Bernoulli **20**(3), 2014, 1292–1343
DOI: 10.3150/13-BEJ522

Optimal alignments of longest common subsequences and their path properties

JÜRI LEMBER[1], HEINRICH MATZINGER[2] and ANNA VOLLMER[3]

[1]*University of Tartu, Liivi 2-513 50409, Tartu, Estonia. E-mail: jyril@ut.ee*
[2]*Georgia Tech, School of Mathematics, Atlanta, GA 30332-0160, USA.*
E-mail: matzing@math.gatech.edu
[3]*University of Plymouth, School of Computation and Mathematics, Plymouth, PL4 8AA, Devon, UK.*
E-mail: anna-lisa.vollmer@plymouth.ac.uk

We investigate the behavior of optimal alignment paths for homologous (related) and independent 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 homologous (having the same common ancestor) sequences have typically the distance between the extremal alignments of much smaller size than independent sequences. In particular, the simulations suggest that for the homologous 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.

Keywords: longest common subsequence; optimal alignments; homologous sequences

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 is a special case of a sequence alignment that is a very important tool in computational biology, used for comparison of DNA and protein sequences (see, e.g., [3,7,9,21,22]). 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 information 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 γ such that

$$\frac{L_n}{n} \to \gamma \qquad \text{a.s. and in } L_1. \tag{1.1}$$

(see, e.g., [1,15,22]). Referring to the celebrated paper of Chvatal and Sankoff [8], the constant γ 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 γ obviously depends on the Bernoulli parameter p. When $p = 0.5$, the various bounds indicate that $\gamma \approx 0.81$ [4,13,19]. For a smaller p, γ 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 [8], it was conjectured that for Bernoulli parameter $p = 0.5$, the variance is of order $o(n^{2/3})$. Using an Efron–Stein type of inequality, Steele [19] proved $\mathrm{Var}[L_n] \le 2p(1-p)n$. In [20], Waterman conjectured that $\mathrm{Var}[L_n]$ grows linearly. In series of papers, Matzinger and others prove the Waterman conjecture for different models [6,12,14,17].

Because of relatively rare knowledge 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., [3,20]), because for these alignments approximate p-values were recently calculated [10,18].

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), 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 or homology of the sequences. To measure the differences between various optimal alignment, we consider so-called extremal alignments and study their differences.

Example. Let us consider a practical example to give an insight in what follows. Let $X = \mathrm{ATAGCGT}$, $Y = \mathrm{CAACATG}$. There are two longest common subsequences: AACG and AACT. Thus, $L_7 = 4$. To every longest common subsequence corresponds two optimal alignments. These optimal alignments can be presented as follows:

	A	T	A	C	C			G	T
C	A		A		C	A	T	G	

	A	T	A	C	C			G	T
C	A		A		C		A	T	G

	A	T	A	C	C		G	T	
C	A		A		C	A		T	G

	A	T	A	C	C		G	T	
C	A		A		C		A	T	G

First, two alignments correspond to optimal subsequence AACG, the last two correspond to AACT. In the following, we shall often consider the alignments as the pairs $\{(i_1, j_1), \ldots, (i_k, j_k)\}$, where $X_{i_t} = Y_{i_t}$ for every $t = 1, \ldots, k$. With this notation, the four optimal alignments above are $\{(1,2),(3,3),(5,4),(6,7)\}$, $\{(1,2),(3,3),(4,4),(6,7)\}$, $\{(1,2),(3,3),(5,4),(7,6)\}$ and $\{(1,2),(3,3),(4,4),(7,6)\}$. We now represent every alignment as two-dimensional plot. For the alignments in our example, the two dimensional plots are

G						$*$	
T							
A							
C					$*$		
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

G						$*$	
T							
A							
C				$*$			
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

G							
T							$*$
A							
C					$*$		
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

G							
T							$*$
A							
C				$*$			
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

Putting all four alignment into one graph, we see that on some regions all alignments are unique, but on some region, they vary:

G						$*$	
T							$*$
A							
C				$*$	$*$		
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

In the picture above, the alignment $(1, 2), (3, 3), (4, 4), (6, 7)$ (corresponding to AACG) lies above all others. This alignment will be called *highest alignment*. Similarly the alignment $(1, 2), (3, 3), (5, 4), (7, 6)$ (corresponding to AACT) lies below all others. This alignment will be called *lowest alignment*. The highest and lowest alignment will be called *extremal alignments*.

Thus, the highest (lowest) alignment is the optimal alignment that lies above (below) all other optimal alignments in two-dimensional representation. For big n, we usually align the dots in the two dimensional representation by lines. 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"). In Figure 1, there are extremal alignments of two independent i.i.d. four letter sequences (with uniform marginal distributions) of length $n = 1000$. It is visible that the extremal alignments are rather far from each other, in particular, the maximum vertical and horizontal distances are relatively big.

We call the sequences X and Y unrelated, if they are independent. There are many ways to model the related sequences, the model in the present paper 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. The sequences with common ancestor are called homologous, detecting homology of given sequences is one of the major tasks in modern computational molecular biology [7]. In this paper, we shall call the homologous sequences *related*.

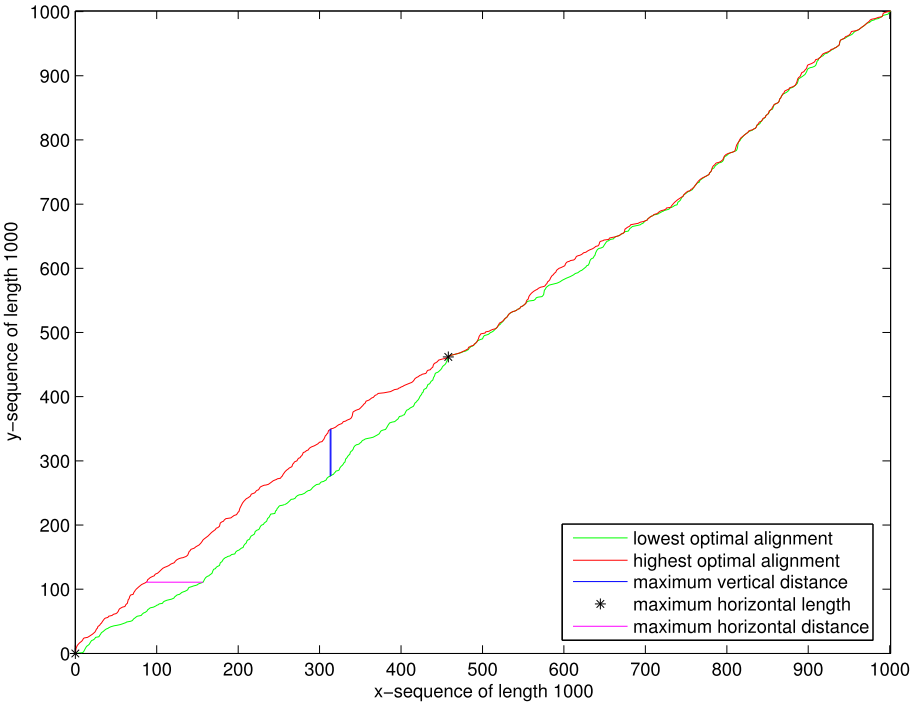


Figure 1. The extremal alignments of two independent i.i.d. four letter sequences.

More precisely, we consider an $\mathcal{A}$-valued i.i.d. process $Z_1, Z_2, \ldots$ that will be referred to as the common ancestor or the ancestor process. A letter Z_i has a probability to mutate according to a transition matrix that does not depend on i. The mutations of the letters are assumed to be independent. After mutations, some letters of the mutated process disappear. The disappearance is modeled via a deletion process $D_1^x, D_2^x, \ldots$ that is assumed to be an i.i.d. Bernoulli sequence with parameter p, that is, $P(D_i^x = 1) = p$. If $D_i^x = 0$, then the ith (possibly mutated letter) disappears. In such a way, a random sequence $X_1, X_2, \ldots$ is obtained. The sequence $Y_1, Y_2, \ldots$ is obtained similarly: the ancestor process $Z_1, Z_2, \ldots$ is the same, but the mutations and deletions (with the same probabilities) are independent of the ones used to generate X-sequence. The formal definition is given in Section 4.1.

Figure 2 presents a typical picture or extremal alignments of two related four-letter sequences (of uniform marginal distribution) of length 949. The sequences in Figure 2, thus, have the same marginal distribution as the ones in Figure 1, but they are not independent any more. Clearly the extremal alignments are close to each other; in particular the maximal vertical and horizontal distance is much smaller than these ones in Figure 1.

Figures 1 and 2 as well as many other similar simulations (see [16]) clearly indicate that for related sequences the differences of optimal alignments are of local nature, whilst for independent sequences they vary much more. This motivates us to find a way to quantify the non-uniqueness

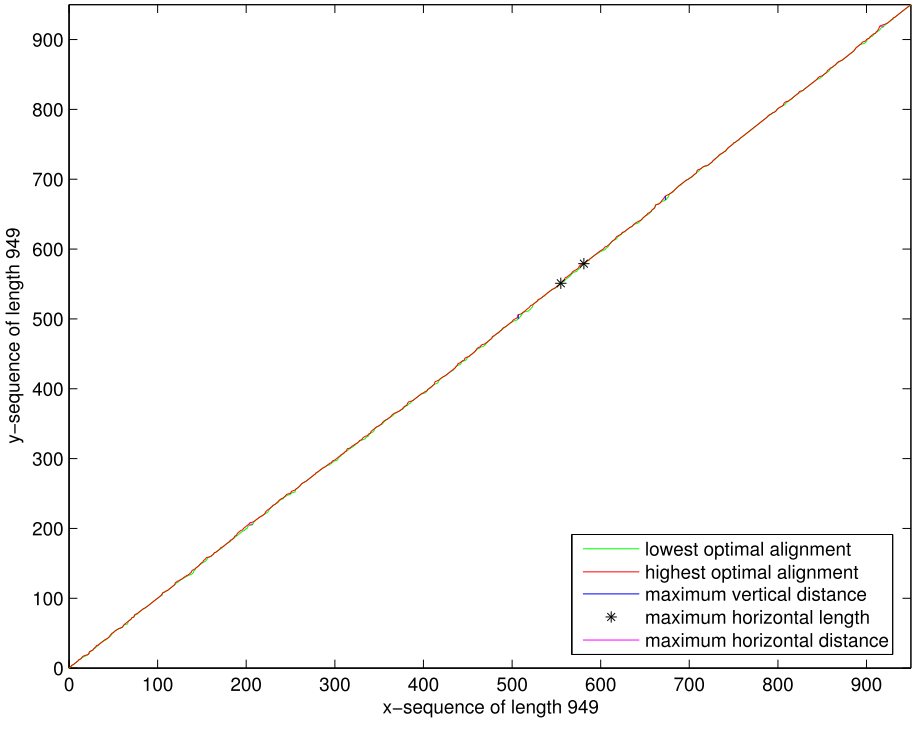


Figure 2. The extremal alignments of two related four letter sequences.

and use the obtained characteristic as a measure of the relatedness. For that we measure the differences of extremal alignments in several ways: the maximal vertical and horizontal distance and Hausdorff's distance (see Section 1.1.2 for formal definition of Hausdorff's distance). The simulations in Section 7 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 main theoretical results about related sequences, Theorems 1.1, 1.2 and 1.3. More specifically, Theorem 1.1 states that under some assumption that never holds for independent sequences, there exist universal constants C and D so that for n big enough,

$$P\big(h_o(H, L) > C \ln n\big) \leq Dn^{-2}.$$

Here, $h_o(H, L)$ stands for a slight modification of Hausdorff's distance between extremal alignments, which we shall call restricted Hausdorff's distance. We conjecture the result also holds for (full) Hausdorff's distance, denoted by h. Note that by Borel–Cantelli lemma, from the inequality above, it follows that

$$P\big(h_o(H, L) \leq C \ln n, \text{eventually } \big) = 1,$$

that is, the ratio $h_o(H, L)/\ln n$ is eventually bounded above by C, a.s. Theorem 1.2 states the similar result with maximal vertical distance instead of Hausdorff's distance.

Theorem 1.3 considers the sequences with random lengths. The expected length of both sequences is n, the randomness comes from the fact that instead of fixing the lengths of both sequences, we fix the length of the common ancestor process. In a sense, this situation is more realistic, since in practice the sequences are hardly of exactly the same lengths; however, when they are related, then the common ancestor must be of the same length for both of the sequences. It turns out that the case of the random lengths the statement of Theorems 1.1 and 1.2 hold with (full) Hausdorff's distance h instead of the restricted Hausdorff's distance h_o. More precisely, Theorem 1.3 states that under the same assumptions as in Theorem 1.1, there exist universal constants C_r and D_r so that

$$P\big(h\big(H^r, L^r\big) > C_r \ln n\big) \le D_r n^{-2},$$

where $h(H^r, L^r)$ stands for (full) Hausdorff's distance between extremal alignments of random-length sequences.

Another measure could be the length of the biggest non-uniqueness stretch, that is, the (horizontal) length between $*$'s. The simulations in Section 7 show that the length of the biggest non-uniqueness stretch behaves similarly: the growth is almost linear for the independent and logarithmic for the related sequences. The latter has not been proven formally in this paper, but we conjecture that it can be done using similar arguments as in the proof of Theorems 1.1, 1.2 and 1.3.

1.1. The organization of the paper and the main results

1.1.1. *Preliminary results*

The paper is organized as follows. In Section 2, the necessary notation is introduced and the extremal alignments are formally defined and proven to exist (Proposition 2.1). Also some properties of the extremal alignments are proven. The section also provides some combinatorial bounds needed later.

Section 3 considers the case, where X and Y are independent. The main result of the section is Theorem 3.1 that states for independent sequences the Chvatal–Sankoff constant γ satisfies the inequality

$$\gamma \log_2 p_o + 2(1-\gamma)\log_2 q + 2h(\gamma) \ge 0, \tag{1.2}$$

where

$$\begin{aligned} p_a &:= P(X_i = a), \qquad q := 1 - \min_a p_a, \\ p_o &:= \sum_{a\in\mathcal{A}} p_a^2, \qquad h(p) := -p\log_2 p - (1-p)\log_2(1-p), \end{aligned} \tag{1.3}$$

that is, h is the binary entropy function. The equality $\gamma \log_2 p_o + 2(1-\gamma)\log_2 q + 2h(\gamma) = 0$ has two solutions, hence, as a byproduct, (1.2) gives (upper and lower) bounds to unknown γ. These

Table 1. Upper bounds to Chvatal–Sankoff constant via inequality (1.2)

K	2	3	4	5	6	7	8
$\bar{\gamma}$	0.866595	0.786473	0.729705	0.686117	0.650983	0.621719	0.596756
$\hat{\gamma}$	0.81	0.72	0.66	0.61	0.57	0.54	0.52

bounds need not to be the best possible bounds, but they are easy and universal in the sense that they hold for any independent model. For example, taking the distribution of X_i and Y_i uniform over the alphabet with K letters (thus $q = 1 - \frac{1}{K}$ and $p_o = \frac{1}{K}$), we obtain the following upper bounds $\bar{\gamma}$ to unknown γ. In the last row of the table, the estimators $\hat{\gamma}$ of unknown γ is obtained via simulations. It is interesting to note that, independently of K, the upper bound overestimates γ about the same amount. We also obtain the lower bounds, but for these model the lower bounds are very close to zero and therefore not informative.

In Section 4, the preliminary results for related (homologous) sequences are presented. In Section 4.1, the formal definition of related sequences are given. Our definition of relatedness is based on the existence of common ancestor. Hence, our model models the homology in most natural way. In our model, the related sequences $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ both consists of i.i.d. random variables, but the sequences are, in general, not independent. Independence is a special case of the model so that all results for related sequences automatically hold for the independent ones. It is also important to note that (unless the sequences are independent), the two dimensional process $(X_1, Y_1), (X_2, Y_2), \ldots$ is not stationary, hence also not ergodic. Hence, for the related sequences ergodic theorems cannot be automatically applied. In particular, Kingsman's subadditive ergodic theorem cannot be applied any more to prove the convergence (1.1). This convergence as well as the corresponding large deviation bound has been proven in Section 4.2. Since we often consider the sequences of unequal length, instead of (1.1), we prove a more general convergence (Proposition 4.1):

$$\frac{L(X_1, \ldots, X_n; Y_1, \ldots, Y_{\lfloor na \rfloor})}{n} \to \gamma_{\mathrm{R}}(a), \qquad \text{a.s.} \tag{1.4}$$

Here $\gamma_{\mathrm{R}}(a)$ is a constant. We shall denote $\gamma_{\mathrm{R}}(1) =: \gamma_{\mathrm{R}}$. From (1.4), it follows that for any $a > 0$,

$$\gamma_{\mathrm{R}}(a) = a\gamma_{\mathrm{R}}\left(\frac{1}{a}\right).$$

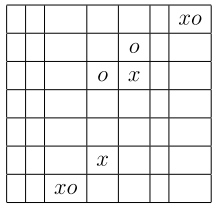
If the sequences are independent and $a = 1$, then $\gamma_{\mathrm{R}}(a) = \gamma$. Corollary 4.1 postulates the corresponding large deviation result, stating that for every $\Delta > 0$ there exists $c > 0$ such that for every n big enough

$$P\left(\left|\frac{L(X_1, \ldots, X_n; Y_1, \ldots, Y_{\lfloor na \rfloor})}{n} - \gamma_{\mathrm{R}}(a)\right| > \Delta\right) \leq \exp[-cn]. \tag{1.5}$$

In the Appendix, it is proven that $\gamma_{\mathrm{R}}(a) > \gamma_{\mathrm{R}}$, if $a > 1$ and $\gamma_{\mathrm{R}}(a) < \gamma_{\mathrm{R}}$, if $a < 1$ (Lemma A.1). That result together with (1.5) (obviously (1.4) follows from (1.5)) are the basic theoretical tools for proving the main results of the paper, Theorems 1.1, 1.2 and 1.3.

1.1.2. (*Restricted*) *Hausdorff's distance and the main results*

Definition of (*restricted*) *Hausdorff's distance.* 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:



(1.6)

To overcome the problem, we measure the distance between two alignments also in terms of Hausdorff's distance. More precisely, let $U, V \subset \{1, \dots, n\}^2$, be two alignments, both represented as sets of two-dimensional points. The *Hausdorff's distance between U and V* is:

$$h(U, V) := \max\Big\{\sup_{u\in U}\inf_{v\in V} d(u, v), \sup_{v\in V}\inf_{u\in U} d(u, v)\Big\},$$

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 Eucledian metric). We remark that Hausdorff's distance is defined for any kind of sets. For the alignments in (1.6), the Hausdorff's distance is obviously 1 (if d were Euclidean, the Hausdorff's distance would be $\sqrt{2}$).

Let now, for every n, $\alpha_n \in (0, 1)$ be fixed, and we define the subset $U_o \subseteq U$ consisting of those elements (i, j) of U that have the first coordinate at least $n\alpha_n$ further from n: $i \le n(1 - \alpha_n)$. Similarly, the subset $V_o \subset V$ is defined. Formally, thus

$$U_o := \big\{(i, j) \in U : i \le n(1 - \alpha_n)\big\}, \qquad V_o := \big\{(i, j) \in V : i \le n(1 - \alpha_n)\big\}.$$

The *restricted Hausdorff's distance* between U and V is defined as follows:

$$h_o(U, V) := \max\Big\{\sup_{u\in U_o}\inf_{v\in V} d(u, v), \sup_{v\in V_o}\inf_{u\in U} d(u, v)\Big\},$$

where d is a distance in $\mathbb{R}^2$. Clearly $h_o(U, V) \le h(U, V)$. Since in our case U and V are alignments so that different points have different coordinates, the definition of h_o can be (somehow loosely) interpreted as a fraction α_n of both alignments are left out when applying maximum in Hausdorff's distance. We shall consider the case $\alpha_n \to 0$. Hence, the proportion of points left out decreases as n grows.

Sequences with fixed length. We now state our main theorems for the sequences of fixed lengths. Recall the definition of p_o and q from (1.3). Let

$$\bar{p} := \max_{a\in\mathcal{A}} p_a, \qquad \bar{q} := 1 - \min_{a,b\in\mathcal{A}} P(X_1 = a | Y_1 = b, p = 1). \tag{1.7}$$

Here $P(X_1 = a|Y_1 = b, p = 1)$ is the conditional probability given that no deletion occurs, or, in other words X_1 and Y_1 have the common ancestor (see Section 4.1 for formal definition). Finally, let

$$\rho := \frac{p_0\bar{q}}{\bar{p}q}.$$

In the following theorems, $h_o(L, H)$ stands for the restricted Hausdorff's distance between alignments L and H, both represented as a set of 2-dimensional points. Recall that Hausdorff's distance could be defined with the help in any metric in $\mathbb{R}^2$. In the following, we shall consider both maximum and l_2-norms. Throughout the paper, we shall use $\wedge$ and $\vee$ for min and max, respectively.

Theorem 1.1. *Let X and Y be related. Let L, H be the (2-dimensional representations of) lowest and highest alignments of X and Y. Assume*

$$\gamma_{\mathrm{R}} \log_2 \bar{p} + (1-\gamma_{\mathrm{R}}) \log_2(q\bar{q}) + \big((1-\gamma_{\mathrm{R}}) \wedge \gamma_{\mathrm{R}}\big) \log_2(\rho \vee 1) + 2h(\gamma_{\mathrm{R}}) < 0. \tag{1.8}$$

Then there exist positive constants $M, C, D < \infty$ such that, for n big enough,

$$P\big(h_o(L, H) > C \ln n\big) \le Dn^{-2}, \tag{1.9}$$

where h_o is defined with

$$\alpha_n := M\sqrt{\frac{16 \ln n}{pn}}$$

and Hausdorff's distance is defined using maximum norm. If h_o is defined with respect to l_2 norm, then (1.9) holds with C replaced by $\sqrt{2}C$.

For independent sequences $\bar{q} = q$, thus $\rho = \frac{p_o}{\bar{p}} \le 1$. Then also $\gamma = \gamma_{\mathrm{R}}$ so that

$$\begin{aligned}&\gamma_{\mathrm{R}} \log_2 \bar{p} + (1-\gamma_{\mathrm{R}}) \log_2(q\bar{q}) + \big((1-\gamma_{\mathrm{R}}) \wedge \gamma_{\mathrm{R}}\big) \log_2(\rho \vee 1)\\ &\quad = \gamma \log_2 \bar{p} + 2(1-\gamma) \log_2 q \ge \gamma \log_2 p_0 + 2(1-\gamma) \log_2 q \ge -2h(\gamma).\end{aligned}$$

The last inequality follows from (1.2) (Theorem 3.1). Hence, for unrelated (independent) sequences the condition (1.8) fails. It does not necessarily mean that in this case (1.9) holds not true, but based on our simulations in Section 7 we conjecture that this is indeed the case.

In Theorem 1.1, 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 two piecewise linear curves. This maximum is called vertical (horizontal) distance between lowest and highest alignment graphs. The following theorem is stated in terms of vertical distance. Clearly the same result holds for horizontal distance as well. In the theorem, we shall also use the letters L and H, but now they stand for extremal alignment graphs rather than for the alignments as the sets of the points. Since an alignment and the corresponding alignment graph are very closely related, we hope that the notation is not too ambiguous and the difference will be clear from the context.

Theorem 1.2. *Let X and Y be related. Let L, H be the lowest and highest alignment graphs of X and Y. Assume* (1.8). *Then for n big enough,*

$$P\Big(\sup_{x\in[0,n(1-\alpha_n)]} H(x) - L(x) > 2C\ln n\Big) \le Dn^{-2}, \tag{1.10}$$

where the constants C, D and α_n are the same as in Theorem 1.1.

Hence, Theorems 1.1 and 1.2 state that when γ_{R} is sufficiently bigger than (corresponding) γ, then the distance between the extremal alignment (either measured with restricted Hausdorff's metric or using alignment graphs) grows no faster than logarithmically in n. Clearly, γ_{R} is the bigger the more X and Y are related. Hence, the inequality (1.8) measures the degree of the relatedness – if this is big enough, then the distance between extremal values grows (at most) logarithmically. Theorem 3.1 states that for independent sequence (1.8) fails, so that the assumptions of Theorems 1.1 and 1.2 hold for related sequences, only.

The fact that the distances between extremal alignments are measured with respect to the restricted Hausdorff's distance, that is, so that a small fraction of the alignments left out is obviously a bit disappointing. Technically, this is due to the requirement that both sequences are of the same length. As we shall see, this is not the case when the lengths of the sequences are random. However, as also the simulations in Section 7 suggest, we believe that the results of Theorems 1.1 and 1.2 hold also when h_o is replaced by the (full) Hausdorff's distance h and supremum is taken over $[0, n]$.

Theorems 1.1 and 1.2 are proven in Section 5. The proof is based on the observation that under (1.8) the probability that the sequences with length about n do not contain any related pairs is exponentially small in n (Lemmas 5.1 and 5.2). Section 5.2 studies the location of the related pairs in two dimensional representation. It turns out that with high probability, the gaps between them are no longer then $A\ln n$, where A is suitable big constant. Applying these properties together with Lemma 5.2, we obtain that every optimal alignment, including the extremal ones, cannot be far away from the related points, since otherwise it would have a long piece without any related pair contradicting Lemma 5.2. This argument is formalized in Lemma 5.3 and Lemma 5.4 in Section 5.3. The formal proof of Theorems 1.1 and 1.2 are given in Sections 5.4 and 5.5, respectively.

The sequences with random length. The related sequences are defined as follows: there is a common ancestor process $Z_1, Z_2, \ldots$ consisting on $\mathcal{A}$-valued i.i.d. random variables. Every letter Z_i has a probability to mutate according to a transition matrix that does not depend on i. The

mutations are independent of each other. After mutations, some of the letters disappears. Thus, to every letter Z_i, there is associated a Bernoulli random variables D_i^x with $P(D_i^x = 1) = p$. When $D_i^x = 0$, then the corresponding (mutated) letter disappears. The deletions $D_1^x, D_2^x, \ldots$ are independent and the remaining letters form the sequence $X_1, X_2, \ldots$. The Y sequence is defined in the same way: every ancestor letter Z_i has another random mutation (independent of the all other mutations including the ones that were used to define the X-sequence), and independent i.i.d. deletions D_i^y with the same probability. For more detailed definition, see Section 4.1.

When dealing with the sequences of random length, we consider exactly m ancestors $Z_1, \ldots, Z_m$. Hence after deletions, the length of obtained X-sequence is $n_x := \sum_{i=1}^m D_i^x$ and the length of Y-sequence is $n_y := \sum_{i=1}^m D_i^y$. The expected length of both sequences is thus mp and we choose $m(n) := \frac{n}{p}$ so that the expected length of the both sequences is n. For simplicity, $m(n)$ is assumed to be integer. Thus, we shall consider the sequences $X := X_1, \ldots, X_{n_x}$ and $Y := Y_1, \ldots, Y_{n_y}$ of random lengths. It turns out that mathematically this case is somehow easier so that the counterpart of Theorem 1.3 holds with full Hausdorff's distance h instead of h_o.

Theorem 1.3. *Let X and Y be the related sequences of random lengths. Let L, H be the* (2-*dimensional representation*) *of the highest and lowest alignment. Assume* (1.8). *Then there exist constants C_r and D_r so that*

$$P\big(h(H, L) > C_r \ln n\big) \le D_r n^{-2}, \tag{1.11}$$

where h is the Hausdoff's distance with respect to maximum norm. If h is defined with respect to l_2 norm, then (1.9) *holds with C_r replaced by $\sqrt{2}C_r$.*

From the proofs, it is easy to see that the random length analogue of Theorem 1.2 with $\alpha_n = 0$ holds as well. Theorem 1.3 is proven in Section 6.

Finally, in Section 7, some simulations about the speed of the convergence are studied. The simulation clearly indicate that for related sequences the growth of Hausdorff's and vertical distance is at order of $O(\ln n)$, hence the simulations fully confirm the main results of the paper.

We would like to mention that to our best knowledge, the idea of considering the extremal alignments as a characterization of the homology has not been exploited, although the optimal and sub-optimal alignments have deserved some attention before [2,5]. 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. In a follow-up paper [11], we apply some extremal-alignments based characteristics to the real DNA-sequences, and compare the results with standard sequence-alignment tools like BLAST.

2. Preliminaries

Let $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_y}$ be two sequences of lengths n_x and n_y from finite alphabet $\mathcal{A} = \{0, 1, \ldots, |\mathcal{A}| - 1\}$. Let there exist two subsets of indices $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n_x\}$ and $\{j_1, \ldots, j_k\} \subset \{1, \ldots, n_y\}$ satisfying $i_1 < i_2 < \cdots < i_k$, $j_1 < j_2 < \cdots < j_k$ and $X_{i_1} = Y_{j_1}, X_{i_2} = Y_{j_2}, \ldots, X_{i_k} = Y_{j_k}$. Then $X_{i_1} \cdots X_{i_k}$ is a common subsequence of X and Y and the pairs

$$\big\{(i_1, j_1), \ldots, (i_k, j_k)\big\} \tag{2.1}$$

are (the 2-dimensional representation of) the corresponding alignment. Let

$$L(X_1, \ldots, X_{n_x}; Y_1, \ldots, Y_{n_y})$$

be the biggest k such that there exist such subsets of indices. The longest common subsequence is any common subsequence with length $L(X_1, \ldots, X_{n_x}; Y_1, \ldots, Y_{n_y})$ and any alignment corresponding to a longest common subsequence is called optimal. In the following, we shall often consider the case, where, for some constants $a, b > 0$, $n_x = \lfloor bn \rfloor$, $n_y = \lfloor an \rfloor$. Let us denote

$$L_{bn,an} = L(X_1, \ldots, X_{\lfloor bn \rfloor}; Y_1, \ldots, Y_{\lfloor an \rfloor}), \qquad L_n := L_{n,n}.$$

Thus L_n is the length of the longest common sequence, when both sequences are of equal length, $n_x = n_y = n$. The random variable L_n is the main objet of interest.

Extremal alignments: Definition and properties

We now formally define the highest (optimal) alignment corresponding to L_n. Let

$$\big\{\big((i_1^1, j_1^1), \ldots, (i_k^1, j_k^1)\big), \ldots, \big((i_1^{|A|}, j_1^{|A|}), \ldots, (i_k^{|A|}, j_k^{|A|})\big)\big\}$$

be the set of all optimal alignments. Hence, $k = L_n$ and $A = \{1, \ldots, |A|\}$ is the index set so that the elements of A will be identified with optimal alignments. For every i_l^α (resp., j_l^α), where $\alpha \in A$ and $l \in \{1, \ldots, k\}$, we shall denote $j(i_l^\alpha) := j_l^\alpha$ (resp., $i(j_l^\alpha) := i_l^\alpha$). We define

$$J := \big\{j_l^\alpha : \alpha \in A, l = 1, \ldots, k\big\}, \qquad I := \big\{i_l^\alpha : \alpha \in A, l = 1, \ldots, k\big\}.$$

Let $j_k^h := \max_\alpha j_k^\alpha = \max J$. There might be many alignments α 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 is smaller than i_k^h. Formally,

$$j_{k-1}^h := \max\big\{j_l^\alpha : i\big(j_l^\alpha\big) < i_k^h, \alpha \in A, l = 1, \ldots, k\big\}.$$

There might be several i's such that corresponding j is j_{k-1}^h. Amongst them, we choose the minimum. Thus,

$$i_{k-1}^h := \min\big\{i_l^\alpha : j\big(i_l^\alpha\big) = j_{k-1}^h, \alpha \in A, l = 1, \ldots, k\big\}.$$

Proceeding so, we obtain an alignment. We call this the *highest alignment procedure*. We now prove that the procedure can be repeated k times, that is, the obtained alignment is optimal.

Proposition 2.1. *The highest alignment procedure produces an optimal alignment*

$$\big\{\big(i_1^h, j_1^h\big), \ldots, \big(i_k^h, j_k^h\big)\big\},$$

where (i_t^h, j_t^h) *can be obtained as follows*

$$j_t^h := \max\big\{j_t^\alpha : \alpha \in A\big\}, \qquad i_t^h := \min\big\{i_t^\alpha : j\big(i_t^\alpha\big) = j_t^h, \alpha \in A\big\}, \qquad t = 1, \ldots, k. \tag{2.2}$$

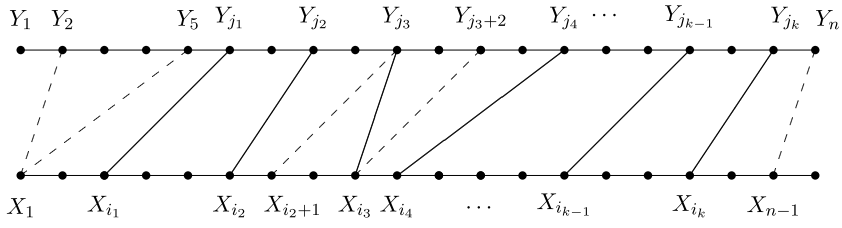


Figure 3. An example of the highest alignment.

Proof. Clearly the pair (i_k^h, j_k^h) is the last pair of an optimal alignment, that is, there exists $\alpha \in A$ such that $(i_k^h, j_k^h) = (i_k^\alpha, j_k^\alpha)$. So (2.2) 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 β 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 $(i_1^\beta, j_1^\beta), \ldots, (i_k^\beta, j_k^\beta), (i_k^h, j_k^h)$ would be an alignment of length $k+1$. Suppose $l \le k-2$. Since $j_l^\beta < j_{k-1}^\beta < j_k^\beta \le j_k^h = \max J$, by definition of j_{k-1}^h, it must be that $i_k^h \le i_{k-1}^\beta$. 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, $j_{k-1}^\alpha \le j_{k-1}^h = j_l^\beta < j_{k-1}^\beta$ implying that $j_{k-1}^\alpha < j_{k-1}^\beta < j_k^\beta$. Hence, $(i_1^\alpha, j_1^\alpha), \ldots, (i_{k-1}^\alpha, j_{k-1}^\alpha), (i_{k-1}^\beta, j_{k-1}^\beta), (i_k^\beta, j_k^\beta)$ would be an alignment of length $k+1$. Therefore, $j_{k-1}^h = \max\{j_{k-1}^\alpha : i_{k-1}^\alpha < i_k^h, \alpha \in A\}$. Let us now prove that (2.2) 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 β so that $j_{k-1}^\beta > j_{k-1}^h$ and $i_{k-1}^\beta \ge i_k^h > i_{k-1}^h$. But as we saw, those inequalities would give an alignment with the length $k+1$. This concludes the proof of (2.2) with $t = k-1$. For $t = k-2, \ldots, 1$ proceed similarly. □

Figure 3 is an example of an highest alignment. The solid lines are aligned pairs (the upper-index h is dropped from the notation). If $Y_{j_3+2} = Y_{j_3}$, then, as showed by dashed line, X_{i_3} could be aligned with Y_{j_3+2} that contradicts the highest alignment procedure. Thus, every Y_{j_t} in the highest alignment is different from all Y_j that are right after Y_{j_t} and before $Y_{j_{t+1}}$. This observation is postulated as statements (2.4) and (2.5) in the following corollary. Similarly, if $X_{i_2+1} = X_{i_3}$, then, as showed by dashed line, X_{i_2+1} could be aligned with Y_{j_3} that also contradicts the highest alignment procedure. Thus, in the highest alignment all X_i-s right after $X_{i_{t-1}}$ and before X_{i_t} must differ from X_{i_t}. This observation is formulated as the statements (2.6) and (2.7) in the following corollary. In the highest alignment, typically, $i_1 < j_1$ and $j_k > i_k$. If X_1 is not aligned, then it must be that $Y_i \ne X_1$ for $i = 1, \ldots, j_1 - 1$, otherwise they could be aligned (as showed by dashed line) contradicting the optimality. These observations are statements (2.8) and (2.9) in the following corollary. Similarly, if Y_n is not aligned, it should be different from all $X_{i_k+1}, \ldots, X_n$. These observations are statements (2.10) and (2.11) in the following corollary.

Corollary 2.1. *The highest alignment has the following properties*:

$$X_{i_t^h} = Y_{j_t^h}, \qquad t = 1, \ldots, k; \tag{2.3}$$

$$Y_{j_t^h} \ne Y_j, \qquad j_t^h < j < j_{t+1}^h, \qquad t = 1, \ldots, k-1; \tag{2.4}$$

$$Y_{j_k^h} \neq Y_j, \qquad j_k^h < j \leq n; \tag{2.5}$$

$$X_{i_t^h} \neq X_i, \qquad i_t^h > i > i_{t-1}^h, \qquad t = 2, \ldots, k; \tag{2.6}$$

$$X_{i_1^h} \neq X_i, \qquad 1 \leq i < i_1^h; \tag{2.7}$$

$$\text{if } i_1^h > 1, \text{ then } X_1 \neq Y_j, \qquad 1 \leq j < j_1^h; \tag{2.8}$$

$$\text{if } j_1^h > 1, \text{ then } Y_1 \neq X_i, \qquad 1 \leq i < i_1^h; \tag{2.9}$$

$$\text{if } n > j_k^h, \text{ then } Y_n \neq X_i, \qquad i_k^h < i \leq n; \tag{2.10}$$

$$\text{if } n > i_k^h, \text{ then } X_n \neq Y_j, \qquad j_k^h < j \leq n. \tag{2.11}$$

Proof. The equalities (2.3) are obvious. Suppose that for a $t = 1, \ldots, k-1$ there exists an index j such that $j_t^h < j < j_{t+1}^h$ and $Y_{j_t^h} = Y_j$. Then the pairs

$$\big\{(i_1^h, j_1^h), \ldots, (i_{t-1}^h, j_{t-1}^h), (i_t^h, j), (i_{t+1}^h, j_{t+1}^h), \ldots, (i_k^h, j_k)\big\}$$

would correspond to an optimal alignment, say β, satisfying

$$j_t^\beta = j > j_t^h = \max\big\{j_t^\alpha : \alpha \in A\big\}.$$

Thus, (2.4) holds. The same argument proves (2.5), (2.6) and (2.7). If one of the inequalities in (2.8)–(2.11) is not fulfilled, then it would be possible to align one more pair without disturbing already existing aligned pairs. This contradicts the optimality. □

One can also think of the left-most or nord-west alignment. It could be defined as an alignment $\{(i_1^w, j_1^w), \ldots, (i_k^w, j_k^w)\}$, where $i_1^w = \min I$, $j_1^w = \max\{j_l^\alpha : i(j_l^\alpha) = i_1^w, \alpha \in A, l = 1, \ldots, k\}$ and for every $t = 2, \ldots, k$,

$$i_t^w := \min\big\{i_l^\alpha : j\big(i_l^\alpha\big) > j_{t-1}^w, \alpha \in A, l = 1, \ldots, k\big\},$$
$$j_t^w := \max\big\{j_l^\alpha : i\big(j_l^\alpha\big) = i_t^w, \alpha \in A, l = 1, \ldots, k\big\}.$$

Here the superscript "w" stands for west. By the analogue of Proposition 2.1,

$$i_t^w = \min\big\{i_t^\alpha : \alpha \in A\big\}, \qquad j_t^w = \max\big\{j_t^\alpha : i\big(j_t^\alpha\big) = i_t^w, \alpha \in A\big\}, \qquad t = 1, \ldots, k. \tag{2.12}$$

Using (2.2) and (2.12), it is easy to see that the left-most and highest alignments actually coincide. Indeed, by (2.2) and (2.12), $j_t^h \geq j_t^w$ and $i_t^w \leq i_t^h$, $\forall t$. If, for a t, $(i_t^h, j_t^h) \neq (i_t^w, j_t^w)$, then, by the definitions, both inequalities have to be strict, that is, $i_t^w < i_t^h$ and $j_t^w < j_t^h$. To see this, suppose $i_t^w = i_t^h$. This means that there exists an alignment α, such that $i_t^\alpha = i_t^w$ and $j_t^\alpha = j_t^h$. This, in turn, implies that

$$\max\big\{j_t^\alpha : i\big(j_t^\alpha\big) = i_t^w, \alpha \in A\big\} = \max\big\{j_t^\alpha, \alpha \in A\big\},$$

that is, $j_t^w = j_t^h$. The same argument shows that if $j_t^w = j_t^h$, then also $i_t^w = i_t^h$. Thus $(i_t^h, j_t^h) \neq (i_t^w, j_t^w)$ implies that $i_t^w < i_t^h$ and $j_t^w < j_t^h$. These inequalities, however, would imply the existence of an alignment with the length $k+1$.

The lowest (the right-most) alignment $\{(i_1^l, j_1^l), \ldots, (i_k^l, j_k^l)\}$ will be defined similarly:

$$j_t^l := \min\{j_t^\alpha : \alpha \in A\}, \qquad i_t^l := \max\{i_t^\alpha : j(i_t^\alpha) = j_t^l, \alpha \in A\}, \qquad t = 1, \ldots, k. \tag{2.13}$$

Remark. Note that the left-most alignment equals the lowest alignment of $(Y_n, \ldots, Y_1)$ and $(X_n, \ldots, X_1)$ implying that the highest alignment of $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ equals to the lowest alignment of $(Y_n, \ldots, Y_1)$ and $(X_n, \ldots, X_1)$. Thus, the lowest alignment between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ can be defined as the highest alignment between $(Y_n, \ldots, Y_1)$ and $(X_n, \ldots, X_1)$.

Combinatorics

Another way to study an alignment of $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_y}$ is to present it as a strictly increasing mapping

$$v : \{1, \ldots, n_x\} \hookrightarrow \{1, \ldots, n_y\}. \tag{2.14}$$

Notation (2.14) means: There exists $I(v) \subset \{1, \ldots, n_x\}$ and a mapping

$$v : I \to \{1, \ldots, n_y\}$$

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 $|v|$. In the notation of previous sections, thus, $j_t := v(i_t)$, $t = 1, \ldots, |v|$.

Consider now the case $n_x = n_y = n$, that is, both sequences are of length n. Let then V_k be the set of all alignments with length k. Formally,

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

Fix $\Delta > 0$, $\gamma \in (0, 1)$ and let

$$W_n(\gamma, \Delta) := \bigcup_{k=(\gamma-\Delta)n}^{(\gamma+\Delta)n} V_k. \tag{2.15}$$

Hence, W_n consists of these alignments that have length not smaller that $(\gamma - \Delta)n$ and not bigger that $(\gamma + \Delta)n$. In the subsequent sections, we shall show that there exists a constant γ (depending on the model) so that for n big enough all optimal alignments belong to W_n with high probability. Thus, in a sense the set W_n contains all alignments of interest. We are interested in bounding the size of that set. For that, we use the binary entropy function

$$h(p) := -p \log_2 p - (1-p) \log_2 (1-p).$$

Let, for $\gamma, \Delta \in (0, 1)$ such that $0 < \gamma - \Delta, \gamma + \Delta < 1$

$$H(\gamma, \Delta) := \max_{\alpha \in [\gamma-\Delta, \gamma+\Delta]} h(\alpha). \tag{2.16}$$

Since

$$\binom{n}{pn} \le 2^{h(p)n},$$

for every

$$(\gamma - \Delta)n \le k \le (\gamma + \Delta)n, \tag{2.17}$$

it holds

$$|V_k| = \binom{n}{\frac{k}{n}n}^2 \le 2^{2H(\gamma,\Delta)n}.$$

Hence, the number of alignments in W_n can be bounded as follows:

$$\left|W_n(\gamma, \Delta)\right| \le 2\Delta n 2^{2H(\gamma,\Delta)n}. \tag{2.18}$$

Let us consider now a more general case $n_y > n_x$. Denote $m = n_y > n_x = n$. Assume that $m \le n(1 + \Delta)$. Then

$$|V_k| = \binom{n}{\frac{k}{n}n}\binom{m}{\frac{k}{m}m} \le 2^{h(k/n)n + h(k/m)n(1+\Delta)}.$$

Instead of (2.17), we assume k to satisfy

$$\gamma - \Delta \le \frac{k}{n} \le \gamma + 2\Delta. \tag{2.19}$$

Then

$$\gamma - 2\Delta \le \frac{\gamma - \Delta}{1 + \Delta} \le \frac{k}{m} \le \frac{k}{n} \le \gamma + 2\Delta$$

and

$$2^{h(k/n)n + h(k/m)n(1+\Delta)} \le 2^{H(\gamma,2\Delta)n + H(\gamma,2\Delta)n(1+\Delta)} = 2^{H(\gamma,2\Delta)n(2+\Delta)}.$$

In this case, defining

$$W_{n,m}(\gamma, \Delta) := \bigcup_{k=(\gamma-\Delta)n}^{(\gamma+2\Delta)n} V_k,$$

it holds

$$|W_{n,m}| \le 3\Delta n 2^{(2+\Delta)H(\gamma,2\Delta)n}.$$

3. Independent sequences

In this section, only, let $X = X_1, \ldots, X_n$ and $Y = Y_1, \ldots, Y_n$ be two independent i.i.d. sequences from the alphabet $\mathcal{A}$. Recall that for any $a > 0$, $L_{an,n} = L(X_1, \ldots, X_{\lfloor an \rfloor}; Y_1, \ldots, Y_n)$ and $L_n =$

$L_{n,n}$. By the Kingman's subbadditive ergodic theorem, there exists a constant $\gamma(a) \in (0,1]$ so that

$$\frac{L_{an,n}}{n} \to \gamma(a) \qquad \text{a.s. and in } L_1.$$

We shall denote $\gamma := \gamma(1)$, the constant γ is often called the *Chvatal–Sankoff* constant. In the Appendix, it will be shown that when $a < 1$, then $\gamma(a) < \gamma$ (Lemma A.1).

Note that $L_{an,n}$ is a function of $n(1+a)$ i.i.d. random variables. Clearly, changing one of the variables changes the value of L_n at most by one, so that by McDiarmid inequality (see, e.g., [15]), for every $\Delta > 0$

$$P\big(|L_{an,n} - EL_{an,n}| > n\Delta\big) \le 2\exp\left[-\frac{2\Delta^2}{(1+a)}n\right]. \tag{3.1}$$

Take $n_o(\Delta, a)$ so big that $|\frac{EL_{an,n}}{n} - \gamma(a)| < \frac{\Delta}{2}$. Then

$$\begin{aligned} &P\big(\big|L_{an,n} - \gamma(a)n\big| \ge n\Delta\big) \\ &\quad\le P\big(|L_{an,n} - EL_{an,n}| + \big|EL_{an,n} - \gamma(a)n\big| \ge n\Delta\big) \\ &\quad\le P\left(|L_{an,n} - EL_{an,n}| \ge n\frac{\Delta}{2}\right) \le 2\exp\left[-\frac{\Delta^2}{2(1+a)}n\right], \qquad n > n_o. \end{aligned} \tag{3.2}$$

Taking $a = 1$, we see the existence of $n_o(\Delta)$ so that for $n > n_o$ with high probability all optimal alignments are contained in the set $W_n(\gamma, \Delta)$ as defined in (2.15).

Recall that for any optimal alignment v, (i_1^h, j_1^h) and $(i_{|v|}^h, j_{|v|}^h)$ are the first and last pairs of indexes of the highest alignment of X and Y. We consider the random variables $S := j_1^h - 1$, $T := n - i_{|v|}^h$. Clearly S and T have the same law. The following proposition states that for any $c \in (0,1)$, the probabilities $P(S > cn) = P(T > cn)$ decrease exponentially fast.

Proposition 3.1. *Let $c \in (0,1)$. Then there exists constant $d(c) > 0$, so that, for n big enough, $P(T > cn) = P(S > cn) \le \exp[-dn]$.*

Proof. Note that $\{T > cn\} \subset \{L_{(1-c)n,n} = L_n\}$ and for any $\bar\gamma$,

$$\{L_{(1-c)n,n} = L_n\} \subset \{L_{(1-c)n,n} \ge \bar\gamma n\} \cup \{L_n \le \bar\gamma n\}.$$

Let $a := 1 - c$. By Lemma A.1, $\gamma > \gamma(a)$. Let

$$\bar\gamma := \frac{\gamma + \gamma(a)}{2}, \qquad \Delta := \gamma - \bar\gamma = \bar\gamma - \gamma(a).$$

Use (3.2) to see that for n big enough,

$$\begin{aligned} P(T > cn) &\le P(L_{an,n} \ge \bar\gamma n) + P(L_n \le \bar\gamma n) \\ &= P\big(L_{an,n} \ge \big(\gamma(a) + \Delta\big)n\big) + P\big(L_n \le (\gamma - \Delta)n\big) \\ &\le 2\exp\left[-\frac{\Delta^2}{2(1+a)}n\right] + 2\exp\left[-\frac{\Delta^2}{4}n\right]. \end{aligned}$$

This concludes the proof. □

Recall the definition of q and p_o in (1.3). Note that for independent sequences, $p_o = P(X_i = Y_i)$. The following lemma bounds the probability that an alignment $v \in V_k$ is the highest optimal alignment.

Lemma 3.1. *Let $v \in V_k$. Let $B(v)$ be the event that v is the highest optimal alignment of X and Y. Then*

$$P\big(B(v)\big) \le p_o^k q^{2(n-k)-(j_1-1)-(n-i_k)}. \tag{3.3}$$

Proof. Let $v \in V_k$ be an alignment. We denote by $i_1, \dots, i_k$ the elements of $I(v)$ and we define $j_t := v(i_t)$, $t = 1, \dots, k$.

Since all random variables $X_1, \dots, X_n, Y_1, \dots, Y_n$ are independent, by Corollary 2.1 the probability of $B(v)$ could be estimated as follows

$$P\big(B(v)\big) \le \prod_{t=1}^{k} P\big(B_t(v)\big),$$

where, for $t = 2, \dots, k-1$

$$B_t(v) := \{X_{i_t} = Y_{j_t}; Y_j \ne Y_{j_t}, j_t < j < j_{t+1}; X_i \ne X_{i_t}, i_{t-1} < i < i_t\}$$

and

$$B_1(v) := \begin{cases} \{X_{i_1} = Y_{j_1}; X_i \ne X_{i_1}, i < i_1; X_1 \ne Y_j, \\ \quad j < j_1; Y_j \ne Y_{j_1}, j_1 < j < j_2\}, & \text{if } i_1 > 1; \\ \{X_{i_1} = Y_{j_1}; Y_j \ne Y_{j_1}, j_1 < j < j_2\}, & \text{if } i_1 = 1. \end{cases}$$

$$B_k(v) := \begin{cases} \{X_{i_k} = Y_{j_k}; X_i \ne X_{i_k}, i_{k-1} < i < i_k; \\ \quad Y_j \ne Y_{j_k}, j_k < j; X_i \ne Y_n, i_k < i\}, & \text{if } j_k > n; \\ \{X_{i_k} = Y_{j_k}; X_i \ne X_{i_k}, i_{k-1} < i < i_k\}, & \text{if } j_k = n. \end{cases}$$

By independence, clearly for $t = 2, \dots, k-1$,

$$P\big(B_t(v)\big) = \sum_a p_a^2 (1-p_a)^{i_t - i_{t-1} - 1 + j_{t+1} - j_t - 1} \le p_o q^{i_t - i_{t-1} + j_{t+1} - j_t - 2}.$$

For the events $B_1(v)$ and $B_k(v)$, we estimate

$$P\big(B_1(v)\big) \le \begin{cases} p_o q^{j_2 - j_1 - 1}, & \text{if } i_1 = 1; \\ p_o q^{j_2 - j_1 - 1 + j_1 - 1 + i_1 - 1}, & \text{if } i_1 > 1. \end{cases}$$

$$P\big(B_k(v)\big) \le \begin{cases} p_o q^{i_k - i_{k-1} - 1}, & \text{if } j_k = n; \\ p_o q^{i_k - i_{k-1} - 1 + n - j_k + n - i_k}, & \text{if } j_k < n. \end{cases}$$

These equations yield (3.3). Note that in (3.3), the term $(n - i_k)$ disappears when $j_k < n$ and the term $(j_1 - 1)$ disappears, when $i_1 > 1$. □

Our first main result is a bound to the unknown Chvatal–Sankoff constant γ.

Theorem 3.1. *Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be two independent i.i.d. sequences with the same distribution. Let γ be the corresponding Chvatal–Sankoff constant. Then the following condition holds*

$$\gamma \log_2 p_o + 2(1-\gamma)\log_2 q + 2h(\gamma) \geq 0. \tag{3.4}$$

Proof. The proof is based on the contradiction: assuming that (3.4) fails leads to the existence of constants $c > 0, b > 0$ (independent of n) such that for n big enough, $P(S + T \leq 2cn) \leq \exp[-bn]$. Then, for big n,

$$1 - \exp[-bn] \leq P(S + T > 2cn) \leq P(S > cn) + P(T > cn)$$

contradicting Proposition 3.1.

If (3.4) is not fulfilled, then it is possible to find constants $\Delta > 0, c > 0$ so small that

$$-b_1 := (\gamma - \Delta)\log_2 p_o + 2(1 - \gamma - \Delta - c)\log_2 q + 2H(\gamma, \Delta) < 0. \tag{3.5}$$

Fix now $\Delta > 0, c > 0$ so small that (3.5) holds. Let

$$E_\Delta := \{|L_n - n\gamma| < n\Delta\}.$$

When E_Δ holds, then all optimal alignments belong to the set $W_n := W_n(\gamma, \Delta)$. By Lemma 3.1, for every $v \in W_n$

$$P\big(B(v)\big) \leq p_o^{n(\gamma-\Delta)} q^{2n(1-\gamma-\Delta)-(n-i_k)-(j_1-1)}. \tag{3.6}$$

Note that $\bigcup_{v\in W_n} B(v) = E_\Delta$. Let, for every v, $s(v) := j_1 - 1$ and $t(v) := n - i_{|v|}$. Then by (3.2) and (2.18), there exists $b > 0$ (independent of n) so that for n big enough

$$\begin{aligned} P(S + T \leq 2cn) &\leq \sum_{v\in W_n : s(v)+t(v)\leq 2cn} P\big(B(v)\big) + P\big(E_\Delta^c\big) \\ &\leq 2\Delta n 2^{n(2H(\gamma,\Delta)+(\gamma-\Delta)\log_2 p_o + 2(1-\gamma-\Delta-c)\log_2 q)} + P\big(E_\Delta^c\big) \\ &\leq 2\Delta n 2^{-b_1 n} + 2\exp\left[-\frac{\Delta^2}{4}n\right] \leq \exp[-bn]. \end{aligned}$$

□

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 $Z_1, Z_2, \ldots$. 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 $f : \mathcal{A} \times \mathbb{R} \to \mathcal{A}$ is a mapping and ξ_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 $\xi_1, \xi_2, \ldots$ or the random mappings $f_1, f_2, \ldots$ are independent (and identically distributed). After mutations, the sequence is $f_1(Z_1), f_2(Z_2), \ldots$. Some of its elements disappear. This is modeled via a deletion process $D_1^x, D_2^x, \ldots$ that is assumed to be an i.i.d. Bernoulli sequence with parameter p that is, $P(D_i^x = 1) = p$. 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

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

Similarly, the sequence Y is obtained from Z. For mutations, fix an i.i.d. standard normal sequence $\eta_1, \eta_2, \ldots$ so that the mutated sequence is $g_1(Z_1), g_2(Z_2), \ldots$ with $g_i(\cdot) := f(\cdot, \eta_i)$. Note that the transition matrix corresponding to Y-mutations equals the one corresponding to X-mutations implying that the random mappings g_i and f_i have the same distribution. Since the mutations of X and Y are supposed to be independent, we assume the sequences ξ and η or the random mappings sequences $f_1, f_2, \ldots$ and $g_1, g_2, \ldots$ are independent. Note that then the pairs $(f_1(Z_1), g_1(Z_1)), (f_2(Z_2), g_2(Z_2)), \ldots$ are independent, but $f_i(Z_i)$ and $g_i(Z_i)$, 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, $D_1^y, D_2^y, \ldots$ 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.

In the following, we shall call the sequences $X = X_1 \ldots X_n$ and $Y = Y_1 \ldots Y_n$ *related*, if they follow the model described above. Note that for the related sequences, the random variables $X_1, X_2, \ldots$ as well as $Y_1, Y_2, \ldots$ are still i.i.d., but these two sequences are, in general, not independent any more. As mentioned above, the process $(X_1, Y_1), (X_2, Y_2), \ldots$ is not stationary, hence also not ergodic. It is, however, a regenerative process. We shall also call the random variables X_i and Y_j *related*, if they have the same ancestor. However, the definition of the related sequences does not exclude the case, when the functions f and g do not depend on Z_i so that the sequences X and Y are independent. Thus, in what follows, all results for related sequences automatically hold for independent sequences as well.

With this notation (recall (1.7)), $\bar{q} = 1 - \min_{a,b \in \mathcal{A}} P(f(\xi, Z) = a | g(\eta, Z) = b)$. Note that $P(f(\xi, Z) = a | g(\eta, Z) = b) = P(X_i = a | Y_j = b)$ given X_i and Y_j are related.

4.2. Limits and large deviation inequalities for related sequences

In this subsection, we consider the random variables $L_{n,an}$, where $a > 0$. By symmetry, for any n, the random variable $L_{n,an}$ has the same law as $L_{an,n}$; moreover, the processes $\{L_{an,n}\}$ and $\{L_{n,an}\}$ have the same distribution so that in what follows, everything holds for $L_{an,n}$ as well.

The existence of $\gamma_{\mathrm{R}}(a)$

At first, we shall prove the convergence (1.4). As mentioned in the Section 3, for independent sequences, this follows from subadditive ergodic theorem. The same holds, if the sequences are related, but no deletion occurs, that is, $p = 1$. In the presence of deletion, however, an additional argument is needed.

Proposition 4.1. *Let $a > 0$. Then there exists a constant $\gamma_{\mathrm{R}}(a)$ such that (1.4) holds.*

Proof. At first note that without loss of generality, we may assume $a \le 1$. Indeed, with $m := \lfloor na \rfloor$,

$$\begin{aligned} L(X_1, \dots, X_n; Y_1, \dots, Y_{\lfloor na \rfloor}) &= L(X_1, \dots, X_{\lceil m/a \rceil}; Y_1, \dots, Y_m) \\ &= L(X_1, \dots, X_m; Y_1, \dots, Y_{\lceil m/a \rceil}), \end{aligned}$$

where the last equality follows from the symmetry of the model. Hence, the limit in (1.4) exists if and only if the limit of $\frac{1}{m} L(X_1, \dots, X_m; Y_1, \dots, Y_{\lceil m/a \rceil})$ exists. The latter is equivalent to the existence of limit $\frac{1}{m} L_{m,m/a}$. Hence, to the end of the proof, let $0 < a \le 1$.

We consider the sequence of i.i.d. random vectors $U_1, U_2, \dots$, where

$$U_i := \big(f_i(Z_i), g_i(Z_i), D_i^x, D_i^y\big). \tag{4.1}$$

Let, for any positive integer m, $n_x(m) := \sum_{i=1}^m D_i^x$ and $n_y(m) := \sum_{i=1}^{\lfloor am \rfloor} D_i^y$. Thus $X_1, \dots, X_{n_x}$ and $Y_1, \dots, Y_{n_y}$ are both determined by i.i.d. random vectors $U_1, \dots, U_m$. Let

$$L(U_1, \dots, U_m) := L(X_1, \dots, X_{n_x}; Y_1, \dots, Y_{n_y}).$$

By subadditivity, there exists constant γ_{U} such that

$$\lim_{m \to \infty} \frac{L(U_1, \dots, U_m)}{m} = \gamma_{\mathrm{U}}, \qquad \text{a.s. and in } L_1. \tag{4.2}$$

Let $\underline{n}(m) := n_x(m) \wedge \frac{n_y(m)}{a}$ and $\overline{n}(m) := n_x(m) \vee \frac{n_y(m)}{a}$. Thus,

$$\frac{\underline{n}}{m} \frac{L(X_1, \dots, X_{\underline{n}}; Y_1, \dots, Y_{\lfloor a\underline{n} \rfloor})}{\underline{n}} \le \frac{L(U_1, \dots, U_m)}{m} \le \frac{\overline{n}}{m} \frac{L(X_1, \dots, X_{\overline{n}}, Y_1, \dots, Y_{\lceil \overline{n}a \rceil})}{\overline{n}}. \tag{4.3}$$

By SLLN, $\frac{\overline{n}(m)}{m} \to p$, a.s. and $\frac{\underline{n}(m)}{m} \to p$, a.s. Since

$$\limsup_n \frac{L_{n,an}}{n} = \limsup_m \frac{L_{\underline{n}(m),a\underline{n}(m)}}{\underline{n}(m)}, \qquad \liminf_n \frac{L_{n,an}}{n} = \liminf_m \frac{L_{\overline{n}(m),a\overline{n}(m)}}{\overline{n}(m)}$$

and

$$\liminf_m \frac{L_{\overline{n}(m),a\overline{n}(m)}}{\overline{n}(m)} = \liminf_m \frac{1}{\overline{n}(m)} L(X_1,\dots,X_{\overline{n}(m)};Y_1,\dots,Y_{\lceil a\overline{n}(m)\rceil}),$$

from (4.3), it follows

$$\begin{aligned}&\limsup_n \frac{1}{n} L(X_1,\dots,X_n;Y_1,\dots,Y_{\lfloor an\rfloor})p\\ &\quad\le \gamma_{\mathrm{U}} \le \liminf_n \frac{1}{n} L(X_1,\dots,X_n;Y_1,\dots,Y_{\lfloor an\rfloor})p, \qquad \text{a.s.}\end{aligned}$$

This is the a.s. convergence in (1.4) with $\gamma_{\mathrm{R}}(a) = \frac{\gamma_{\mathrm{U}}}{p}$. The convergence in L_1 follows by dominated convergence theorem. □

Large deviation inequalities

Next, we prove a large deviation lemma for related sequences.

Lemma 4.1. *Assume X and Y are related. Then, for every $\Delta > 0$ and $0 < a \le 1$,*

$$P\big(|L_{n,an} - EL_{n,an}| \ge n\Delta\big) \le 4\exp\left[-\frac{p}{8}\Delta^2 an\right]. \tag{4.4}$$

Proof. As we saw in Section 3, for independent sequence, this type of inequality (3.1) trivially follows from McDiarmid inequality. In the present case, we have to add an extra control over the deletion process.

Fix positive integer m and consider the vectors $U_1,\dots,U_m$ defined in (4.1). Recall $n_x(m)$ and $n_y(m)$. Fix n and let

$$\tilde{L}_m := L(X_1,\dots,X_{n\wedge n_x};Y_1,\dots,Y_{\lfloor an\rfloor\wedge n_y}).$$

Note that $\tilde{L}_m$ is a function of $5m$ independent random variables:

$$\tilde{L}_m = \tilde{L}_m\big(Z_1,\dots,Z_m,\xi_1,\dots,\xi_m,\eta_1,\dots,\eta_m,D_1^x,\dots,D_m^x,D_1^y,\dots,D_m^y\big).$$

Changing Z_i (given all other variables are fixed) corresponds to possible change of an element of X and an element of Y. A change of one element of X (or Y) causes the change of $\tilde{L}_m$ at most by 1. Hence, the maximum change of $\tilde{L}_m$ induced by changing of Z_i (given all other variables are fixed) is 2. Similarly, the maximum change of $\tilde{L}_m$ due to the change of ξ_i or η_i (given all other variables are fixed) is 1. Changing D_i^x from 1 to 0 corresponds to removing one element of X-side and, in the case $n_x > n$ adding one more X to the end. Changing D_i^x from 0 to 1 corresponds to adding one element to X-side and, perhaps, removing the last X (when $n_x \ge n$). This, again, changes the value of $\tilde{L}_m$ at most by 1. Any change of η_i has the same effect. Denoting by r_i, $i = 1,\dots,5m$ the maximum change of $\tilde{L}_m$ induced by the ith variable, we have that $r_i = 2$ if

$i = 1, \dots, m$ and $r_i = 1$ for $i = m+1, \dots, 5m$ so that $\sum_{i=1}^{5m} r_i^2 = 8m$. Therefore, by McDiarmid inequality,

$$P\big(|\tilde{L}_m - E\tilde{L}_m| \geq m\Delta\big) \leq 2\exp\bigg[-\frac{\Delta^2}{4}m\bigg]. \tag{4.5}$$

Let E_m be the event that $n_x \geq n$ and $n_y \geq an$. Formally, $E_m := E_y(m) \cap E_x(m)$, where $E_x(m) := \{\sum_{i=1}^{m} D_i^x \geq n\}$ and $E_y(m) := \{\sum_{i=1}^{\lfloor am \rfloor} D_i^y \geq an\}$. When E_m holds, then $\tilde{L}_m = L_{n,an}$, so that

$$\big\{|L_{n,an} - EL_{n,an}| < m\Delta\big\} \supset \big\{|\tilde{L}_m - E\tilde{L}_m| < m\Delta\big\} \cap E_m$$

and

$$P\big(|L_{n,an} - EL_{n,an}| \geq m\Delta\big) \leq P\big(|\tilde{L}_m - E\tilde{L}_m| \geq m\Delta\big) + P\big(E_m^c\big). \tag{4.6}$$

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

$$P\bigg(|\tilde{L}_m - E\tilde{L}_m| \geq \frac{2}{p}n\Delta\bigg) \leq 2\exp\bigg[-\frac{\Delta^2}{2p}n\bigg]. \tag{4.7}$$

To estimate $P(E_m^c) \leq P(E_x^c) + P(E_y^c)$, use Hoeffding inequality (with $m = \frac{2n}{p}$)

$$\begin{aligned} P\big(E_y^c\big) &= P\bigg(\sum_{i=1}^{am} D_i^y < an\bigg) = P\bigg(\sum_{i=1}^{am} D_i^y - amp < an - amp\bigg) \\ &\leq P\bigg(\sum_{i=1}^{am} D_i^y - amp < -\frac{amp}{2}\bigg) \leq \exp\bigg[-\frac{p^2}{2}am\bigg] = \exp[-pan], \\ P\big(E_x^c\big) &= P\bigg(\sum_{i=1}^{m} D_i^x < n\bigg) \leq \exp[-pn] \leq \exp[-pan]. \end{aligned}$$

Thus, with $m = \frac{2n}{p}$, $P(E_m^c) \leq 2\exp[-pan]$ and plugging it together with (4.7) into (4.6) entails

$$P\bigg(|L_{n,an} - EL_{n,an}| \geq \frac{2n}{p}\Delta\bigg) \leq 2\exp\bigg[-\frac{\Delta^2}{2p}n\bigg] + 2\exp[-pan]. \tag{4.8}$$

Take $\Delta' = \frac{2\Delta}{p}$. Then (4.8) is

$$P\big(|L_{n,an} - EL_{n,an}| \geq \Delta' n\big) \leq 2\exp\bigg[-\frac{(\Delta')^2 p}{8}n\bigg] + 2\exp[-pan].$$

If $\Delta' \leq 1$, then $2\exp[-apn] \leq 2\exp[-(\Delta')^2 apn]$, implying that the right-hand side is bounded by $4\exp[-\frac{(\Delta')^2}{8}apn]$. This proves (4.4) for $\Delta \leq 1$. Since $L_{n,an} \leq n$, for $\Delta > 1$, (4.4) trivially holds. □

The following corollary states an inequality similar to that of (3.2) for related sequences.

Corollary 4.1. *Assume X and Y are related, $0 < a \le 1$. Then, for every $\Delta > 0$ there exists $n_o(\Delta, a)$ big enough so that*

$$P\big(\big|L_{n,an} - \gamma_{\mathrm{R}}(a)n\big| \ge n\Delta\big) \le 4\exp\bigg[-\frac{p}{32}\Delta^2 an\bigg], \qquad n > n_o. \tag{4.9}$$

Proof. Let n be so big that $|EL_{n,an}/n - \gamma_{\mathrm{R}}(a)| < \Delta/2$. Then $|EL_{n,an} - \gamma_{\mathrm{R}}(a)n| \le (\Delta/2)n$ and

$$\begin{aligned} P\big(\big|L_{n,an} - \gamma_{\mathrm{R}}(a)n\big| \ge n\Delta\big) &\le P\big(|L_{n,an} - EL_{n,an}| + \big|EL_{n,an} - \gamma_{\mathrm{R}}(a)n\big| \ge n\Delta\big) \\ &\le P\bigg(|L_{n,an} - EL_{n,an}| \ge n\frac{\Delta}{2}\bigg) \le 4\exp\bigg[-\frac{p}{32}a\Delta^2 n\bigg], \end{aligned}$$

where the last inequality follows from (4.4) □

5. Proofs of main results for related sequences

5.1. Every highest alignment contains a related pair

5.1.1. *The key lemma*

The following lemma is the cornerstone of what follows.

Lemma 5.1. *Assume that $X = X_1 \dots X_n$ and $Y = Y_1 \dots Y_n$ are related and satisfy (1.8). Then there exists a constant $b_2 > 0$ such that for every n big enough,*

$$P(\textit{highest alignment of } X \textit{ and } Y \textit{ alignes no related letters}) \le \mathrm{e}^{-nb_2}. \tag{5.1}$$

Proof. Let $v \in V_k$ be an alignment. Let $I = I(v) = \{i_1, \dots, i_k\}$ and let $j_t := v(i_t)$. Hence $X_{i_t} = Y_{j_t}$, for every $t = 1, \dots, k$. We denote by J the set $\{j_1, \dots, j_k\}$.

We are bounding the probability that v is the highest optimal alignment of X and Y and that the random variables X_{i_t} and Y_{j_t} are not related for every $t = 1, \dots, k$. Let us introduce some notations and events. Let, for every $j = j_1 + 1, \dots, n$, $b(j)$ be the last element of J strictly smaller than j. Formally, $b(j) := \max\{j_t : j_t < j\}$. Similarly, for every $i = 1, \dots, i_k - 1$, let $c(i)$ be the first element of I strictly larger than i. Formally, $c(i) := \min\{i_t : i_t > i\}$. Also denote

$$a^x = \big(a^x(i_1), \dots, a^x(i_k)\big), \qquad a^y = \big(a^y(j_1), \dots, a^y(j_k)\big)$$

and let $a^x \ne a^y$ be $a^x(j_t) \ne a^y(j_t)$ for every $t = 1, \dots, k$. We now define the following events

$$\begin{aligned} A(v) &:= \{X_{i_1} = Y_{j_1}, \dots, X_{i_k} = Y_{j_k}\}, \\ B(v) &:= \big\{Y_j \ne Y_{b(j)}, j \in \{j_1 + 1, \dots, n\} \setminus J\big\}, \\ C(v) &:= \big\{X_i \ne X_{c(i)}, i \in \{1, \dots, i_k - 1\} \setminus I\big\}, \\ D(v) &:= \big\{a^x \ne a^y\big\}, \qquad E(v) := A(v) \cap B(v) \cap C(v) \cap D(v). \end{aligned}$$

By Corollary 2.1, it holds

$$\{v \text{ is the highest optimal alignment and no aligned pair of } v \text{ is related}\} \subset E(v).$$

Note that the vectors a^x and a^y depend on the deletion processes D^x and D^y, only. Thus, given a^x and a^y, the events $A(v)$, $B(v)$ and $C(v)$ depend on the ancestor process Z and on the random mappings g and f, only. In particular, given a^x and a^y, the dependence structure (related pairs) is fixed as well. In the following, we shall consider the case $a^x \neq a^y$. This means that there exists no $t = 1, \dots, k$ such that X_{i_t} is related to Y_{j_t}.

We shall bound the probability $P(E(v)|a^x, a^y)$.

At first, let us bound the probability $P(A(v)|a^x, a^y)$, $a^x \neq a^y$. Thus, in what follows, we assume a^x and a^y satisfying $a^x \neq a^y$ are fixed. For any two indexes $s, t \in \{1, \dots, k\}$, let $s \leftrightarrow t$ denote that either X_{i_t} and Y_{j_s} are related (i.e., they have the same ancestor) or X_{i_s} and Y_{j_t} are related. We call a subset $G = \{t_1, \dots, t_l\} \subset \{1, \dots, k\}$ a *dependence group*, if:

1. $t_i \leftrightarrow t_{i+1}$ for every $i = 1, \dots, l-1$;
2. there is no index in $\{1, \dots, k\} \setminus G$ that is related to t_1 or t_l.

Note that a group with $|G|$ elements contains $|G| - 1$ related pairs. Let $\{t_1, \dots, t_l\}$ be a dependence group. Without loss of generality, assume that $X_{i_{t_1}}$ is related to $Y_{j_{t_2}}$. Then $X_{i_{t_2}}$ is related to $Y_{j_{t_3}}$ and so on. In particular, $X_{i_{t_k}}$ is independent of $Y_{j_{t_l}}$, $l \leq k$. Recall the definition of p_o, $\overline{p}$ q and $\overline{q}$ from (1.3) and (1.7). Hence,

$$\begin{aligned} &P(X_{i_t} = Y_{j_t}; t \in G) \\ &\quad = P(X_{i_{t_1}} = Y_{j_{t_1}}) \prod_{k=2}^{l} P(X_{i_{t_k}} = Y_{j_{t_k}} | X_{i_{t_1}} = Y_{j_{t_1}}, \dots, X_{i_{t_{k-1}}} = Y_{j_{t_{k-1}}}) \\ &\quad = p_o \prod_{k=2}^{l} \Bigg(\sum_{a \in \mathcal{A}} P(X_{i_{t_k}} = a) P(Y_{j_{t_k}} = a | X_{i_{t_1}} = Y_{j_{t_1}}, \dots, X_{i_{t_{k-1}}} = Y_{j_{t_{k-1}}}) \Bigg) \leq p_o (\overline{p})^{l-1}. \end{aligned}$$

By 2., the random variables $\{X_{i_t}, Y_{j_t} : t \in G\}$ are all independent of the random variables $\{X_{i_t}, Y_{j_t} : t \in \{1, \dots, k\} \setminus G\}$. Let $G_1, \dots, G_u$ be all dependence groups. Let $G = G_1 \cup \dots \cup G_u$. Thus, $r := |G| - u$ is the number of related pairs amongst $X_{i_1}, \dots, X_{i_k}$ and $Y_{j_1}, \dots, Y_{j_k}$. By independence of the groups,

$$\begin{aligned} P\big(A(v)|a^x, a^y\big) &= \prod_{s=1}^{u} P(X_{i_t} = Y_{j_t}; t \in G_s) \prod_{t \notin G} P(X_{i_t} = Y_{j_t}) \\ &\leq p_o^u (\overline{p})^{|G|-u} p_o^{k-|G|} = p_o^{k-r} (\overline{p})^r. \end{aligned} \tag{5.2}$$

Let us now bound the probability $P(B(v)|A(v), a^x, a^y)$, where, as previously, $a^x \neq a^y$. Recall the sets I and J. Let

$$I^c := \{1, \dots, i_k - 1\} \setminus I, \qquad J^c := \{j_1 + 1, \dots, n\} \setminus J.$$

Let J_1^c be the set of indexes in J^c with the property that the corresponding Y-s are related to an element in $X_i, i \in I$. Formally, $j \in J_1^c$ if and only if there exists a $i \in I$ so that Y_j is related to X_i. Let $J_2^c = J^c \setminus J_1^c$. It means, if $j \in J_2^c$, then Y_j is either related to an X_i with $i \notin I$ or not related to any other random variable at all. In particular, the random variables $\{Y_j : j \in J_2^c\}$ are independent of the event $A(v)$. Since $Y_1, Y_2, \dots$ are independent, we obtain (let us omit the fixed a^x and a^y from the notations)

$$\begin{aligned} P\big(B(v)|A(v)\big) &= P\big(Y_j \neq Y_{b(j)},\, j \in J_2^c \cup J_1^c | A(v)\big) \\ &= P\big(Y_j \neq Y_{b(j)},\, j \in J_2^c\big) P\big(Y_j \neq Y_{b(j)},\, j \in J_1^c | A(v)\big). \end{aligned}$$

Clearly,

$$P\big(Y_j \neq Y_{b(j)},\, j \in J_2^c\big) \leq q^{|J_2^c|}. \tag{5.3}$$

Let us estimate $P(Y_j \neq Y_{b(j)},\, j \in J_1^c | A(v))$. Note

$$\begin{aligned} &P\big(Y_j \neq Y_{b(j)},\, j \in J_1^c | A(v)\big) \\ &= \sum_{(y_1,\dots,y_k)\in\mathcal{A}^k} P\big(Y_j \neq Y_{b(j)},\, j \in J_1^c | Y_{j_t} = X_{i_t} = y_t, \forall t\big) P\big(X_{i_t} = y_t, \forall t | A(v)\big). \end{aligned}$$

Given $(y_1, \dots, y_k)$, let $y_{b(j)}$ be the value of $Y_{b(j)}$. Let us estimate

$$\begin{aligned} P\big(Y_j \neq Y_{b(j)},\, j \in J_1^c | Y_{j_t} = X_{i_t} = y_t, \forall t\big) &= P\big(Y_j \neq y_{b(j)},\, j \in J_1^c | Y_{j_t} = X_{i_t} = y_t, \forall t\big) \\ &= P\big(Y_j \neq y_{b(j)},\, j \in J_1^c | X_{i_t} = y_t, \forall t\big) \\ &= \frac{P(Y_j \neq y_{b(j)},\, j \in J_1^c;\, X_{i_t} = y_t, \forall t)}{\prod_{t=1}^k P(X_{i_t} = y_t)}. \end{aligned}$$

The last two equalities follow from the fact that $Y_1, Y_2, \dots$ are independent and $X_1, X_2, \dots$ are independent. When $j \in J_1^c$, then Y_j is related to a X_{i_t}. Denote $J_1^c := \{j^1, \dots, j^s\}$. Clearly,

$$s := |J_1^c| \leq |J^c| \wedge k = (n - j_1 + 1 - k) \wedge k. \tag{5.4}$$

Without loss of generality, assume that the random variables in J_1^c are related to the $X_{i_1}, \dots, X_{i_s}$. Then the pairs of related random variables $(Y_{j^1}, X_{t_1}), \dots, (Y_{j^s}, X_{t_s})$ are independent so that

$$\begin{aligned} \frac{P(Y_j \neq y_{b(j)},\, j \in J_1^c;\, X_{i_t} = y_t, \forall t)}{\prod_{t=1}^k P(X_{i_t} = y_t)} &= \frac{\prod_{t=1}^s P(Y_{j^t} \neq y_{b(j^t)}, X_{i_t} = y_t)}{\prod_{t=1}^s P(X_{i_t} = y_t)} \\ &= \prod_{t=1}^s P(Y_{j^t} \neq y_{b(j^t)} | X_{i_t} = y_t) \leq (\bar{q})^s. \end{aligned}$$

Therefore,

$$P\big(Y_j \neq Y_{b(j)};\, j \in J_1^c | A(v)\big) \leq (\bar{q})^s. \tag{5.5}$$

By entirely similar argument, we estimate $P(C(v)|A(v)\cap B(v))$. Indeed, given $(y_1,\dots,y_k)\in \mathcal{A}^k$,

$$\begin{aligned}&P\big(C(v)|A(v)\cap B(v), Y_{j_t}=y_t,\forall t\big)\\ &\quad= P\big(X_i\neq X_{c(i)}, i=I^c|Y_{j_t}=X_{i_t}=y_t,\forall t; Y_j\neq Y_{b(j)}, j=J^c\big)\\ &\quad= P\big(X_i\neq a_{c(i)}, i=I^c|Y_{j_t}=y_t,\forall t; Y_j\neq y_{b(j)}, j=J^c\big).\end{aligned}$$

Every X_i is related to at most one Y_j. Let I_0^c, I_1^c, I_2^c be mutually exclusive set of indexes so that:

- If $i\in I_0^c$, then X_i is not related to any Y_j from $J\cup J^c$.
- If $i\in I_1^c$, then X_i is related to a Y_j so that $j\in J$. Let $t(i)\in\{1,\dots,k\}$ be the corresponding index.
- If $i\in I_2^c$, then X_i is related to a Y_j so that $j\in J^c$. Let $j_r(i)\in J$ be the corresponding index.

Then, just like previously, using the independence of related pairs, we obtain

$$\begin{aligned}&P\big(X_i\neq y_{c(i)}, i=I^c|Y_{j_t}=y_t,\forall t; Y_j\neq y_{b(j)}, j=J^c\big)\\ &=\prod_{i\in I_0^c}P(X_i\neq y_{c(i)})\prod_{i\in I_1^c}P(X_i\neq y_{c(i)}|Y_{t(i)}=y_{t(i)})\prod_{i\in I_2^c}P(X_i\neq y_{c(i)}|Y_{j_r(i)}\neq y_{b(j_r(i))})\\ &\le q^{|I_0^c|}(\bar q)^{|I_1^c|+|I_2^c|}\le(\bar q)^{|I^c|},\end{aligned}$$

where the second last inequality follows from the fact that given X_i and Y_j are related, for any $a,b\in\mathcal{A}$

$$\begin{aligned}P(X_i\neq a|Y_j\neq b)&=\sum_{c\neq b}P(X_i\neq a|Y_j=c)P(Y_j=c|Y_j\neq b)\\ &=\sum_{c\neq b}\big(1-P(X_i=a|Y_j=c)\big)P(Y_j=c|Y_j\neq b)\le\bar q\end{aligned}$$

and the last inequality follows from the fact that $q\le\bar q$. Therefore,

$$P\big(C(v)|A(v)\cap B(v)\big)\le(\bar q)^{|I^c|}=(\bar q)^{i_k-k}. \tag{5.6}$$

By (5.2), (5.3), (5.5), (5.6) with $\rho=(p_o\bar q)/(\bar p q)$ and $r+s\le k$, we have

$$\begin{aligned}P\big(E(v)|a^x,a^y\big)&\le p_o^{k-r}(\bar p)^r q^{n-j_1+1-k-s}(\bar q)^{s+i_k-k}=p_o^k\left(\frac{\bar p}{p_o}\right)^r\left(\frac{\bar q}{q}\right)^s q^{n-j_1+1-k}(\bar q)^{i_k-k}\\ &\le p_o^k\left(\frac{\bar p}{p_o}\right)^{k-s}\left(\frac{\bar q}{q}\right)^s q^{n-j_1+1-k}(\bar q)^{i_k-k}\le(\bar p)^k\rho^s q^{n-j_1+1-k}(\overline{q})^{i_k-k}.\end{aligned}$$

By (5.4), it holds $0\le s\le k\wedge(n-j_1+1-k)\le k\wedge(n-k)$, so that

$$\max_s\rho^s\le\begin{cases}\rho^{k\wedge(n-k)}, & \text{if } \rho\ge 1;\\ 1, & \text{if } \rho<1.\end{cases}$$

Hence,

$$\begin{aligned} P\big(E(v)\big) &\le \sum_{a^x,a^y:a^x\neq a^y} P\big(E(v)|a^x,a^y\big)P\big(D^x=a^x,D^y=a^y\big) \\ &\le (\bar{p})^k(\rho\vee 1)^{k\wedge(n-k)}(q\bar{q})^{n-k}q^{1-j_1}(\overline{q})^{i_k-n}. \end{aligned} \tag{5.7}$$

Recall that (1.8) is

$$\gamma_{\mathrm{R}}\log_2\bar{p}+(1-\gamma_{\mathrm{R}})\log_2(q\bar{q})+\big((1-\gamma_{\mathrm{R}})\wedge\gamma_{\mathrm{R}}\big)\log_2(\rho\vee 1)+2h(\gamma_{\mathrm{R}})<0.$$

When this holds, then it is possible to find $\Delta>0$ so small that

$$\begin{aligned} -b :=\ & (\gamma_{\mathrm{R}}-\Delta)\log_2\bar{p}+(1-\gamma_{\mathrm{R}}-\Delta)\log_2(q\bar{q}) \\ &+\big((1-\gamma_{\mathrm{R}}+\Delta)\wedge(\gamma_{\mathrm{R}}+\Delta)\big)\log_2(\rho\vee 1)-\Delta\log_2(q\bar{q})+2H(\gamma_{\mathrm{R}},\Delta)<0. \end{aligned}$$

Let

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

When E_Δ holds, then all optimal alignments belong to the set $W_n := W_n(\gamma_{\mathrm{R}},\Delta)$. For every $v\in W_n$, with $|v|=k$, it holds

$$n(\gamma_{\mathrm{R}}-\Delta)\le k\le n(\gamma_{\mathrm{R}}+\Delta). \tag{5.8}$$

Let, for every v, $s(v)=j_1-1$ and $t(v)=n-i_k$. Let

$$U_n(\gamma_{\mathrm{R}},\Delta) := \big\{v\in W_n : s(v)\le\Delta n, t(v)\le\Delta n\big\}.$$

Using these two inequalities together with (5.8), we have that for every $v\in U_n$,

$$\begin{aligned} \log_2 P\big(E(v)\big) \le n\Big[&(\gamma_{\mathrm{R}}-\Delta)\log_2\bar{p}+(1-\gamma_{\mathrm{R}}-\Delta)\log_2(q\bar{q}) \\ &+\big((1-\gamma_{\mathrm{R}})\wedge\gamma_{\mathrm{R}}+\Delta\big)\log_2(\rho\vee 1)-\frac{s(v)}{n}\log_2 q-\frac{t(v)}{n}\log_2\bar{q}\Big] \\ \le\ &\big(-b-2H(\gamma_R,\Delta)\big)n. \end{aligned}$$

Let

$$E := \{\exists \text{ highest alignment of} X \text{ and } Y \text{ alignes no related letters}\}.$$

Recall that $S=j_1^h-1$, $T=n-i_k^h$ and, by (2.18), it holds

$$|U_n|\le\big|W_n(\gamma_{\mathrm{R}},\Delta)\big|\le 2\Delta n2^{2H(\gamma_{\mathrm{R}},\Delta)}.$$

Then by Corollary A.1 and Corollary 4.1, for n big enough

$$\begin{aligned} P(E) &\le \sum_{v\in U_n} P\big(E(v)\big) + P(S > \Delta n) + P(T > \Delta n) + P\big(E_\Delta^c\big) \\ &\le |U_n| 2^{(-b-2H(\gamma_{\mathrm{R}},\Delta))} + P(S > \Delta n) + P(T > \Delta n) + P\big(E_\Delta^c\big) \\ &\le 2\Delta n 2^{-bn} + 4\exp\left[-\frac{\Delta^2}{32}n\right] + 2\exp\big[-d(\Delta)n\big]. \end{aligned}$$

Hence, for big n, the inequality (5.1) holds. □

Sequences with unequal lengths. 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 $m \ge n$. We know that if (1.8) holds, then there exists $\Delta > 0$ so small that

$$\begin{aligned} &(\gamma_{\mathrm{R}} - \Delta)\log_2 \bar{p} + (1 - \gamma_{\mathrm{R}} - 2\Delta)\log_2(q\bar{q}) + \big((1-\gamma_{\mathrm{R}}) \wedge \gamma_{\mathrm{R}} + 2\Delta\big)\log_2(\rho \vee 1) \\ &\quad - 2\Delta \log_2(q\bar{q}) + 2H(\gamma_{\mathrm{R}}, 2\Delta) < 0. \end{aligned} \tag{5.9}$$

The restriction for m is: $m \le (1+\Delta)n$.

Lemma 5.2. *Let $n \le m \le (1+\Delta)n$, where $\Delta > 0$ satisfies (5.9). Assume that X^n and Y^m are related. Then there exists a constant $b_3(\Delta) > 0$ such that for every $n > n_o$,*

$$P\big(\textit{the highest alignment of } X^n \textit{ and } Y^m \textit{ aligns no related letters}\big) \le \mathrm{e}^{-nb_3}.$$

Proof. The proof follows the one of Lemma 5.1; Δ is now taken from the assumptions, so it satisfies (5.9). This Δ defines the set E_Δ 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. Clearly $L_n \le L_{n,m} \le L_n + n\Delta$. Hence, if E_Δ holds, then

$$\gamma_{\mathrm{R}} - \Delta \le \frac{L_n}{n} \le \frac{L_{n,m}}{n} \le \frac{L_n}{n} + \Delta \le \gamma_{\mathrm{R}} + 2\Delta,$$

that is, all optimal alignments belong to the set $W_{n,m}(\gamma_{\mathrm{R}}, \Delta)$. The set $U_{n,m}$ is defined as follows

$$U_{n,m}(\gamma_{\mathrm{R}}, \Delta) := \big\{v \in W_{n,m}(\gamma_{\mathrm{R}}, \Delta) : s(v) \le 2\Delta n, t(v) \le 2\Delta n\big\}.$$

The upper bound (5.7) holds with n replaced by m:

$$P\big(E(v)\big) \le (\bar{p})^k (\rho \vee 1)^{k \wedge (m-k)} (q\bar{q})^{m-k} q^{1-j_1} (\overline{q})^{i_k - m}.$$

Using the bounds $(\gamma_{\mathrm{R}} - \Delta)n \le |u| \le (\gamma_{\mathrm{R}} + 2\Delta)n$ and $n \le m \le n(1+\Delta)$, for every $v \in U_{n,m}$, we obtain the following estimate

$$\begin{aligned}\log_2 P\big(E(v)\big) &\le n\Big[(\gamma_{\mathrm{R}} - \Delta)\log_2 \bar{p} + \big((1-\gamma_{\mathrm{R}}) \wedge \gamma_{\mathrm{R}} + 2\Delta\big)\log_2(\rho \vee 1) \\ &\qquad + (1 - \gamma_{\mathrm{R}} - 2\Delta)\log_2(q\bar{q}) - \frac{s(v)}{n}\log q_2 - \frac{t(v)}{n}\log_2 \bar{q}\Big] \\ &\le -\big(b + 2H(\gamma_{\mathrm{R}}, \Delta)\big)n,\end{aligned}$$

where $b > 0$ by the assumption (5.9) on Δ. The rest of the proof goes as the one of Lemma 5.1 with $P(S > 2\Delta n)$ and $P(T > 2\Delta n)$ instead of $P(S > \Delta n)$ and $P(T > \Delta n)$ and $3\Delta n 2^{-bn}$ instead of $2\Delta n 2^{-bn}$. □

5.1.2. *Applying Lemma 5.2 repeatedly*: *The B-events*

Regenerativity. Let $\tau_0^x = \tau_0^y = 0$ and let τ_k^x (τ_k^y), $k = 1, 2, \ldots$ be the indexes of the kth 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 kth related pair, that is,

$$a_k = a^x\big(\tau_k^x\big) = a^y\big(\tau_k^y\big).$$

We shall use the fact that the process $(X_1, Y_1), (X_2, Y_2), \ldots$ is regenerative with respect to the times (τ_k^x, τ_k^y), i.e.

$$(X_{\tau_k^x+1}, Y_{\tau_k^y+1}), (X_{\tau_k^x+2}, Y_{\tau_k^y+2}), \ldots \tag{5.10}$$

has the same law as $(X_1, Y_1), (X_2, Y_2), \ldots$. The Z-process for (5.10) is $Z_{a_k+1}, Z_{a_k+2}, \ldots$.

Definition of B-events. In what follows, let $\Delta > 0$ and $0 < A < \infty$. Denote $n' := A \ln n$. We shall consider the following events:

$$\begin{aligned}B_k(\tilde{n}, \tilde{m}) &:= \{\text{the highest alignment of } X_{\tau_k^x+1}, \ldots, X_{\tau_k^x+\tilde{n}} \text{ and } Y_{\tau_k^y+1}, \ldots, Y_{\tau_k^y+\tilde{m}} \\ &\qquad \text{contains a related pair}\}, \\ B_k^1\big(n', \Delta\big) &:= \bigcap_{n' \le \tilde{n} \le \tilde{m} \le \tilde{n}(1+\Delta)} B_k(\tilde{n}, \tilde{m}), \qquad B_k^2\big(n', \Delta\big) := \bigcap_{n' \le \tilde{m} \le \tilde{n} \le \tilde{m}(1+\Delta)} B_k(\tilde{n}, \tilde{m}), \\ B_k^h\big(n', \Delta\big) &:= B_k^1\big(n', \Delta\big) \cap B_k^2\big(n', \Delta\big).\end{aligned}$$

Let $B_k^l(n', \Delta)$ be defined similarly, with "lowest" instead of "highest" in the definition of $B_k(\tilde{n}, \tilde{m})$. Finally, let

$$B\big(k, n', \Delta\big) := B_k^l\big(n', \Delta\big) \cap B_k^h\big(n', \Delta\big).$$

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

$$B_n\bigl(n',\Delta\bigr):=\bigcup_{i=0}^{n}\Biggl(\{K=i\}\cap\Biggl(\bigcap_{k=0}^{i}B\bigl(k,n',\Delta\bigr)\Biggr)\Biggr),$$

where

$$K:=\arg\max_{k=0,1,\ldots}\bigl\{\max\bigl\{\tau_k^x,\tau_k^y\bigr\}\le n\bigr\}. \tag{5.11}$$

The bound on $P(B_n(n',\Delta))$ for small Δ. We aim to bound $P(B_n(n',\Delta))$ from below. We use the regenerativity described above: for every k, the event $B_k(\tilde n,\tilde m)$ has the same probability as $B_0(\tilde n,\tilde m)$ so that for every k, Lemma 5.2 applies:

$$P\bigl(B_k(\tilde n,\tilde m)\bigr)\ge 1-\exp[-\tilde n b_3],$$

provided $\tilde n(1+\Delta)\ge\tilde m\ge\tilde n\ge n_o$ and $\Delta>0$ is small enough to satisfy the assumptions (5.9). Thus, for small enough Δ and big enough n', we have

$$P\bigl(B_k^2\bigl(n'\bigr)\bigr)=P\bigl(B_k^1\bigl(n'\bigr)\bigr)\ge 1-\sum_{\tilde n\ge n'}\sum_{\tilde n(1+\Delta)\ge\tilde m\ge\tilde n}\mathrm{e}^{-b_3\tilde n}=1-\sum_{\tilde n\ge n'}(\Delta\tilde n+1)\mathrm{e}^{-b_3\tilde n}. \tag{5.12}$$

Clearly, for every $0<b_4<b_3$, for every n big enough, $(n+\Delta^{-1})\mathrm{e}^{-b_3 n}\le\mathrm{e}^{-b_4 n}$ for every $n>n_1$. Let $0<b_4<b_3$ and without loss of generality assume n_o being so big that for every $n>n_o$ the inequality above holds. Then (5.12) can be bounded as follows

$$P\bigl(B_k^1\bigl(n',\Delta\bigr)\bigr)\ge 1-\Delta\sum_{\tilde n\ge n'}\mathrm{e}^{-b_4\tilde n}\ge 1-\frac{B}{4}\mathrm{e}^{-b_4 n'},\qquad n'\ge n_o, \tag{5.13}$$

where B is a constant depending on Δ. Hence,

$$P\bigl(B\bigl(k,n',\Delta\bigr)\bigr)\ge 1-B\mathrm{e}^{-b_4 n'},\qquad n'\ge n_o.$$

Finally, since $\bigcap_{k=0}^{n}B(k,n')\subset B_n(n')$, we have that (recall $n'=A\ln n$)

$$\begin{aligned}P\bigl(B_n^c\bigl(n',\Delta\bigr)\bigr)&\le(n+1)P\bigl(B_k^c\bigl(n',\Delta\bigr)\bigr)\\&\le B(n+1)\exp\bigl[-b_4 n'\bigr]\le 2Bn\exp\bigl[-b_4 n'\bigr]=2Bn^{1-b_4 A}.\end{aligned} \tag{5.14}$$

5.2. The location of the related pairs

We consider the related sequences $X_1,X_2,\ldots$ and $Y_1,Y_2,\ldots$. Recall the definition of τ_k^x, τ_k^y and a_k. As previously, we take $n'=A\ln n$.

5.2.1. *The location of the first and last related pair*: *G-events*

The location of last related pair: *Definition of* $G_n(\Delta)$. Let $i(n)$ and $j(n)$ be the biggest τ_k^x and τ_k^y before n, that is,

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

Clearly $i(n) = n$ if and only if the ancestor of X_n is also an ancestor of a Y_j that is, $D^y_{a^x(n)} = 1$. Similarly, $i(n) = u < n$ if and only if

$$D^y_{a^x(u)} = 1, \qquad D^y_{a^x(u+1)} = \cdots = D^y_{a^x(n)} = 0.$$

Since the process D^y is independent of D^x and, therefore, also independent of the random variables $a^x(i)$, $i = 1, 2\ldots$, we have that for every $u = 1, 2, \ldots, n$

$$P\big(i(n) = u\big) = (1-p)^{n-u}p, \qquad P\big(i(n) = 0\big) = (1-p)^n.$$

Hence, for any $\Delta > 0$,

$$\begin{aligned} P\big(n - i(n) \ge \Delta n\big) &= P\big(n - i(n) \ge \lceil \Delta n \rceil\big) = (1-p)^{\lceil \Delta n \rceil} \\ &\le (1-p)^{\Delta n} = \exp\big[\Delta n \ln(1-p)\big]. \end{aligned} \tag{5.15}$$

Hence, the probability that the last related X_i before n is further that Δn from n is exponentially small in n. The same obviously holds for $j(n)$ so that

$$P\big(n - i(n) < \Delta n, n - j(n) < \Delta n\big) \ge 1 - 2\exp\big[\ln(1-p)\Delta n\big]. \tag{5.16}$$

However, the event $\{(n - j(n)) \vee (n - i(n)) < \Delta n\}$ does not necessarily imply that the last related pair, let that be (i', j') is necessarily such that $\{(n - j') \vee (n - i') < \Delta n\}$. Indeed, if (i', j') is the last related pair, then either $i' = i(n)$ or $j' = j(n)$ but the both inequalities need not hold simultaneously. We shall now show that also the event $\{(n - j') \vee (n - i')\} \le \Delta n\}$ holds with great probability. Let us first define the last related pair formally as follows

$$\begin{aligned} &i'(n) := \tau^x_{l(n)}, \qquad j'(n) := \tau^y_{l(n)}, \\ &\quad \text{where } l(n) := \max\big\{l = 0, 1, 2, \ldots : \tau_k^x \le n, \tau_k^y \le n\big\}. \end{aligned} \tag{5.17}$$

Let $0 < \Delta < 1$, $r(n) := (1 - \frac{3}{4}\Delta)\frac{n}{p}$ and consider the event

$$G_n^x := \left\{ n(1-\Delta) \le \sum_{j=1}^{r} D_j^x \le n\left(1 - \frac{\Delta}{2}\right)\right\}, \qquad G_n^y := \left\{ n(1-\Delta) \le \sum_{j=1}^{r} D_j^y \le n\left(1 - \frac{\Delta}{2}\right)\right\}.$$

To simplify the calculations, let us assume without the loss of generality that r is an integer. Assume that $G_n^x \cap G_n^y$ holds, $i(n) > n(1 - \frac{\Delta}{2})$ and let Y_j be related to $X_{i(n)}$. Let a be their common

ancestor, that is, $a := a^y(j) = a^x(i(n))$. Since $i(n) > n(1 - \frac{\Delta}{2})$, the event G_n^x guarantees that $a > r$. Indeed, if $a \le r$, then we reach to the contradiction, since

$$i(n) = \sum_{j=1}^{a} D_j^x \le \sum_{j=1}^{r} D_j^x \le n\left(1 - \frac{\Delta}{2}\right).$$

Thus $a > r$ and because of G_n^y, it holds

$$j = \sum_{j=1}^{a} D_j^y \ge \sum_{j=1}^{r} D_j^y \ge n(1 - \Delta).$$

Hence, if $\tau_{l(n)}^x \ge \tau_{l(n)}^y$ (i.e., $i(n) = i'(n) \ge j$), then we have that $\tau_{l(n)}^y = j'(n) \ge n(1 - \Delta)$. The roles of X and Y can be changed so that

$$\begin{aligned} &\left\{\big(n - j(n)\big) \vee \big(n - i(n)\big) < \frac{\Delta}{2}n\right\} \cap G_n^x \cap G_n^y \\ &\quad \subset \left\{\big(n - j'(n)\big) \vee \big(n - i'(n)\big) \le \Delta n\right\} =: G_n(\Delta). \end{aligned} \tag{5.18}$$

When $G_n(\Delta)$ holds, then 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 Δn in the upper-right corner.

The bound on $P(G_n(\Delta))$. By Hoeffding's inequality,

$$\begin{aligned} P\big(\big(G_n^x(\Delta)\big)^c\big) &= P\left(\left|\sum_{j=1}^{r} D_j^x - rp\right| > \frac{\Delta}{4}n\right) = P\left(\left|\sum_{j=1}^{r} D_j^x - rp\right| > \frac{p\Delta}{4 - 3\Delta}r\right) \\ &\le 2\exp\left[-2\left(\frac{p\Delta}{4 - 3\Delta}\right)^2 r\right] = 2\exp\left[-\frac{p\Delta^2}{2(4 - 3\Delta)}n\right]. \end{aligned}$$

Therefore, for Δ small enough,

$$\begin{aligned} P\big(G_n^c(\Delta)\big) &\le P\big(\big(G_n^x(\Delta)\big)^c\big) + P\big(\big(G_n^y(\Delta)\big)^c\big) + P\left(n - i(n) \ge \frac{\Delta}{2}n\right) + P\left(n - j(n) \ge \frac{\Delta}{2}n\right) \\ &\le 4\exp\left[-\frac{p\Delta^2}{2(4 - 3\Delta)}n\right] + 2\exp\left[\frac{\Delta}{2}n\ln(1 - p)\right] \\ &\le 4\exp\left[-\frac{p\Delta^2}{8}n\right] + 2\exp\left[\frac{\Delta}{2}n\ln(1 - p)\right] \\ &\le 6\exp\left[-\frac{p\Delta^2}{8}n\right]. \end{aligned}$$

Finally, we shall apply the event $G_n(\Delta_n)$ with $\Delta_n := \sqrt{\frac{16\ln n}{pn}}$. Then

$$P\big(G_n^c(\Delta_n)\big) \le 6\exp\left[-\frac{p}{8}\Delta_n^2 n\right] = 6n^{-2}. \tag{5.19}$$

5.2.2. *The location of the rest of the related pairs*: F*-events*

Fix $\Delta > 0$ and denote $\alpha := 1 + \frac{\Delta}{2}$, $\beta := 1 + \Delta$ and $l(n) := \frac{\alpha}{p}n$. Again, to simplify the technicalities, let us assume that $l(n)$ is an integer.

F*-events*: *The definition.* At first, we consider the events

$$F_n^x := \left\{ n < \sum_{i=1}^{l} D_i^x \le \beta n \right\}, \qquad F_n^y := \left\{ n < \sum_{i=1}^{l} D_i^y \le \beta n \right\}, \qquad F_n := F_n^x \cap F_n^y.$$

These events are similar to the events G_n^x and G_n^y defined in the previous section and we shall argue similarly. Suppose X_i and Y_j are related and $i \le n$. When F_n^x holds, then the ancestor of X_i is at most l, that is, $a^x(i) \le l$. Since X_i and Y_j are related, $a^x(i) = a^y(j) =: a$. If F_n^y holds, we have $\sum_{i=1}^{a} D_i^y \le \sum_{i=1}^{l} D_i^y \le \beta n$, implying that $j \le \beta n = (1+\Delta)n$. By symmetry, the roles of i and j can be changed. Thus, when the event F_n holds and (i, j) is a related pair, then the following implication holds true: if $\min\{i, j\} \le n$, then $\max\{i, j\} \le (1+\Delta)n$.

We now consider more refined events

$$F\big(k, n'\big) := \bigcap_{m \ge n'} \left\{ a_k + m < \sum_{i=1}^{l(m)} D_{a_k+i}^x, \sum_{i=1}^{l(m)} D_{a_k+i}^y \le a_k + (1+\Delta)m \right\}, \qquad k = 0, 1, 2, \ldots.$$

The event $F(k, n')$ states that for any other related pair $X_{\tau_l^x}, Y_{\tau_l^y}$, $l > k$, the following holds: if $\tau_l^x - \tau_k^x \le n'$, then $\tau_l^y - \tau_k^y \le n'(1+\Delta)$. If $\tau_l^x - \tau_k^x = m > n'$, then $\tau_l^y - \tau_k^y \le m(1+\Delta) = (\tau_l^x - \tau_k^x)(1+\Delta)$. The roles of X and Y can be changed, so that the statements above can be restated as follows:

$$\max\big\{\big(\tau_l^x - \tau_k^x\big) \vee n', \big(\tau_l^y - \tau_k^y\big) \vee n'\big\} \le \min\big\{\big(\tau_l^x - \tau_k^x\big) \vee n', \big(\tau_l^y - \tau_k^y\big) \vee n'\big\}(1+\Delta). \tag{5.20}$$

Finally, let $F_n(n', \Delta)$ denote the event that for every k that satisfies $\max\{\tau_k^x, \tau_k^y\} \le n$, $F(k, n')$ holds. Formally,

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

where K is as in (5.11). The event $F_n(n', \Delta)$ ensures that (5.20) holds for every $k \le K$. In particular, if (i, j) is a related pair such that $i \le n$ and $j \le n$ and (i', j') is another related pair, then

$$\max\big\{\big|i - i'\big| \vee n', \big|j - j'\big| \vee n'\big\} \le \min\big\{\big|i - i'\big| \vee n', \big|j - j'\big| \vee n'\big\}(1+\Delta). \tag{5.21}$$

The bound on $P(F_n(n', \Delta))$. Let us first estimate from below the probability of F_n. Since

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

by Hoeffding's inequality (recall that $l = \frac{\alpha}{p}n$)

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

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

Since $\frac{(\beta-\alpha)^2}{\alpha} = \frac{(1-\alpha)^2}{\alpha} = \frac{\Delta^2}{2(2+\Delta)} =: \frac{d(\Delta)}{2}$, it holds

$$P\left(F_n^c\right) \le 2\exp[-pdn]. \tag{5.22}$$

For estimating $P(F_n(n', \Delta))$, we use the regenerativity argument to see that for every k, the event $F(k, n')$ has the same probability as $\bigcap_{m\ge n'} F_m$ so that by (5.22), there exist constant $R(\Delta, p) < \infty, b_6(\Delta, p) > 0$

$$P\left(F^c\left(k, n'\right)\right) \le \sum_{m\ge n'} P\left(F_m^c\right) \le 2\sum_{m\ge n'} \exp[-pdm] \le R\exp\left[-b_6 n'\right].$$

Finally, since $\bigcap_{k=0}^{n} F(k, n') \subset F_n(n', \Delta)$, we have ($n' = A\ln n$)

$$\begin{aligned} P\left(F_n^c(n, \Delta)\right) &\le (n+1)P\left(F^c\left(k, n'\right)\right) \le M(n+1)\exp\left[-b_6 n'\right] \\ &\le 2Rn\exp\left[-b_6 n'\right] = 2Rn^{1-Ab_6}. \end{aligned} \tag{5.23}$$

5.3. The related pairs in extremal alignments

In previous subsection, we showed that with the high probability the related pairs are rather uniformly located almost in the main diagonal of the two-dimensional representation of alignments (the F-event). We also know that with high probability every piece of length $A\ln n$ of extremal alignments contains at least one related pair (the B-event). Hence, both extremal alignments cannot diverge from the main diagonal too much and therefore they cannot be too far from each other. The following lemma postulates this observation.

In the following, let K^h and K^l be the random number of related pairs of the highest and lowest alignment, respectively. We shall denote by $(i_1^{*h}, j_1^{*h}), \ldots, (i_{K^h}^{*h}, j_{K^h}^{*h})$ the related pairs of the highest alignment and $(i_1^{*l}, j_1^{*l}), \ldots, (i_{K^l}^{*l}, j_{K^l}^{*l})$ the related pairs of the lowest alignment. Let

$$\overline{i} := i_{K^h}^{*h} \wedge i_{K^l}^{*l}.$$

We also agree that $i_0^{*h} =: j_0^{*h} =: i_0^{*l} =: j_0^{*l} := 0$ and with some abuse of terminology, we shall call also the pair $(0,0)$ related of both highest and lowest alignments.

Lemma 5.3. *Let $\Delta > 0$ and assume that $B_n(n', \Delta) \cap F_n(n', \frac{\Delta}{2})$ holds. Let (i^h, j^h) be a pair of the highest alignment of X and Y such that $i^h \le \overline{i}$. Then there exists a related pair (i_u^{*l}, j_u^{*l}), $u \in \{0, \dots, K^l\}$ of the lowest alignment such that*

$$\left|i^h - i_u^{*l}\right| \vee \left|j^h - j_u^{*l}\right| \le n'(1+\Delta). \tag{5.24}$$

Moreover, there exists a related pair (i_l^{*l}, j_l^{*l}), $l \in \{0, \dots, K^l\}$ of the lowest alignment such that

$$i_l^{*l} \le i^h \quad \text{and} \quad \left|j^h - j_l^{*l}\right| \le 2n'(1+\Delta). \tag{5.25}$$

Similarly, for every pair (i^l, j^l) of the lowest alignment of X and Y such that $i^l \le \overline{i}$, there exists a related pair (i_u^{*h}, j_u^{*h}), $u \in \{0, \dots, K^h\}$ such that

$$\left|i^l - i_u^{*h}\right| \vee \left|j^l - j_u^{*h}\right| \le n'(1+\Delta). \tag{5.26}$$

Moreover, there exists a related pair (i_l^{*h}, j_l^{*h}) of the highest alignment such that

$$i_l^{*h} \le i^l \quad \text{and} \quad |j^l - j_l^{*h}| \le 2n'(1+\Delta). \tag{5.27}$$

Proof. At first, we shall see that for every $0 \le t \le K^h - 1$,

$$\left(i_{t+1}^{*h} - i_t^{*h}\right) \wedge \left(j_{t+1}^{*h} - j_t^{*h}\right) \le n', \tag{5.28}$$

$$\left(i_{t+1}^{*h} - i_t^{*h}\right) \vee \left(j_{t+1}^{*h} - j_t^{*h}\right) \le n'(1+\Delta). \tag{5.29}$$

Suppose there exists t such that (5.28) fails. The pairs (i_t^{*h}, j_t^{*h}) and $(i_{t+1}^{*h}, j_{t+1}^{*h})$ are both in the highest alignment, let it be v. Since v is highest, the restriction of v between

$$X_{i_t^{*h}+1}, \dots, X_{i_{t+1}^{*h}-1}, \quad \text{and} \quad Y_{j_t^{*h}+1}, \dots, Y_{j_{t+1}^{*h}-1}$$

must be highest as well. Denote $\tilde{n} = i_{t+1}^{*h} - 1 - i_t^{*h}$ and $\tilde{m} = j_{t+1}^{*h} - 1 - j_t^{*h}$. If (5.28) does not hold, then $\tilde{m}, \tilde{n} \ge n'$. Suppose, without loss of generality that $\tilde{m} \ge \tilde{n}$. Since $F_n(n', \frac{\Delta}{2})$ holds, then (5.21) states that $(\tilde{m}+1) \le (\tilde{n}+1)(1+\frac{\Delta}{2})$ implying that $\tilde{m} \le \tilde{n}(1+\Delta)$. Therefore, we have that the sequences

$$X_{i_t^{*h}+1}, \dots, X_{i_t^{*h}+\tilde{n}}, \quad \text{and} \quad Y_{j_t^{*h}+1}, \dots, Y_{j_t^{*h}+\tilde{m}}$$

with $n' \le \tilde{n} \le \tilde{m} \le \tilde{n}(1+\Delta)$ have an optimal alignment that contains no related pair. This contradicts $B_n(n', \Delta)$. Hence, (5.28) holds. Since $t < K^h$, then (5.21) proves (5.29) (recall that (5.21) also holds for $i = j = 0$).

Consider an arbitrary (not necessarily related) pair (i^h, j^h) of the highest alignment so that $i^h \le \bar{i} \le i^{*h}_{K^h}$. By (5.29), there exists $0 \le k < K^h$ such that $i^{*h}_k \le i^h \le i^{*h}_{k+1}$ and

$$\left(i^{*h}_{k+1} - i^{*h}_k\right) \vee \left(j^{*h}_{k+1} - j^{*h}_k\right) \le n'(1+\Delta). \tag{5.30}$$

Similarly, since $i^h \le \bar{i} \le i^{*l}_{K^l}$, by applying (5.29) to the lowest alignment, there exists $0 \le l < K^l$ such that $i^{*l}_l \le i^h \le i^{*l}_{l+1}$ and

$$\left(i^{*l}_{l+1} - i^{*l}_l\right) \vee \left(j^{*l}_{l+1} - j^{*l}_l\right) \le n'(1+\Delta). \tag{5.31}$$

Hence $i^{*l}_l \le i^h$ and $|i^{*l}_u - i^h| \le n'(1+\Delta)$, for $u = l, l+1$. For (5.25), it suffices to show that $|j^h - j^{*l}_l| \le 2n'(1+\Delta)$. For (5.24), it suffices to show that $\min_{u=l,l+1} |j^h - j^{*l}_u| \le n'(1+\Delta)$. For that, we consider three cases separately:

(1) Suppose $i^{*h}_k \le i^{*l}_l$. Because (i^{*h}_k, j^{*h}_k), (i^{*l}_l, j^{*l}_l), $(i^{*h}_{k+1}, j^{*h}_{k+1})$ are related pairs and $i^{*l}_l \le i^h \le i^{*h}_{k+1}$, we have $j^{*l}_l \le j^{*h}_{k+1}$ so that by $i^{*h}_k \le i^{*l}_l$, it holds $j^{*h}_k \le j^{*l}_l \le j^{*h}_{k+1}$. Clearly at least one inequality is strict. Since (i^{*h}_k, j^{*h}_k), (i^h, j^h), $(i^{*h}_{k+1}, j^{*h}_{k+1})$ are aligned pairs, we have $j^{*h}_k \le j^h \le j^{*h}_{k+1}$ (with at least one of the inequalities being strict). By (5.30), we have $j^{*h}_{k+1} - j^{*h}_k \le n'(1+\Delta)$, implying that $|j^h - j^{*l}_l| \le n'(1+\Delta)$. Thus, (5.24) holds with $u = l$ and then (5.25) trivially holds.

(2) Suppose $i^{*l}_{l+1} \le i^{*h}_{k+1}$. The pairs (i^{*h}_k, j^{*h}_k), $(i^{*l}_{l+1}, j^{*l}_{l+1})$, $(i^{*h}_{k+1}, j^{*h}_{k+1})$ are related. Since $i^{*h}_k \le i^h \le i^{*l}_{l+1} \le i^{*h}_{k+1}$, we have that $j^{*h}_k \le j^{*l}_{l+1} \le j^{*h}_{k+1}$ (again, at least one inequality is strict). Since (i^{*h}_k, j^{*h}_k), (i^h, j^h), $(i^{*h}_{k+1}, j^{*h}_{k+1})$ are aligned pairs, we have $j^{*h}_k \le j^h \le j^{*h}_{k+1}$. By (5.30), we have $j^{*h}_{k+1} - j^{*h}_k \le n'(1+\Delta)$, implying that $|j^h - j^{*l}_{l+1}| \le n'(1+\Delta)$. Therefore, (5.24) holds for $u = l+1$. For (5.25), use the inequalities (5.31) together with the inequalities $|j^h - j^{*l}_l| \le |j^h - j^{*l}_{l+1}| + |j^{*l}_l - j^{*l}_{l+1}| \le 2n'(1+\Delta)$.

(3) Suppose $i^{*l}_l < i^{*h}_k$ and $i^{*h}_{k+1} < i^{*l}_{l+1}$. Since all pairs, except perhaps (i^h, j^h) are related, we have that $i^{*l}_l < i^{*h}_k \le i^h \le i^{*h}_{k+1} < i^{*l}_{l+1}$ and $j^{*l}_l < j^{*h}_k \le j^h \le j^{*h}_{k+1} < j^{*l}_{l+1}$. By (5.31), $|j^h - j^{*l}_l| \le j^{*l}_{l+1} - j^{*l}_l \le n'(1+\Delta)$. Hence, (5.24) holds with $u = l$ and then (5.25) trivially holds.

By symmetry, the second statement of the lemma holds by the same argument. □

Recall the definition of $\Delta_n := \sqrt{\frac{16}{p}\frac{\ln n}{n}}$.

Lemma 5.4. *Let $1 > \Delta > 0$ and assume that $B_n(n', \Delta) \cap G_n(\Delta_n) \cap F_n(n', \frac{\Delta}{2})$ holds. Then there exists $n_1(\Delta) < \infty$ and $M(\Delta) < \infty$ so that for every $n > n_1$, $n - \bar{i} \le M\Delta_n n$.*

Proof. Let $(i^*, j^*) := (i^{*h}_{K^h}, j^{*h}_{K^h})$. Since $G_n(\Delta_n)$ holds, there exists a related pair (i', j') so that $i', j' \ge (1-\Delta_n)n$. Without loss of generality, we can take (i', j') the last related pair satisfying $i' \le n$ and $j' \le n$ so that $i' \ge i^*$ and $j' \ge j^*$. Let now $M(\Delta)$ be so big that

$$1 < (M-1)\frac{\Delta}{2+\Delta}. \tag{5.32}$$

First, we shall show that $n - i^* \le M\Delta_n n$. If not, then for n big enough,

$$i' - i^* > n(M\Delta_n - \Delta_n) = n\Delta_n(M-1) > n'. \tag{5.33}$$

Then, by the definition of M

$$n\Delta_n \le n\Delta_n(M-1)\frac{\Delta/2}{1+\Delta/2} \le \big(i'-i^*\big)\frac{\Delta/2}{1+\Delta/2} \le \big(i'-i^*\big)\frac{\Delta}{2}. \tag{5.34}$$

We shall now show that when (5.33) holds, then

$$\big(n-i^*\big) \le \big(n-j^*\big)(1+\Delta), \qquad \big(n-j^*\big) \le \big(n-i^*\big)(1+\Delta). \tag{5.35}$$

Consider two cases separately:

(a) $i' - i^* \le j' - j^*$. Since $F_n(n', \frac{\Delta}{2})$ holds, we have that $j' - j^* \le (i' - i^*)(1 + \frac{\Delta}{2})$ so that

$$n - j^* = \big(n-j'\big) + \big(j'-j^*\big) \le n\Delta_n + \big(i'-i^*\big)\left(1+\frac{\Delta}{2}\right) \le \big(i'-i^*\big)(1+\Delta) \le \big(n-i^*\big)(1+\Delta),$$

where the second last inequality holds due to (5.34). We also have that

$$n - i^* = \big(n-i'\big) + \big(i'-i^*\big) \le n\Delta_n + \big(j'-j^*\big) \le \big(i'-i^*\big)\frac{\Delta}{2} + \big(j'-j^*\big) \le \big(j'-j^*\big)\left(1+\frac{\Delta}{2}\right).$$

(b) $i' - i^* \ge j' - j^*$. By $F_n(n', \frac{\Delta}{2})$ we have again that $i' - i^* \le (j' - j^*)(1 + \frac{\Delta}{2})$ so that by (5.34), we have

$$n\Delta_n \le \big(i'-i^*\big)\frac{\Delta/2}{1+\Delta/2} \le \big(j'-j^*\big)\frac{\Delta}{2} \tag{5.36}$$

and arguing similarly as in the case (a), we now obtain

$$n - j^* \le \big(i'-i^*\big)\left(1+\frac{\Delta}{2}\right), \qquad n - i^* \le \big(j'-j^*\big)(1+\Delta).$$

We are now applying the same argument as in the previous lemma. Recall that (i^*, j^*) belongs to the highest alignment. The restriction of the highest alignment between

$$X_{i^*+1}, \ldots, X_n \quad \text{and} \quad Y_{j^*+1}, \ldots, Y_n$$

must be highest as well. Moreover, the restriction contains no related pairs. The lengths of $X_{i^*+1}, \ldots, X_n$ and $Y_{j^*+1}, \ldots, Y_n$ are $n - i^*$ and $n - j^*$, respectively. By (5.33) and (5.35), for n big enough, both lengths are bigger than n'; by (5.35) their lengths are comparable, so that by event $B_n(n', \Delta)$, they have to contain a related pair. This contradicts the assumption that (i^*, j^*) is the last related pair of the highest alignment. The contradiction is due to assumption $n - i^* > M\Delta_n n$. Hence, $n - i^* \le M\Delta_n n$, eventually. The same argument holds for the lowest alignment, hence $\overline{i} \ge n - M\Delta_n n$, eventually. □

5.4. Proof of Theorem 1.1

Choose $1 > \Delta > 0$ so small that (5.9) holds. Let $M(\Delta)$ be defined as in (5.32) and $\alpha_n = M\Delta_n$. Clearly $\alpha_n \to 0$, in particular, $\alpha_n < 1$ for n big enough. Recall the definition of Hausdorff's distance between extremal alignments both represented as a set of 2-dimensional points. More precisely, let H and L be the highest and lowest alignments, both represented as the set of two-dimensional points. Clearly, $|H| = |L| = L_n$. In the statement of Theorem 1.1, the subsets of H and L, where the pairs (i, j) satisfying $i > n - \alpha_n n$ are left out, are considered. More precisely, we consider the consider the points

$$H_o := \{(i^h, j^h) \in H : i^h \le n(1-\alpha_n)\}, \qquad L_o := \{(i^l, j^l) \in L : i^l \le n(1-\alpha_n)\}.$$

If for an arbitrary element (i^h, j^h) of H_o, there exists an element (i^l, j^l) of L such that $|i^h - i^l| \vee |j^h - j^l| \le (1+\Delta)n'$, then $\max_{(i,j)\in H_o} \min_{(i^l,j^l)\in L} |i - i^l| \vee |j - j^l| \le (1+\Delta)n'$. If, in addition, for an arbitrary element (i^l, j^l) of L_o, there exists an element (i^h, j^h) of H such that $|i^h - i^l| \vee |j^h - j^l| \le (1+\Delta)n'$, then the restricted Hausdorff's distance with respect to the maximum norm between H and L is at most $(1+\Delta)n'$. The restricted Hausdorff's distance between H and L with respect to the l_2-norm is then $\sqrt{2}n'(1+\Delta)$.

Proof of Theorem 1.1. Choose $1 > \Delta > 0$ so small that (5.9) holds. Now, let $M := M(\Delta)$ as in (5.32), $b_4 := b_4(\Delta) > 0$ and $n_o(\Delta)$ be as in the bound (5.14), $b_6 := b_6(\frac{\Delta}{2}) > 0$ be as in (5.23). Let, moreover, $n > n_1 \vee n_o$, where $n_1(\Delta)$ is as in Lemma 5.4 and let h_o be the restricted Hausdorff's distance with respect to the maximum norm and $\alpha_n := M\Delta_n = M\sqrt{\frac{16\ln n}{pn}}$. Since $n > n_1$, $\alpha_n < 1$ so that h_o is correctly defined. Finally, choose A so big that $\min\{Ab_4, Ab_6\} \ge 3$. We aim to bound the probability of the event $E_n := \{h_o(L, H) \le 2n'\}$, where $n' = A\ln n$. If the event $B_n(n', \Delta) \cap G_n(\Delta_n) \cap F_n(n', \frac{\Delta}{2})$ holds, then by Lemma 5.4, $\bar{i} \ge n(1-\alpha_n)$ so that for every $(i^h, j^h) \in H_o$ and $(i^l, j^l) \in L_o$ Lemma 5.3 applies. Since $(1+\Delta) < 2$, (5.24) and (5.26) of Lemma 5.3 ensure that $h_o(H, L) \le 2n'$. Therefore,

$$B_n(n', \Delta) \cap G_n(\Delta_n) \cap F_n\left(n', \frac{\Delta}{2}\right) \subset E_n.$$

Hence, from (5.14), (5.19) and (5.23), for $n > n_o \vee n_1$,

$$\begin{aligned} P(E_n^c) &\le P(B_n^c(n', \Delta)) + P\left(G_n^c\left(n', \frac{\Delta}{2}\right)\right) + P\left(F_n^c\left(n', \frac{\Delta}{2}\right)\right) \\ &\le 2Bn^{1-Ab_4} + 6n^{-2} + 2Rn^{1-Ab_6} \\ &\le 2(R+B+3)n^{-2}. \end{aligned}$$

Thus, the theorem holds with $D = 2(R+B+3)$ and $C = 2A$. □

5.5. Proof of Theorem 1.2

In Theorem 1.1, 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.

Proof of Theorem 1.2. From Lemma 5.4 and (5.25) of Lemma 5.3, it follows that on the event $F_n(n', \frac{\Delta}{2}) \cap G(\Delta_n) \cap B_n(n', \Delta)$ the following holds: for every pair (i^h, j^h) of the highest alignment such that $i^h \le \bar{i}$, there exists a pair (i^l_k, j^l_k) (including the possibility that $k = 0$) of the lowest alignment such that $i^l_k \le i^h$ and $|j^h - j^l_k| \le 2n'(1 + \Delta)$. Recall that L and H are the lowest and highest alignment graphs, respectively. Since L in non-decreasing, it follows that $H(i^h) - L(i^h) = j^h - L(i^h) \le j^h - j^l_k \le 2n'(1 + \Delta)$. By (5.27) of Lemma 5.3, we obtain (using the same argument) that for every pair (i^l, j^l) of the lowest alignment such that $i^l \le \bar{i}$, the following inequality holds: $H(i^h) - L(i^h) \le 2n'(1 + \Delta)$. Since the function $H - L$ is piecewise linear, we obtain that $\sup_{x\in[0,\bar{i}]}(H(x) - L(x)) \le 2n'(1 + \Delta)$.

The rest of the proof is the same as the one of Theorem 1.1. □

6. Proof of Theorem 1.3

When dealing with the sequences of random lengths, it is more convenient to consider the locations of ancestors. Recall the i.i.d. vectors U_i as defined in (4.1). Thus, given $k, l \in \mathbb{N}$ $(k < l)$, with some abuse of terminology, we shall call the highest (lowest) alignment of $U_{k+1}, \dots, U_{k+l}$ the highest (lowest) alignment between these X and Y sequences that have the ancestors in the interval $[k + 1, k + l]$. Note that these sequences as well as corresponding optimal alignments are all functions of $U_{k+1}, \dots, U_{k+l}$, only. This justifies the terminology. Hence, the highest alignment of the random lengths sequences X and Y (as defined above) is the highest alignment of $U_1, \dots, U_{m(n)}$.

Let now $k = 0$ and let $l \ge 1$ be fixed. We shall consider the vectors $U_1, \dots, U_l$ and the corresponding X and Y sequences. Thus, $n_x(l) := \sum_{j=1}^{l} D^x_j$, $n_y(l) := \sum_{j=1}^{l} D^y_j$ are their lengths. Let us define the events

$$A^x_l := \left\{ \left| n_x(l) - lp \right| < \frac{p}{2} l \right\}, \qquad A^y_l := \left\{ \left| n_y(l) - lp \right| < \frac{p}{2} l \right\}, \qquad A_l := A^x_l \cap A^y_l.$$

For a fixed $\Delta > 0$, let

$$C_l(\Delta) := A_l \cap \left\{ \left| n_x(l) - n_y(l) \right| < \frac{p}{2} \Delta l \right\}.$$

Hence, on C_l the following inequalities hold true:

$$l\frac{p}{2} < n_x(l) \wedge n_y(l) \le n_x(l) \vee n_y(l) < \frac{3p}{2} l, \qquad n_x(l) \vee n_y(l) \le \big(n_x(l) \wedge n_y(l)\big)(1 + \Delta).$$

Let

$$\mathcal{B}_l(\Delta) := \left\{(\tilde{n},\tilde{m}) \in \mathbb{N}^2 : l\frac{p}{2} < \tilde{n}\wedge\tilde{m} \le \tilde{n}\vee\tilde{m} < \frac{3p}{2}l, \tilde{n}\vee\tilde{m} \le (\tilde{n}\wedge\tilde{m})(1+\Delta)\right\}.$$

Thus,

$$C_l \subset \bigcup_{(\tilde{n},\tilde{m})\in\mathcal{B}_l(\Delta)} \left\{n_x(l) = \tilde{n}, n_y(l) = \tilde{m}\right\}. \tag{6.1}$$

With applying Hoeffding's inequality three times, it is easy to see the existence of a constant c_1 (depending on Δ and p) so that

$$P(C_l) \ge 1 - 6\exp[-c_1 l]. \tag{6.2}$$

The B-event for the sequences of random lengths

We shall now study the random lengths analogue of the B-events. Recall that the event $B_0(\tilde{n},\tilde{m})$ states that the highest alignment between the sequences $X_1,\dots,X_{\tilde{n}}$ and $Y_1,\dots,Y_{\tilde{m}}$ contains a related pair. We shall define now the event

$$E_k(l) := \{\text{the highest alignment of } U_{k+1},\dots,U_{k+l} \text{ contains a related pair}\}.$$

We shall bound the probability of $P(E_k(l))$. Clearly $P(E_k(l)) = P(E_0(l))$ for every $k = 1,2,\dots$, hence we shall consider the event $E_0(l)$. Obviously,

$$\bigcup_{(\tilde{n},\tilde{m})\in\mathcal{B}_l(\Delta)} \left(B_0(\tilde{n},\tilde{m}) \cap \left\{n_x(l) = \tilde{n}, n_y(l) = \tilde{m}\right\}\right) \subset E_0(l).$$

Since

$$P\left(B_0(\tilde{n},\tilde{m}) \cap \left\{n_x(l) = \tilde{n}, n_y(l) = \tilde{m}\right\}\right) \ge P\left(n_x(l) = \tilde{n}, n_y(l) = \tilde{m}\right) - P\left(B_0^c(\tilde{n},\tilde{m})\right).$$

Since $\tilde{n}$ and $\tilde{m}$ belong to $\mathcal{B}_l(\Delta)$, by Lemma 5.2, we obtain

$$P\left(B_0^c(\tilde{n},\tilde{m})\right) \le \exp\left[-b_3\frac{p}{2}l\right],$$

provided that l is big and Δ small enough. Thus, by (6.1) and (6.2), we obtain

$$\begin{aligned} P\left(E_0(l)\right) &\ge \sum_{(\tilde{n},\tilde{m})\in\mathcal{B}_l(\Delta)} \left(P\left(n_x(l) = \tilde{n}, n_y(l) = \tilde{m}\right) - P\left(B_0^c(\tilde{n},\tilde{m})\right)\right) \\ &\ge P(C_l) - \left|\mathcal{B}_l(\Delta)\right| \exp\left[-b_3\frac{p}{2}l\right] \ge 1 - 4\exp[-c_1 l] - (pl)^2\exp\left[-b_3\frac{p}{2}l\right]. \end{aligned}$$

Hence, there exists l_o and a constant $c_2 > 0$ (both depending on Δ and p) so that for any $l > l_o$,

$$P\left(E_0(l)\right) \ge 1 - \mathrm{e}^{-c_2 l}. \tag{6.3}$$

Let now $\underline{l}(n) := \frac{2}{p} A \ln n$ and we assume n to be fixed and so big that $\underline{l} > l_o$ so that (6.3) holds for any $l \geq \underline{l}$. Let, for any $k = 0, 1, \ldots$

$$E_k := \bigcup_{l \geq \underline{l}} E_k(l), \qquad E^h := \bigcup_{k=0}^{m-\underline{l}} E_k.$$

When the event E^h holds, then the following is true: the highest alignment of $U_{k+1}, \ldots, U_{k+l}$ contains a related pair whenever $l \geq \underline{l}$ and $k + l \leq m$. By (6.3), we obtain the estimate

$$P\big(E_k^c\big) \leq \sum_{l \geq \underline{l}} P\big(E_k^c(l)\big) \leq \sum_{l \geq \underline{l}} \mathrm{e}^{-c_2 l} \leq K \mathrm{e}^{-c_2 \underline{l}} = K n^{-2c_2 A/p}, \tag{6.4}$$

where K is a constant. Thus,

$$P\big(\big(E^h\big)^c\big) \leq m(n) K n^{-2c_2 A/p} = \frac{n}{p} K n^{-2c_2 A/p} = \frac{K}{p} n^{1-2c_2 A/p}.$$

The event E^h was defined for the highest alignment. Similar event, let it be E^l can be defined for the lowest alignment. The bound (6.4) holds also for E^l. Hence, with $E := E^h \cap E^l$, we obtain that $P(E) \geq 1 - 2Kp^{-1}n^{-2c_2 A/p}$. Now we are ready to prove Theorem 1.3.

Proof of Theorem 1.3. Choose $1 > \Delta > 0$ so small that (5.9) holds. Now let $c_2(\Delta)$ be as in (6.3) and choose A so big that $\frac{2c_2(\Delta)A}{p} > 3$. By (6.4), the event E holds then with probability at least $1 - 2Kp^{-1}n^{-2}$. Now proceed as in the proof of Lemma 5.3. Let $a_1^h, \ldots, a_{K^h}^h$ the ancestors of all related pairs in the highest alignment. Let $a_0^h := 0$ and $a_{K^h+1}^h := m + 1$. Assume that E holds. Then we have that for every $k = 0, \ldots, K^h$, $a_{k+1}^h - a_k^h < \underline{l}$. Hence, for every pair of the highest alignment (i^h, j^h), there exists $k \in \{0, \ldots, K^h\}$ such that $a_k^h \leq a^x(i^h) \leq a_{k+1}^h$ so that $|a^x(i^h) - a_k^h| \vee |a^x(i^h) - a_{k+1}^h| \leq \underline{l} = \frac{2}{p} A \ln n$. Clearly,

$$\big(i_{k+1}^{*h} - i_k^{*h}\big) \vee \big(j_{k+1}^{*h} - j_k^{*h}\big) \leq \big|a_{k+1}^h - a_k^h\big| \leq \frac{2}{p} A \ln n$$

and also $|i_k^{*h} - i^h| \vee |i_{k+1}^{*h} - i^h| \leq \frac{2}{p} A \ln n$.

Similarly, by E^l, there exists $0 \leq l \leq K^l$ so that $a_l^l \leq a^x(i^h) \leq a_{l+1}^l$, where a_l^l, $k = 1, \ldots, K^l$ are the ancestors of the related pairs in the lowest alignment, $a_0^l := 0$ and $a_{K^l+1}^l := m + 1$. Thus,

$$\big(i_{l+1}^{*l} - i_l^{*l}\big) \vee \big(j_{l+1}^{*l} - j_l^{*l}\big) \leq \big|a_{l+1}^l - a_l^l\big| \leq \frac{2}{p} A \ln n$$

and also $|i_l^{*l} - i^h| \vee |i_{l+1}^{*l} - i^h| \leq \frac{2}{p} A \ln n$. Hence, the inequalities (5.30) and (5.31) hold. Now proceed as in the proof of Lemma 5.3 and Theorem 1.1 to see that

$$P\left(h(H, L) > \frac{2}{p} A \ln n\right) \leq P\big(E^c\big) \leq 2Kp^{-1}n^{-2}$$

so that (1.11) holds with $C_r := \frac{2}{p}A$ and $D_r := 2Kp^{-1}$, where K is as in (6.4). □

7. Simulations

We now present some simulations about the growth of the distance between the extremal alignments as well as another statistics. In simulations, for different n-s up to 10 000, 100 pairs of i.i.d. sequences of length n with were generated. Half of them were independent i.i.d. sequences with X_1 and Y_1 distributed uniformly over four letter alphabet. Another half of the sequences were related with following parameters: the common ancestor process $Z_1, Z_2, \ldots$ is i.i.d. with Z_1 being uniformly distributed over four letter. The mutation matrix for generating X and Y sequences were the following:

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

The deletion probability $1 - p = 0.05$. Thus, the mutation matrix is such that $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ were, as for unrelated case, i.i.d. sequences with X_1 and Y_1 distributed uniformly over four letter alphabet, but the sequences X and Y are clearly not independent any more. The same models were used in generating Figures 1 and 2. Since, the marginal distributions of X and Y are uniform, we have $p_o = \overline{p} = \frac{1}{4}$ and $q = \frac{3}{4}$. From the mutation matrix, it follows that for related sequences $\overline{q} = 1 - 0.02 = 0.98$. Hence, for related sequences $\rho = \frac{\overline{q}}{q} > 1$ and the left-hand side of (1.8) is (clearly $\gamma_{\mathrm{R}} > 0.5$)

$$\begin{aligned} &\gamma_{\mathrm{R}} \log_2 \overline{p} + (1 - \gamma_{\mathrm{R}})\big(\log_2(\overline{q}q) + \log_2\big(\overline{q}q^{-1}\big)\big) + 2h(\gamma_{\mathrm{R}}) \\ &\quad = -2\gamma_{\mathrm{R}} + 2(1 - \gamma_{\mathrm{R}}) \log_2(0.98) + 2h(\gamma_{\mathrm{R}}). \end{aligned}$$

Hence, (1.8) in this case is

$$h(\gamma_{\mathrm{R}}) < \gamma_{\mathrm{R}} - (1 - \gamma_{\mathrm{R}}) \log_2(0.98). \tag{7.1}$$

The condition (7.1) holds, if γ_{R} is big enough. Since the solution of $h(x) = x - (1-x)\log_2(0.98)$ is about 0.770481, (7.1) holds if and only if $\gamma_{\mathrm{R}} > 0.770481$. From Figure 2, we estimate γ_{R} as $\frac{747}{949} = 0.787$. Another simulations confirm that γ_{R} is somewhere around 0.79. Thus, it is reasonable believe that for our model, the condition (1.8) holds true. Recall that for independent sequences, (1.8) always fails.

In both cases – related and unrelated sequences – the average of following statistics were found: L_n, the horizontal length of the maximum non-uniqueness stretch, the maximum vertical distance and the maximum (full) Hausdorff's distance.

The top-plot in Figure 4 shows the growth of L_n as n grows. 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

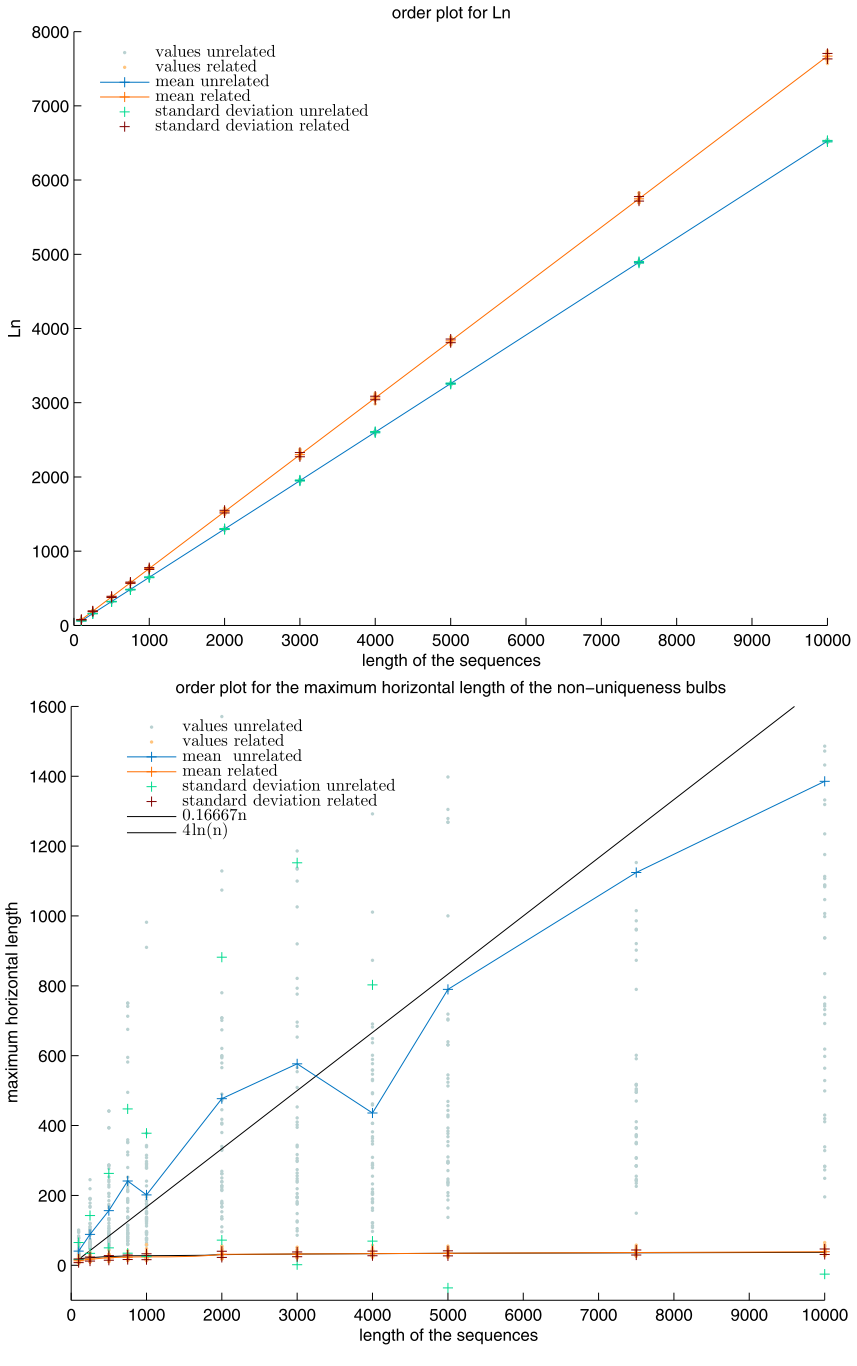


Figure 4. Growth of L_n (top). Growth of non-uniqueness stretch (bottom).

both cases, the slope, however, is different: the upper line corresponds to the related sequences, the lower line is for independent sequences.

The bottom-plot in Figure 4 shows the horizontal length of maximum non-uniqueness stretch. For independent sequences (upper curve), the growth is, perhaps, smaller than linear but considerably faster than logarithmic. The straight line is, in some sense, the best linear approximation. The +-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 (lower 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 remarkably smaller and this only confirms the logarithmic growth.

In Figure 5, the maximum vertical distance (top) and (full) Hausdorff's distance with respect to the maximum-norm (bottom) are plotted. Both pictures are similar to the bottom picture of Figure 4 and can be interpreted analogously. For the related case, the growth is clearly logarithmic (the best approximation is $1.25 \ln n$ for maximum vertical distance, and $0.65 \ln n$ for Hausdorff's distance) and that is a full correspondence with Theorems 1.1, 1.2 and 1.3. Note that we have used the full Hausdorff's distance instead of the restricted one so that the simulations confirm the conjecture that Theorems 1.1 and 1.2 also hold with h instead of h_o.

In Figure 6, there is a zoomsection fragment of two extremal alignment of the related sequences. Recall the definition of related pairs – the corresponding sites have the same ancestor. It does not necessarily mean that they have the color of the common ancestor, but often it is so. In the last picture, the related pars *with the color of the common ancestor* are marked with dots. Note that in some small region, there are relatively many those pairs, on same other region, there are less those pairs. The picture (and other similar simulations) also shows that in the regions with many these pairs, the extremal alignments coincide with them. This means that in both sequences, there are parts that relatively less mutated and the behavior of the extremal alignments indicate the existence of such region rather well. In the area with relatively few dots, the extremal alignments fluctuate indicating that in this part (at least in one sequence) many mutations have been occurred. Hence, based on these simulations, we can conclude the extremal alignments are rather good tools for finding the less mutated regions and obtaining information about the common ancestor.

Appendix

In the following, let $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ be related sequences. Recall, that our model for related sequences incorporates the independent case. Recall the convergence (1.4):

$$\frac{1}{n} L(X_1, \ldots, X_{\lfloor na \rfloor}; Y_1, \ldots, Y_n) \to \gamma_{\mathrm{R}}(a), \qquad \text{a.s.}$$

Lemma A.1. *For every* $0 < a < 1$, $\gamma_{\mathrm{R}}(a) < \gamma_{\mathrm{R}}$, *for every* $a > 1$, $\gamma_{\mathrm{R}}(a) > \gamma_{\mathrm{R}}$.

Proof. Clearly the function $a \mapsto \gamma_{\mathrm{R}}(a)$ is nondecreasing in a and there exists a $K \in \mathbb{N}$ so big that $\gamma_{\mathrm{R}}(K) > \gamma_{\mathrm{R}}$.

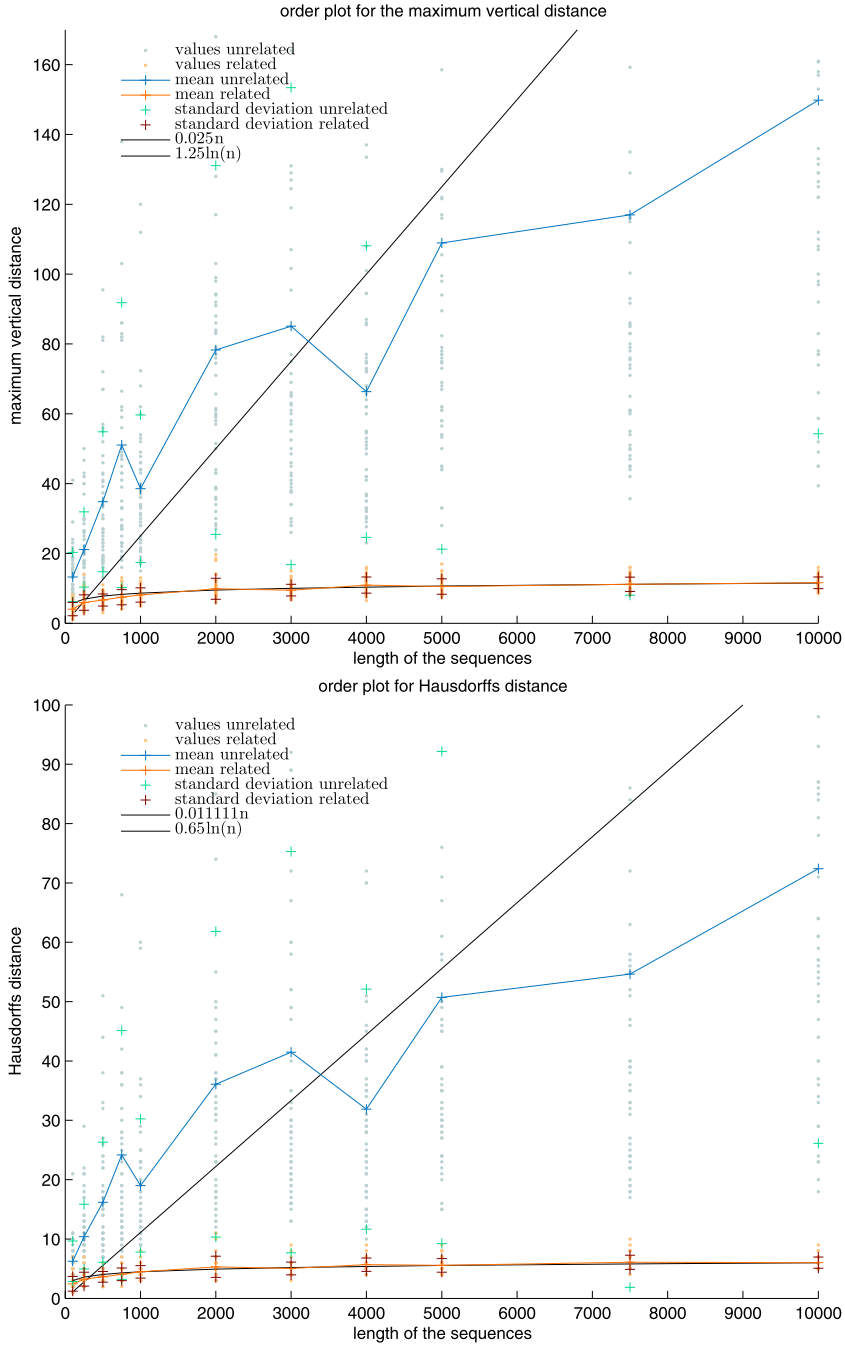


Figure 5. Growth of maximal vertical distance (top). Growth of Hausdorff's distance (bottom).

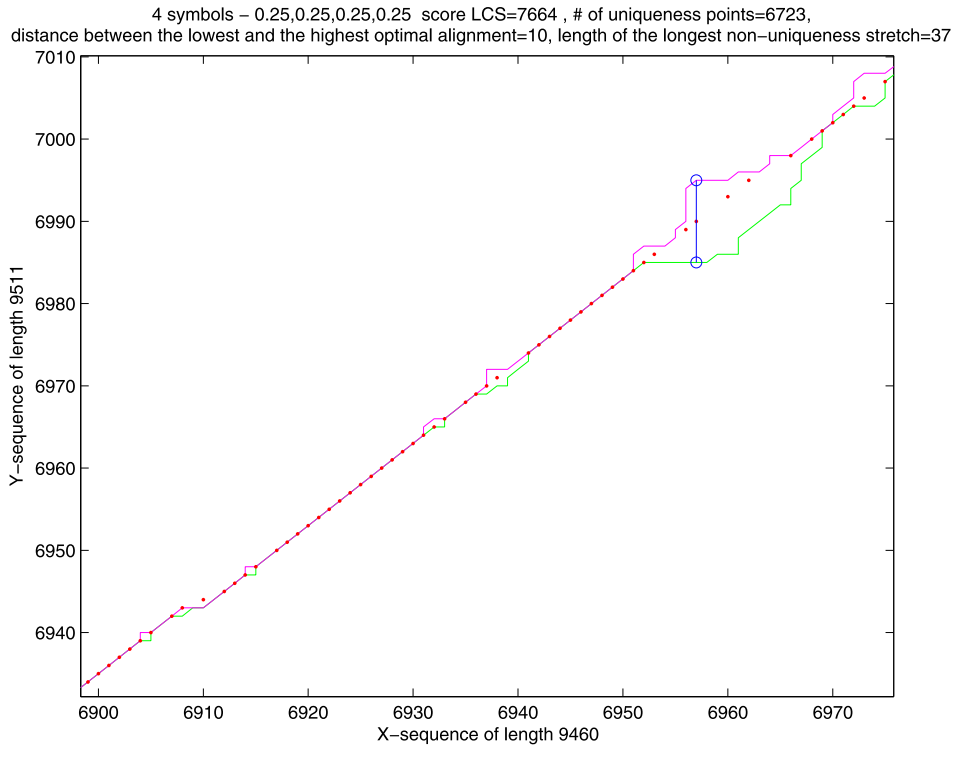


Figure 6. The related pairs with the color of the common ancestor (dots) and extremal alignments.

Fix $0 < a < 1$ and choose $\varepsilon > 0$ be so small that $\frac{1-a}{\varepsilon} > K$. For every $m \in \mathbb{Z}$, $a, b > 0$ let

$$L_{m:an,bn} := L(X_{m+1}, \ldots, X_{\lfloor na \rfloor}; Y_{m+1}, \ldots, Y_{\lfloor nb \rfloor}).$$

Let $c := 1 - a$. By superadditivity,

$$L_{n(1+c),n} \geq L_{n(1-\varepsilon),n(1-\varepsilon)} + L_{\lfloor n(1-\varepsilon) \rfloor : n(1+c),n}.$$

Let $m = \lfloor n(1-\varepsilon) \rfloor$. Since, $c > K\varepsilon$ and for every $c \geq 0$,

$$\lfloor n(1-\varepsilon) \rfloor + \lfloor n(c+\varepsilon) \rfloor \leq \lfloor \lfloor n(1-\varepsilon) \rfloor + n(c+\varepsilon) \rfloor \leq \lfloor n(1+c) \rfloor,$$

it holds

$$L_{m:n(1+c),n} \geq L_{m:m+n(c+\varepsilon),m+n\varepsilon} \geq L_{m:m+\lfloor n\varepsilon \rfloor K,m+\lfloor n\varepsilon \rfloor}.$$

Clearly

$$\lim_n \frac{1}{n} L_{n(1-\varepsilon),n(1-\varepsilon)} = (1-\varepsilon) \lim_{u\to\infty} \frac{L_u}{u} = (1-\varepsilon)\gamma_{\mathrm{R}} \qquad \text{a.s.} \tag{A.1}$$

Let us now show that

$$\lim_n \frac{1}{n} L_{m:m+\lfloor n\varepsilon\rfloor K, m+\lfloor n\varepsilon\rfloor} = \lim_u \frac{\varepsilon}{u} L_{Ku,u} = \gamma_{\mathrm{R}}(K)\varepsilon \qquad \text{a.s.} \tag{A.2}$$

For independent sequences, (A.2) follows from (3.2). Indeed, for i.i.d. sequences, the random variables $L_{m:m+\lfloor n\varepsilon\rfloor K, m+\lfloor n\varepsilon\rfloor}$ and $L_{\lfloor n\varepsilon\rfloor K, \lfloor n\varepsilon\rfloor}$ are identically distributed. By (3.2), thus, for any $\Delta > 0$ (and ignoring $\lfloor\cdot\rfloor$, for simplicity)

$$\begin{aligned}
P\big(\big|L_{m:m+n\varepsilon K, m+n\varepsilon} - \gamma(K)\varepsilon n\big| > \Delta n\big) &= P\big(\big|L_{n\varepsilon K, n\varepsilon} - \gamma(K)\varepsilon n\big| > \Delta n\big) \\
&= P\left(\left|L_{n\varepsilon K, n\varepsilon} - \frac{1}{K}\gamma(K)K\varepsilon n\right| > \Delta n\right) \\
&= P\left(\left|L_{n\varepsilon K, n\varepsilon} - \gamma\left(\frac{1}{K}\right)K\varepsilon n\right| > \Delta n\right) \\
&= P\left(L_{uk,k} - \gamma(u)k > \Delta\frac{uk}{\varepsilon}\right) \\
&\le 2\exp\left[-\frac{\Delta^2 u^2}{2\varepsilon^2(1+u)}k\right] = 2\exp\left[-\frac{\Delta^2}{\varepsilon(K+1)}n\right].
\end{aligned}$$

Here $u = \frac{1}{K}$ and $k = \varepsilon n K$. In the third equality, the relation $K\gamma(\frac{1}{K}) = \gamma(K)$ is used. For related sequences, the random variables $L_{m:m+\lfloor n\varepsilon\rfloor K, m+\lfloor n\varepsilon\rfloor}$ and $L_{\lfloor n\varepsilon\rfloor K, \lfloor n\varepsilon\rfloor}$ are not necessarily identically distributed, hence another argument should be used. Let $\overline{a}(m) := a^x(m) \vee a^y(m)$ and $\underline{a}(m) := a^x(m) \wedge a^y(m)$. Let Y_{m+k^y} (resp., X_{m+k^x}) be the smallest element in Y (resp., in X) that has ancestor at least $\overline{a}(m)$. Similarly, let X_{m-l^x} (resp., Y_{m-l^y}) be the smallest element in X (resp., in Y) that has ancestor at least $\underline{a}(m)$. If $a^x(m) \ge a^y(m)$, then $k^x = l^y = 0$ and if $a^x(m) \le a^y(m)$, then $k^y = l^x = 0$. Hence,

$$\begin{aligned}
&L(X_{m-l^x+1}, \dots, X_{m+K\lfloor n\varepsilon\rfloor}; Y_{m-l^y+1}, \dots, Y_{m+\lfloor n\varepsilon\rfloor}) \\
&\quad \ge L_{m:m+\lfloor n\varepsilon\rfloor K, m+\lfloor n\varepsilon\rfloor} \qquad (A.3) \\
&\quad \ge L(X_{m+k^x+1}, \dots, X_{m+K\lfloor n\varepsilon\rfloor}; Y_{m+k^y+1}, \dots, Y_{m+\lfloor n\varepsilon\rfloor}). \qquad (A.4)
\end{aligned}$$

Note that the random variables $X_{m+k^x+1}, X_{m+k^x+2}, \dots$ and $Y_{m+k^y+1}, Y_{m+k^y+2}, \dots$ depend on i.i.d. random vectors $U_{\overline{a}(m)+1}, U_{\overline{a}(m)+2}, \dots$, where U_i is defined as in (4.1). Similarly $X_{m-l^x+1}, X_{m-l^x+2}, \dots$ and $Y_{m-l^y+1}, Y_{m-l^y+2}, \dots$ depend on i.i.d. random vectors $U_{\underline{a}(m)+1}$, $U_{\underline{a}(m)+2}, \dots$. Hence, the random variables

$$L(X_{m+k^x+1}, \dots, X_{m+k^x+K\lfloor n\varepsilon\rfloor}; Y_{m+k^y+1}, \dots, Y_{m+k^y+\lfloor n\varepsilon\rfloor})$$

and

$$L(X_1, \dots, X_{K\lfloor n\varepsilon\rfloor}; Y_1, \dots, Y_{\lfloor n\varepsilon\rfloor})$$

have the same distribution so that (as in the independent case) by (4.9)

$$P\big(\big|L(X_{m+k^x+1},\dots,X_{m+k^x+K\lfloor n\varepsilon\rfloor};Y_{m+k^y+1},\dots,Y_{m+k^y+\lfloor n\varepsilon\rfloor})-\varepsilon\gamma_{\mathrm{R}}(K)n\big|>\Delta n\big)$$
$$=P\big(\big|L(X_1,\dots,X_{K\lfloor n\varepsilon\rfloor};Y_1,\dots,Y_{\lfloor n\varepsilon\rfloor})-\varepsilon\gamma_{\mathrm{R}}(K)n\big|>\Delta n\big)\le 4\exp\bigg[-\frac{1}{32}\frac{\Delta^2}{\varepsilon K^2}n\bigg].$$

Thus, as n grows,

$$\frac{1}{n}L(X_{m+k^x+1},\dots,X_{m+k^x+K\lfloor n\varepsilon\rfloor};Y_{m+k^y+1},\dots,Y_{m+k^y+\lfloor n\varepsilon\rfloor})\to\varepsilon\gamma_{\mathrm{R}}(K)\quad\text{a.s.}$$

The random variables $k:=k^x\vee k^y$ and $l:=l^x\vee l^y$ satisfy

$$k\vee l\le\overline{a}(m)-\underline{a}(m).$$

Note that

$$\sum_{i=1}^{a^x(m)}D_i^x=\sum_{i=1}^{a^y(m)}D_i^y=m.$$

Now, using Hoeffding inequality for i.i.d. random variables D_i^x and D_i^y, it is easy to see that

$$\frac{a^x(m)}{m}\to\frac{1}{p}\quad\text{a.s.},\qquad\frac{a^y(m)}{m}\to\frac{1}{p}\quad\text{a.s.}$$

Therefore,

$$\frac{k\vee l}{m}\le\frac{\overline{a}(m)-\underline{a}(m)}{m}\to 0\quad\text{a.s.}$$

so that $\frac{k(n)}{n}\to 0$ a.s. and $\frac{l(n)}{n}\to 0$ a.s. Since

$$\big|L(X_{m+k^x+1},\dots,X_{m+K\lfloor n\varepsilon\rfloor};Y_{m+k^y+1},\dots,Y_{m+\lfloor n\varepsilon\rfloor})$$
$$-L(X_{m+k^x+1},\dots,X_{m+k^x+K\lfloor n\varepsilon\rfloor};Y_{m+k^y+1},\dots,Y_{m+k^y+\lfloor n\varepsilon\rfloor})\big|\le k(n),$$

from $\frac{k(n)}{n}\to 0$ a.s., we get that

$$\lim_n\frac{1}{n}L(X_{m+k^x+1},\dots,X_{m+K\lfloor n\varepsilon\rfloor};Y_{m+k^y+1},\dots,Y_{m+\lfloor n\varepsilon\rfloor})$$
$$=\lim_u\frac{1}{u}\varepsilon L_{Ku,u}=\gamma_{\mathrm{R}}(K)\varepsilon\quad\text{a.s.}\tag{A.5}$$

By similar argument,

$$\lim_n\frac{1}{n}L(X_{m-l^x+1},\dots,X_{m+K\lfloor n\varepsilon\rfloor};Y_{m-l^y+1},\dots,Y_{m+\lfloor n\varepsilon\rfloor})$$
$$=\lim_u\frac{1}{u}\varepsilon L_{Ku,u}=\gamma_{\mathrm{R}}(K)\varepsilon\quad\text{a.s.}\tag{A.6}$$

The inequalities (A.5) and (A.6) together with (A.3) and(A.4) imply (A.2). The convergences (A.1) and (A.2) imply

$$\lim_n \frac{1}{n} L_{n(1+c),n} = \gamma_{\mathrm{R}}(1+c) > \gamma_{\mathrm{R}} \qquad \text{a.s.} \tag{A.7}$$

The limit in (A.7) exists by Proposition 4.1, the inequality $\gamma_{\mathrm{R}}(1+c) > \gamma_{\mathrm{R}}$ follows from (A.1) and (A.2), since $\varepsilon\gamma_{\mathrm{R}}(K) + (1-\varepsilon)\gamma_{\mathrm{R}} > \gamma_{\mathrm{R}}$. This proves that $\gamma_{\mathrm{R}}(a) > \gamma_{\mathrm{R}}$, when $a > 1$.

Finally,

$$\frac{1}{2n} L_{2n,2n} \geq \frac{1}{2n} L_{n(1+c),n} + \frac{1}{2n} L(X_{\lfloor n(1+c)\rfloor+1}, \ldots, X_{2n}; Y_{n+1}, \ldots, Y_{2n}).$$

Since $\frac{1}{2n} L_{2n,2n} \to \gamma_{\mathrm{R}}$, a.s. and, using the same argument as proving (A.2), we get

$$\lim_n \frac{1}{2n} L(X_{\lfloor n(1+c)\rfloor+1}, \ldots, X_{2n}; Y_{n+1}, \ldots, Y_{2n}) = \frac{1}{2} \lim_n \frac{1}{n} L_{(1-c)n,n} = \frac{\gamma_{\mathrm{R}}(1-c)}{2} \qquad \text{a.s.},$$

by (A.7), we have $\gamma_{\mathrm{R}} \geq \frac{\gamma_{\mathrm{R}}(1+c)}{2} + \frac{\gamma_{\mathrm{R}}(1-c)}{2} > \frac{\gamma_{\mathrm{R}}+\gamma_{\mathrm{R}}(1-c)}{2}$. This implies that $\gamma_{\mathrm{R}}(a) = \gamma_{\mathrm{R}}(1-c) < \gamma_{\mathrm{R}}$. □

The following corollary generalizes Proposition 3.1 for related sequences. Moreover, we allow the sequences to be unequal length. Hence, we consider the case $X = X_1, \ldots, X_n$, $Y = Y_1, \ldots, Y_m$, $n \leq m \leq n(1+\Delta)$, where $\Delta \geq 0$. The case $\Delta = 0$ corresponds to the case $m = n$. Recall the random variables $S := j_1^h - 1$ and $T := n - i_k^h$, that obviously are the functions of X and Y. The proof of the following corollary is very similar to that one of Proposition 3.1.

Corollary A.1. *Let $1 > c > \Delta$. Then there exists constant $d(c) > 0$, so that, for n big enough, $P(T > cn) \leq \exp[-dn]$, $P(S > cn) \leq \exp[-dn]$.*

Proof. As in the proof of Proposition 3.1, note that for any $\bar{\gamma}$,

$$\{S > cn\} \subset \{L_{n,m-cn} = L_{n,m}\} \subset \{L_{n,m-cn} \geq \bar{\gamma} n\} \cup \{L_{n,m} \leq \bar{\gamma} n\}.$$

By Lemma A.1, $\gamma_{\mathrm{R}} > \gamma_{\mathrm{R}}(1+\Delta-c)$. Let $\bar{\gamma} := \frac{1}{2}(\gamma_{\mathrm{R}} + \gamma_{\mathrm{R}}(1+\Delta-c))$. Let $\varepsilon := \gamma_{\mathrm{R}} - \bar{\gamma}$. Since $L_{n,m-cn} \leq L_{n,(1+\Delta-c)n}$ and $L_n = L_{n,n} \leq L_{n,m}$, Corollary 4.1 states that for n big enough,

$$\begin{aligned} P(S > cn) &\leq P(L_{n,(1+\Delta-c)n} \geq \bar{\gamma} n) + P(L_n \leq \bar{\gamma} n) \\ &= P\big(L_{n,(1+\Delta-c)n} \geq \big(\gamma_{\mathrm{R}}(1+\Delta-c) + \varepsilon\big)n\big) + P\big(L_n \leq (\gamma_{\mathrm{R}} - \varepsilon)n\big) \\ &\leq 8 \exp\left[-\frac{p}{32}(1+\Delta-c)\varepsilon^2 n\right]. \end{aligned}$$

This concludes the proof. □

Acknowledgements

Supported by the Estonian Science Foundation Grant nr. 9288; SFB 701 of Bielefeld University and targeted financing project SF0180015s12.

References

[1] Alexander, K.S. (1994). The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.* **4** 1074–1082. MR1304773

[2] Amsalu, S., Matzinger, H. and Popov, S. (2007). Macroscopic non-uniqueness and transversal fluctuation in optimal random sequence alignment. *ESAIM Probab. Stat.* **11** 281–300. MR2320822

[3] Arratia, R. and Waterman, M.S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.* **4** 200–225. MR1258181

[4] Baeza-Yates, R.A., Gavaldà, R., Navarro, G. and Scheihing, R. (1999). Bounding the expected length of longest common subsequences and forests. *Theory Comput. Syst.* **32** 435–452. MR1693383

[5] Barder, S., Lember, J., Matzinger, H. and Toots, M. (2012). On suboptimal LCS-alignments for independent Bernoulli sequences with asymmetric distributions. *Methodol. Comput. Appl. Probab.* **14** 357–382. MR2912343

[6] Bonetto, F. and Matzinger, H. (2006). Fluctuations of the longest common subsequence in the asymmetric case of 2- and 3-letter alphabets. *ALEA Lat. Am. J. Probab. Math. Stat.* **2** 195–216. MR2262762

[7] Chao, K.M. and Zhang, L. (2009). *Sequence Comparison*: *Theory and Methods*. *Computational Biology*. London: Springer MR2723076

[8] Chvatal, V. and Sankoff, D. (1975). Longest common subsequences of two random sequences. *J. Appl. Probability* **12** 306–315. MR0405531

[9] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis*: *Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge Univ. Press.

[10] Hansen, N.R. (2006). Local alignment of Markov chains. *Ann. Appl. Probab.* **16** 1262–1296. MR2260063

[11] Hirmo, E., Lember, J. and Matzinger, H. (2012). Detecting the homology of DNA-sequence based on the variety of optimal alignments: A case study. Available at arXiv:1210.3771.

[12] Houdré, C., Lember, J. and Matzinger, H. (2006). On the longest common increasing binary subsequence. *C. R. Math. Acad. Sci. Paris* **343** 589–594. MR2269870

[13] Kiwi, M., Loebl, M. and Matoušek, J. (2005). Expected length of the longest common subsequence for large alphabets. *Adv. Math.* **197** 480–498. MR2173842

[14] Lember, J. and Matzinger, H. (2009). Standard deviation of the longest common subsequence. *Ann. Probab.* **37** 1192–1235. MR2537552

[15] Lember, J., Matzinger, H. and Torres, F. (2012). The rate of the convergence of the mean score in random sequence comparison. *Ann. Appl. Probab.* **22** 1046–1058. MR2977985

[16] Lember, J., Matzinger, H. and Vollmer, A. (2007). Path properties of LCS-optimal alignments. SFB 701 Preprintreihe, Univ. Bielefeld (07 - 77).

[17] Matzinger, H., Lember, J. and Durringer, C. (2007). Deviation from mean in sequence comparison with a periodic sequence. *ALEA Lat. Am. J. Probab. Math. Stat.* **3** 1–29. MR2324746

[18] Siegmund, D. and Yakir, B. (2000). Approximate p-values for local sequence alignments. *Ann. Statist.* **28** 657–680. MR1792782

[19] Steele, J.M. (1986). An Efron-Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758. MR0840528

[20] Waterman, M.S. (1994). Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B* **344** 383–390.
[21] Waterman, M.S. (1995). *Introduction to Computational Biology*. London: Chapman & Hall.
[22] Waterman, M.S. and Vingron, M. (1994). Sequence comparison significance and Poisson approximation. *Statist. Sci.* **9** 367–381. MR1325433

Received December 2011 and revised November 2012