optimal if it corresponds to the longest common subsequence (LCS). We prove the existence of lowest and highest optimal alignments and study their differences. High differences between the extremal alignments imply the high variety of all optimal alignments. We present several simulations indicating that the related sequences have typically the distance between the extremal alignments of much smaller size than independent (unrelated) sequences. In particular, the simulations suggest that for the related sequences, the growth of the distance between the extremal alignments is logarithmical. The main theoretical results of the paper prove that (under some assumptions) this is the case, indeed. The paper suggests that the properties of the optimal alignment paths characterize the relatedness of the sequences.

---

\*Supported by the Estonian Science Foundation Grant nr. 7553 and SFB 701 of Bielefeld University

# 1 Introduction

Let  $\mathcal{A}$  be a finite alphabet. In everything that follows,  $X = X_1 \dots X_n \in \mathcal{A}^n$  and  $Y = Y_1 \dots Y_n \in \mathcal{A}^n$  are two strings of length  $n$ . A common subsequence of  $X$  and  $Y$  is a sequence that is a subsequence of  $X$  and at the same time of  $Y$ . We denote by  $L_n$  the length of the longest common subsequence (LCS) of  $X$  and  $Y$ . LCS's are a very important tool in computational biology, where they are used for comparing DNA- and protein-alignments (see, e.g. [17, 18, 3, 7]). They are also used in computational linguistics, speech recognition and so on. In all these applications, two strings with a relatively long LCS, are deemed related. Hence, to distinguish related pairs of strings from unrelated via the length of LCS (or other similar optimality measure), it is important to have some knowledges about the (asymptotical) distribution of  $L_n$ . Unfortunately, although studied for a relatively long time, not much about the statistical behavior of  $L_n$  is known even when the sequences  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  are both i.i.d. and independent of each other. Using the subadditivity, it is easy to see the existence of a constant  $\gamma$  such that

$$\frac{L_n}{n} \rightarrow \gamma \text{ a.s and in } L_1. \quad (1.1)$$

(see, e.g. [1, 18]). Referring to the celebrated paper of Chvatal and Sankoff [6], the constant  $\gamma$  is called the Chvatal-Sankoff constant; its value is unknown for even as simple cases as i.i.d. Bernoulli sequences. In this case, the value of  $\gamma$  obviously depends on the Bernoulli parameter  $p$ . When  $p = 0.5$ , the various bounds indicate that  $\gamma \approx 0.81$  [15, 10, 4]. For a smaller  $p$ ,  $\gamma$  is even bigger. Hence, a common subsequence of two independent Bernoulli sequences typically makes up large part of the total length, if the sequences are related, LCS is even larger. As for the mean of  $L_n$ , not much is also known about the variance of  $L_n$ . In [6], it was conjectured that for Bernoulli parameter  $p = 0.5$ , the variance is of order  $o(n^{\frac{2}{3}})$ . Using an Efron-Stein type of inequality, Steele [15] proved  $Var[L_n] \leq 2p(1-p)n$ . In [16], Waterman conjectured that  $Var[L_n]$  grows linearly. In series of papers, Matzinger and others prove the Waterman conjecture for different models [5, 12, 9, 11].

Because of relatively rare knowledges about its asymptotics, it is rather difficult to build any statistical test based on  $L_n$  or any other global optimality criterion. The situation is better for local alignments (see e.g. [3, 16]), because for these alignments approximate  $p$ -values were recently calculated [14, 8].

In the present paper, we propose another approach – instead of studying the length of LCS, we investigate the properties and behavior of the optimal alignments. Namely, even for moderate  $n$ , the LCS is hardly unique. Every LCS corresponds to an optimal alignment (not necessarily vice versa, but for time being these notions can be considering equivalent), so in general, we have several optimal alignments. The differences can be of the local nature meaning that the optimal alignments do not vary much, or they can be of global nature. We conjecture that the variation of the optimal alignments characterizes the relatedness of the sequences. The simulations in Section 3 clearly indicate that for related sequences the differences of optimal alignments are of local nature, whilst for independent sequences they

vary much more. Those simulations motivate to find a way to quantify the non-uniqueness and use the obtained characteristic as a measure of the relatedness. For that we define the lowest and highest alignment and measure their distance in terms of maximal vertical difference or Hausdorff's distance. The simulations in Section 3 show that for independent sequences, the growth of both of them is almost linear; for related sequence, however, it is logarithmic. Under some assumptions, the latter is confirmed by the the main theoretical results, Theorems 5.1 and 5.2. We would like to mention that to our best knowledge, such an approach has not been exploited before, although the optimal alignments have been deserved some attention before [2]. Therefore, the present paper as the first step does not aim to minimize the assumptions or propose any ready-made tests. These are the issues of the further research. Instead, we present several simulation results to motivate the study in this direction.

We finish the introduction with an example giving the insight in what follows.

**Example.** Let us look at a practical example. Take the two related words: the English  $X = mother$  and the German  $Y = mutter$ . The longest common subsequence is  $mter$  and hence  $L_6 = 4$ . This relatively large value of 4 indicates that the words are related. We represent any common subsequence as an alignment (possibly a collection of alignments) of  $X$  with  $Y$ . An alignment can contain gaps. The letters appearing in the common subsequence are aligned one on top of the other. The letters which are not aligned with the same letter in the other text get aligned with a gap. In the case of the present numerical example the common subsequence  $mter$  corresponds to the following alignment:

$$\begin{array}{c|c|c|c|c|c|c} m & o & & t & & h & e & r \\ \hline m & & u & t & t & & e & r \end{array} \quad (1.2)$$

Every common subsequence can be represented by such an alignment with gaps. An alignment corresponding to a LCS is called an *optimal alignment*. The optimal alignment is, in general, not unique. For example, to the same common subsequence  $mter$  corresponds also the following optimal alignment:

$$\begin{array}{c|c|c|c|c|c|c} m & o & & & t & h & e & r \\ \hline m & & u & t & t & & e & r \end{array} \quad (1.3)$$

In the following, we represent alignments in 2 dimensions. For this we view alignments as subsets of  $\mathbb{R}^2$ , in the following manner:

if the  $i$ -th letter of  $X$  gets aligned with the  $j$ -th letter of  $Y$ , then the set representing the alignment is to contain  $(i, j)$ . For example, the alignment (1.2) can be represented as follows:  $(1, 1), (3, 3), (5, 5), (6, 6)$  with the corresponding plot

$$\begin{array}{c|c|c|c|c|c|c} r & & & & & & x \\ \hline e & & & & & & x \\ \hline t & & & & & & \\ \hline t & & & x & & & \\ \hline u & & & & & & \\ \hline m & x & & & & & \\ \hline m & o & t & h & e & r & \end{array} \quad (1.4)$$

Here, the symbol  $x$  indicates pairs of aligned letters. The alignment (1.3) has the following

2-dimensional representation:  $(1, 1), (3, 4), (5, 5), (6, 6)$  with the corresponding plot

$r$						$x$
$e$						$x$
$t$			$x$			
$t$						
$u$						
$m$	$x$					
	$m$	$o$	$t$	$h$	$e$	$r$

(1.5)

Since  $Y_3 = Y_4 = t$ , both alignments correspond to the same LCS  $mter$ . In our present example, there exists no other 2-dimensional representation of a LCS. In both cases, the letter  $m$  in the two words are aligned with each other. Hence,  $(1, 1)$  is part of the optimal alignment in any case (more precisely,  $(1, 1)$  is part of the 2-dimensional representation of any optimal alignment of  $X$  and  $Y$ ). Such a point is called *uniqueness point* of the optimal alignment of  $X$  and  $Y$ . In the present example  $(5, 5)$  and  $(6, 6)$  are other uniqueness points of the optimal alignment. Between  $(1, 1)$  and  $(5, 5)$  there are no further uniqueness points. We call this stretch a *non-uniqueness stretch*. We describe it by giving its  $x$ -coordinate:  $[1, 5]$  is a non-uniqueness stretch. We measure its length by projection on the  $x$ -coordinate: in our example this means that  $[1, 5]$  is a non uniqueness stretch of *length* 4.

Note the following: all the points of the alignment-representation (1.4) are below the points of the alignment-representation (1.5). Hence we say that (1.4) is the *lowest optimal alignment* and (1.5) is the *highest optimal alignment*.

## 2 Notation and preliminaries

Recall that  $X = X_1 \dots X_n$  and  $Y = Y_1 \dots Y_n$  are two strings of length  $n$  from alphabet  $\mathcal{A}$ . Let there exist two subsets of indices  $\{i_1, \dots, i_k\}, \{j_1, \dots, j_k\} \subset \{1, \dots, n\}$  satisfying  $i_1 < i_2 < \dots < i_k, j_1 < j_2 < \dots < j_k$  and  $X_{i_1} = Y_{j_1}, X_{i_2} = Y_{j_2}, \dots, X_{i_k} = Y_{j_k}$ . Then  $X_{i_1} \dots X_{i_k}$  is a common subsequence of  $X$  and  $Y$  and the pairs  $\{(i_1, j_1), \dots, (i_k, j_k)\}$  are (the 2-dimensional representation of) the corresponding alignment.  $L_n$  is the biggest  $k$  such that there exist such subsets of indices.

We now formally define the highest alignment. Let  $\{(i_1^\alpha, j_1^\alpha), \dots, (i_k^\alpha, j_k^\alpha)\}_\alpha$  be the set of all optimal alignments. Hence,  $k = L_n$  and  $\alpha \in A$  runs over all possible optimal alignments. We denote

$$J := \{j_l^\alpha : \alpha \in A, l = 1, \dots, k\}, \quad I := \{i_l^\alpha : \alpha \in A, l = 1, \dots, k\}.$$

Let  $j_k^h := \max_\alpha j_k^\alpha = \max J$ . There might be many alignments  $\alpha$  such that  $j_k^\alpha = j_k^h$ . Among such alignments take  $i_k^h$  to be minimum. Formally,  $i_k^h = \min\{i_k^\alpha : j_k^\alpha = j_k^h\}$ . After fixing  $(i_k^h, j_k^h)$ , we take  $j_{k-1}^h$  as the biggest  $j \in J$  such that the corresponding  $i$ , let it be  $i(j)$ , is smaller than  $i_k^h$ . Formally,  $j_{k-1}^h = \max\{j \in J : i(j) < i_k^h\}$ . There might be several  $i$ 's such that corresponding  $j$  is  $j_{k-1}^h$ . Amongst them, we choose the minimum. Thus  $i_{k-1}^h = \min\{i : j(i) = j_{k-1}^h\}$ . Proceeding so, we obtain an alignment. We call this the highest alignment procedure. We now prove that the procedure can be repeated  $k$ -times, i.e. the obtained alignment is optimal.

**Proposition 2.1** *The highest alignment procedure produces an optimal alignment  $\{(i_1^h, j_1^h), \dots, (i_k^h, j_k^h)\}$ , where  $(i_t^h, j_t^h)$  can be obtained as follows*

$$j_t^h = \max\{j_t^\alpha : \alpha \in A\}, \quad i_t^h = \min\{i_t^\alpha : j(i_t^\alpha) = j_t^h\}, \quad t = 1, \dots, k. \quad (2.1)$$

**Proof.** Clearly the pair  $(i_k^h, j_k^h)$  is the last pair of an optimal alignment, i.e. there exists  $\alpha \in A$  such that  $(i_k^h, j_k^h) = (i_k^\alpha, j_k^\alpha)$ . So (2.1) holds with  $t = k$ . Similarly, there exists a  $\beta \in A$  such that  $j_{k-1}^h = j_{k-1}^\beta$ . Let us show this. There exists a  $\beta$  such that  $j_{k-1}^h = j_l^\beta$ , we have to show that  $l = k - 1$ . Note that  $l$  cannot be  $k$ , since otherwise  $(i_1^\beta, j_1^\beta), \dots, (i_k^\beta, j_k^\beta), (i_k^h, j_k^h)$  would be an alignment of length  $k + 1$ . Suppose  $l = k - 2$ . Since  $j_{k-2}^\beta < j_{k-1}^\beta < j_k^\beta \leq j_k^h = \max J$ , by definition of  $j_{k-1}^h$ , it must be that  $i_k^h \leq i_{k-1}^\beta$ . Since  $i_k^h = i_k^\alpha > i_{k-1}^\alpha$ , we have that  $i_{k-1}^\alpha < i_{k-1}^\beta < i_k^\beta$ . On the other hand,  $j_{k-1}^\alpha \leq j_{k-1}^h = j_{k-2}^\beta$  implying that  $j_{k-1}^\alpha < j_{k-1}^\beta < j_k^\beta$ . Hence  $(i_1^\alpha, j_1^\alpha), \dots, (i_{k-1}^\alpha, j_{k-1}^\alpha), (i_{k-1}^\beta, j_{k-1}^\beta), (i_k^\beta, j_k^\beta)$  would be an alignment of length  $k + 1$ . Hence  $j_{k-1}^h = \max\{j_{k-1}^\alpha : \alpha \in A, i_{k-1}^\alpha < i_k^h\}$ . Let us now prove that (2.1) with  $t = k - 1$  holds. If this were not the case, then  $j_{k-1}^h < \max\{j_{k-1}^\alpha : \alpha \in A\}$ . This implies the existence of  $\beta$  so that  $j_{k-1}^\beta > j_{k-1}^h$  and  $i_{k-1}^\beta \geq i_k^h$ . But as we saw, those inequalities would give an alignment with the length  $k + 1$ . This concludes the proof of (2.1) with  $t = k - 1$ . For  $t = k - 2, \dots, 1$  proceed similarly. ■

One can also think of the right-most alignment. The right-most alignment could be defined as an alignment  $\{(i_1^r, j_1^r), \dots, (i_k^r, j_k^r)\}$ , where  $i_1^r = \min I$ ,  $j_1^r = \max\{j \in J : i(j) = i_1^r\}$  and  $i_t^r := \min\{i \in I : j(i) > j_{t-1}^r\}$ ,  $j_t^r = \max\{j \in J : i(j) = i_t^r\}$ ,  $t = 2, \dots, k$ . By the analogue of Proposition 2.1,

$$i_t^r = \min\{i_t^\alpha : \alpha \in A\}, \quad j_t^r = \max\{j_t^\alpha : i(j_t^\alpha) = i_t^r\}, \quad t = 1, \dots, k. \quad (2.2)$$

Using (2.1) and (2.2), it is easy to see that the right-most and highest alignments actually coincide. Indeed, by (2.1) and (2.2),  $j_t^h \geq j_t^r$  and  $i_t^r \leq i_t^h$ ,  $\forall t$ . If, for a  $t$ ,  $(i_t^h, j_t^h) \neq (i_t^r, j_t^r)$ , then, by the definitions, both inequalities have to be strict, i.e.  $i_t^r < i_t^h$  and  $j_t^r < j_t^h$ . This would imply the existence of an alignment with the length  $k + 1$ .

The lowest (the left-most) alignment  $\{(i_1^l, j_1^l), \dots, (i_k^l, j_k^l)\}$  will be defined similarly:  $j_1^l := \min J$ ,  $i_1^l = \max\{i \in I : j(i) = j_1^l\}$ ,

$$j_u^l := \min\{j \in J : j > j_{u-1}^l\}, \quad i_u^l = \max\{i \in I : j(i) = j_u^l\}, \quad l = 2, \dots, k.$$

By the analogue of Proposition 2.1, the lowest alignment  $\{(i_1^l, j_1^l), \dots, (i_k^l, j_k^l)\}$ , satisfies

$$j_t^l = \min\{j_t^\alpha : \alpha \in A\}, \quad i_t^l = \max\{i_t^\alpha : j(i_t^\alpha) = j_t^l\}, \quad t = 1, \dots, k. \quad (2.3)$$

Finally note that the right most alignment equals the highest alignment of  $(Y_n, \dots, Y_1)$  and  $(X_n, \dots, X_1)$  implying that the latter equals to the highest alignment of  $(X_1, \dots, X_n)$  and  $(Y_1, \dots, Y_n)$ . Similarly, the lowest alignment can be defined as the highest alignment between  $(X_n, \dots, X_1)$  and  $(Y_n, \dots, Y_1)$ ; the left-most alignment is the highest

alignment between  $(Y_1 \dots Y_n)$  and  $(X_1 \dots X_n)$  implying that they are equal.

We are interested in measuring the distance between the lowest and highest alignment. One possible measure would be the maximum vertical or horizontal distance (provided they are somehow defined). However those distances need not match the intuitive meaning of the closeness of the alignment. For example, the following two alignments (marked with  $x$  and  $o$ , respectively) have a relatively long maximal vertical distance (3), though they are intuitively rather close:

						$xo$
				$o$		
			$o$	$x$		
			$x$			
	$xo$					

(2.4)

To overcome the problem, we measure the distance between two alignments also in terms of Hausdorff's distance. More precisely, let  $A = \{a_1, \dots, a_k\}$  and  $B = \{b_1, \dots, b_l\}$  be two alignments, both represented as sets of two-dimensional points. The *Hausdorff's distance between A and B* is:

$$h(A, B) := \max\left\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\right\},$$

where  $d$  is a distance in  $\mathbb{R}^2$ . In our case, we take  $d$  as the maximum-distance (but one can also consider the usual Euclidian metric). We remark that Hausdorff's distance is defined for any kind of sets. For the alignments in (2.4), the Hausdorff's distance is obviously 1 (if  $d$  were Euclidean, the Hausdorff's distance would be  $\sqrt{2}$ ).

In the following, we consider long sequences, hence the optimal alignments are long as well. When representing such optimal alignments in two dimensions, the dots appear like a line. To make the pictures more illustrative, we connect the points of such representations with a line. Then, to every alignment corresponds a curve. We shall call this curve the *alignment graph* (when it is obvious from the context, we skip "graph"). Given two alignment graphs, it is easy to find the maximal vertical (horizontal) distance between them. Besides the Hausdorff's distance, we shall also use this quantity, called the vertical (horizontal) distance, as a measure of the closeness of the graphs. Note that the minimum of the vertical and horizontal distances gives an upper bound to the Hausdorff's distance.

### 3 Simulation study: motivation of the research

The main purpose of the paper is to study the difference (distance) between the lowest and highest optimal alignment graphs. We start with the simulation study by generating random sequences and finding the highest and lowest alignment graphs. In all simulations, the alphabet  $\mathcal{A}$  consists of four symbols. The sequences  $X$  and

$Y$  are i.i.d. with uniform distribution over  $\mathcal{A}$ . We begin with the case, when  $X$  and  $Y$  are independent of each other.

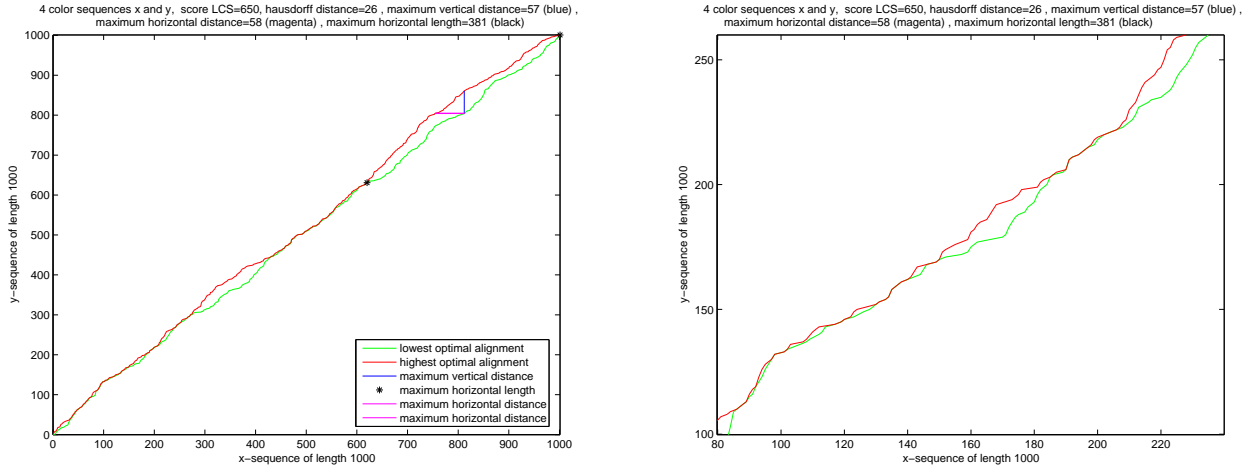


Figure 1:  $X$  and  $Y$  are independent,  $n = 1000$ . Right: zoomsection.

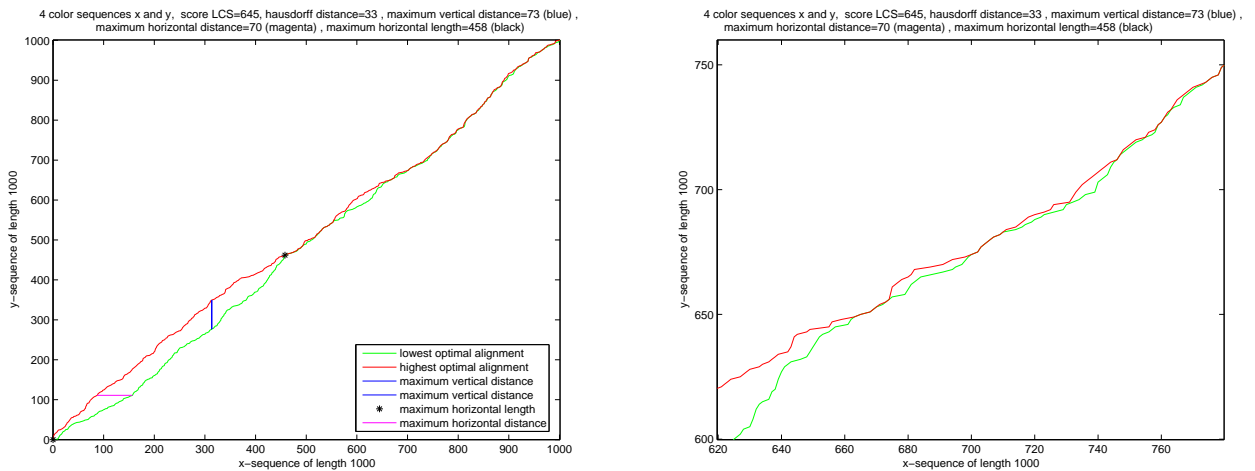


Figure 2:  $X$  and  $Y$  are independent,  $n = 1000$ . Right: zoomsection.

Figure 1 shows the highest and lowest alignment graphs (red and green curve, respectively) for a random draw of independent sequences of length  $n = 1000$ . Hence all optimal alignments are located between the red and the green path. In those places where the red and the green path coincide all optimal alignments are identical. As already mentioned, such places are called uniqueness places of the optimal alignment. Figure 2 represents the results of another draw of the same setup. Both pictures are provided with a zoomsection picture to illustrate the local behavior. Note that all over the path there are uniqueness points with some non-uniqueness stretches between them. The longest non-uniqueness stretch is marked with \*'s, as well as the maximum vertical and horizontal distance. From the pictures, one gets

the impression that the non-uniqueness stretches are in size of strictly smaller order than linear in  $n$  but many are typically larger than logarithmic order in  $n$ . To see the long-run behavior better, we increase  $n$  and perform the same simulations with  $n = 10000$ . The results are presented in figures 3 and 4.

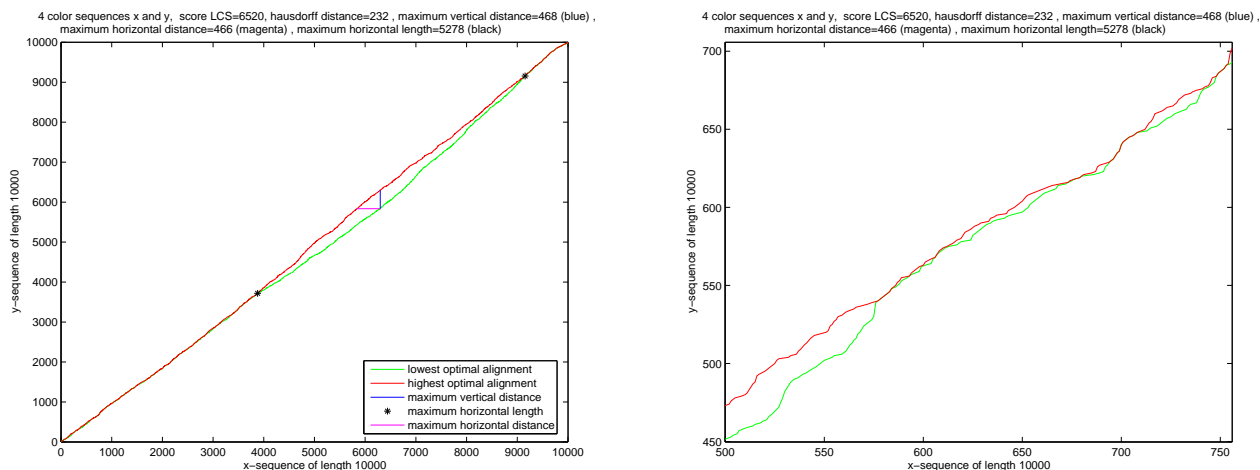


Figure 3:  $X$  and  $Y$  are independent,  $n = 10000$ . Right: zoomsection.

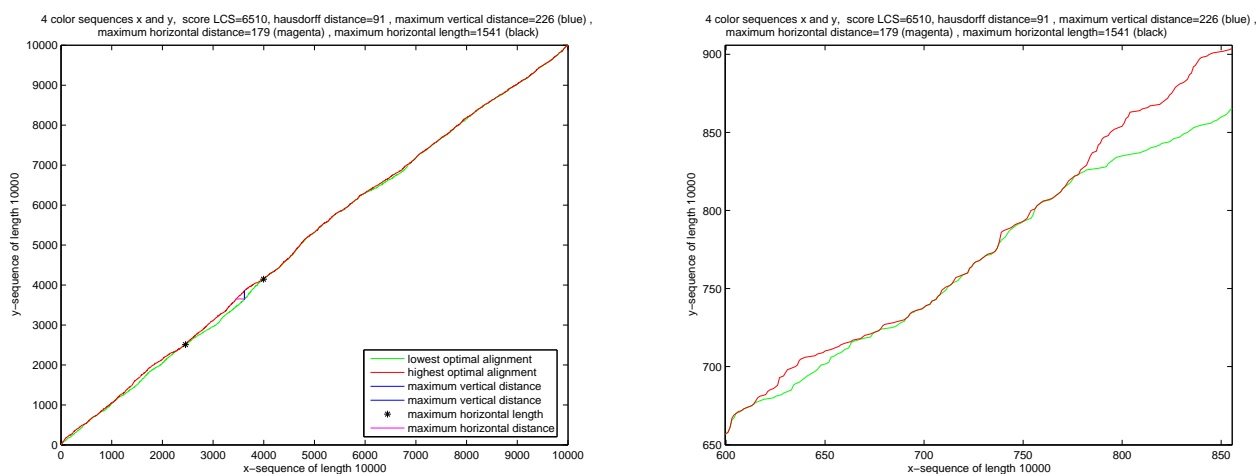


Figure 4:  $X$  and  $Y$  are independent,  $n = 10000$ . Right: zoomsection.

The main aim of this paper is to investigate the difference in the geometric structure of optimal alignments between the case where  $X$  and  $Y$  are independent and the case where they are not. By 'geometric structure' we mean, among others, size and frequency of the non-uniqueness bulbs (stretches) and uniqueness stretches. Let us look next at the picture of the highest and lowest optimal alignment when the sequences  $X$  and  $Y$  are dependent of each other. In Section 4, we formally define the notion of relatedness in our setup (condition R). Let us right now just mention that  $X$  and  $Y$  are related, if there is a sequence of common ancestors  $Z_1, Z_2, \dots$ ,

from which both  $X$  and  $Y$  are obtained by random mutations and deletions. The sequences  $X$  and  $Y$  are both still i.i.d. with uniform marginal distribution, but they are not independent of each other any more. As previously, we simulate the sequences of length 1000 (Figures 5 and 6) as well as 10000 (Figures 7 and 8). The red dots in the zoomsection pictures correspond to the pairs that are equal (have the same color) and have the same ancestor of the same color. Hence, a pair  $(i, j)$  is marked with a red dot, if  $X_i$  and  $Y_j$  have the common ancestor, say  $Z_k$ , and there are no mutations, i.e.  $X_i = Y_j = Z_k$ .

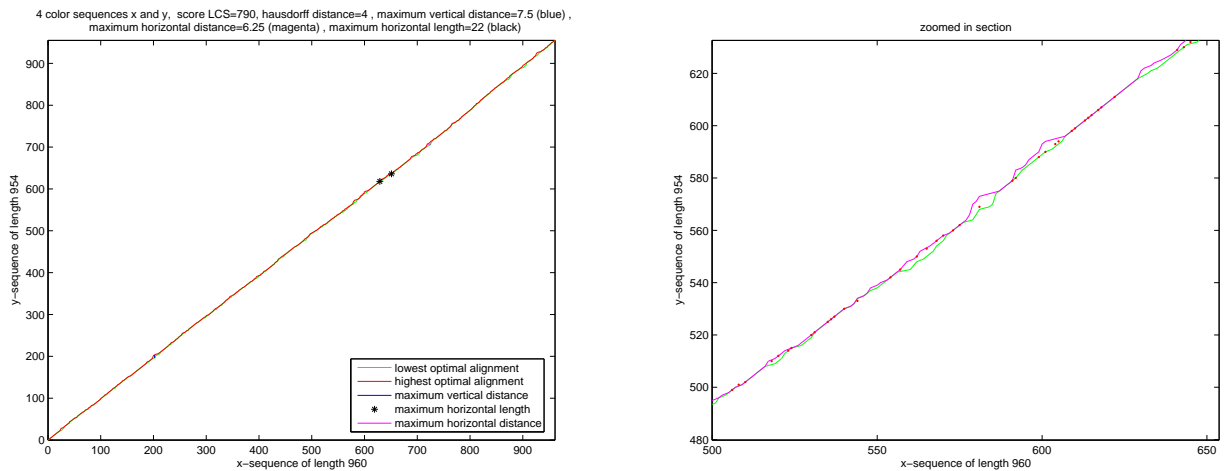


Figure 5:  $X$  and  $Y$  are related,  $n = 1000$ . Right: zoomsection.

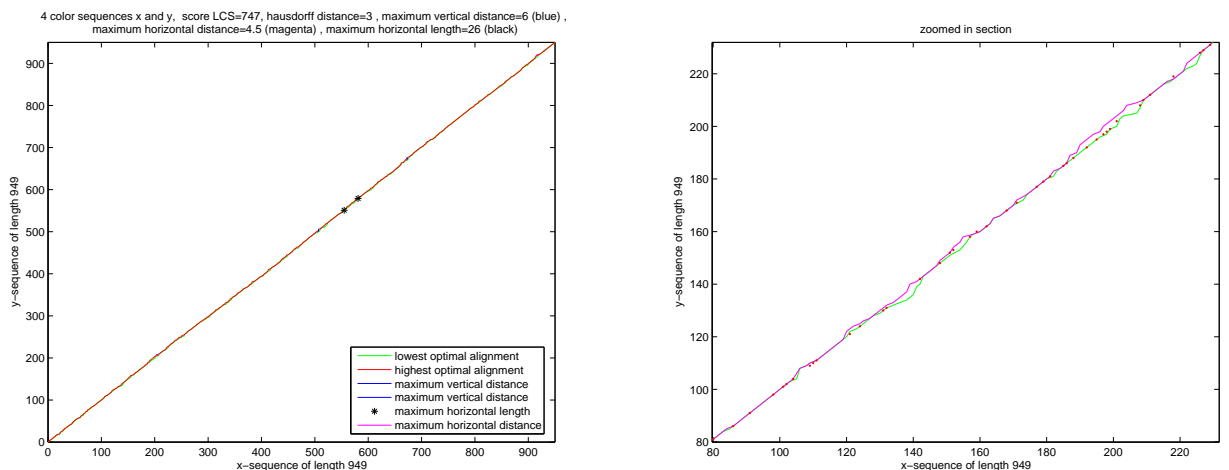


Figure 6:  $X$  and  $Y$  are related,  $n = 1000$ . Right: zoomsection.

From the pictures we see the clear difference between related and unrelated case. We are interested, however, in finding some statistics that are sensible with respect to that difference, and, therefore, can be used for measuring the relatedness.

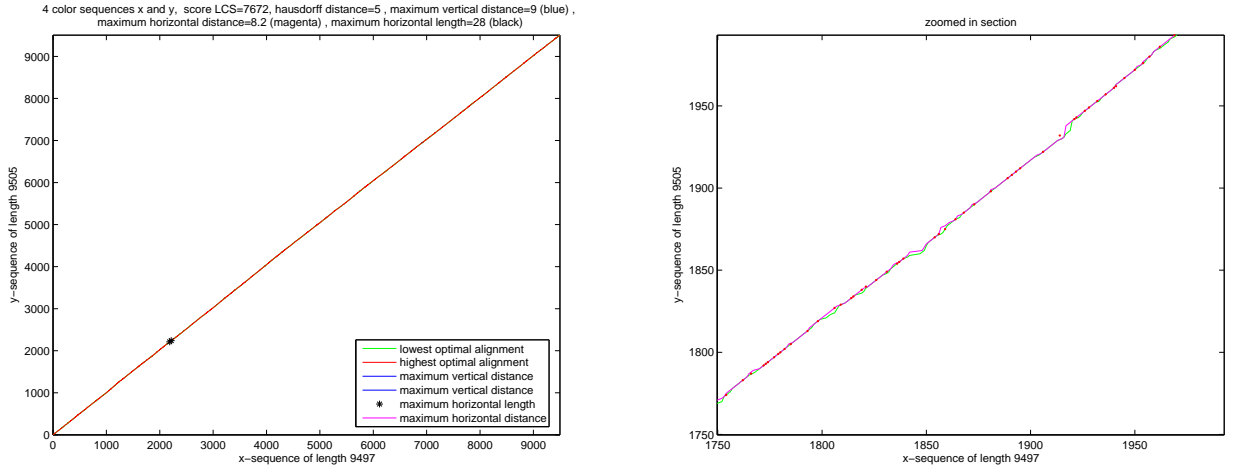


Figure 7:  $X$  and  $Y$  are related,  $n = 10000$ . Right: zoomsection.

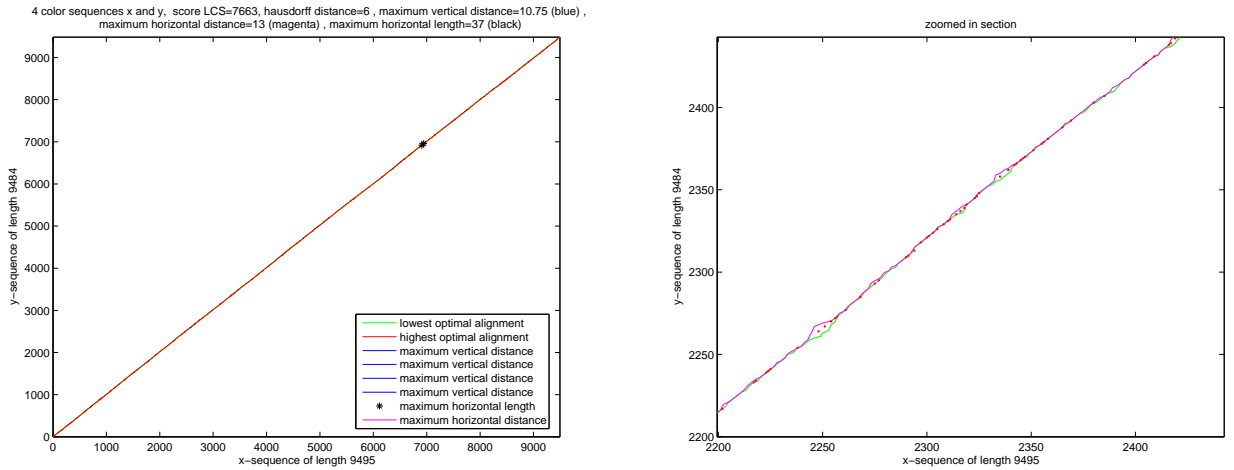


Figure 8:  $X$  and  $Y$  are related,  $n = 10000$ . Right: zoomsection.

A classical measure is the relative length of LCS:  $L_n/n$ . In our simulations,  $L_n/n$  for independent sequences has values 0.65, 0.645, 0.652, 0.651 (figures 1,2,3,4, resp.), while for related sequences  $L_n/n$  is 0.79, 0.747, 0.7672, 0.7663 (figures 5,6,7,8, resp.) The difference is noticeable.

Another measure could be the length of the biggest non-uniqueness stretch i.e. the (horizontal) length between \*'s. For independent sequences, those numbers are  $381 = 0.381 \times 1000$ ,  $458 = 0.458 \times 1000$ ,  $5278 = 0.5278 \times 10000$ ,  $1541 = 0.1541 \times 10000$  (figures 1,2,3,4, resp.) – linear order of  $n$ . For related sequences, those numbers are  $22 = 0.022 \times 1000$ ,  $26 = 0.026 \times 1000$ ,  $28 = 0.0028 \times 10000$ ,  $37 = 0.0037 \times 10000$  (figures 5,6,7,8, resp.) – about logarithmic order of  $n$ . The difference between related and unrelated case is very noticeable.

Although the length of the biggest non-uniqueness stretch seems to be a good measure of relatedness, it certainly has some disadvantages. The main problem

is that it is not robust against shifts. For example, the sequences  $ATATATAT$  and  $TATATATA$  are rather related ( $L_8 = 7$ ), however because of the shift, the highest and lowest optimal alignment run parallel, so there are no uniqueness points (biggest non-uniqueness stretch has length 7).

Next, we compare the vertical distances between highest and lowest alignment graph. For independent sequences, those numbers are  $57 = 0.057 \times 1000$ ,  $73 = 0.073 \times 1000$ ,  $468 = 0.0468 \times 10000$ ,  $226 = 0.0226 \times 10000$  (figures 1,2,3,4, resp.) Although the increase seems to be slower as linear in  $n$ , it certainly is much faster than logarithmic growth. For related sequences the vertical distances are  $7.5 = 0.0075 \times 1000$ ,  $6 = 0.006 \times 1000$ ,  $9 = 0.0009 \times 10000$ ,  $10.75 = 0.001075 \times 10000$  (figures 5,6,7,8, resp.) The growth seems to be logarithmic.

The same holds for horizontal distance between the highest and lowest alignment graph. For independent sequences, those numbers are  $58 = 0.058 \times 1000$ ,  $70 = 0.07 \times 1000$ ,  $466 = 0.0466 \times 10000$ ,  $179 = 0.0179 \times 10000$  (figures 1,2,3,4, resp.) For related sequences the horizontal distances are  $6.25 = 0.00625 \times 1000$ ,  $4.5 = 0.0045 \times 1000$ ,  $8.2 = 0.00082 \times 10000$ ,  $13 = 0.0013 \times 10000$  (figures 5,6,7,8, resp.) The growth, again, is logarithmic.

Finally, we measure the Hausdorff's distance between the highest and lowest alignments with respect to the maximum norm. Since the Hausdorff's distance is well-defined for the alignments as the sets of (2-dimensional) dots, we do not actually need to join the dots by the lines (make the alignment graphs). Hence, the Hausdorff's distance is measured between the alignments not between the graphs. (Of course, one could also measure the Hausdorff's distance between the graphs, that would be bigger). So, for independent sequences, we have the numbers  $26 = 0.026 \times 1000$ ,  $33 = 0.033 \times 1000$ ,  $232 = 0.0232 \times 10000$ , and  $91 = 0.0091 \times 10000$  (figures 1,2,3,4, resp.). Whereas for related sequences the numbers are  $4 = 0.004 \times 1000$ ,  $3 = 0.003 \times 1000$ ,  $5 = 0.0005 \times 10000$ , and  $6 = 0.0006 \times 10000$  (figures 5,6,7,8, resp.).

The zoomsection pictures for related sequences (Figures 5,6,7,8) indicate another interesting phenomenon – the local differences between the highest and lowest alignment graphs are relatively big in the regions that contain less common ancestors (red dots), while in the regions which are relatively less mutated, the highest and lowest alignment coincide or are close to each other. Since finding the (unobservable) common ancestors is often of interest, the described phenomenon suggests that the uniqueness stretches of optimal alignments can provide more reliable information about the common ancestor, while the regions where the highest and lowest alignment are far from each other indicate many mutations.

Figures 9 and 10 complement the previous study. The aim of these simulations is to find out more about the order of growth of all considered statistics. In those simulations, for different  $n$ -s up to 10000, 100 pairs of i.i.d sequences of length  $n$  with uniform marginals were generated. Half of them were independent and another half were related. In both cases, the average of considered statistics:  $L_n$ , the horizontal of the maximum non-uniqueness bulb, then length of maximum vertical distance, Hausdorff's distance are plotted against  $n$ .

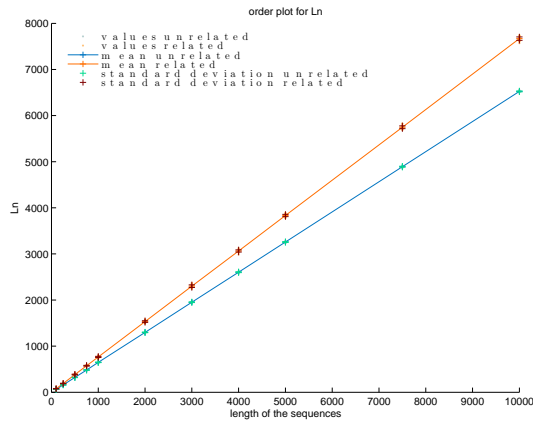
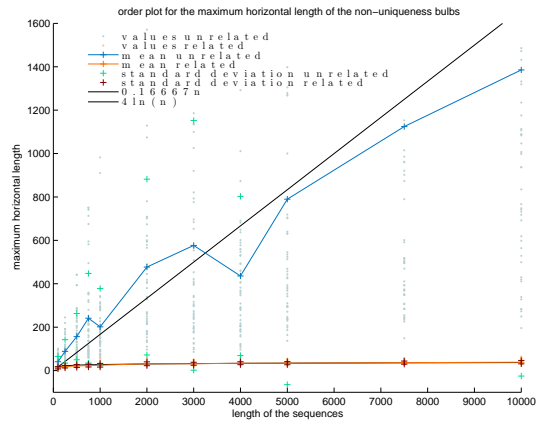


Figure 9: Growth of  $L_n$



Growth of non-uniqueness stretch

The left-plot in Figure 9 shows the growth of  $L_n$ . The standard deviation around the means are marked with crosses. For independent case the crosses are almost overlapping implying that the deviation is relatively small. As the picture shows, the growth of  $L_n$  is linear in both cases, the slope, however, is different: the upper line corresponds to the related sequences, the lower line is for independent sequences. The right-plot in Figure 9 shows the horizontal length of maximum non-uniqueness stretch. For independent sequences (blue curve), the growth is, perhaps, smaller than linear but considerably faster than logarithmic. The black line is, in some sense, the best linear approximation. The blue +-signs mark the standard deviation around the mean that in this case is rather big, meaning that these simulations do not give enough evidence to conclude the non-linear growth. For related sequences (brown curve), the growth is clearly logarithmic because it almost overlaps with the  $4 \ln n$ -curve. We also point out that the standard deviation for this case is remarkable smaller and this only confirms the conjecture of logarithmic growth.

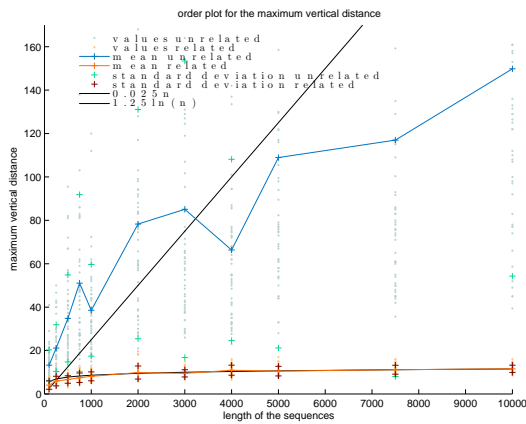
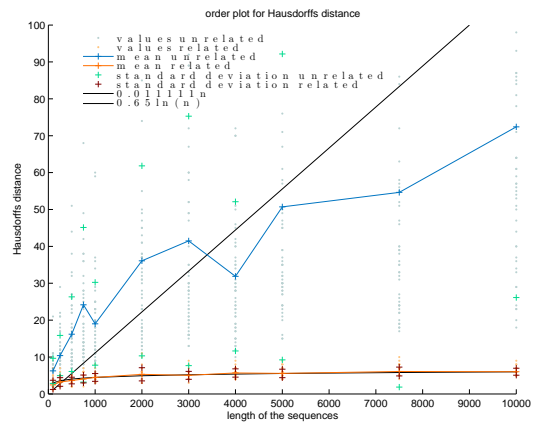


Figure 10: Growth of maximal vertical distance



Growth of Hausdorff's distance

In Figure 10, the maximum vertical distance (left) and Hausdorff's distance (right) are plotted. Both pictures are similar to the right picture of Figure 9 and can be interpreted analogously. For the related case, the growth is clearly logarithmic and that is a full correspondence with our theoretical result in Section 5, where we prove that all optimal alignments lie within a narrow band with breadth of logarithmic order in  $n$  (Theorems 5.1 and 5.2).

## 4 Related sequences: definition and theory

### 4.1 Definition of relatedness

Let us now define the relatedness of the sequences  $(X, Y)$ . Our concept of relatedness is based on the assumption that there exists a common ancestor, from which both sequences  $X$  and  $Y$  are obtained by independent random mutations and deletions. In the following, the common ancestor is an  $\mathcal{A}$ -valued i.i.d. process  $Z_1, Z_2, \dots$ . We could imagine that  $X$  and  $Y$  is the genome of two species whilst  $Z$  is the genome of a common ancestor. In computational linguistics  $X$  and  $Y$  could be words from two languages which both evolved from the word  $Z$  in an ancient language.

A letter  $Z_i$  has a probability to mutate according to a transition matrix that does not depend on  $i$ . Hence, a mutation of the letter  $Z_i$  can be formalized as  $f(Z_i, \xi_i)$ , where  $f : \mathcal{A} \times \mathbb{R} \rightarrow \mathcal{A}$  is a mapping and  $\xi_i$  is a standard normal random variable. The mapping  $f_i(\cdot) := f(\cdot, \xi_i)$  from  $\mathcal{A}$  to  $\mathcal{A}$  will be referred as the random mapping. The mutations of the letters are assumed to be independent. This means that the random variables  $\xi_1, \xi_2, \dots$  or the random mappings  $f_1, f_2, \dots$  are independent (and identically distributed). After mutations, the sequence is  $f_1(Z_1), f_2(Z_2), \dots$ . Some of its elements disappear. This is modeled via a deletion process  $D_1^x, D_2^x, \dots$  that is assumed to be an i.i.d. Bernoulli sequence. If  $D_i^x = 0$ , then  $f_i(Z_i)$  is deleted. The resulting sequence, let it be  $X$ , is, therefore, the following:  $X_i = f_j(Z_j)$  if and only if  $D_j^x = 1$  and  $\sum_{k=1}^j D_k^x = i$ . We call the index  $j$  the *ancestor of  $i$* , it shall be denoted by  $a_x(i)$ . The mapping  $a_x$  depends on the deletion process  $D^x$ , only. Now

$$X_i = f_{a^x(i)}(Z_{a^x(i)}), \quad i = 1, \dots, n.$$

Similarly, the sequence  $Y$  is obtained from  $Z$ . For mutations, fix an i.i.d. standard normal sequence  $\eta_1, \eta_2, \dots$  so that the mutated sequence is  $h_1(Z_1), h_2(Z_2), \dots$  with  $h_i(\cdot) := f(\cdot, \eta_i)$ . Note that the transition matrix corresponding to  $Y$ -mutations equals the one corresponding to  $X$ -mutations implying that the random mappings  $h_i$  and  $f_i$  have the same distribution. Since the mutations of  $X$  and  $Y$  are supposed to be independent, we assume the sequences  $\xi$  and  $\eta$  or the random mappings sequences  $f_1, f_2, \dots$  and  $h_1, h_2, \dots$  are independent. Note that then the pairs  $(f_1(Z_1), h_1(Z_1)), (f_2(Z_2), h_2(Z_2)), \dots$  are independent, but  $f_i(Z_i)$  and  $h_i(Z_i)$ , in general, are not. Finally,

$$Y_i = f_{a^y(i)}(Z_{a^y(i)}),$$

where, as previously,  $a^y(i) = j$  if and only if  $D_j^y = 1$  and  $\sum_{k=1}^j D_k^y = i$ . Here,  $D_1^y, D_2^y, \dots$  is an i.i.d. Bernoulli sequence with the same parameter as  $D^x$  but

independent of  $D^x$ . Hence the deletions of  $Y$  and  $X$  are independent. The formal definition goes as follows.

**Condition R** We say that  $X$  and  $Y$  satisfy condition **R** if and only if all of the following five conditions are satisfied:

1. There exist  $Z_1, Z_2, \dots, h_1, h_2, \dots, f_1, f_2, \dots, D_1^x, D_2^x, \dots$  and  $D_1^y, D_2^y, \dots$  five i.i.d. sequences independent of each other.
2. The  $Z_i$ 's are elements of  $\mathcal{A}$ ;  $h_i$ 's and the  $f_i$ 's are maps from  $\mathcal{A}$  to  $\mathcal{A}$  and the sequences  $D_1^x, D_2^x, \dots$  and  $D_1^y, D_2^y, \dots$  are Bernoulli sequences with the same parameter  $p$ .
3. The sequence  $h_1, h_2, \dots$  has same distribution as the sequence  $f_1, f_2, \dots$ .
4. The sequence  $X_1, X_2, \dots$  is obtained by the following rule:  $X_i = f_j(Z_j)$  iff  $D_j^x = 1$  and  $\sum_{k=1}^j D_k^x = i$ .
5. The sequence  $Y_1, Y_2, \dots$  is obtained by the following rule:  $Y_i = h_j(Z_j)$  iff  $D_j^y = 1$  and  $\sum_{k=1}^j D_k^y = i$ .

In the simulations of Section 3, the related sequences satisfied the condition **R** with the following parameters: the common ancestor process  $Z_1, Z_2, \dots$  was i.i.d. with uniform marginal distribution. The mutation matrix is the following

$$(P(f_1(Z_1) = a_j | Z_1 = a_i))_{i,j=1,\dots,4} = \begin{pmatrix} 0.9 & 0.02 & 0.02 & 0.06 \\ 0.02 & 0.9 & 0.06 & 0.02 \\ 0.02 & 0.06 & 0.9 & 0.02 \\ 0.06 & 0.02 & 0.02 & 0.9 \end{pmatrix}$$

With such a matrix, the marginal distribution of  $Y$  and  $X$  is uniform. The deletion probability  $1 - p = 0.05$ .

## 4.2 Properties

When condition **R** is satisfied,  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  are i.i.d. sequences which depend on each other only through  $Z_1, Z_2, \dots$ . Note that because of the deletion, the pairs

$$(X_1, Y_1), (X_2, Y_2), \dots \tag{4.1}$$

are not any more independent, so the 2-dimensional process (4.1) is not i.i.d. However, it is a stationary process and, as the following simple observation shows, it is still an ergodic process.

**Proposition 4.1** *Let the processes  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  satisfy condition **R**. Then the 2-dimensional process (4.1) is mixing and, therefore, ergodic.*

**Proof.** Recall that a stochastic process  $U_1, U_2, \dots$ , taking values on a countable set  $\mathcal{U}$ , is mixing, if for every  $k, l \in \mathbb{N}$  and  $u^k \in \mathcal{U}^k, u^l \in \mathcal{U}^l$ , it holds

$$\lim_{n \rightarrow \infty} P((U_1, \dots, U_k) = u^k, (U_{n+1}, \dots, U_{n+l}) = u^l) = P((U_1, \dots, U_k) = u^k) P((U_1, \dots, U_l) = u^l) \tag{4.2}$$

(see e.g. [13]). Take  $U_i = (X_i, Y_i)$ ,  $\mathcal{U} = \mathcal{A} \times \mathcal{A}$ . Fix  $k, l, u^k, u^l$ . Let  $U_1^k := (U_1, \dots, U_k)$  and  $U_{n+1}^{n+l} := (U_{n+1}, \dots, U_{n+l})$ . Note that when  $\sum_{i=1}^n D_i^x \geq k$  and  $\sum_{i=1}^n D_i^y \geq k$ , then

$$P(U_1^k = u^k, U_{n+1}^{n+l} = u^l | D^x, D^y) = P(U_1^k = u^k)P(U_1^l = u^l).$$

Define the events

$$E_x(n) := \left\{ \sum_{i=1}^n D_i^x \geq k \right\}, \quad E_y(n) := \left\{ \sum_{i=1}^n D_i^y \geq k \right\}, \quad E_n := E_y(n) \cap E_x(n).$$

Clearly

$$\begin{aligned} P(U_1^k = u^k, U_{n+1}^{n+l} = u^l | E_n) P(E_n) &\leq P(U_1^k = u^k, U_{n+1}^{n+l} = u^l) \\ &\leq P(U_1^k = u^k, U_{n+1}^{n+l} = u^l | E_n) P(E_n) + P(E_n^c). \end{aligned}$$

Since  $P(U_1^k = u^k, U_{n+1}^{n+l} = u^l | E_n) = P(U_1^k = u^k)P(U_1^l = u^l)$ , we have that (4.2) holds, if  $P(E_n) \rightarrow 1$ . The latter follows easily from the large deviation. Indeed, recall  $p = P(D_i^x = 1)$ . If  $n$  is so big that  $\frac{k}{n} \leq \frac{p}{2}$ , then, by Hoeffding's inequality,

$$\begin{aligned} P(E_n^c) &= P\left(\sum_{i=1}^n D_i^x < k\right) = P\left(\sum_{i=1}^n D_i^x - np < k - np\right) \\ &\leq P\left(\sum_{i=1}^n D_i^x - np < -\frac{p}{2}n\right) \leq \exp\left[-\frac{p^2}{2}n\right] \rightarrow 0. \end{aligned}$$

■

Having the ergodicity, we can apply the Kingman's subadditivity theorem to deduce the existence of a constant  $\gamma_{\mathbb{R}}$  such that

$$\lim_{n \rightarrow \infty} \frac{L_n}{n} = \gamma_{\mathbb{R}}, \quad \text{a.s. and in } L_1. \quad (4.3)$$

The convergence in  $L_1$  implies

$$\frac{EL_n}{n} \rightarrow \gamma_{\mathbb{R}}.$$

We say that  $X = X_1 X_2 \dots X_n$  and  $Y := Y_1 Y_2 \dots Y_n$  are *related* if they satisfy condition **R**. For related sequences, we say that letters  $X_i$  and  $Y_j$  are *related* if and only if they have the same common ancestor. In other words,  $X_i$  and  $Y_j$  are related, if and only if  $a^x(i) = a^y(j)$ . This means that there exists  $Z_k$  such that  $X_i = h_k(Z_k)$  and  $Y_j = f_k(Z_k)$ .

Next we prove a large deviation lemma for related sequences similar to the one proven for independent sequences by Waterman and Arratia [3].

**Lemma 4.1** *Assume  $X$  and  $Y$  are related. Let  $L_n$  be the length of the LCS of  $X$  and  $Y$ . Then, for every  $\Delta > 0$  and  $n$  big enough*

$$P(|L_n - E[L_n]| \geq n\Delta) \leq 4 \exp\left[-\frac{p}{16} \Delta^2 n\right]. \quad (4.4)$$

**Proof.** For independent sequence, (4.4) trivially follows from McDiarmid's inequality (see, e.g. [12]). In the present case, we have to add an extra control over the deletion process, exactly as in the proof of Proposition 4.1. Let

$$E_x(m) := \left\{ \sum_{i=1}^m D_i^x \geq n \right\}, \quad E_y(m) := \left\{ \sum_{i=1}^m D_i^y \geq n \right\}, \quad E_m := E_y(m) \cap E_x(m).$$

If  $E_x(m)$  holds, then  $X_1, \dots, X_n$  is a function of

$$(f_1(Z_1), D_1^x), \dots, (f_m(Z_m), D_m^x),$$

where  $(f_i(Z_i), D_i^x)$  are i.i.d.. Similarly, if  $E_y(m)$  holds, then  $Y_1, \dots, Y_n$  is a function of  $(h_1(Z_1), D_1^y), \dots, h_m(Z_m), D_m^y)$ . Hence,  $L_n$  is a function of

$$(f_1(Z_1), h_1(Z_1), D_1^x, D_1^y), \dots, (f_m(Z_m), h_m(Z_m), D_m^x, D_m^y),$$

where  $(f_i(Z_i), h_i(Z_i), D_i^x, D_i^y)$  are i.i.d. The proof is based on the fact that changing  $(f_i(Z_i), h_i(Z_i), D_i^x, D_i^y)$  changes the value of  $L_n$  at most by 2. Indeed, changing  $f_i(Z_i)$  corresponds to the change of an element of  $X$  (if  $D_i^x = 1$ ) and this changes the value of  $L_n$  at most by 1. Changing  $D_i^x$  corresponds to removing one element of  $X$  and adding another element (somewhere else). This, again, changes the value of  $L_n$  at most by 1. It is easy to see that changing  $f_i(Z_i)$  as well as  $D_i^x$  has the same effect – the value of  $L_n$  changes at most by 1. Thus, the maximum change of  $L_n$  that we could obtain by changing all elements in  $(f_i(Z_i), h_i(Z_i), D_i^x, D_i^y)$  is at most 2. Hence, by McDiarmid's inequality, conditioning on  $E_m$ , we get for every  $\Delta > 0$

$$P(|L_n - E[L_n|E_m]| > m\Delta | E_m) \leq 2 \exp\left[-\frac{\Delta^2}{2} m\right]. \quad (4.5)$$

Take  $m = \frac{2}{p}n$ . Then (4.5) is

$$P(|L_n - E[L_n|E_m]| > \frac{2\Delta}{p}n | E_m) \leq 2 \exp\left[-\frac{\Delta^2}{p}n\right]. \quad (4.6)$$

Since  $EL_n = E[L_n|E_m]P(E_m) + E[L_n|E_m^c]P(E_m^c)$  and  $0 \leq L_n \leq n$ , we have

$$\begin{aligned} |E[L_n|E_m] - EL_n| &= |E[L_n|E_m](1 - P(E_m)) - E[L_n|E_m^c]P(E_m^c)| \\ &= P(E_m^c)|E[L_n|E_m^c] - E[L_n|E_m]| \leq nP(E_m^c) =: \alpha(n). \end{aligned}$$

As in the proof of Proposition 4.1, we have

$$P(E_m^c) \leq 2 \exp\left[-\frac{p^2}{2}m\right] = 2 \exp[-pn].$$

Hence, if  $n$  is so big that  $\alpha(n) < \frac{\Delta}{p}n$ , from (4.6), we have

$$\begin{aligned}
P(|L_n - EL_n| \geq \frac{2\Delta}{p}n | E_m) &\leq P(|L_n - E[L_n | E_m]| + |EL_n - E[L_n | E_m]| \geq \frac{2\Delta}{p}n | E_m) \\
&\leq P(|L_n - E[L_n | E_m]| + \alpha(n) \geq \frac{2\Delta}{p}n | E_m) \\
&= P(|L_n - E[L_n | E_m]| \geq \frac{2\Delta}{p}n - \alpha(n) | E_m) \\
&\leq P(|L_n - E[L_n | E_m]| \geq \frac{\Delta}{p}n | E_m) \\
&\leq 2 \exp[-\frac{\Delta^2}{4p}n].
\end{aligned}$$

Since

$$P(|L_n - EL_n| \geq \Delta n) \leq P(|L_n - EL_n| \geq \Delta n | E_m) + P(E_m^c),$$

we have that with  $\Delta' = \frac{2\Delta}{p}$ ,

$$P(|L_n - EL_n| \geq \Delta' n) \leq 2 \exp[-\frac{(\Delta')^2 p}{16}n] + 2 \exp[-pn].$$

If  $\Delta' \leq 1$ , then  $2 \exp[-pn] \leq 2 \exp[-(\Delta')^2 pn]$ , implying that the right side is bounded by  $4 \exp[-\frac{(\Delta')^2}{16}pn]$ . This proves (4.4) for  $\Delta \leq 1$ . Since  $L_n \leq n$ , for  $\Delta > 1$ , (4.4) trivially holds. ■

**Corollary 4.1** *Assume  $X$  and  $Y$  are related. Let  $L_n$  be the length of the LCS of  $X$  and  $Y$ . Then, for every  $\Delta > 0$  there exists  $n_o(\Delta)$  big enough*

$$P(|L_n - \gamma_{\mathbb{R}} n| \geq n\Delta) \leq 4 \exp[-\frac{p}{64}\Delta^2 n], \quad n > n_o \quad (4.7)$$

**Proof.** Let  $n$  be so big that  $|EL_n/n - \gamma_{\mathbb{R}}| < \Delta/2$ . Then  $|EL_n - \gamma_{\mathbb{R}} n| \leq (\Delta/2)n$  and

$$\begin{aligned}
P(|L_n - \gamma_{\mathbb{R}} n| \geq n\Delta) &\leq P(|L_n - EL_n| + |EL_n - \gamma_{\mathbb{R}} n| \geq n\Delta) \\
&\leq P(|L_n - EL_n| \geq n\frac{\Delta}{2}) \leq 4 \exp[-\frac{p}{64}\Delta^2 n].
\end{aligned}$$

■

### 4.3 Combinatorics

Let us introduce some notation. Let  $x_1, \dots, x_n$  and  $y_1, \dots, y_m$  be two fixed finite sequences. An *alignment* of  $x_1, \dots, x_n$  and  $y_1, \dots, y_m$  is a strictly increasing mapping

$$v : \{1, \dots, n\} \hookrightarrow \{1, \dots, m\}. \quad (4.8)$$

Notation (4.8) means: There exists  $I \subset \{1, \dots, n\}$  and a mapping

$$v : I \rightarrow \{1, \dots, m\}$$

such that  $y_{v(i)} = x_i, \forall i \in I$  and  $v$  is strictly increasing:  $v(i_2) > v(i_1)$ , if  $i_2 > i_1$ . The length of  $v$  is denoted as  $|v|$ .

Consider now the case  $m = n$ , i.e. both sequences are of length  $n$ . Let then  $V_k$  be the set of all alignments with length  $k$ . Formally,

$$V_k := \{v : \{1, \dots, n\} \hookrightarrow \{1, \dots, n\} \mid |v| = k\}.$$

Fix  $\Delta > 0$  and let

$$V_n := \bigcup_{k=(\gamma_R-\Delta)n}^{(\gamma_R+\Delta)n} V_k.$$

From Corollary 4.1, it follows that with high probability, all optimal alignments of  $X$  and  $Y$  belong to the set  $V_n$ , provided  $n$  is big enough (depending on  $\Delta$ ). Let  $H$  be the binary entropy function:

$$H(p) := -p \log_2 p - (1-p) \log_2 (1-p)$$

and let

$$H_1 := \max_{\alpha \in [\gamma_R - \Delta, \gamma_R + \Delta]} H(\alpha). \quad (4.9)$$

Since

$$C_n^{pn} \leq 2^{H(p)n},$$

for every

$$(\gamma_R - \Delta)n \leq k \leq (\gamma_R + \Delta)n, \quad (4.10)$$

it holds

$$|V_k| = (C_n^{\frac{k}{n}n})^2 \leq 2^{2H_1 n}$$

and, therefore,

$$|V_n| \leq 2\Delta n 2^{2H_1 n}. \quad (4.11)$$

Let us consider now a more general case  $m > n$ . Assume that  $m \leq n(1 + \Delta)$ . Then

$$|V_k| = (C_n^{\frac{k}{n}n})(C_m^{\frac{k}{m}m}) \leq 2^{H(\frac{k}{n})n + H(\frac{k}{m})n(1+\Delta)}.$$

Instead of (4.10), we assume  $k$  to satisfy

$$\gamma_R - \Delta \leq \frac{k}{n} \leq \gamma_R + 2\Delta. \quad (4.12)$$

Then

$$\gamma_R - 2\Delta \leq \frac{\gamma_R - \Delta}{1 + \Delta} \leq \frac{k}{m} \leq \frac{k}{n} \leq \gamma_R + 2\Delta.$$

Let

$$H_2 := \max_{\alpha \in [\gamma_R - 2\Delta, \gamma_R + 2\Delta]} H(\alpha). \quad (4.13)$$

Then

$$2^{H(\frac{k}{n})n + H(\frac{k}{m})n(1+\Delta)} \leq 2^{H_2 n + H_2 n(1+\Delta)} = 2^{H_2 n(2+\Delta)}$$

In this case, defining

$$V_{n,m} := \bigcup_{k=(\gamma_{\mathbb{R}}-\Delta)n}^{(\gamma_{\mathbb{R}}+2\Delta)n} V_k,$$

it holds

$$|V_{n,m}| \leq 3\Delta n 2^{(2+\Delta)H_2 n}.$$

## 5 Main results

### 5.1 Every optimal alignment contains a related pair

In this section, we prove that for related sequences typically all optimal alignments lie in a narrow band of breadth having logarithmic order in  $n$ .

**Lemma 5.1** *Assume that  $X$  and  $Y$  are related and*

$$2H(\gamma_{\mathbb{R}}) + \gamma_{\mathbb{R}} \log_2(\max_{a \in \mathcal{A}} P(Y_1 = a)) < 0. \quad (5.1)$$

*Then there exists a constant  $k_2 > 0$  such that for every  $n$  big enough,*

$$P(\exists \text{ optimal alignment of } X \text{ and } Y \text{ aligning no related letters}) \leq e^{-nk_2}.$$

**Proof.** Let  $v$  be a mapping (4.8) of length  $k$ . More precisely, there exists indexes  $1 \leq i_1 < i_2 < \dots < i_k \leq n$  such that  $X_{i_l} = Y_{v(i_l)}$ ,  $l = 1, \dots, k$ . To keep the notations simple, without loss of generality, we assume that  $i_1 = 1, i_2 = 2, \dots, i_k = k$ .

Denote

$$a^x = (a^x(1), \dots, a^x(k)), \quad a^y = (a^y(v(1)), \dots, a^y(v(k)))$$

and

$$a^x \neq a^y \quad \Leftrightarrow \quad a^x(1) \neq a^y(v(1)), \dots, a^x(k) \neq a^y(v(k)).$$

Define an event

$$A(v) := \{X_1 = Y_{v(1)}, \dots, X_k = Y_{v(k)}, a^x \neq a^y\}.$$

Thus,  $A(v)$  is the event that  $v$  is an alignment that aligns no related letters. We calculate

$$P(A(v)) = \sum_{a^x, a^y: a^x \neq a^y} P(A(v)|a^x, a^y) P(D^x = a^x, D^y = a^y).$$

$$\begin{aligned} P(A(v)|a^x, a^y) &= P(X_1 = Y_{v(1)}|a^x, a^y) \times P(X_2 = Y_{v(2)}|X_1 = Y_{v(1)}, a^x, a^y) \\ &\quad \times \dots \times P(X_k = Y_{v(k)}|X_1 = Y_{v(1)}, \dots, X_{k-1} = Y_{v(k-1)}, a^x, a^y). \end{aligned}$$

Note,

$$\begin{aligned} P(X_i = Y_{v(i)}|X_1 = Y_{v(1)}, \dots, X_{i-1} = Y_{v(i-1)}, a^x, a^y) &= \\ P(f_{a^x(i)}(Z_{a^x(i)}) = h_{a^y(v(i))}(Z_{a^y(v(i))})|f_{a^x(j)}(Z_{a^x(j)}) = h_{a^y(v(j))}(Z_{a^y(v(j))}), j = 1, \dots, i-1) \end{aligned}$$

When  $a^x \neq a^y$ , then for every  $i$ ,  $a^x(i) \neq a^y(v(i))$ . Suppose without loss of generality that  $a^x(i) > a^y(v(i))$ . Then  $a^x(j) < a^x(i)$  and  $a^y(v(j)) < a^x(i)$   $j = 1, \dots, i-1$ . This means that  $Z_{a^x(i)}$  is independent of

$$Z_{a^y(v(i))}, Z_{a^x(i-1)}, Z_{a^y(v(i-1))}, \dots, Z_{a^x(1)}, Z_{a^y(v(1))}.$$

Then also  $X_i = f_{a^x(i)}(Z_{a^x(i)})$  is independent of

$$h_{a^y(v(i))}(Z_{a^y(v(i))}), f_{a^x(i-1)}(Z_{a^x(i-1)}), h_{a^y(v(i-1))}(Z_{a^y(v(i-1))}), \dots, f_{a^x(1)}(Z_{a^x(1)}), h_{a^y(v(1))}(Z_{a^y(v(1))}).$$

Thus,  $X_i$  is independent of  $Y_{v(i)}, X_{i-1}, Y_{v(i-1)}, \dots, X_1, Y_{v(1)}$  and, hence, skipping  $a^x$  and  $a^y$  from notations,

$$\begin{aligned} & P\left(X_i = Y_{v(i)} \mid X_{i-1} = Y_{v(i-1)} \dots, X_1 = Y_{v(1)}\right) = \\ & \sum_{a \in \mathcal{A}} P(X_i = a \mid Y_{v(i)} = a, X_{i-1} = Y_{v(i-1)}, \dots, X_1 = Y_{v(1)}) P(Y_{v(i)} = a \mid X_{i-1} = Y_{v(i-1)}, \dots, X_1 = Y_{v(1)}) = \\ & \sum_{a \in \mathcal{A}} P(X_i = a) P(Y_{v(i)} = a \mid X_{i-1} = Y_{v(i-1)}, \dots, X_1 = Y_{v(1)}) \leq \\ & q \sum_{a \in \mathcal{A}} P(Y_{v(i)} = a \mid X_{i-1} = Y_{v(i-1)}, \dots, X_1 = Y_{v(1)}) = q, \end{aligned}$$

where  $q := \max_{a \in \mathcal{A}} P(X_i = a)$ . Therefore,  $P(A(v) \mid a^x, a^y) \leq q^k$  and  $P(A(v)) \leq q^k$ . By assumption,  $2H(\gamma_{\mathbb{R}}) < \gamma_{\mathbb{R}}(-\log_2 q)$ . Since  $H$  is continuous, it is possible to choose  $\Delta$  so small that

$$2H_1 < (\gamma_{\mathbb{R}} - \Delta)(-\log_2 q), \quad (5.2)$$

where  $\gamma_1$  is as in (4.9). Fix now  $\Delta > 0$  so small that (5.2) holds. Let

$$E_{\Delta} := \{|L_n - n\gamma_{\mathbb{R}}| < n\Delta\}.$$

When  $E_{\Delta}$  holds, then all optimal alignments belong to the set  $V_n$ . Finally, let

$$A := \{\exists \text{ optimal alignment of } X \text{ and } Y \text{ aligning no related letters}\}.$$

When  $E_{\Delta}$  holds, then  $\cup_{v \in V_n} A(v) = A$ . So, when  $E_{\Delta}$  holds, then with  $n$  big enough,

$$\begin{aligned} P(A) & \leq \sum_{v \in V_n} P(A(v)) \leq \sum_{v \in V_n} q^{|v|} = \sum_{v \in V_n} 2^{|v| \log_2 q} \leq \sum_{v \in V_n} 2^{(\gamma_{\mathbb{R}} - \Delta)n \log_2 q} \\ & \leq 2\Delta n 2^{2H_1 n} 2^{(\gamma_{\mathbb{R}} - \Delta)n \log_2 q} = 2\Delta n 2^{(2H_1 - (\gamma_{\mathbb{R}} - \Delta)(-\log_2 q))n}. \end{aligned}$$

By assumption,  $2H_1 - (\gamma_{\mathbb{R}} - \Delta)(-\log_2 q)$  is negative. So, there exists a constant  $k_1 > 0$  such that the right hand is bounded above by  $\exp[-k_1 n]$ , provided  $n$  is big enough. Hence, from Corollary 4.1, there exists  $k_2 > 0$  such that for  $n$  big enough

$$P(A) \leq P(E_{\Delta}^c) + \exp[-k_1 n] \leq \exp[-k_2 n].$$

■

In the previous lemma,  $X$  and  $Y$  were of the same length,  $n$ . This lemma can be generalized for the case  $X$  and  $Y$  are of different length, provided that the difference is not too big. Let  $X^n := X_1 \dots X_n$ ,  $Y^m = Y_1 \dots Y_m$ . Without loss of generality, let us assume  $m \geq n$ . We know that if (5.1) holds, then there exists  $\Delta > 0$  such that

$$(2 + \Delta)H_2 < (\gamma_{\mathbb{R}} - \Delta)(-\log_2 q), \quad (5.3)$$

where  $\gamma_2$  is defined in (4.13). The restriction for  $m$  is:  $m \leq (1 + \Delta)n$ .

**Lemma 5.2** *Let  $n \leq m \leq (1 + \Delta)n$ , where  $\Delta > 0$  satisfies (5.3). Assume that  $X^n$  and  $Y^m$  are related. Then there exists a constant  $k_3(\Delta) > 0$  such that for every  $n > n_0$ ,*

$$P(\exists \text{ optimal alignment of } X^n \text{ and } Y^m \text{ aligning no related letters}) \leq e^{-nk_3}.$$

**Proof.** The proof follows the one of Lemma 5.1;  $\Delta$  is now taken from the assumptions, so instead of (5.2), it satisfies (5.3). This  $\Delta$  defines the set  $E_\Delta$  as in the previous lemma. However, by definition,  $L_n$  is the length of the LCS between  $X^n$  and  $Y^n$ , whilst in the present case we are dealing with the LCS between  $X^n$  and  $Y^m$ . Let  $L_{n,m}$  be the length of that LCS. Clearly  $L_n \leq L_{n,m} \leq L_n + n\Delta$ . Hence, if  $E_\Delta$  holds, then

$$\gamma_{\mathbb{R}} - \Delta \leq \frac{L_n}{n} \leq \frac{L_{n,m}}{n} \leq \frac{L_n}{n} + \Delta \leq \gamma_{\mathbb{R}} + 2\Delta.$$

Hence, if  $E_\Delta$  holds, then all optimal alignments belong to the set  $V_{n,m}$ , implying that, when  $E_\Delta$  holds, then with  $n$  big enough,

$$\begin{aligned} P(A) &\leq \sum_{v \in V_{n,m}} P(A(v)) \leq \sum_{v \in V_{n,m}} q^{|v|} = \sum_{v \in V_{n,m}} 2^{|v| \log_2 q} \leq \sum_{v \in V_{n,m}} 2^{(\gamma_{\mathbb{R}} - \Delta)n \log_2 q} \\ &\leq 3\Delta n 2^{(2+\Delta)H_2 n} 2^{(\gamma_{\mathbb{R}} - \Delta)n \log_2 q} \leq 3\Delta n 2^{((2+\Delta)H_2 - (\gamma_{\mathbb{R}} - \Delta)(-\log_2 q))n}. \end{aligned}$$

By (5.3),

$$(2 + \Delta)H_2 - (\gamma_{\mathbb{R}} - \Delta)(-\log_2 q) < 0.$$

This finishes the proof. ■

## 5.2 The properties of related pairs

### 5.2.1 Global properties

Consider the sequences  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  that are related. Let  $\tau_0^x = \tau_0^y = 0$  and let  $\tau_k^x$  ( $\tau_k^y$ ),  $k = 1, 2, \dots$  be the indexes of the  $k$ -th related pair. So,  $(X_{\tau_1^x}, Y_{\tau_1^y})$  is the first related pair,  $(X_{\tau_2^x}, Y_{\tau_2^y})$  is the second related pair and so on. Let  $a_0 = 0$  and  $a_k$  be the common ancestor of the  $k$ -th related pair, i.e.  $a_k = a^x(\tau_k^x) = a^y(\tau_k^y)$ . Note that  $\tau_{k+1}^x - \tau_k^x \leq a_{k+1} - a_k$  and for  $u \in \mathbb{N}$ ,  $P(a_{k+1} - a_k > u) = (1 - p^2)^u$  implying that, for fixed  $n$  and  $1 > \Delta > 0$ , the probability that the last  $\tau_k^x$  before  $n$ , let it be  $i(n)$ , is further than  $\Delta n$ , is smaller than  $\exp[\Delta n \ln(1 - p^2)]$ . Formally,

$$P((G_n^x)^c) \leq \exp[\Delta n \ln(1 - p^2)], \quad (5.4)$$

where

$$G_n^x = \{n - i(n) \leq \Delta n\}, \quad G_n^y = \{n - j(n) \leq \Delta n\}, \quad G_n =: G_n^y \cap G_n^x$$

$$i(n) = \max\{\tau_k^x : \tau_k^x \leq n\}, \quad j(n) = \max\{\tau_k^y : \tau_k^y \leq n\}.$$

When  $G_n$  holds then the  $X$ -side index of the last related pair is at least  $(1 - \Delta)n$ . Formally  $i(n) \geq (1 - \Delta)n$ . Hence, when drawing a line from the origin to  $(i(n), j(i(n)))$ , the slope of this line is at most  $\frac{1}{1-\Delta}$ . Reversing the roles of  $i$  and  $j$ , we have that the  $Y$ -side index of the last related pair is at least  $(1 - \Delta)n$ . This means that the last related pair, say,  $(i', j')$  satisfies:  $(i', j') \in [(1 - \Delta)n, n] \times [(1 - \Delta)n, n]$ . In 2-dimensional representation, this means that the last related pair is located in a square of size  $\Delta n$  in the upper-right corner. If  $n$  is big enough, then  $G_n$  has a big probability even when  $\Delta > 0$  is small, implying that the last related pair is located almost in the upper right corner. By symmetry, with high probability the first related pair is almost in the lower-left corner. Hence, if  $n$  is big, then the curve that joins all the related pairs in 2-dimensional representation, goes from one corner to the opposite one. However, this does not mean that it should go along the shortest way (approximately) i.e. diagonally from one corner to the other. On the other hand, the figures 5 - 8 in Section 3 clearly indicate that this is the case, since the red dots are almost following the diagonal line. In the following, we show that for big  $n$ , this kind of behavior is typical, i.e. the related pairs go from corner to corner also locally.

### 5.2.2 Local properties

Fix  $\Delta > 0$  and denote  $\alpha := 1 + \frac{\Delta}{2}$ ,  $\beta := 1 + \Delta$  and  $l(n) := \frac{\alpha}{p}n$ . We consider the events

$$F_n^x := \{n \leq \sum_{i=1}^l D_i^x \leq \beta n\}, \quad F_n^y := \{n \leq \sum_{i=1}^l D_i^y \leq \beta n\}, \quad F_n = F_n^x \cap F_n^y.$$

Let us estimate  $(F_n^x)^c$ .

$$(F_n^x)^c = \left\{ \sum_{i=1}^l D_i^x < n \right\} \cup \left\{ \sum_{i=1}^l D_i^x > \beta n \right\}.$$

By Hoeffding's inequality

$$P\left(\sum_{i=1}^l D_i^x - pl < n - pl\right) \leq \exp\left[-2p \frac{(1 - \alpha)^2}{\alpha} n\right]$$

$$P\left(\sum_{i=1}^l D_i^x - pl > \beta n - pl\right) \leq \exp\left[-2p \frac{(\beta - \alpha)^2}{\alpha} n\right].$$

Since

$$\frac{(\beta - \alpha)^2}{\alpha} = \frac{(1 - \alpha)^2}{\alpha} = \frac{\Delta^2}{2(2 + \Delta)} =: a(\Delta),$$

it holds

$$P(F_n^c) \leq 4 \exp[-pan]. \quad (5.5)$$

Consider the sequences  $X_1, X_2, \dots, Y_1, Y_2, \dots$  that satisfy condition R. Suppose  $X_i$  and  $Y_j$  are related and  $i \leq n$ . When  $F_n^x$  holds, then the ancestor of  $X_i$  is at most  $l$ , i.e.  $a^x(i) \leq m$ . Since  $X_i$  and  $Y_j$  are related,  $a^x(i) = a^y(j) =: a$ . If  $F_n^y$  holds, we have  $\sum_{i=1}^a D_i^y \leq \sum_{i=1}^l D_i^y \leq \beta n$ , implying that  $j \leq \beta n$ . By symmetry, the roles of  $i$  and  $j$  can be changed.

The 2-dimensional process  $(X_1, Y_1), (X_2, Y_2), \dots$  is regenerative with respect to the times  $(\tau_k^x, \tau_k^y)$ , i.e.

$$(X_{\tau_k^x+1}, Y_{\tau_k^y+1}), (X_{\tau_k^x+2}, Y_{\tau_k^y+2}), \dots \quad (5.6)$$

has the same law as  $(X_1, Y_1), (X_2, Y_2), \dots$ . The  $Z$ -process for (5.6) is  $Z_{a_k+1}, Z_{a_k+2}, \dots$ . In the following  $n' \ll n$ . Let

$$F(k, n') = \bigcup_{n \geq n'} \left\{ n \leq \sum_{i=1}^{l(n)} D_{a_k+i}^x, \sum_{i=1}^{l(n)} D_{a_k+i}^y \leq (1 + \Delta)n \right\}$$

and

$$H(k, n') := F(k, n') \cap \{n' - i_k(n'), n' - j_k(n') \leq \Delta n'\},$$

where  $i_k(n') + \tau_k^x$  (resp.  $j_k(n') + \tau_k^y$ ) is the last  $\tau_l^x$  (resp.  $\tau_l^y$ ) before  $\tau_k^x + n'$  (resp.  $\tau_k^y + n'$ ). Since  $(X_1, Y_1), (X_2, Y_2), \dots$  is regenerative with respect to the times  $(\tau_k^x, \tau_k^y)$ , the event  $H(k, n')$  has the same probability as  $G_{n'} \cap (\cup_{n \geq n'} F_{n'})$ . Hence, by (5.5) and (5.4), there exist constants  $K(p, \Delta), b(\Delta, p)$  that depend on  $\Delta$  and  $p$  such that

$$P(H^c(k, n')) \leq 4 \sum_{n \geq n'} \exp[-an] + 2 \exp[\Delta \ln(1 - p^2)n'] \leq K(\Delta, p) \exp[-b(\Delta, p)n']. \quad (5.7)$$

Let  $H_n(n', \Delta)$  denote the event that for every  $k$  that satisfies  $\max\{\tau_k^x, \tau_k^y\} \leq n$ ,  $H(k, n')$  holds. Formally,

$$H_n(n', \Delta) := \bigcup_{i=0}^n \left( \{K = i\} \cap \left( \bigcap_{k=0}^i H(k, n') \right) \right),$$

where

$$K = \arg \max_{k=0,1,\dots} \{ \max\{\tau_k^x, \tau_k^y\} : \max\{\tau_k^x, \tau_k^y\} \leq n \}. \quad (5.8)$$

**Proposition 5.1** *When  $(X, Y)$  are such that  $H_n(n', \Delta)$  holds, then in 2-dimensional representation the following is true:*

1. *if  $(i, j)$  is a related pair satisfying  $i \leq n, j \leq n$  and  $(i', j')$  is another related pair such that  $i' \leq i + n'$ , then  $j' \leq j + (1 + \Delta)n'$ ;*
2. *if  $(i, j)$  is a related pair satisfying  $i \leq n, n' \leq n$  and  $(i', j')$  is another related pair such that  $j' \leq j + n'$ , then  $i' \leq i + (1 + \Delta)n'$ ;*

3. if  $(i, j)$  is a related pair satisfying  $i \leq n, j \leq n$  and  $(i', j')$  is another related pair such that  $i' \geq i + n'$ , then  $j' \leq j + (1 + \Delta)(i' - i)$ ;
4. if  $(i, j)$  is a related pair satisfying  $i \leq n, j \leq n$  and  $(i', j')$  is another related pair such that  $j' \geq j + n'$ , then  $i' \leq i + (1 + \Delta)(j' - j)$ ;
5. if  $(i, j)$  is a related pair satisfying  $i \leq n, j \leq n$  and  $(i', j')$  is another related pair such that  $j' \geq j + n'$  and  $i' \geq i + n'$ , then

$$\frac{(j' - j)}{(1 + \Delta)} \leq i' - i \leq (1 + \Delta)(j' - j), \quad \frac{(i' - i)}{(1 + \Delta)} \leq j' - j \leq (1 + \Delta)(i' - i)$$

6. if  $(i, j)$  is a related pair satisfying  $n' \leq i \leq n$  and  $n' \leq j \leq n$ , then

$$\frac{i}{1 + \Delta} \leq j \leq (1 + \Delta)i, \quad \frac{j}{1 + \Delta} \leq i \leq (1 + \Delta)j;$$

7. if  $(i, j)$  is a related pair satisfying  $i \leq n, j \leq n$  then there exists another related pair  $(i', j')$  such that  $i' \geq i + n'(1 - \Delta)$ ;
8. if  $(i, j)$  is a related pair satisfying  $i \leq n, j \leq n$  then there exists another related pair  $(i', j')$  such that  $j' \geq j + n'(1 - \Delta)$ .

**Proof.** Properties 1 – 6 follow from the event  $F(k, n')$ . This event states that if  $(i, j)$  and  $(i', j')$  are related pairs such that  $i' - i = n$ , then the following 2 statements hold: 1) If  $n \leq n'$ , then  $(j' - j) \leq (1 + \Delta)n'$ . This proves 1. 2) If  $n \leq n'$ , then  $(j' - j) \leq (1 + \Delta)n$ . This proves 3. The roles of  $i$  and  $j$  can be changed. This proves 2. and 4. Now 5. follows from 3. and 4. and 6. is a special case of 5. with  $i = j = 0$  (recall that  $H_n(n')$  includes the case  $k = 0$ ). The properties 7 and 8 follow from the definition  $H(k, n')$ . ■

Thus, if  $n'$  is relatively small in comparison with  $n$  and  $\Delta$  is small, too, then 6. means that in 2-dimensional graphs the related pairs  $(\tau_k^x, \tau_k^y)$ ,  $k = 1, \dots, K$  are located between the lines  $(1 + \Delta)^{-1}i$  and  $(1 + \Delta)i$  except, perhaps, the first square of size  $n'$ . The property 5. says that the pairs  $(\tau_{k+l}^x - \tau_k^x, \tau_{k+l}^y - \tau_k^y)$   $l = 1, 2, \dots, K - k$  have the same property. The properties 7 and 8 basically say that every  $n' \times n'$ -square (inside the big  $n \times n$  square) that has a related pair in a lower left corner has another related pair (almost) in the upper right corner. So, if  $F_n, G_n$  and  $H_n(n', \Delta)$  all hold (with relatively small  $n'$ ), then the related pairs almost follow the diagonal.

We shall show now that  $P(H_n(n', \Delta)) \rightarrow 1$ , if  $n' = A \ln n$  for a suitable chosen  $A < \infty$ . Indeed, since the random variable  $K$  takes values in  $\{0, 1, \dots, n\}$ , we have

$$H_n(n', \Delta) \supset \bigcup_{i=0}^n \left( \{K = i\} \cap \left( \bigcap_{k=0}^i H(k, n') \right) \right) = \bigcap_{k=0}^n H(k, n').$$

Use (5.7) to see that there exists a constant  $E$  depending on  $\Delta$  and  $p$  such that

$$P(H_n(n', \Delta)^c) \leq (n + 1)P(H^c(k, n')) \leq K(n + 1) \exp[-bn'] = En^{1-bA} \rightarrow 0, \quad (5.9)$$

if  $A(\Delta, p)$  is big enough.

### 5.3 Main theorems

Let  $\Delta' > 0$  and define

$B_k(\tilde{n}, \tilde{m}) := \{\text{every optimal alignment of } X_{\tau_k^x+1}, \dots, X_{\tau_k^x+\tilde{n}} \text{ and } Y_{\tau_k^y+1}, \dots, Y_{\tau_k^y+\tilde{m}}$   
contains a related pair}

$$B_k^1(n') := \bigcap_{n' \leq \tilde{n} \leq \tilde{m} \leq \tilde{n}(1+\Delta')} B_k(\tilde{n}, \tilde{m}), \quad B_k^2(n') := \bigcap_{n' \leq \tilde{m} \leq \tilde{n} \leq \tilde{m}(1+\Delta')} B_k(\tilde{n}, \tilde{m})$$

$$B_k(n') := B_k^1(n') \cap B_k^2(n').$$

By the regenerativity and Lemma 5.2, for every  $k$ ,  $P(B_k(\tilde{n}, \tilde{m})) \geq 1 - \exp[-\tilde{n}k_3]$ , provided  $\tilde{n}(1 + \Delta') \geq \tilde{m} \geq \tilde{n} \geq n_o$  and  $\Delta' > 0$  is small enough to satisfy the assumptions. Thus, there exists a constant  $B$  such that

$$P(B_k^2(n')) = P(B_k^1(n')) \geq 1 - \sum_{\tilde{n} \geq n'} e^{-k_3 \tilde{n}} \geq 1 - \frac{B}{2} e^{-k_3 n'}, \quad n' \geq n_o.$$

implying that

$$P(B_k(n')) \geq 1 - B e^{-k_3 n'}, \quad n' \geq n_o.$$

Let  $B_n(n')$  be the event that for every  $k$  that satisfies  $\max\{\tau_k^x, \tau_k^y\} \leq n$ ,  $B(k, n')$  holds. Formally,

$$B_n(n', \Delta') := \bigcup_{i=0}^n \left( \{K = i\} \cap \left( \bigcap_{k=0}^i B(k, n') \right) \right),$$

where  $K$  is as in (5.8). We know that for  $n' = A' \ln n \geq n_o$ ,

$$P(B_n(n')^c) \leq (n+1)P(B_k^c(n')) \leq B(n+1) \exp[-k_3 n'] = E' n^{1-k_3 A'} \rightarrow 0, \quad (5.10)$$

if  $A'(\Delta')$  is big enough and  $E'(\Delta')$  is a constant.

**Lemma 5.3** *Let  $\Delta > 0$  and assume that  $H_n(n', \Delta) \cap B_n(n', 2\Delta)$  holds. Let  $u, v$  be two arbitrary optimal alignments of  $X$  and  $Y$  that satisfy condition **R**. Then, for every pair  $(i, j)$  of (the 2-dimensional representation of)  $u$ , there exists a pair  $(i^v, j^v)$  of  $v$  such that*

$$\max\{|i - i^v|, |j - j^v|\} \leq n'(1 + \Delta). \quad (5.11)$$

**Proof.** Fix  $\Delta' = 2\Delta > 0$  and let  $v$  be an optimal alignment of  $X$  and  $Y$ . Denote by  $(i_1, j_1), (i_2, j_2), \dots, (i_{K(v)}, j_{K(v)})$  the related pairs of  $v$ . Thus  $(X_{i_k}, Y_{j_k})$  is a related pair and  $v(i_k) = j_k$ , i.e. the pair is included into the alignment  $v$ . Clearly  $i_k \geq \tau_k^x$ ,  $j_k \geq \tau_k^y$  and, unlike  $(\tau_k^x, \tau_k^y)$ ,  $(i_k, j_k)$  depend on  $v$ . Let  $i_0 := j_0 := 0$  and  $i_{K(v)+1} := j_{K(v)+1} := n + 1$ .

By assumption,  $H_n(n', \Delta)$  holds. Then for every  $0 \leq k \leq K(v)$ ,

$$\min\{i_{k+1} - i_k, j_{k+1} - j_k\} \leq n' \quad (5.12)$$

$$\max\{i_{k+1} - i_k, j_{k+1} - j_k\} \leq n'(1 + \Delta) \quad (5.13)$$

Let us show that (5.12) and (5.13) hold. Suppose there exists  $k$  such that (5.12) fails. The pairs  $(i_k, j_k)$  and  $(i_{k+1}, j_{k+1})$  are both in  $v$ . Since  $v$  is optimal, the restriction of  $v$  between

$$X_{i_{k+1}, \dots, X_{i_{k+1}-1}}, \quad \text{and} \quad Y_{i_{k+1}, \dots, Y_{i_{k+1}-1}}$$

must be optimal as well. Denote  $\tilde{n} = i_{k+1} - 1 - i_k$  and  $\tilde{m} = j_{k+1} - 1 - j_k$ . If (5.12) does not hold, then  $\tilde{m}, \tilde{n} \geq n'$ . Suppose, without loss of generality that  $\tilde{m} \geq \tilde{n}$ . Since  $H_n(n')$  holds, then the property 3) of Proposition 5.1 states that  $(\tilde{m} + 1) \leq (\tilde{n} + 1)(1 + \Delta)$  implying that  $\tilde{m} \leq \tilde{n}(1 + \Delta')$ . Therefore, we have that the sequences

$$X_{i_{k+1}, \dots, X_{i_k + \tilde{n}}}, \quad \text{and} \quad Y_{i_{k+1}, \dots, Y_{i_k + \tilde{m}}}$$

with  $n' \leq \tilde{n} \leq \tilde{m} \leq \tilde{n}(1 + \Delta')$  have an optimal alignment that contains no related pair. This contradicts  $B_n(n', \Delta')$ . If (5.12) holds, then the properties 1 and 2 of Proposition 5.1 prove (5.13).

Hence, if  $H_n(n', \Delta) \cap B_n(n', \Delta')$  holds, then (5.13) holds, implying that for every optimal alignment  $v$ , the maximum-norm between two consecutive related pairs in  $v$  is at most  $n'(1 + \Delta)$ . Let  $u$  and  $v$  be now two optimal alignments and let  $(i_k^u, j_k^u)$   $k = 1, \dots, K(u)$  and  $(i_k^v, j_k^v)$ ,  $k = 1, \dots, K(v)$  denote the related pairs of  $u$  and  $v$ , respectively. Let  $(i, j)$  be a (not necessarily related) pair in  $u$  i.e.  $u(i) = j$ . There exists  $0 \leq k \leq K(u)$  such that  $i_k^u \leq i \leq i_{k+1}^u$  and  $i_{k+1}^u - i_k^u \leq n'(1 + \Delta)$ . We consider 2 cases separately:

1) Suppose there exists a related pair of  $v$ ,  $(i_l^v, j_l^v)$  such that  $i_k^u \leq i_l^v \leq i_{k+1}^u$ . This means that  $|i_l^v - i| \leq n'(1 + \Delta)$ . Because  $(i_k^u, j_k^u), (i_l^v, j_l^v), (i_{k+1}^u, j_{k+1}^u)$  are related pairs, we have that  $j_k^u < j_l^v < j_{k+1}^u$ . Since  $(i_k^u, j_k^u), (i, j), (i_{k+1}^u, j_{k+1}^u)$  are aligned, we have  $j_k^u < j < j_{k+1}^u$ . Because  $(i_k^u, j_k^u)$  are related and  $H_n(n', \Delta)$  holds, by (5.12) and (5.13), we have  $j_{k+1}^u - j_k^u \leq n'(1 + \Delta)$ , implying that  $|j - j_l^v| \leq n'(1 + \Delta)$ . Hence, with  $i^v = i_l^v$  and  $j^v = j_l^v$ , (5.11) holds.

2) Suppose there exists no  $i_l^v$  such that  $i_k^u < i_l^v < i_{k+1}^u$ . However, there exists  $0 \leq l \leq K(v)$  such that  $i_l^v \leq i \leq i_{l+1}^v$  and  $i_{l+1}^v - i_l^v \leq n'(1 + \Delta)$ . Hence  $i_l^v < i_k^u < i < i_{k+1}^u < i_{l+1}^v$  and  $j_l^v < j_k^u < j < j_{k+1}^u < j_{l+1}^v$ . By (5.12) and (5.13), again,  $j_{l+1}^v - j_l^v \leq n'(1 + \Delta)$ . Thus,  $i - i_l^v \leq n'(1 + \Delta)$  and (5.11) holds with  $i^v = i_l^v$  and  $j^v = j_l^v$ . This proves the theorem in  $l \neq 0$ . However, it might be that  $l = 0$ . But since  $i_{l+1}^v - i \leq n'(1 + \Delta)$ , we get that (5.11) also holds with  $i^v = i_{l+1}^v$  and  $j^v = j_{l+1}^v$ . ■

Recall the definition of Hausdorff's distance between alignments  $u$  and  $v$ , both represented as a set of 2-dimensional points. If for an arbitrary element  $(i, j)$  of  $u$ , there exists an element  $(i^v, j^v)$  of  $v$  such that (5.11) and vice versa, then the Hausdorff's distance between  $u$  and  $v$  with respect to the maximum norm is at most  $n'(1 + \Delta)$ . The Hausdorff's distance between  $u$  and  $v$  with respect to the  $l_2$ -norm is  $\sqrt{2}n'(1 + \Delta)$ . This gives our first main result.

**Theorem 5.1** *Let  $X$  and  $Y$  be related. Let  $u, v$  be the (2-dimensional representations of) lowest and highest alignments of  $X$  and  $Y$ . Assume (5.1). Then there exists  $C < \infty$  such that, for  $n$  big enough,*

$$P(h(u, v) > C \ln n) \leq Dn^{-1}, \quad (5.14)$$

where  $h$  is Hausdorff's distance with respect to maximum or  $l_2$  norm and  $D$  is a constant.

**Proof.** Choose  $1 > \Delta' > 0$  so small that (5.3) holds and take  $\Delta := \Delta'/2$ . Let  $A(\Delta)$  be such that  $bA = 2$ , where  $b$  is as in (5.9). Let  $A'(\Delta')$  be such that  $k_3A' = 2$ , where  $k_3$  is as in (5.10). Take  $A_1 = \max\{A, A'\}$  and define  $n' := \sqrt{2}A_1 \ln n$ . Let  $E_n := \{h(u, v) > 2n'\}$ . Since  $(1 + \Delta) < 2$ , from lemma 5.3, we know that

$$H_n(n', \Delta) \cap B_n(n', 2\Delta) \subset E_n.$$

Hence, from (5.9) and (5.10), for  $n$  big enough,

$$P(E_n^c) \leq P(H_n^c(n', \Delta)) + P(B_n^c(n', 2\Delta)) \leq En^{1-bA_1} + E'n^{1-k_3A_1} \leq (E + E')n^{-1}.$$

So, with  $C = 2\sqrt{2}A_1$ , the lemma holds. ■

In Theorem 5.1, we used the 2-dimensional representation of alignments, so an alignment were identified with a finite set of points. In the alignment graph, these points are joined by a line. We consider the highest and lowest alignment graphs, and we are interested in the maximal vertical (horizontal) distance between these 2 piecewise linear curves. This maximum is called vertical (horizontal) distance between lowest and highest alignment graphs. The next lemma shows that under the assumptions of lemma 5.3, the vertical distance between 2 alignment graphs is bounded above by  $2h(u, v)$ .

We need some notations and conventions. Let  $u$  be an alignment between  $X$  and  $Y$ . Thus  $u$  is a set of pairs  $u = \{(i_1^u, j_1^u), \dots, (i_L^u, j_L^u)\}$ , where  $i_1^u < \dots < i_L^u$  and  $j_1^u < \dots < j_L^u$ ,  $j_l^u = u(i_l^u)$ , (here  $(i_l^u, j_l^u)$  stands for any pair of  $u$ , not necessarily related, so  $L$  is the length of LCS of  $X$  and  $Y$ ). Let  $U$  be the corresponding graph. Formally,  $U$  is a piecewise linear mapping from  $[i_1^u, i_L^u] \rightarrow [1, n]$  such that  $U(i_l) = j_l$ ,  $l = 1, \dots, L$ .

**Lemma 5.4** *Let  $\Delta > 0$  and assume that  $H_n(n', \Delta) \cap B_n(n', 2\Delta)$  holds. Let  $u, v$  be two arbitrary optimal alignments of  $X$  and  $Y$  that satisfy condition R. Let  $U, V$  be the corresponding alignment graphs defined on  $[i_1^u, i_L^u]$  and  $[i_1^v, i_L^v]$ , respectively. Then*

$$\max_{x \in [i_1, i_L]} |U(x) - V(x)| \leq 2n'(1 + \Delta), \quad (5.15)$$

where  $i_1 := \max\{i_1^u, i_1^v\}$ ,  $i_L := \min\{i_L^u, i_L^v\}$ .

**Proof.** The proof goes along the same line as the one of lemma 5.3. Let  $(i, j)$  be an arbitrary pair of  $u$ , i.e.  $(i, j) = (i_s^u, j_s^u)$  for some  $1 \leq s \leq k$ . Then there exists two related pairs of  $u$ ,  $(i_k, j_k)$  and  $(i_{k+1}, j_{k+1})$  such that  $i_k^u < i < i_{k+1}^u$  and  $j_{k+1}^u - j_k^u \leq n'(1 + \Delta)$ . Similarly, there exists two related pairs of  $v$ ,  $(i_l^v, j_l^v)$  and  $(i_{l+1}^v, j_{l+1}^v)$  such that  $i_l^v \leq i \leq i_{l+1}^v$  and  $j_{l+1}^v - j_l^v \leq n'(1 + \Delta)$ . It might be that  $i_k^u = 0$  or  $i_{k+1}^u = n + 1$ . It might also be that  $i_l^v = 0$  or  $i_{l+1}^v = n + 1$ . We consider 3 cases separately:

1) Suppose  $i_k^u \leq i_l^v$ . This is the first case as in the proof of lemma 5.3. We know

that then  $|j - j_l^v| \leq n'(1 + \Delta)$ . On the other hand, we know that  $i_k^u \leq i_l^v \leq i \leq i_{l+1}^v$  and  $i \leq i_k^u$  implying that  $j_k^u \leq j_l^v \leq j \leq j_{k+1}^u$  and  $j_l^v \leq j_{l+1}^v$ . Then,  $|j - j_{l+1}^v| \leq (j_{l+1}^v - j_l^v) + (j_{k+1}^u - j_k^u) \leq 2n'(1 + \Delta)$ .

2) Suppose  $i_{l+1}^v \leq i_{k+1}^u$ . This case is similar to the previous one.

3) Suppose  $i_l^v < i_k^u < i < i_{k+1}^u < i_{l+1}^v$ . This corresponds to the case 2) in the proof of lemma 5.3. Then  $|j - j_l^v| \leq n'(1 + \Delta)$  and  $|j - j_{l+1}^v| \leq n'(1 + \Delta)$ .

There are no more options. Recall that  $(i, j)$  is an arbitrary point of  $u$ . The points  $(i_l^v, j_l^v)$  and  $(i_{l+1}^v, j_{l+1}^v)$  are the neighbors of  $(i, j)$ , were the neighborhood is measured along  $i$ -axis. We just showed that the the maximum distance along  $j$ -axis (this is the vertical distance) between  $(i, j)$  and these neighbors is  $2n'(1 + \Delta)$ . Since  $V$  is linear between  $(i_l^v, j_l^v)$  and  $(i_{l+1}^v, j_{l+1}^v)$ , we get that  $|U(i) - V(i)| = |j - V(i)| \leq \max\{|j - j_l^v|, |j - j_{l+1}^v|\} \leq 2n'(1 + \Delta)$ . Therefore,  $\max_{l=1, \dots, L} |U(i_l^u) - V(i_l^u)| \leq 2n'(1 + \Delta)$ . Similarly,  $\max_{l=1, \dots, L} |U(i_l^v) - V(i_l^v)| \leq 2n'(1 + \Delta)$ . Finally, since the vertical distance between two piecewise linear graphs achieves its maximum at the knots  $i_1^u, \dots, i_L^u$  or  $i_1^v, \dots, i_L^v$ , we obtain (5.15). ■

**Theorem 5.2** *Let  $X$  and  $Y$  be related. Let  $U, V$  be the lowest and highest alignment graphs of  $X$  and  $Y$ . Assume (5.1). Then, for  $n$  big enough,*

$$P\left(\sup_{x \in [i_1, i_L]} V(x) - U(x) > 2C \ln n\right) \leq Dn^{-1}, \quad (5.16)$$

where  $[i_1, i_L]$  is as in Lemma 5.4 and the constant  $C$  is the same as in Theorem 5.1.

**Proof.** The proof follows from Lemma 5.4 along the same line as the proof of Theorem 5.1. ■

Theorem 5.2 states that with big probability, the vertical distance between the highest and lowest alignment graphs is at most  $2C \ln n$ , provided  $n$  is big enough. The same holds for horizontal distance. Recall that, according to Theorem 5.1, for Hausdorff's distance between the alignments (but not alignment graphs!) the bound is twice as small.

Theorems 5.1 and 5.2 hold under (5.1). If the marginal distribution of  $X$  and  $Y$  is uniform over a four-letter alphabet, as in Section 3, the condition (5.1) is simply  $H(\gamma_R) < \gamma_R$ . This clearly holds for sufficiently large  $\gamma_R$ . The simulation in Section 3 imply that the value of  $\gamma_R$  is about 0.76. Since  $H(0.76) \approx 0.795$ , this particular  $\gamma_R$  is a bit to small for (5.1) to hold. On the other hand, the simulations clearly indicate that the statements of Theorems 5.1 and 5.2 also hold for this case implying that the assumption (5.1) can be improved. This is matter of the further research.

## References

- [1] Kenneth S. Alexander. The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.*, 4(4):1074–1082, 1994.
- [2] Saba Amsalu, Heinrich Matzinger, and Sergei Popov. Macroscopic non-uniqueness and transversal fluctuation in optimal random sequence alignment. *ESAIM: Probability and Statistics*, 2007. (to appear).

- [3] Richard Arratia and Michael S. Waterman. A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, 4(1):200–225, 1994.
- [4] R.A. Baeza-Yates, R. Gavaldà, G. Navarro, and R. Scheihing. Bounding the expected length of longest common subsequences and forests. *Theory Comput. Syst.*, 32(4):435–452, 1999.
- [5] Federico Bonetto and Heinrich Matzinger. Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets. *Latin American Journal of Probability and Statistics*, 2:195–216, 2006.
- [6] Václáv Chvatal and David Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probability*, 12:306–315, 1975.
- [7] R. Durbin, S. Eddy, Krogh A., and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
- [8] Niels Hansen. Local alignment of markov chains. *Ann. App. Prob.*, 16(3):1262–1296, 2000.
- [9] Christian Houdre, Jüri Lember, and Heinrich Matzinger. On the longest common increasing binary subsequence. *Comptes Rendus Mathématique*, 343(9):589–594, 2006.
- [10] Marcos A. Kiwi, Martin Loeb, and Jirí Matousek. Expected length of the longest common subsequence for large alphabets. *Advances in Mathematics*, 197(2):480–498, 2005.
- [11] Jüri Lember and Heinrich Matzinger. Standard deviation of the longest common subsequence. 2007. submitted.
- [12] Jüri Lember, Heinrich Matzinger, and Clemont Durringer. Deviation from mean in sequence comparison with a periodic sequence. *Latin American Journal of Probability and Statistics*, 3:1–29, 2007.
- [13] Paul Shields. *The ergodic theory of discrete sample paths*. AMS, Providence, 1996.
- [14] David Siegmund and Behjamen Yakir. Approximate p-values for local sequence alignments. *Ann. Stat.*, 28(3):657–680, 2000.
- [15] Michael J. Steele. An Efron-Stein inequality for non-symmetric statistics. *Annals of Statistics*, 14:753–758, 1986.
- [16] Michael S. Waterman. Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 344:383–390, 1994.
- [17] Michael S. Waterman. *Introduction to Computational Biology*. Chapman & Hall, 1995.
- [18] Michael S. Waterman and M. Vingron. Sequence comparison significance and Poisson approximation. *Statistical Science*, 9(3):367–381, 1994.