Closeness to the diagonal for longest common subsequences in random words

Christian Houdré[*]   Heinrich Matzinger[†]

Abstract

The nature of the alignment with gaps corresponding to a longest common subsequence (LCS) of two independent iid random sequences drawn from a finite alphabet is investigated. It is shown that such an optimal alignment typically matches pieces of similar short-length. This is of importance in understanding the structure of optimal alignments of two sequences. Moreover, it is also shown that any property, common to two subsequences, typically holds in most parts of the optimal alignment whenever this same property holds, with high probability, for strings of similar short-length. Our results should, in particular, prove useful for simulations since they imply that the re-scaled two dimensional representation of a LCS gets uniformly close to the diagonal as the length of the sequences grows without bound.

1   Introduction

Let x and y be two finite strings. A common subsequence of x and y is a subsequence of both x and y, while a longest common subsequence (LCS) of x and y is a common subsequence of maximal length.

It is well known that common subsequences can be represented via *alignments with gaps* as illustrated, next, on some examples: First take the binary strings $x = 0010$ and $y = 0110$. A common subsequence is 01, which can be represented as an alignment with gaps as follows: The common letters are aligned together, while each letter not appearing in the common subsequence is aligned with a gap. Several alignments can thus represent the same common subsequence and in this first example, an alignment corresponding to the common subsequence 01 is given by

$$
\begin{array}{c|c|c|c|c|c|c|c}
x & & 0 & 0 & 1 & 0 & & \\
\hline
y & & & 0 & 1 & & 1 & 0
\end{array}
\tag{1.1}
$$

---

[*]Georgia Institute of Technology, School of Mathematics, Atlanta, GA, 30332-0160. E-mail: houdre@math.gatech.edu. Research partially supported by the grant # 246283 from the Simons Foundation and by a Simons Fellowship grant # 267336.

[†]Georgia Institute of Technology, School of Mathematics, Atlanta, GA, 30332-0160. E-mail: matzi@math.gatech.edu

while another one is given by

$$
\begin{array}{c|c|c|c|c|c|c|c|}
x & & 0 & 0 & 1 & 0 & & & \\
\hline
y & & 0 & & 1 & & 1 & 0 & \\
\end{array}
\tag{1.2}
$$

Above, the LCS is not 01 but rather 010. An alignment corresponding to a LCS is called an *optimal alignment* (OA) or is said to be *optimal*. Neither (1.1) nor (1.2) represent optimal alignments, but an optimal alignment is given by:

$$
\begin{array}{c|c|c|c|c|c|}
x & & 0 & 0 & 1 & & 0 \\
\hline
y & & & 0 & 1 & 1 & 0 \\
\end{array}
$$

which, again, is clearly not unique. Here the LCS of x and y is $LCS(x; y) = 010$, which has length three, a fact denoted by $|LCS(x; y)| = 3$. Here is another example: let $x = christian$ and $y = krystyaan$. Then, $LCS(x; y) = rstan$ and an alignment with gaps representing the LCS is:

$$
\begin{array}{c|c|c|c|c|c|c|c|c|c|c|}
c & & h & r & i & & s & t & i & & a & & n \\
\hline
& k & & r & & y & s & t & & y & a & a & n \\
\end{array}
\tag{1.3}
$$

Again, all the letters which are part of the LCS are aligned with one another, while the other letters are aligned with gaps. In the above alignment of x and y, $x_5 x_6 x_7 x_8 x_9 = stian$ is aligned (with gaps) with $y_4 y_5 y_6 y_7 y_8 y_9 = styaan$, and we say that the integer interval $[5, 9]_{\mathbb{N}}$ is aligned with $[4, 9]_{\mathbb{N}}$. Alternatively, we say that $[5, 9]_{\mathbb{N}}$ gets mapped to $[4, 9]_{\mathbb{N}}$ by the alignment we consider, meaning that the following two conditions are satisfied:

i) The letters $x_5 x_6 x_7 x_8 x_9$ are all aligned exclusively with gaps or with letters from the string $y_4 y_5 y_6 y_7 y_8 y_9$.

ii) The letters from $y_4 y_5 y_6 y_7 y_8 y_9$ are all aligned with gaps or with letters from the substring $x_5 x_6 x_7 x_8 x_9$.

To emphasize our terminology, we see that in the alignment (1.3), $[1, 4]_{\mathbb{N}}$ is aligned with $[1, 2]_{\mathbb{N}}$ ($[1, 4]_{\mathbb{N}}$ is also aligned with $[1, 3]_{\mathbb{N}}$). In other words, in an alignment a piece of x gets aligned with a piece of y if and only if the letters from the piece of x which get aligned to letters get only aligned with letters from the piece of y and vice versa. Longest common subsequences and optimal alignments are important tools used in Computational Biology and Computational Linguistics for strings matching, e.g., see [6], [20], [21], and [25].

Throughout this paper, $X = X_1 \cdots X_n$ and $Y = Y_1 \cdots Y_n$ are two random strings/words where $(X_i)_{i \geq 1}$ and $(Y_i)_{i \geq 1}$ are two independent sequences of iid random variables with values in a finite alphabet $\mathcal{A}$ and having the same law. (No other assumption, besides its non triviality, is made on this common law.) Further, and again throughout, LC_n denotes the length of the LCSs of X and Y, i.e., $LC_n := |LCS(X_1 X_2 \cdots X_n; Y_1 Y_2 \cdots Y_n)|$.

To further put our framework in context, also note that $2(n - LC_n)$ is a version of the edit/Levenshtein distance used, e.g., in computer science. It is equal to the minimal number of insertions and deletions to change either string/word into the other.

A well known result of Chvátal and Sankoff [7] asserts that, when scaled by n, $\mathbb{E}LC_n$ converges to a constant $\gamma^* \in (0, 1)$ (which depends on the size of the alphabet and on the law of X_1) given, via superadditivity, by $\gamma^* = \sup_{n \geq 1} \mathbb{E}LC_n / n$. However, to this day, even in the uniform binary case, the exact value of γ^* is unknown. Moreover, Alexander [1] determined the rate of convergence to γ^* showing that there exists an absolute constant K_A, independent of n, of the size of the finite alphabet $\mathcal{A}$, (and of the law of X_1), such that for all $n \geq 1$,

$$
\gamma^* n - K_A \sqrt{n \log n} \leq \mathbb{E}LC_n \leq \gamma^* n.
\tag{1.4}
$$

This law of large numbers (with rate) gives the first order behavior of LC_n, and the next problem of interest is to study the order of the variance (or more generally, of the centered absolute moments) of LC_n. This order has been found to be linear in the length of the sequences in various instances (e.g., [10], [14], [18], [13]). This is, in particular, the case for iid binary sequences with zeros and ones having very different probabilities or for some classes of models "as close as one wants" to the uniform iid one [2]. In all the known instances, it turns out that the order of the variance of LC_n is thus linear in n which is the order conjectured by Waterman [24], for which Steele [23] had previously obtained a generic linear upper bound. However, the most important equiprobable iid (say, binary) case remains open.

In the present paper our purpose is different and we study optimal alignments at a macroscopic level. We present a general methodology showing that a property of optimal alignments holds true, provided this property typically holds true for short-strings alignments. To prove these results, we partition X into pieces of *fixed* length k and, as n goes to infinity, show that typically, and further with probability exponentially close to one, in any optimal alignment most of these pieces get aligned with pieces of Y of similar length k. In other words, with high probability, there are no macroscopic gaps in the optimal alignments.

Let us explain, on a further example, how alignments have to stay close to the diagonal. Take the two related, English and German, words: $X = mother$ and $Y = mutter$. The longest common subsequence is $mter$, hence $LC_6 = 4$ and the common subsequence $mter$ corresponds to the following two alignments:

$$
\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
m & o & & t & & h & e & r \\
\hline
m & & u & t & t & & e & r \\
\hline
\end{array}
\tag{1.5}
$$

and

$$
\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
m & o & & & t & h & e & r \\
\hline
m & & u & t & t & & e & r \\
\hline
\end{array}
\tag{1.6}
$$

Next, view alignments as subsets of $\mathbb{R}^2$ as follows: If the i-th letter of X gets aligned with the j-th letter of Y, then the set representing the alignment is to contain (i,j). For example, the alignment (1.5) can be represented as: $(1,1),(3,3),(5,5),(6,6)$ with the corresponding plot

$$
\begin{array}{c|c|c|c|c|c|c|}
r & & & & & & \bullet \\
\hline
e & & & & & \bullet & \\
\hline
t & & & & & & \\
\hline
t & & & \bullet & & & \\
\hline
u & & & & & & \\
\hline
m & \bullet & & & & & \\
\hline
 & m & o & t & h & e & r \\
\end{array}
\tag{1.7}
$$
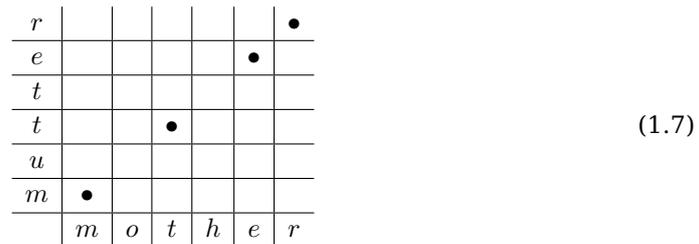
while the alignment (1.6) can be represented as: $(1,1),(3,4),(5,5),(6,6)$ with the corresponding plot

$$
\begin{array}{c|c|c|c|c|c|c|}
r & & & & & & \bullet \\
\hline
e & & & & & \bullet & \\
\hline
t & & & \bullet & & & \\
\hline
t & & & & & & \\
\hline
u & & & & & & \\
\hline
m & \bullet & & & & & \\
\hline
 & m & o & t & h & e & r \\
\end{array}
\tag{1.8}
$$
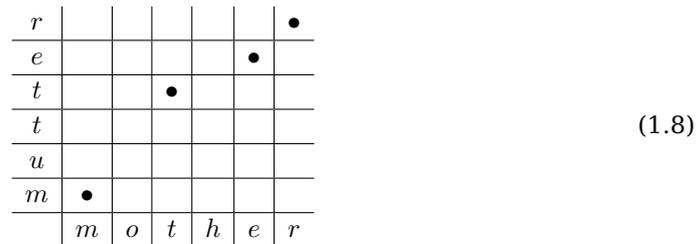
Above, the symbol $\bullet$ indicates the aligned letter-pairs, and these points are said to represent the optimal alignment. An optimal alignment can then be viewed as the graph

of a function f defined as follows: Let $f(0) = 0$ and for $i, j = 1, \ldots, n$, let $f(i) = j$ if the ith letter of the first sequence is aligned with the jth letter of the second sequence, while between these values and till (n, n) let f be defined via a linear interpolation. Then as shown in Section 4, with its notation, the function $f = f_{p_1, p_2} : [0, n] \to [0, n]$ stays between the two lines with respective slope p_1 and p_2. Moreover, under a strict concavity assumption on the limiting shape $\tilde{\gamma}$ (see the next section), f_{p_1, p_2} becomes uniformly close to the identity as n tends to infinity. More precisely, let $g_n : [0, 1] \to [0, 1]$ be defined via $g_n(x) = f(nx)/n$, then $\lim_{n \to +\infty} \sup_{0 \le x \le 1} |g_n(x) - x| = 0$. In other words, there are no macroscopic gaps in any optimal alignment and any such alignment must remain close to the main diagonal.

This closeness to the diagonal property has proved crucial in obtaining the first result on the limiting law of LC_n, under a lower bound on the order of the variance, see [9]. Broadly speaking, when not close to the diagonal, many terms contribute to our CLT estimation but the corresponding set of random alignments has exponentially small probability; while when alignments are close to the diagonal, the estimation comes from only a few terms. The balance between these two cases leads to a central limit theorem. This CLT contrasts with the case of two independent uniform random permutations of $\{1, 2, \cdots, n\}$, where the limiting distribution of the length of the longest common subsequences is the Tracy-Widom distribution, see [9]. The other, more pathological, instances, we are aware of and where a CLT holds true in a related problem, is for the length of the longest increasing subsequences of a single random word, potentially Markovian, where only one letter is attained with maximal probability, see [15], [16], [11], [12]. (However, for two or more random words the limiting law is no longer Gaussian [5].) This one word result contrasts, once more, with the single word permutation result of Baik, Deift and Johansson [4] where the limiting law is the Tracy-Widom one.

Note that the LCS problem can be represented as a directed last passage percolation (LPP) problem with dependent weights. Indeed, let the set of vertices be

$$V := \{0, 1, 2, \ldots, n\} \times \{0, 1, 2, \ldots, n\},$$

and let the set of oriented edges $\mathcal{E} \subset V \times V$ contain horizontal, vertical and diagonal edges. The horizontal edges are oriented to the right, while the vertical edges are oriented upwards, both having unit length. The diagonal edges point up-right at a 45-degree angle and have length $\sqrt{2}$. Hence,

$$\mathcal{E} := \{(v, v + e_1), (v, v + e_2), (v, v + e_3) : v \in V\},$$

where $e_1 := (1, 0)$, $e_2 := (0, 1)$ and $e_3 := (1, 1)$. With the horizontal and vertical edges, we associate a weight of 0. With the diagonal edge from (i, j) to $(i + 1, j + 1)$ we associate the weight 1 if $X_{i+1} = Y_{j+1}$ and 0 (or $-\infty$) otherwise. In this manner, we obtain that LC_n is equal to the total weight of the heaviest paths going from $(0, 0)$ to (n, n). (Another directed LPP representation can be obtained via $LC_n = \max_{\pi \in SI} \sum_{(i,j) \in \pi} \mathbf{1}_{\{X_i = Y_j\}}$, where SI refers to the set of all paths with *strictly* increasing steps, i.e., paths with *both* coordinates strictly increasing from a step to another, from $(0, 0)$ to the East, $x = n$, or North, $y = n$, boundary. A third representation would be as above but where now the paths going from $(0, 0)$ to (n, n) have either strictly increasing steps or North or East unit steps. Again to the strictly increasing steps the associated weight is $\mathbf{1}_{\{X_i = Y_j\}}$ while to the North as well as to the East unit steps is associated a weight value of 0. As a final representation one could still proceed with strictly increasing paths but with the requirement that one ends the paths with a 1.)

Note that the weights in our percolation representations are not "truly 2-dimensional" and, in our opinion, this is the reason for the order of magnitude of the mean, variance

as well as the limiting law in the LCS problem to be different from other first/last passage-related models. To return to our specific results, they are the first studying the transversal fluctuations of the maximal paths of LCSs. Such questions have been of much interest in other percolation models. Let us only mention that Johansson [17] showed that for a Poisson points model in the plane, typical deviations of a maximal path from the diagonal is of order $n^{2/3}$, and $n^{2/3}$ is also the order of the transversal fluctuations in the directed polymer model studied in Seppäläinen [22]. (We refer to the respective bibliography of [17] and [22] for a much more complete and detailed picture on transversal fluctuations. We finally also note that in view of the results of [9], the transversal fluctuations of the LCSs of two independent random permutations of $\{1, 2, \ldots, n\}$, one uniform and one arbitrary, are exactly the same as those of the longest increasing subsequences of a single uniform random permutation of $\{1, 2, \ldots, n\}$.)

To finish this introductory section, let us briefly describe the rest of the paper. The next section presents some preliminary results, examples, and states the main result of the paper which is, in turn, proved in Section 3. Section 4 settles the closeness to the diagonal result and shows that each maximal path corresponding to the LCSs stays close to the diagonal. The last section explores the generic nature of short-strings alignments and some of its computational consequences.

2 Preliminaries

Throughout, let $n = km$ and let the integers

$$r_0 = 0 < r_1 < r_2 < r_3 < \cdots < r_{m-1} < r_m = n, \tag{2.1}$$

be such that

$$LC_n = \sum_{i=1}^{m} |LCS(X_{k(i-1)+1}X_{k(i-1)+2} \cdots X_{ki}; Y_{r_{i-1}+1}Y_{r_{i-1}+2} \cdots Y_{r_i})|, \tag{2.2}$$

where $|LCS(\cdot; \cdot)|$ is the length of the corresponding longest common substrings. In words, (2.2) asserts that there exists an optimal alignment aligning $[k(i-1)+1, ki]_{\mathbb{N}}$ with $[r_{i-1}+1, r_i]_{\mathbb{N}}$, for all $i = 1, 2, \ldots, m$.

The first goal of the present paper is to show that for k large but fixed, and n large enough, any such generic optimal alignment is such that the vast majority of the intervals $[r_{i-1}+1, r_i]_{\mathbb{N}}$ have length close to k. Building on this, a second goal is to show (see Section 5) that if a property $\mathcal{P}$ holds, with high probability, for string-pairs of (short) length order k, then typically a large proportion of aligned string-pairs satisfy the property $\mathcal{P}$.

Let us deal with our first goal and show that with high probability the optimal alignments satisfying (2.2), are such that most of their lengths $r_i - r_{i-1}$ are close to k. Of course, we need to quantify what is meant by "close to k". To do so, we first provide a definition. For $p > 0$, let

$$\tilde{\gamma}(p) := \lim_{n \to \infty} \frac{\mathbb{E}|LCS(X_1 X_2 \cdots X_n; Y_1 Y_2 \cdots Y_{np})|}{n(1+p)/2}, \tag{2.3}$$

where, when not integers, the indices np are understood to be rounded-up to the nearest positive integers. This function $\tilde{\gamma}$ is just a re-parametrization of "the usual function"

$$\gamma(q) = \lim_{n \to \infty} \frac{\mathbb{E}|LCS(X_1 X_2 \cdots X_{n-nq}; Y_1 Y_2 \cdots Y_{n+nq})|}{n},$$

$q \in (-1, 1)$, i.e.,

$$\tilde{\gamma}(p) = \gamma\left(q(p)\right), \tag{2.4}$$

with $q(p) = (p-1)/(p+1)$.

A superadditivity argument, as in Chvátal and Sankoff [7], shows that the above limits do exist (and depend, for example, on the size of the alphabet but this is of no importance for our purposes). For X and Y identically distributed, the function γ is symmetric about the origin, while a further superadditivity argument shows that it is concave there and so it reaches its maximum at $q = 0$ (see [3] for details). It is not known whether or not the function γ is strictly concave around $q = 0$. From simulations this appears to be the case but, at present, a proof is elusive. (Again, the LCS problem is a last passage percolation problem, and for first/last passage percolation proving that the shape of the wet region is strictly concave seems difficult and in many cases has not been done.) Since $q(p) = (p-1)/(p+1) = 1 - 2/(p+1)$ is strictly increasing in p, with $q(1) = 0$, if γ were strictly concave around $q = 0$, then it would reach a strict maximum there. Thus, $\tilde{\gamma}$ would also reach a strict maximum at $p = 1$ but, without the strict concavity of γ, $p = 1$ might not be the unique point of maximal value.

However, strict concavity is not needed, concavity is enough, for our results to hold. Indeed, (see Lemma 3.1) $\tilde{\gamma}$ is non-decreasing on $[0,1]$ and non-increasing on $[1,\infty)$ and thus ad hoc methods will show that $\tilde{\gamma}(p)$ is strictly smaller than $\tilde{\gamma}(1) = \gamma^*$, as soon as p is farther away from 1 than a small given quantity. (If the unproven strict concavity is valid then, below, p_1 and p_2 could be chosen as close to 1 as one wishes to.)

Before coming to the proof of the main theorem, it is thus important to show the existence of, and provide estimates on, the aforementioned p_1 and p_2 chosen as close to 1 as possible. Let us first explore this question in the binary equiprobable case. Let γ_ℓ be any strict lower bound on $\gamma^* = \lim_{n\to\infty} \mathbb{E}|LCS(X_1 \cdots X_n; Y_1 \cdots Y_n)|/n$. Then, if $x > 0$ is such that

$$(1-x)\left(H_2\left(1 - \frac{\gamma_\ell(2-x)}{2(1-x)}\right) - 1\right) + H_2\left(x + (1-x)\left(1 - \frac{\gamma_\ell(2-x)}{2(1-x)}\right)\right)$$
$$= (1-x)\left(H_2\left(1 - \frac{\gamma_\ell}{2}\left(1 + \frac{1}{1-x}\right)\right) - 1\right) + H_2\left(1 - \frac{\gamma_\ell}{2}(1 + (1-x))\right) < 0, \tag{2.5}$$

where $H_2(x) = -x\log_2 x - (1-x)\log_2(1-x)$, $0 < x < 1$, is the binary entropy function, we claim that $p_1 := 1 - x$ and $p_2 := 1/(1-x)$ are such that

$$\tilde{\gamma}(p_1) < \tilde{\gamma}(1) = \gamma^*, \quad \tilde{\gamma}(p_2) < \tilde{\gamma}(1) = \gamma^*. \tag{2.6}$$

Indeed, an easy upper bound on the probability that the length of the LCSs of $X_1 \cdots X_n$ and $Y_1 \cdots Y_{n(1-x)}$ is larger than $n(1-\eta)(1-x)$, $0 < \eta < 1$ is found as follows: First, take a non-random string s of length $n(1-\eta)(1-x)$. The probability that s is a subsequence of $Y_1 \cdots Y_{(1-x)n}$ does not depend on s (but only on the length of s), and so this probability would be the same if s would consist only of ones. Therefore, the probability that s is a subsequence of $Y_1 \cdots Y_{n(1-x)}$ is the same as the probability that there are at least $n(1-x)(1-\eta)$ ones in $Y_1 \cdots Y_{n(1-x)}$, which, in turn, is nothing but the probability for a binomial random variable, with parameters $(1-x)n$ and $1/2$, to be greater or equal to $n(1-x)(1-\eta)$. But, via classical exponential inequalities, this last probability is bounded above by

$$2^{n(1-x)(H_2(\eta)-1)}. \tag{2.7}$$

Now, the number of subsequences of $X_1 \cdots X_n$ of length $n(1-\eta)(1-x)$ is given by:

$$\binom{n}{n(\eta + x - \eta x)} \le 2^{nH_2(\eta + x - x\eta)}.$$

Combining this last inequality with (2.7) leads to

$$\mathbb{P}\left(|LCS(X_1\cdots X_n; Y_1\cdots Y_{n(1-x)})| \geq (1-x)(1-\eta)n\right) \leq 2^{n((1-x)(H_2(\eta)-1)+H_2(\eta+x-\eta x))}.$$
(2.8)

Therefore, from (2.8), as soon as

$$(1-x)(H_2(\eta)-1) + H_2(\eta + x - \eta x) < 0,$$
(2.9)

then, the probability that $|LCS(X_1\cdots X_n; Y_1\cdots Y_{n(1-x)})|$ is at least $(1-x)(1-\eta)n$ is exponentially small in n. In other words, the probability that the rescaled (by the average length of the two strings) LCS is at least:

$$\frac{2(1-x)(1-\eta)}{2-x},$$
(2.10)

is exponentially small in n. Now, to require the rescaled LCS value to be equal to γ_ℓ, choose γ_ℓ equal to the quantity in (2.10), i.e., let

$$\eta = 1 - \frac{\gamma_\ell(2-x)}{2(1-x)}.$$
(2.11)

So, for x given, (2.11) gives the value of η corresponding to γ_ℓ. We next find values of x for which the probability, that the rescaled LCS of $X_1\cdots X_n$ and $Y_1\cdots Y_{n(1-x)}$ is at least γ_ℓ, is exponentially small in n. Indeed, in (2.9) it is enough to replace η by the value given in (2.11). This leads to the condition (2.5), and when (2.5) is satisfied, the probability that the LCS of $X_1\cdots X_n$ and $Y_1\cdots Y_{n(1-x)}$ has a rescaled value of at least γ_ℓ is exponentially small in n. Therefore, in this case, the rescaled limit, that is

$$\tilde{\gamma}(1-x) = \lim_{n\to\infty} \frac{\mathbb{E}|LCS(X_1\cdots X_n; Y_1\cdots Y_{n(1-x)})|}{n},$$

has to be at most γ_ℓ. Thus, in this case, if γ_ℓ is a lower bound on $\tilde{\gamma}(1)$, then

$$\tilde{\gamma}(1-x) \leq \gamma_\ell < \tilde{\gamma}(1) = \gamma^*,$$

so that $p_1 = 1 - x$ and $p_2 = 1/(1-x)$ satisfy (2.6).

Using $\gamma_\ell = 0.7880$, which is a lower bound on γ^* obtained, in the binary case, by Lueker [19], it is easily seen that (2.5) is negative for $x = 0.28$. So $p_1 = 0.72$ and $p_2 = 1.39$ satisfy the needed conditions. Using these values as well as the upper bound 0.8263 (see [19]) provide an estimate for the fixed length k for the pieces one would divide the strings into (see, below, the statement of our first theorem).

The entropic method, on obtaining bound on p_1 and p_2, presented above carries over, beyond the binary case, to arbitrary-size alphabets. However, in the non-uniform case such bounds might be far from optimal and require the knowledge of the probability associated with each letter. To further address this question, let us present a lemma, with a somehow easier approach, to deal with the generic case.

Lemma 2.1. *Let $0 < p_1 = \gamma^*/(2 - \gamma^*) < 1$ and let $1 < p_2 = (2 - \gamma^*)/\gamma^*$, then*

$$\tilde{\gamma}(p_1) < \tilde{\gamma}(1) = \gamma^*, \quad \tilde{\gamma}(p_2) < \tilde{\gamma}(1) = \gamma^*.$$
(2.12)

Moreover, (2.12) continue to hold by taking $p_1 = \gamma_\ell/(2 - \gamma_\ell)$ and $p_2 = (2 - \gamma_\ell)/\gamma_\ell$ where γ_ℓ is any positive lower bound, such as $\sum_{\alpha\in\mathcal{A}}(\mathbb{P}(X_1 = \alpha))^2$, on γ^*.

Proof. First, note that when one sequence has length 0, then the LCS also has length 0, and thus

$$\gamma(-1) = \gamma(1) = 0.$$

Recall that the function γ is concave and symmetric about 0. Next, consider the LCS of the string $X = X_1 X_2 \cdots X_{2n}$ and the empty string Y. Then, take ϵn letters away from $X_1 \cdots X_{2n}$ and add that many letters to the Y string, so as to now have, instead of the empty string, the string $Y_1 Y_2 \cdots Y_{\epsilon n}$. Then, provided $0 < \epsilon < 1$ is a small enough constant, it follows that with very high (in n) probability, the string $Y_1 Y_2 \cdots Y_{\epsilon n}$ is a substring of $X_1 \cdots X_{2n - \epsilon n}$. Indeed, let $T_1 = \inf\{k \geq 0 : X_1 = Y_k\}$, $T_2 = \inf\{k \geq T_1 : X_2 = Y_k\}$, $T_3 = \inf\{k \geq T_2 : X_3 = Y_k\}, \ldots$ Then, conditionally on $X = x$, $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are independent geometric random variables with individual parameter depending on the probabilities associated to the letters. But,

$$|LCS(X_1 \cdots X_{2n - \epsilon n}; Y_1 \cdots Y_{\epsilon n})| = \epsilon n,$$

if and only if $T_{\epsilon n} \leq 2n - \epsilon n$. Thus, since the geometric property holds for any $x \in \mathcal{A}^n$,

$$\mathbb{P}(|LCS(X_1 \cdots X_{2n - \epsilon n}; Y_1 \cdots Y_{\epsilon n})|) = \epsilon n) \geq 1 - \exp(-cn), \tag{2.13}$$

where $c > 0$ is a constant depending neither on n nor on $\epsilon > 0$ (provided $\epsilon > 0$ is small enough) but depending on the minimal parameter of the geometric random variables (hence on the probabilities associated to the letters). Therefore, from the above, at $q = -1$, the slope is 1, i.e., $\gamma'((-1)^+) = 1$ and similarly $\gamma'(1^-) = -1$. By concavity and symmetry, for any q_1 with

$$q_1 < -(1 - \gamma(0)), \tag{2.14}$$

it therefore follows that

$$\gamma(q_1) < \gamma(0),$$

and similarly for any q_2 with

$$q_2 > 1 - \gamma(0), \tag{2.15}$$

it follows that

$$\gamma(q_2) < \gamma(0).$$

The bounds obtained above rely on the value of $\gamma(0) = \gamma^*$ which is unknown, but for which upper and lower bounds (which depend on the distributions of the letters) do exist (and are rather accurate for uniform distributions). In our case, a lower bound on γ^* is what is needed. The most trivial lower bound is obtained when aligning the two strings, without gaps, and just counting the number of correctly aligned letter pairs, more precisely, $LC_n \geq \sum_{i=1}^n \mathbf{1}_{\{X_i = Y_i\}}$. Hence, by the iid and independence assumptions,

$$\mathbb{E}LC_n \geq \sum_{i=1}^n \mathbb{P}(X_i = Y_i) = n\mathbb{P}(X_1 = Y_1) = n \sum_{\alpha \in \mathcal{A}} \mathbb{P}((X_1 = \alpha))^2. \tag{2.16}$$

Now, converting to $\tilde{\gamma}$, and recalling that $p(q) = (1 + q)/(1 - q)$, (2.14) becomes

$$p_1 < \frac{\gamma(0)}{2 - \gamma(0)}, \tag{2.17}$$

and in such a case, $\tilde{\gamma}(p_1) < \tilde{\gamma}(1)$. Similarly, (2.15) becomes

$$p_2 > \frac{2 - \gamma(0)}{\gamma(0)}, \tag{2.18}$$

and then $\tilde{\gamma}(p_2) < \tilde{\gamma}(1)$. Finally, in both (2.17) and (2.18), one can replace $\gamma(0)$ by any of its positive lower bound, for example the lower bound resulting from (2.16). $\qquad\square$

So let us now assume that $0 < p_1 < 1 < p_2$ are such that

$$\tilde{\gamma}(p_1) < \tilde{\gamma}(1) = \gamma^*, \quad \tilde{\gamma}(p_2) < \tilde{\gamma}(1) = \gamma^*. \tag{2.19}$$

The first result we next set to state, asserts that for k fixed and n large enough then typically in any optimal alignment (2.2) most of the intervals $[r_{i-1} + 1, r_i]_{\mathbb{N}}$ (for $i = 1, 2, \ldots, m$) have their length, $r_i - r_{i-1}$, between kp_1 and kp_2. By most, it is meant that by

taking k large, *but fixed*, and n large enough, the proportion of such intervals gets as close to 1 as one wishes to.

To do so, let us introduce some notation. Let $\epsilon > 0$, $p_1 > 0$ and $p_2 > 0$ be constants. Let A^n_{ϵ,p_1,p_2} be the (random) set of optimal alignments of $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$ satisfying (2.2), for which a proportion of at least $1 - \epsilon$ of the intervals $[r_{i-1} + 1, r_i]_{\mathbb{N}}$, $i = 1, 2, \ldots, m$, have their length between kp_1 and kp_2. More precisely, A^n_{ϵ,p_1,p_2} is the event that for all integer vectors $(r_0, r_1, \ldots, r_m)$ satisfying (2.1) and for which (2.2) holds,

$$\text{Card}\left(\{i \in 1, 2, \ldots, m : kp_1 \leq r_i - r_{i-1} \leq kp_2\} \right) \geq (1 - \epsilon)m. \tag{2.20}$$

As stated next, A^n_{ϵ,p_1,p_2} holds with high probability.

Theorem 2.2. *Let $\epsilon > 0$. Let $0 < p_1 < 1 < p_2$ be such that $\tilde{\gamma}(p_1) < \tilde{\gamma}(1) = \gamma^*$ and $\tilde{\gamma}(p_2) < \tilde{\gamma}(1) = \gamma^*$, and let $\delta \in (0, \min(\gamma^* - \tilde{\gamma}(p_1), \gamma^* - \tilde{\gamma}(p_2)))$. Fix the integer k to be such that $(1 + \ln k)/k \leq \delta^2 \epsilon^2 / 16$, then*

$$\mathbb{P}\left(A^n_{\epsilon,p_1,p_2} \right) \geq 1 - \exp\left(-n \left(-\frac{1 + \ln k}{k} + \frac{\delta^2 \epsilon^2}{16} \right) \right), \tag{2.21}$$

for all $n = n(\epsilon, \delta)$ large enough.

In words, and broadly, Theorem 2.2 asserts that for any $\epsilon > 0$, there exists k large enough, but fixed, such that if X is divided into segments of length k then, typically (at least a fraction $1 - \epsilon$ of segments), and with high probability, the LCSs match these segments to segments of similar length in Y.

3 Proof of the main theorem

The proof of Theorem 2.2 requires the introduction of a few more definitions. So far we have looked at the integer intervals which are mapped by optimal alignments to the integer intervals $[k(i-1) + 1, ki]_{\mathbb{N}}$. The opposite stand is now taken: given (non-random) integers $r_0 = 0 < r_1 < r_2 < \cdots < r_m = n$, we request that the alignment aligns $[k(i-1) + 1, ki]_{\mathbb{N}}$ with $[r_{i-1} + 1, r_i]_{\mathbb{N}}$, for every $i = 1, 2, \ldots, m$. In general, such an alignment is not optimal and, the best score an alignment can attain under the above constraint is:

$$L_n(\vec{r}) := L_n(r_0, r_1, \ldots, r_m)$$
$$:= \sum_{i=1}^{m} |LCS(X_{k(i-1)+1} X_{k(i-1)+2} \cdots X_{ki}; Y_{r_{i-1}+1} Y_{r_{i-1}+2} \cdots Y_{r_i})|.$$

Therefore, $L_n(\vec{r})$ represents the maximum number of aligned identical letter pairs under the constraint that $X_{(i-1)k+1} X_{(i-1)k+2} \cdots X_{ik}$ gets aligned with $Y_{r_{i-1}+1} Y_{r_{i-1}+2} \cdots Y_{r_i}$, for all $i = 1, 2, \ldots, m$. Note moreover that for a non-random $(r_0, r_1, \ldots, r_m)$, the partial scores

$$|LCS(X_{(i-1)k+1} X_{(i-1)k+2} \cdots X_{ik}; Y_{r_{i-1}+1} Y_{r_{i-1}+2} \cdots Y_{r_i})|,$$

are independent of each other and, in this context, concentration inequalities will prove useful when dealing with $L_n(\vec{r})$. Next, let $\mathcal{R}_{\epsilon,p_1,p_2}$, be the (non-random) set of all integer vectors $(r_0, r_1, \ldots, r_m)$ satisfying (2.1) and (2.20), while $\overline{\mathcal{R}}_{\epsilon,p_1,p_2}$, denotes the (non-random) set of all integer vectors $\vec{r} = (r_0, r_1, \ldots, r_m)$ satisfying (2.1) but not (2.20).

Let us begin with a lemma.

Lemma 3.1. *Let $\epsilon > 0$. Let $0 < p_1 < 1 < p_2$ be such that $\tilde{\gamma}(p_1) < \gamma^*$ and $\tilde{\gamma}(p_2) < \gamma^*$, and let $\delta > 0$ be such that $\delta < \min(\gamma^* - \tilde{\gamma}(p_1), \gamma^* - \tilde{\gamma}(p_2))$. Let $\vec{r} = (r_0, \ldots, r_m) \in \overline{\mathcal{R}}_{\epsilon,p_1,p_2}$, then*

$$\mathbb{E}\left(L_n(\vec{r}) - LC_n \right) \leq -\frac{\delta \epsilon n}{2}, \tag{3.1}$$

for all $n = n(\epsilon, \delta)$ large enough.

Proof. Let $p > 0$, and let

$$\tilde{\gamma}(p) := \lim_{n \to \infty} \frac{\mathbb{E}|LCS(X_1 X_2 \cdots X_n; Y_1 Y_2 \cdots Y_{np})|}{n(1+p)/2},$$

where, by superadditivity, this limit exists with, moreover,

$$\frac{2\mathbb{E}|LCS(X_1 X_2 \cdots X_n; Y_1 Y_2 \cdots Y_{np})|}{n(1+p)} \le \tilde{\gamma}(p), \tag{3.2}$$

for any $n \ge 1$. Now, $\gamma : q \in (-1, 1) \to \gamma(q) \in (0, \infty)$, defined via

$$\gamma(q) := \lim_{n \to \infty} \frac{\mathbb{E}|LCS(X_1 X_2 \cdots X_{n-nq}; Y_1 Y_2 \cdots Y_{n+nq})|}{n},$$

is symmetric about $q = 0$ and, as already mentioned, is also concave (see [3]). Hence,

$$\tilde{\gamma}(p) = \gamma\left(\frac{p-1}{p+1}\right),$$

is non-decreasing up to $p = 1$ and non-increasing afterwards. Thus, choosing the interval $[p_1, p_2]$ to contain $p = 1$, it follows that for $p \notin [p_1, p_2]$,

$$\tilde{\gamma}(p) \le \max(\tilde{\gamma}(p_1), \tilde{\gamma}(p_2)). \tag{3.3}$$

Hence, for any $p \notin [p_1, p_2]$, combining (3.2) and (3.3) leads to:

$$\frac{2\mathbb{E}|LCS(X_1 X_2 \cdots X_k; Y_1 Y_2 \cdots Y_{kp})|}{k(1+p)} \le \max(\tilde{\gamma}(p_1), \tilde{\gamma}(p_2)), \tag{3.4}$$

and therefore,

$$\frac{2\mathbb{E}|LCS(X_1 X_2 \cdots X_k; Y_1 Y_2 \cdots Y_{kp})|}{k(1+p)} \le \tilde{\gamma}(1) - \delta^* = \gamma^* - \delta^*, \tag{3.5}$$

where $\delta^* := \min(\gamma^* - \tilde{\gamma}(p_1), \gamma^* - \tilde{\gamma}(p_2))$.

Since the sequences $(X_i)_{i \ge 1}$ and $(Y_i)_{i \ge 1}$ are stationary, and assuming that $r_i - r_{i-1} = kp$, the left-hand side of (3.5) becomes

$$\frac{2\mathbb{E}|LCS(X_{(i-1)k+1} X_{(i-1)k+2} \cdots X_{ik}; Y_{r_{i-1}+1} Y_{r_{i-1}+2} \cdots Y_{r_i})|}{k + r_i - r_{i-1}}.$$

Thus, from (3.5), when $(r_i - r_{i-1})/k \notin [p_1, p_2]$, then

$$\gamma^* - \frac{2\mathbb{E}|LCS(X_{(i-1)k+1} X_{(i-1)k+2} \cdots X_{ik}; Y_{r_{i-1}+1} Y_{r_{i-1}+2} \cdots Y_{r_i})|}{k + r_i - r_{i-1}} \ge \delta^*. \tag{3.6}$$

Hence, from (3.6),

$$\gamma^*\left(\frac{k + r_i - r_{i-1}}{2}\right) - \mathbb{E}|LCS(X_{(i-1)k+1} X_{(i-1)k+2} \cdots X_{ik}; Y_{r_{i-1}+1} Y_{r_{i-1}+2} \cdots Y_{r_i})|$$

$$\ge \delta^*\left(\frac{k + r_i - r_{i-1}}{2}\right) \ge \delta^* \frac{k}{2}.$$

Letting $\mathcal{M} := \{i : [k(i-1) + 1, ki] \text{ gets matched with strings of length not in } [kp_1, kp_2]\}$, we then have

$$\sum_{i \in \mathcal{M}} \left(\gamma^*\left(\frac{k + r_i - r_{i-1}}{2}\right) - \mathbb{E}|LCS(X_{(i-1)k+1} X_{(i-1)k+2} \cdots X_{ik}; Y_{r_{i-1}+1} Y_{r_{i-1}+2} \cdots Y_{r_i})|\right)$$

$$\ge \sum_{i \in \mathcal{M}} \delta^* \frac{k}{2} \ge \delta^* \frac{k}{2} \epsilon m = \frac{n\delta^* \epsilon}{2}. \tag{3.7}$$

On the other hand,

$$\sum_{i \in \mathcal{M}} \left(\gamma^* \left(\frac{k + r_i - r_{i-1}}{2} \right) - \mathbb{E}|LCS(X_{(i-1)k+1}X_{(i-1)k+2} \cdots X_{ik}; Y_{r_{i-1}+1}Y_{r_{i-1}+2} \cdots Y_{r_i})| \right)$$

$$\leq \sum_{i=1}^{m} \left(\gamma^* \left(\frac{k + r_i - r_{i-1}}{2} \right) - \mathbb{E}|LCS(X_{(i-1)k+1}X_{(i-1)k+2} \cdots X_{ik}; Y_{r_{i-1}+1}Y_{r_{i-1}+2} \cdots Y_{r_i})| \right)$$

$$= \gamma^* n - \mathbb{E}L_n(\vec{r}). \tag{3.8}$$

Therefore, combining (3.7) and (3.8) leads to

$$\gamma^* n - \mathbb{E}L_n(\vec{r}) \geq \frac{n\delta^*\epsilon}{2}, \tag{3.9}$$

as soon as $(r_0, r_1, \ldots, r_m) \in \overline{\mathcal{R}}_{\epsilon, p_1, p_2}$. Now $\lim_{n \to +\infty} \mathbb{E}LC_n/n = \gamma^*$, while by hypothesis $\delta^* - \delta > 0$,

$$0 \leq \gamma^* - \frac{\mathbb{E}LC_n}{n} \leq \frac{(\delta^* - \delta)\epsilon}{2}. \tag{3.10}$$

for all n large enough. (At this last stage, a more quantitative bound, depending on n, could also be obtained using (1.4).) Combining (3.9) and (3.10) yields that for any $\vec{r} \in \overline{\mathcal{R}}_{\epsilon, p_1, p_2}$,

$$\mathbb{E}\left(LC_n - L_n(\vec{r})\right) \geq \frac{n\delta\epsilon}{2},$$

for all n large enough. The proof of the lemma is now complete. $\qquad\square$

Proof of Theorem 2.2. Clearly,

$$\text{Card}\left(\overline{\mathcal{R}}_{\epsilon, p_1, p_2}\right) \leq \binom{n}{m} \leq \frac{n^m}{m!} \leq \left(\frac{en}{m}\right)^m = (ek)^m, \tag{3.11}$$

by a well known and simple bound on the binomial coefficients.

Now, let $\delta < \delta^* := \min(\gamma^* - \tilde{\gamma}(p_1), \gamma^* - \tilde{\gamma}(p_2))$. By definition $LC_n \geq L_n(\vec{r})$, and so for $\vec{r}$ to define an optimal alignment requires:

$$L_n(\vec{r}) \geq LC_n. \tag{3.12}$$

Hence, for the event A_{ϵ, p_1, p_2}^n not to hold (see (2.20)), there needs to exist at least one $\vec{r} \in \overline{\mathcal{R}}_{\epsilon, p_1, p_2}$ for which (3.12) is satisfied. Thus,

$$(A_{\epsilon, p_1, p_2}^n)^c = \bigcup_{\vec{r} \in \overline{\mathcal{R}}_{\epsilon, p_1, p_2}} \{L_n(\vec{r}) - LC_n \geq 0\},$$

and

$$\mathbb{P}((A_{\epsilon, p_1, p_2}^n)^c) \leq \sum_{\vec{r} \in \overline{\mathcal{R}}_{\epsilon, p_1, p_2}} \mathbb{P}(L_n(\vec{r}) - LC_n \geq 0). \tag{3.13}$$

When $\vec{r} \in \overline{\mathcal{R}}_{\epsilon, p_1, p_2}$, it follows from Lemma 3.1 that:

$$\mathbb{E}\left(L_n(\vec{r}) - LC_n\right) \leq -\frac{\delta\epsilon n}{2},$$

and so

$$\mathbb{P}(L_n(\vec{r}) - LC_n \geq 0) \leq \mathbb{P}\left(L_n(\vec{r}) - LC_n - \mathbb{E}\left(L_n(\vec{r}) - LC_n\right) \geq \frac{\delta\epsilon n}{2}\right), \tag{3.14}$$

for all n large enough. Now, the difference $L_n(\vec{r}) - LC_n$ changes by at most plus or minus two, when any one of the iid entries $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$ is changed. Therefore, Hoeffding's martingale inequality, applied to the right-hand side of (3.14), gives

$$\mathbb{P}\left(L_n(\vec{r}) - LC_n \geq 0\right) \leq \mathbb{P}\left(L_n(\vec{r}) - LC_n \ - \ \mathbb{E}\left(L_n(\vec{r}) - LC_n\right) \geq \frac{\delta \epsilon n}{2}\right) \leq \exp\left(-\frac{\delta^2 \epsilon^2}{16}n\right).$$

(Recall that Hoeffding's martingale inequality asserts that if f is a function of j variables, such that changing any single of its entries changes f by at most $\pm\Delta$, $\Delta > 0$, and if $Z_1, Z_2, \ldots, Z_j$ are independent random variables, then

$$\mathbb{P}\left(f(Z_1, Z_2, \ldots, Z_j) - \mathbb{E}f(Z_1, Z_2, \ldots, Z_j) \geq z\right) \leq \exp\left(-\frac{2z^2}{j\Delta^2}\right),$$

provided the expectation exists.) Combining this last inequality with (3.13), one obtains:

$$\mathbb{P}((A_{\epsilon,p_1,p_2}^n)^c) \leq \mathrm{Card}(\overline{\mathcal{R}}_{\epsilon,p_1,p_2}) \exp\left(-\frac{\delta^2 \epsilon^2}{16}n\right). \tag{3.15}$$

But, from (3.11),

$$\mathbb{P}((A_{\epsilon,p_1,p_2}^n)^c) \leq (ek)^m \exp\left(-\frac{\delta^2 \epsilon^2}{16}n\right) = \exp\left(-n\left(-\frac{1 + \ln k}{k} + \frac{\delta^2 \epsilon^2}{16}\right)\right). \tag{3.16}$$

Therefore, the proof of Theorem 2.2 is complete. $\qquad\square$

4 Closeness to the diagonal

Let us begin with a definition. Let D_{ϵ,p_1,p_2}^n be the event that all the points representing any optimal alignment of $X_1 X_2 \cdots X_n$ with $Y_1 Y_2 \cdots Y_n$ are above the line $y = p_1 x - p_1 n\epsilon - p_1 k$, and below the line $y = p_2 x + p_2 n\epsilon + p_2 k$.

Theorem 4.1. *Let $\epsilon > 0$. Let $0 < p_1 < 1 < p_2$ be such that $\tilde{\gamma}(p_1) < \gamma^*$ and $\tilde{\gamma}(p_2) < \gamma^*$, and let $0 < \delta < \min\left(\gamma^* - \tilde{\gamma}(p_1), \gamma^* - \tilde{\gamma}(p_2)\right)$. Fix the integer k to be such that $(1 + \ln k)/k \leq \delta^2 \epsilon^2/16$, then*

$$\mathbb{P}(D_{\epsilon,p_1,p_2}^n) \geq 1 - 2\exp\left(-n\left(-\frac{1 + \ln k}{k} + \frac{\delta^2 \epsilon^2}{16}\right)\right), \tag{4.1}$$

for all $n = n(\epsilon, \delta)$ large enough.

Proof. Let D_a^n be the event that any optimal alignment of $X_1 X_2 \cdots X_n$ with $Y_1 Y_2 \cdots Y_n$ is above the line $y_1 := y_1(x) = p_1 x - p_1 n\epsilon - p_1 k$; and let D_b^n be the event that any optimal alignment of $X_1 X_2 \cdots X_n$ with $Y_1 Y_2 \cdots Y_n$ is below the line $y_2 := y_2(x) = p_2 x + p_2 n\epsilon + p_2 k$. Clearly, $D_a^n \cap D_b^n = D_{\epsilon,p_1,p_2}^n$, hence

$$\mathbb{P}((D_{\epsilon,p_1,p_2}^n)^c) \leq \mathbb{P}((D_a^n)^c) + \mathbb{P}((D_b^n)^c), \tag{4.2}$$

and the result will be a consequence of the following two inclusions:

$$A_{\epsilon,p_1,p_2}^n \subset D_a^n, \quad A_{\epsilon,p_1,p_2}^n \subset D_b^n, \tag{4.3}$$

where A_{ϵ,p_1,p_2}^n is as in Theorem 2.2. Let us prove the first inclusion in (4.3). To start, assume that x is an integer multiple of k, i.e., let $x = uk$, $u \in \mathbb{N}$. Next, and at first, let us consider the case where $x \leq n\epsilon$, i.e., that $p_1 x - p_1 n\epsilon \leq 0$. Any alignment (and, in particular, any optimal alignment) we consider, aligns any $x \in [0, n]_{\mathbb{N}}$ with $[0, n]_{\mathbb{N}}$. Hence, for every $x \leq n\epsilon$, the condition is always verified, that is any optimal alignment aligns x with a y which is at least equal to $p_1 x - p_1 n\epsilon$. Let us now consider the case where $x \geq n\epsilon$.

When the event A_{ϵ,p_1,p_2}^n holds, then any optimal alignment aligns all but a proportion ϵ of the interval $[(i-1)k+1, ik]_{\mathbb{N}}$, $i \in \{1, \ldots, m\}$ to integer intervals of length greater or equal to kp_1. The maximum number of integer intervals which could be matched with integer intervals of length less than kp_1 is thus ϵm. In the interval $[0, x]_{\mathbb{N}}$ there are u intervals from the partition $[(i-1)k+1, ik]_{\mathbb{N}}$, $i \in \{1, \ldots, m\}$. Therefore, at least $u - \epsilon m$ of these intervals are matched to intervals of length no less than kp_1, implying that, when the event A_{ϵ,p_1,p_2}^n holds, x gets matched by the optimal alignment with a value no less than $(u - \epsilon m)kp_1 = p_1 x - p_1 \epsilon n$, since $x = uk$ and $n = mk$. This finishes the case where x is an integer multiple of k. If x is not an integer multiple of k, let x_1 denote the largest integer multiple of k which is smaller than x. By definition,

$$x - x_1 < k. \tag{4.4}$$

But, the two-dimensional alignment curve cannot go down, hence x gets aligned with a point which cannot be below the point where x_1 gets aligned to. But, since x_1 is an integer multiple of k, it gets aligned with a point which is greater or equal to $p_1 x_1 - p_1 \epsilon n$. Using (4.4), it follows that

$$p_1 x_1 - p_1 \epsilon n \geq p_1 x - p_1 \epsilon n - p_1 k,$$

and this shows that when the event A_{ϵ,p_1,p_2}^n holds, then x gets aligned above or on $p_1 x - p_1 \epsilon n - p_1 k$. This finishes proving that the event A_{ϵ,p_1,p_2}^n is a sub-event of D_a^n. Therefore, by (2.21),

$$\mathbb{P}\left((D_a^n)^c\right) \leq \mathbb{P}\left((A_{\epsilon,p_1,p_2}^n)^c\right) \leq \exp\left(-n\left(-\frac{1 + \ln k}{k} + \frac{\delta^2 \epsilon^2}{16}\right)\right).$$

Similarly, and symmetrizing the above arguments,

$$\mathbb{P}\left((D_b^n)^c\right) \leq \mathbb{P}\left((A_{\epsilon,p_1,p_2}^n)^c\right) \leq \exp\left(-n\left(-\frac{1 + \ln k}{k} + \frac{\delta^2 \epsilon^2}{16}\right)\right),$$

finishing, via (4.2), the proof of the theorem. □

Theorem 4.1 should prove useful in reducing the time to compute the LCS of two random sequences. Indeed, first by (4.1), when rescaled by n, the two-dimensional representation of an optimal alignment is, with high probability and up to a distance of order $\epsilon > 0$, above the line $x \to p_1 x$ and below the line $x \to p_2 x$. Moreover, $\epsilon > 0$ can be taken as small as we want, leaving it fixed though when n goes to infinity. Next, simulations seem to indicate that the mean curve $\tilde{\gamma}$ is strictly concave at $p = 1$. If strict concavity indeed hold, then p_1 and, say, $p_2 = 1/p_1$, can be taken as close to 1 as we want, and still satisfy the conditions of the theorem. That is, taking ϵ as close to 0 as we want and p_1 as close to 1 as we want, the re-scaled two-dimensional representation of the optimal alignments would get uniformly as close to the diagonal as we want, as n grows without bound.

Figure 1 is the graph of a simulation with two iid binary sequences of length $n = 1000$. All the optimal alignments are contained between the two graphs below and are thus seen to all stay extremely close to the diagonal. The maximal vertical distance between two optimal paths is 26 and, for this vertical distance, the maximal horizontal stretch between which the two optimal paths split and then meet again is 112.

5 Short string-lengths properties are generic

Often, a desirable property we want string-pairs to verify, e.g., a similar number of a given symbol or pattern, the presence of dominant matches, only holds with high

score LCS=838 , longest vertical distance=26, longest horizontal length=112
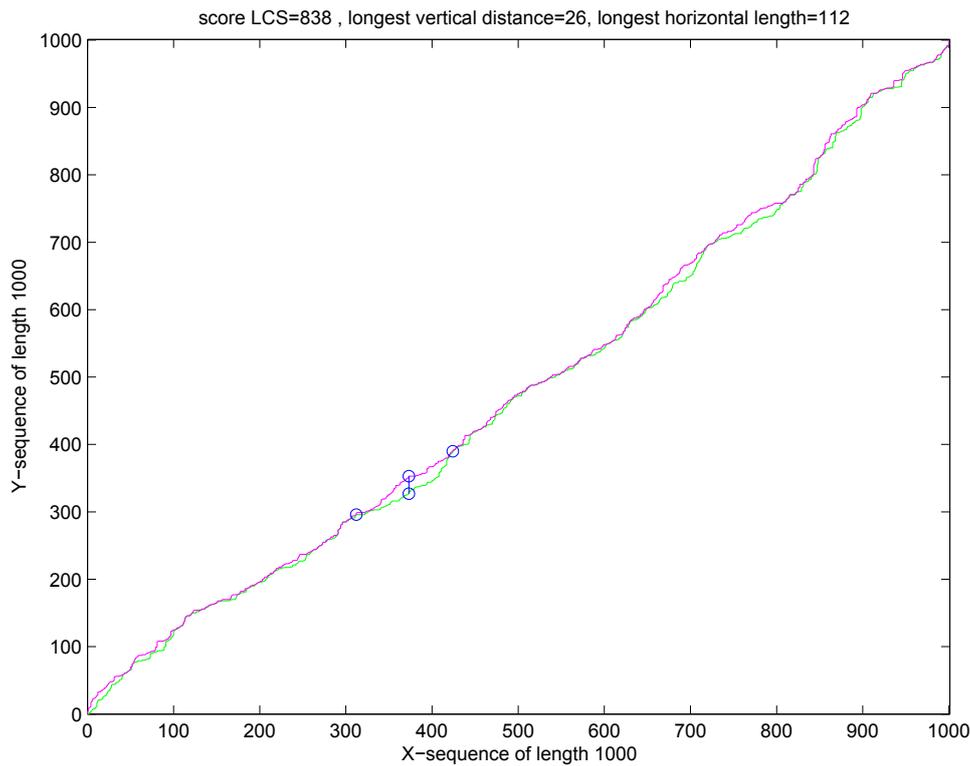
Figure 1: $n = 1000$, uniform Bernoulli sequences

probability and if the two strings have their lengths not too far away from each other. Moreover, short strings are also often used as "seeds" to find longer or more global similarities and homologous properties ([21] contains many examples of such instances in applied problems). It is our purpose now to attempt to quantify such a generic phenomenon.

To to so, let $\mathcal{P}$ be a relation assigning to every pair of strings (x, y) the value 1 if the pair (x, y) has a certain property, and 0 otherwise. Hence, if $\mathcal{A}$ is the alphabet we consider,

$$\mathcal{P} : (\cup_k \mathcal{A}^k) \times (\cup_k \mathcal{A}^k) \to \{0, 1\},$$

and if $\mathcal{P}(x, y) = 1$, the string pair (x, y) is said to have the property $\mathcal{P}$.

Let now $\epsilon > 0$, be fixed, and let $\vec{r} = (r_0, r_1, \ldots, r_m)$ satisfy the condition (2.1). Let also $B_{\mathcal{P}}^n(\vec{r}, \epsilon)$ be the event that there is a proportion of at least $1 - \epsilon$ of the string pairs

$$(X_{(i-1)k+1} \cdots X_{ik}; Y_{r_{i-1}+1} \cdots Y_{r_i}) \tag{5.1}$$

satisfying the property $\mathcal{P}$, i.e.,

$$B_{\mathcal{P}}^n(\vec{r}, \epsilon) = \left\{ \sum_{i=1}^m \mathcal{P}(X_{(i-1)k+1} \ldots X_{ik}; Y_{r_{i-1}+1} \cdots Y_{r_i}) \geq (1 - \epsilon)m \right\}. \tag{5.2}$$

Next, let $B_{\mathcal{P}}^n(\epsilon)$ be the event that for every optimal alignment the proportion of aligned string pairs (5.1) satisfying the property $\mathcal{P}$ is at least $1 - \epsilon$, i.e., $B_{\mathcal{P}}^n(\epsilon)$ holds if and only if for every $\vec{r} = (r_0, r_1, \ldots, r_m)$ satisfying (2.1) and such that $LC_n = L_n(\vec{r})$, the event $B_{\mathcal{P}}^n(\vec{r}, \epsilon)$ holds. Finally, assume that as soon as $r_i - r_{i-1} \in [kp_1, kp_2]$, the probability that

the string-pairs (5.1) have the required property is at least $1-q$, $q \in [0,1]$. Hence, assume that for every integer $\ell \in [kp_1, kp_2]$:

$$\mathbb{P}\left(\mathcal{P}(X_1 \cdots X_k; Y_1 Y_2 \cdots Y_\ell) = 1 \right) \geq 1 - q.$$

We investigate, now, how small $q = q(k)$ needs to be in order to ensure that a large proportion of the aligned string pairs (5.1) has the property $\mathcal{P}$ (for every optimal alignment). Recall that A^n_{ϵ, p_1, p_2} is the event that every optimal alignment aligns a proportion of at least $1 - \epsilon$ of the sub-strings $X_{(i-1)k+1} \cdots X_{ik}$ with sub-strings of Y with length in $[kp_1, kp_2]$. Recall also that $\mathcal{R}_{\epsilon, p_1, p_2}$ is the set of integer vectors $\vec{r} = (r_0, r_1, \ldots, r_m)$, satisfying (2.1) and such that there is at least $(1-\epsilon)m$ of the differences $r_i - r_{i-1}$ in $[kp_1, kp_2]$.

Below, we deal with a small modification of the event $B^n_{\mathcal{P}}(\vec{r}, \epsilon)$. For this, let $\tilde{B}^n_{\mathcal{P}}(\vec{r}, \epsilon)$ be the event that among the aligned string pieces (5.1) there are no more than $m\epsilon$ which do not satisfy the property $\mathcal{P}$ and have their length $r_i - r_{i-1}$ in $[kp_1, kp_2]$. Clearly, for $\epsilon_1 > 0$, $\epsilon_2 > 0$,

$$A^n_{\epsilon_1, p_1, p_2} \cap \left(\bigcap_{\vec{r} \in \mathcal{R}_{\epsilon_1, p_1, p_2}} \tilde{B}^n_{\mathcal{P}}(\vec{r}, \epsilon_2) \right) \subset B^n_{\mathcal{P}}(\epsilon_1 + \epsilon_2),$$

and so

$$\mathbb{P}((B^n_{\mathcal{P}}(\epsilon_1 + \epsilon_2))^c) \leq \mathbb{P}((A^n_{\epsilon_1, p_1, p_2})^c) + \sum_{\vec{r} \in \mathcal{R}_{\epsilon_1, p_1, p_2}} \mathbb{P}((\tilde{B}^n_{\mathcal{P}}(\vec{r}, \epsilon_2))^c). \tag{5.3}$$

Next,

$$\mathbb{P}((\tilde{B}^n_{\mathcal{P}}(\vec{r}, \epsilon_2))^c) \leq \binom{m}{\epsilon_2 m} q^{\epsilon_2 m} \leq \exp(H_e(\epsilon_2)m) q^{\epsilon_2 m},$$

where H_e is the base e entropy function, given by $H_e(x) = -x \ln x - (1-x)\ln(1-x), 0 < x < 1$. Hence,

$$\mathbb{P}((\tilde{B}^n_{\mathcal{P}}(\vec{r}, \epsilon_2))^c) \leq q^{\epsilon_2 m} \exp(H_e(\epsilon_2)m). \tag{5.4}$$

Using (5.4) into (5.3) and, proceeding as in (3.11), noting that $\mathcal{R}_{\epsilon, p_1, p_2}$ has at most $(ek)^m$ elements, lead to

$$\mathbb{P}((B^n_{\mathcal{P}}(\epsilon_1 + \epsilon_2))^c) \leq \mathbb{P}((A^n_{\epsilon_1, p_1, p_2})^c) + (ek)^m q^{\epsilon_2 m} \exp(H_e(\epsilon_2)m). \tag{5.5}$$

Taking $q(k) = 1/(2ek)^{1/\epsilon_2}$, finally yields

$$\mathbb{P}((B^n_{\mathcal{P}}(\epsilon_1 + \epsilon_2))^c) \leq \mathbb{P}((A^n_{\epsilon_1, p_1, p_2})^c) + \exp((H_e(\epsilon_2) - \ln 2)m). \tag{5.6}$$

But, for $\epsilon_2 < 1/2$, $H_e(\epsilon_2) < \ln 2$, and then $\exp((H_e(\epsilon_2) - \ln 2)m)$ is exponentially small in m. Now, our main theorem provides an exponentially small lower bound on $\mathbb{P}((A^n_{\epsilon_1, p_1, p_2})^c)$. Therefore, (5.6) asserts that a high proportion of the aligned string pairs (5.1) has property $\mathcal{P}$, in any optimal alignment, as soon as for pairs (5.1) with similar length, $q(k) \leq 1/(2ek)^{1/\epsilon_2}$, where

$$q(k) := \max_{\ell \in [kp_1, kp_2]} \mathbb{P}(\text{the pair } (X_1 \cdots X_k; Y_1 \cdots Y_\ell) \text{ does not satisfy property } \mathcal{P}).$$

These assertions are summarized in the next theorem, which is obtained by letting, above, $\epsilon_1 = \epsilon_2 = \epsilon/2$, using also Theorem 2.2.

Theorem 5.1. *Let $0 < \epsilon < 1$. Let $0 < p_1 < 1 < p_2$ be such that $\tilde{\gamma}(p_1) < \gamma^*$ and $\tilde{\gamma}(p_2) < \gamma^*$, and let $0 < \delta < \min(\gamma^* - \tilde{\gamma}(p_1), \gamma^* - \tilde{\gamma}(p_2))$. Finally, let the integer $k \geq 1$ be such that*

$$\max_{\ell \in [kp_1, kp_2]} \mathbb{P}((X_1 \cdots X_k; Y_1 \cdots Y_\ell) \text{ does not satisfy property } \mathcal{P}) \leq \frac{1}{(2ek)^{2/\epsilon}}.$$

Then, for any optimal alignment $\vec{r}$ (i.e., such that $LC_n = L_n(\vec{r})$), the proportion of string pairs $(X_{(i-1)k+1} \cdots X_{ik}; Y_{r_{i-1}+1} \cdots Y_{r_i})$ satisfying property $\mathcal{P}$ is at least $1 - \epsilon$ with probability at least equal to:

$$1 - \mathbb{P}((A^n_{\epsilon/2, p_1, p_2})^c) - \exp\left(\left(H_e\left(\frac{\epsilon}{2}\right) - \ln 2\right) m\right),$$

and thus at least equal to:

$$1 - \exp\left(-n\left(-\frac{1 + \ln k}{k} + \frac{\delta^2 \epsilon^2}{64}\right)\right) - \exp\left(\frac{n}{k}\left(H_e\left(\frac{\epsilon}{2}\right) - \ln 2\right)\right),$$

for all $n = n(\epsilon, \delta)$ large enough.

Hence, from the above statement, the probability that less than a proportion $1 - \epsilon$ of string pairs (5.1) have property $\mathcal{P}$ in every optimal alignment is exponentially small in n (while holding k, ϵ and δ fixed) as soon as

$$k > \frac{64(1 + \ln k)}{\epsilon^2 \delta^2}, \tag{5.7}$$

and

$$\max_{\ell \in [kp_1, kp_2]} \mathbb{P}((X_1 \cdots X_k; Y_1 \cdots Y_\ell) \text{ does not satisfy property } \mathcal{P}) \leq \frac{1}{(2ek)^{2/\epsilon}}. \tag{5.8}$$

The above theorem is very useful for showing that when a property holds for aligned string pairs with similar lengths, say of order k, then the property typically holds in most parts of the optimal alignment. From our experience, for most properties one is interested in, such as the study of dominant matches in optimal alignments, when p_1 and p_2 are close to 1, but fixed, then the probability that

$$(X_1 \cdots X_k; Y_1 Y_2 \cdots Y_\ell)$$

does not satisfy this property is approximately the same for all $\ell \in [kp_1, kp_2]$. In other words, the behavior of the alignment of $X_1 \cdots X_k$ with $Y_1 \cdots Y_\ell$, does not depend much on ℓ as soon as ℓ is close to k and k is fixed. From (5.8), what is needed there is a bound, on the left hand-side probability, smaller than any inverse polynomial-order in k. (At least to be able to take ϵ as close to 0 as one wants to.) If instead $\epsilon > 0$ is chosen small but fixed, then an inverse polynomial bound with a very large exponent will do). So, if this probability is, for example, of order $k^{-\ln k}$ or e^{-k^α} for some constant $\alpha > 0$, the condition (5.8) is satisfied by taking k large enough. Similarly, condition (5.7) is always satisfied for k large enough.

We could also envision using Monte Carlo simulation to find a bound for the probability on the left of (5.8). For that purpose, assume that $\epsilon = 0.2$ and take $\delta = 0.1$. Then, by (5.7), k must be at least 2518253. The probability that strings of length approximately k do not satisfy property $\mathcal{P}$ must be at most $(2ek)^{-10} \approx (13690642)^{-10}$, so a probability smaller than 10^{-70}. However, this is hardly feasible, indeed, to show that a probability is as small as 10^{-70}, one would need to run an order of 10^{70} simulations.

Further Improvements

There are several ways to improve our various bounds. First, we took $\binom{n}{m}$ as an upper bound for the cardinality of $\mathcal{R}_{\epsilon, p_1, p_2}$ and this can be improved. Indeed, note that if $\vec{r} = (r_0, r_1, \ldots, r_m) \in \mathcal{R}_{\epsilon, p_1, p_2}$, then at least $(1 - \epsilon)m$ of the lengths $r_{i+1} - r_i$ are in the interval $[kp_1, kp_2]$. To determine these lengths we have at most

$$((p_2 - p_1)k)^m \tag{5.9}$$

choices. Then, there can be as many as ϵm of the lengths $r_{i+1} - r_i$, which are not in $[kp_1, kp_2]$. Choosing those lengths is like choosing at most ϵm points from a set of at most n elements. Hence, we get as upper bound $\binom{n}{\epsilon m}$ which, in turn, can be upper bounded by, say,

$$\left(\frac{ek}{\epsilon}\right)^{\epsilon m}, \tag{5.10}$$

or via the entropy bound $\exp(nH_e(\epsilon/k))$. Finally, we have to decide which among the m lengths $r_i - r_{i-1}$ have their length in $[kp_1, kp_2]$ and which have not. That choice is further bounded via:

$$\binom{m}{\epsilon m} \leq \exp(H_e(\epsilon)m). \tag{5.11}$$

Combining the bounds (5.9), (5.10) and (5.10), yields

$$\mathrm{Card}\left(\mathcal{R}_{\epsilon, p_1, p_2}\right) \leq \left((p_2 - p_1)k \left(\frac{ek}{\epsilon}\right)^{\epsilon} \exp(H_e(\epsilon))\right)^m. \tag{5.12}$$

With this better bounding of the cardinality of $\mathcal{R}_{\epsilon, p_1, p_2}$, the inequality (5.5) becomes:

$$\mathbb{P}((B_{\mathcal{P}}^n(\epsilon_1 + \epsilon_2))^c) \leq \mathbb{P}((A_{\epsilon_1, p_1, p_2}^n)^c)$$
$$+ \left((p_2 - p_1)k \left(\frac{ek}{\epsilon_1}\right)^{\epsilon_1} \exp(H_e(\epsilon_1))\right)^m q^{\epsilon_2 m} \exp(H_e(\epsilon_2)m), \tag{5.13}$$

which when combined with Theorem 2.2 yields that

$$\mathbb{P}((B_{\mathcal{P}}^n(\epsilon_1 + \epsilon_2))^c) \leq \exp\left(-n\left(-\frac{1 + \ln k}{k} + \frac{\delta^2 \epsilon_1^2}{16}\right)\right)$$
$$+ \left((p_2 - p_1)k \left(\frac{ek}{\epsilon_1}\right)^{\epsilon_1} \exp(H_e(\epsilon_1))q^{\epsilon_2} \exp(H_e(\epsilon_2))\right)^m. \tag{5.14}$$

Again, this last expression is exponentially small in n (assuming k fixed) if the following two conditions are satisfied:
(i)

$$k > \frac{16(1 + \ln k)}{\epsilon_1^2 \delta^2}, \tag{5.15}$$

(the smallest integer k satisfying (5.15) with $\epsilon_1 = 0.2$ and $\delta = 0.1$ is now 570146) and
(ii)

$$q(k) < \frac{1}{\left((p_2 - p_1)k \left(\frac{ek}{\epsilon_1}\right)^{\epsilon_1} \exp(H_e(\epsilon_1) + H_e(\epsilon_2))\right)^{1/\epsilon_2}}; \tag{5.16}$$

and combining these last two conditions yields:

$$q(k) < \left(\frac{\epsilon_1^2 \delta^2}{(p_2 - p_1)16(1 + \ln k) \left(\frac{ek}{\epsilon_1}\right)^{\epsilon_1} \exp(H_e(\epsilon_1) + H_e(\epsilon_2))}\right)^{1/\epsilon_2}. \tag{5.17}$$

Typically $\epsilon_1 + \epsilon_2$ should be of a given order. So, let us maximize the right-hand side of (5.17) under the constraint $\epsilon = \epsilon_1 + \epsilon_2$. To do so, note that the power $1/\epsilon_2$ has a much more minimizing influence than the expression ϵ_1^2 in the numerator, while $1 \leq \exp(H_e(\epsilon_1) + H_e(\epsilon_2)) \leq 2$ and so this last quantity does not have much of an influence. Also, note $(ek/\epsilon_1)^{\epsilon_1}$ is somewhat negligible compared to ek. So, at first, let us disregard the quantities $(ek/\epsilon_1)^{\epsilon_1}$ and $\exp(H_e(\epsilon_1)) + H_e(\epsilon_2))$, and let

$$g(k, \epsilon_1, \epsilon_2) := \left(\frac{\epsilon_1^2 \delta^2}{(p_2 - p_1)16(1 + \ln k)}\right)^{1/\epsilon_2}.$$

Clearly, $g(k, \epsilon_1, \epsilon_2)$ is larger than the bound on the right-hand side of (5.17) and when all the parameters p_1, p_2 and δ are held fixed, $g(k, \epsilon_1, \epsilon_2)$ is decreasing in both ϵ_1 and ϵ_2. However, ϵ_2 "has more decreasing influence" than ϵ_1. Therefore, given ϵ and given that all the parameters are fixed (including k), maximizing $g(k, \epsilon_1, \epsilon_2)$ under the constraint $\epsilon_1 + \epsilon_2 = \epsilon$, $\epsilon_1, \epsilon_2 > 0$ lead to a quantity where ϵ_2 is quite a bit larger than ϵ_1.

Could Monte Carlo simulations be realistic with $\epsilon = 0.1$ and the bounds which we have? The answer is no. Indeed, at first, $\delta/(p_2 - p_1)$ gets better when $|p_2 - p_1|$ increases since the derivative of $\tilde{\gamma}$ at $p = 1$ is zero. When the interval $[kp_1, kp_2]$ becomes too large however, then the property might no longer hold with high probability for all pairs $(X_1 \cdots X_k, Y_1 \cdots Y_\ell)$, with $\ell \in [kp_1, kp_2]$. So, we will take $[p_1, p_2]$, as large as possible, so this property still holds with high probability for all the string pairs mentioned before. With such a choice, $\delta/(p_2 - p_1)$ can be treated as a constant. Somewhat, optimistically, say that the constant is less than $1/3$. Now if $\epsilon = 0.1$, then $\epsilon_1, \epsilon_2 \leq 0.1$. In that case,

$$g(k, \epsilon_1, \epsilon_2) \leq g(k, 0.1, 0.1) \leq \left(\frac{0.01\delta}{3 \times 16(1 + \ln k)} \right)^{10}.$$

Returning to (5.15) and taking $\delta = 0.2$, we find that k must be larger than 10^{10}, so that $\ln k$ is bigger than 20. With this in mind, and in the present case where $\epsilon_1 + \epsilon_2 = 0.1$, we find that $g(k, \epsilon_1, \epsilon_2)$ is smaller 10^{-56}, so there is still little hope to perform Monte Carlo simulation here.

Monte Carlo simulation with $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.2$: Take also $\delta = 0.2$ and $\delta/(p_2 - p_1) = 1/2$. With these values, and using (5.15), then k must be somewhat larger than $10^2 \cdot 10 \cdot 8 \ln(24000) \approx 10^5$. Then, by (5.16), $q(k)$ should also be less than

$$\left(10^2 \cdot 10 \cdot 11 \cdot 8 \right)^{-5} \approx 10^{-25}.$$

This is still a difficult order for Monte Carlo simulation and if we had $\epsilon_2 = 0.3$ instead, then we would get a bound 10^{-15} which remains a difficult order.

When only dealing with the inequality (5.16), things look somewhat better. Take $k = 1000$ and $(p_1 - p_1)k = 100$, then the bound on $q(k)$ is of order about 10^{-5} which is feasible with Monte Carlo. So, if we could find another method than the one described here to make sure that most of the pieces of strings $X_{(i-1)k+1} X_{(i-1)k+2} \cdots X_{ik}$ are aligned with pieces of similar length we would end up in a favorable setting.

References

[1] Alexander, K. S. The rate of convergence of the mean length of the longest common subsequence, Annals of Applied Probability 4, (1994), 1074–1082. MR-1304773

[2] Amsalu, S., Houdré, C. and Matzinger, H., Sparse long blocks and the variance of the length of the longest common subsequences in random words, arXiv:1204.1009 [math.PR], (2012).

[3] Amsalu, S., Houdré, C. and Matzinger, H., Sparse long blocks and the micro-structure of the longest common subsequences in random words, J. Stat. Phys., **154**, 6, (2014), 1516–1549.

[4] Baik, J., Deift, P. and Johansson, K. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999. MR-1682248

[5] Breton, J.-C. and Houdré, C., On the limiting law of the length of the longest common and increasing subsequences in random words. arXiv:1505.06164 [Math.PR], (2015).

[6] Capocelli, R.M., Sequences: Combinatorics, Compression, Security, and Transmission, Springer-Verlag New York, (1989). MR-1040295

[7] Chvátal, V. and Sankoff, D., Longest common subsequences of two random sequences, J. Appl. Probability, **12**, (1975), 306–315. MR-0405531

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

[9] Houdré, C. and Işlak, Ü., A central limit theorem for the length of the longest common subsequences in random words, arXiv:1408.1559v3 [math.PR], (2015).

[10] Houdré, C., Lember, J. and Matzinger, H., On the longest common increasing binary subsequence, C.R. Acad. Sci. Paris Ser. I 343, (2006), 589–594. MR-2269870

[11] Houdré, C. and Litherland, T. L. On the longest increasing subsequence for finite and countable alphabets. *HDPV: The Luminy Volume, IMS Collection*, 5:185–212, 2009. MR-2797948

[12] Houdré, C. and Litherland, T. L. On the limiting shape of Young diagrams associated with Markovian random words. arXiv:1110.4570 [math.PR], (2011). MR-2797948

[13] Houdré, C. and Ma, J., On the order of the central moments of the length of the longest common subsequences in random words, To appear: Progress in Probability: Birkhauser (2016).

[14] Houdré, C. and Matzinger, H., On the variance of the optimal alignments score for binary random words and an asymmetric scoring function, Under Revision: J. Stat. Phys (2016).

[15] Its, A. R., Tracy, C. and Widom, H. Random words, Toeplitz determinants, and integrable systems. I. Random matrix models and their applications. *Math. Sci. Res. Inst. Publ.*, 40, 2001. MR-1842789

[16] Its, A. R., Tracy, C. and Widom, H. Random words, Toeplitz determinants, and integrable systems. II. Advances in nonlinear mathematics and science. *Phys. D*, 152–153:199–224, 2001. MR-1837910

[17] Johansson, K. Transversal fluctuations for increasing subsequences on the plane, Probab. Theory Related Fields 116, (2000) 445–456. MR-1757595

[18] Lember, J. and Matzinger, H., Standard deviation of the LCS when zero and one have different probabilities, Annals of Probability 37, (2009), 1192–1235. MR-2537552

[19] Lueker, G. S., Improved bounds on the average length of longest common subsequences Journal of the ACM, **56**, Art. 17, (2009). MR-2536132

[20] Robin, S., Rodolphe, F., and Schbath, S., ADN, mots et modèles, Belin, Paris, 2003.

[21] Sankoff, D., and Kruskal, J., Time warps, string edits and macromolecules: The theory and practice of sequence comparison Center for the Study of Language and Information, 1999.

[22] Seppäläinen, T., Scaling for a one-dimensional directed polymer with boundary conditions, Ann. Probab. 40, (2012) 19–73. MR-2917766

[23] Steele, J. M., Long common subsequences and the proximity of two random strings, SIAM J. Appl. Math. 42 (1982), 731–737. MR-0665382

[24] Waterman, M. S., Estimating statistical significance of sequence alignments Phil. Trans. R. Soc. Lond. B. (1994), 383–390.

[25] Waterman, M. S., Introduction to Computational Biology: Maps, Sequences and Genomes (Interdisciplinary Statistics), CRC Press (2000).

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards

- Free for authors, free for readers

- Quick publication (no backlog)

- Secure publication (LOCKSS[1])

- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4], ProjectEuclid[5]

- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)

- Submit your best articles to EJP-ECP

- Choose EJP-ECP over for-profit journals

---

[1]LOCKSS: Lots of Copies Keep Stuff Safe http://www.lockss.org/
[2]EJMS: Electronic Journal Management System http://www.vtex.lt/en/ejms.html
[3]IMS: Institute of Mathematical Statistics http://www.imstat.org/
[4]BS: Bernoulli Society http://www.bernoulli-society.org/
[5]Project Euclid: https://projecteuclid.org/
[6]IMS Open Access Fund: http://www.imstat.org/publications/open.htm