Macroscopic non-uniqueness and transversal fluctuation in optimal random sequence alignment

Saba Amsalu[1] Heinrich Matzinger[2] Serguei Popov[*,3]

January 1, 2007

[1]University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany
e-mail: saba@uni-bielefeld.de
[2]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332–0160, USA
e-mail: matzi@math.gatech.edu
[3]Instituto de Matemática e Estatística, Universidade de São Paulo, rua do Matão 1010, CEP 05508–090, São Paulo SP, Brasil
e-mail: popov@ime.usp.br

Abstract

We investigate the optimal alignment of two independent random sequences of length n. We provide a polynomial lower bound for the probability of the optimal alignment to be macroscopically non-unique. We furthermore establish a connection between the transversal fluctuation and macroscopic non-uniqueness.

1 Introduction

In computational genetics and computational linguistics one of the basic problem is to find an optimal alignment between two given sequences $X := X_1 \dots X_n$ and $Y := Y_1 \dots Y_n$. This requires a scoring system which can rank the alignments. Typically a substitution matrix gives the score for each possible pair of letters. The total score of an alignment is the sum of terms for each aligned pair of residues, plus terms for each gap.

*Partially supported by SFB 701 grant and CNPq (302981/02–0)

In this paper we take the texts X and Y to be i.i.d. and independent of each other. One may immediately remark that, for most of the applications in computer science and biology, one normally assumes a much more complicated relationship between X and Y than this i.i.d. setup. However, the mathematical problems arising in the sequence alignment theory are usually very difficult, and even the theory of the i.i.d. case is far from being complete. Therefore, we study the i.i.d. case since it can be considered as the first step in understanding sequence alignments for more complex situations, which one may encounter in practice.

One of the main purposes of this paper is to try to understand what causes the optimal alignment to be non-unique on portions of the texts of length of order n (we call that *macroscopic non-uniqueness*). In Theorem 2.1, we prove that macroscopic non-uniqueness happens with probability at least $1/(n+1)$. This seems to suggest that typically the optimal alignment is non-unique on stretches of order at least $n^{0.5-\epsilon}$, see Remark 2.1. For two sequences which have been obtained from one common ancestor by transformation, we expect the non-uniqueness stretches to be of order at most $\ln n$. This difference could prove to be useful to determine if the sequences are related or not. Our present result sheds some new light on the question of non-uniqueness. Previously, Hauser and Matzinger [7] proved a result that goes in the opposite direction. They showed that typically, for large gap penalty, the optimal alignment is unique in most points.

The second main result of this paper is that a small probability of macroscopic non-uniqueness implies a large transversal fluctuation. More precisely, a lower bound (3.2) for the interquartile distance of the transversal fluctuation is derived. Roughly speaking, this lower bound equals the inverse of the probability of macroscopic non-uniqueness. Macroscopic non-uniqueness is present when we observe simultaneously two close-to-optimal alignments differing on a stretch of order n.

Optimal alignment can be viewed formally as a Last Passage Percolation (LPP) problem with correlated weights. The weights depend on the one-dimensional texts X and Y. This confers a different structure than in standard LPP and many of the techniques of percolation (BK and FKG inequalities do not work for example). For standard LPP the question of the order of the transversal fluctuation has long been open [8].

Let us present an example to illustrate the practical usefulness of optimal sequence alignment.

Let us explain how an automatic spell-checker works. The spell-checker has to identify misspelled words. For each misspelled word it should give a list of similar words from a lexicon. In this list, one hopes to find the word which was originally meant to be written. Take, for example, the word $Y =$ *probability* and the misspelled word $X =$ *prbobilite*. One could try to align the two words to detect similarities. We obtain the alignment

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c}
p & r & o & b & a & b & i & l & i & t & y \\ \hline
p & r & b & o & b & i & l & i & t & e &
\end{array}$$

The computer counts two matching letters. These are the first two letters: *pr*. Here, the computer is unable to detect the great degree of similarity between X and Y by this simple alignment.

A better approach consists in aligning the words X and Y while allowing the existence of gaps in the alignment. We search for the alignment with gaps yielding the maximum number of matching letters. In our example, such an alignment with gaps maximizing the number of matches is given by:

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c|c|c}
p & r & o & b & & a & b & i & l & i & t & y & \\ \hline
p & r & & b & o & & b & i & l & i & t & & e
\end{array} \tag{1.1}$$

We count 8 matched letters. These 8 letters form the Longest Common Subsequence (LCS) of X and Y (the longest sequence which is a subsequence of X and of Y). That is, the LCS is *prbbilit*. The length of the LCS is a *score* which measures the degree of similarity between X and Y. We call the alignment (1.1) *optimal* since it gives the maximal number of coinciding letters. Note that (1.1) is not the only optimal alignment. Another optimal alignment is:

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c|c|c}
p & r & & o & b & a & b & i & l & i & t & y & \\ \hline
p & r & b & o & b & & & i & l & i & t & & e
\end{array} \tag{1.2}$$

The alignment (1.2) matches 8 letters correctly. Note that the fifth letter in X is aligned to different letters of text X depending on the optimal alignment we chose. We say that the optimal alignment is *non-unique* in the fifth letter of X. In this paper, we investigate the possibility for two long independent random texts to have long stretches where the optimal alignment is non-unique. The type of alignment scores we consider are fairly general. We also allow for the alignment of non-identical letters. In our example, we could reward the alignment of identical, respectively, similar letters with a score of 1, respectively, 0.5. Using this scoring scheme and assuming that y and e are similar letters, the optimal alignment becomes

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c|c}
p & r & & o & b & a & b & i & l & i & t & y \\ \hline
p & r & b & o & b & & & i & l & i & t & e
\end{array} \tag{1.3}$$

with a score of 8.5.

At present let us mention a little more on the history of mathematical problems related to optimal sequence alignments. The optimal alignment problem, as we mentioned, belongs to the area of First Passage and Last Passage percolation. In general, the exact order of transversal fluctuation is not known for standard First Passage Percolation (FPP) and LPP. In standard FPP the properties of the optimal path were also studied. Typically the interesting object of research there is, for example, the total number of edges in an optimal path. The convergence of this number properly rescaled is an open problem [8]. For optimal alignment an interesting question about the optimal path is to which extend it is unique. This is one of the questions investigated in this paper.

Let L_n designate the length of the LCS of two independent i.i.d. sequences of length n. Using a subadditivity argument, Chvàtal and Sankoff [5] proved that the limit

$$\gamma := \lim_{n\to\infty} \frac{E[L_n]}{n}$$

exists. They consider two binary sequences (this is the standard setting for this problem). The constant γ is called the Chvàtal-Sankoff constant and its value is unknown. Neither is the exact order of the fluctuation of the LCS length known. Steele [14] proved that $\mathrm{Var}[L_n] \leq n$. In [15], Waterman conjectured that in many cases the variance of L_n grows linearly. Matzinger and Lember [11] proved that indeed this is the right order in an important case.

In LPP language, the fluctuation of L_n is called longitudinal fluctuation (the fluctuation under investigation in this paper is the transversal fluctuation).

The determination of the Chvàtal-Sankoff constant and the order of the longitudinal and transversal fluctuations for the LCS problem are long standing open problems. Montecarlo simulations lead Chvàtal and Sankoff to conjecture that $\mathrm{Var}[L_n] = o(n^{\frac{2}{3}})$. This order of magnitude is similar to the order for the longest increasing subsequence (LIS) of random permutations (see Baik, Deift and Johansson [4] and also Aldous and Diaconis [1]). This similarity of the order of magnitudes is not a complete surprise. The LIS problem is asymptotically equivalent to a Poisson-graph based LPP model. For standard LPP the order of magnitude of these fluctuations has been open for decades despite LPP being one of the central research areas in discrete probability. The only case for which the order of magnitude of these fluctuations is exactly known, is for the Poisson-based model used for the solution of the LIS-problem. There, the precise order of the transversal fluctuation has been established by Johansson [10].

Waterman and Arratia [3] derive a law of large deviations for L_n for fluctuations on scales larger than $\sqrt{n}$. In their ground breaking article [3], they show the existence of critical phenomena (i.e., whether L_n is positive linear in n or not).

For a general discussion on the relevance of string comparison for biology and on other similar problems in computational biology the reader can refer to the standard texts [13] and [6].

This paper is organized as follows. In Section 2, first, we give some formal definitions necessary to formulate our results. Then, we derive a lower bound for the probability of the macroscopical non-uniqueness (Theorem 2.1) and after that, we improve this bound, which requires however a plausible, but still unproven, assumption (Theorem 2.2). In Section 3 we establish a relation between the transversal fluctuation and the probability of macroscopic non-uniqueness (Theorem 3.1). Proofs are given in Section 4.

2 A lower bound for probability of macroscopical non-uniqueness

Now, let us proceed to the formal definitions. In everything that follows, $\{X_i\}_{i\in\mathbb{N}}$ and $\{Y_i\}_{i\in\mathbb{N}}$ are two processes independent of each other. We assume that the X_i's are i.i.d. and that the Y_i's are i.i.d., and they are all drawn from a finite alphabet $\mathcal{A}$.

Definition 2.1 *An* alignment of length k *is a couple* (π,η) *consisting of two increasing sequences of length* k *each, such that* $1 \leq \pi(1) < \pi(2) < \ldots < \pi(k) \leq n$ *and* $1 \leq \eta(1) < \eta(2) < \ldots < \eta(k) \leq n$.

The interpretation is the following: the alignment (π,η) aligns the $\pi(i)$-th letter of the the first text with the $\eta(i)$-th letter of the second text, $i = 1,\ldots,k$.

Let us give an example of an alignment. In this example we align the English word "think" with its German translation "denke". A possible alignment is

t	h	i	n	k	$-$
d	_	e	n	k	e

the alignment is there to show which letters are related. The letters which are aligned with a gap are supposed to be missing in one of the two words. The "t" of "think" is aligned with the "d" of "denke". The is the first letter of the first string is aligned with the first letter of the second string. According to our notation, this means that $\pi(1) = \eta(1) = 1$. Next, the "i"of "think" is aligned with the first "e" of "denke". Hence, the third letter of the first string is aligned with the second letter of the second string. This implies that $\pi(2) = 3$ and $\eta(2) = 2$. Eventually, the third aligned letter, that is the "n"'s appear in the fourth position, respectively, third position. Hence, $\pi(3) = 4$ and $\eta(3) = 3$. The last letter to get aligned are the "k"'s in both texts. We find $\pi(4) = 5$ and $\eta(4) = 4$.

The total number of aligned pairs of letters (letters that are not aligned with gaps) is 4. Hence the length of the alignment (π, η) is equal to 4.

In many cases, it can be useful to view an alignment as a two dimensional table. For this the x entries are given by the first string whilst the second string gives the y entries. A "•" sign shows every pair of aligned letters. The alignment considered in this example then becomes:

e					
k					•
n				•	
e			•		
d	•				
	t	h	i	n	k

Let $s(a, b)$ be the score obtained by aligning a with b. Denote by $s = (s(a, b))_{a,b\in\mathcal{A}}$ the substitution matrix and by q the gap penalty. Let (π, η) be an alignment of length k.

Definition 2.2 *The score $S(\pi, \eta)$ of the alignment (π, η) is defined to be equal to:*

$$S(\pi, \eta) := \sum_{i=1}^{k} s(X_{\pi(i)}, Y_{\eta(i)}) - 2q(n - k).$$

In the above definition, $s(X_{\pi(i)}, Y_{\eta(i)})$ is the contribution which we obtain for aligning letter $X_{\pi(i)}$ from the first text with letter $Y_{\eta(i)}$ of the second text. The quantity $2q(n-k)$ is the total gap penalty since there are $2(n-k)$ letters which are not aligned.

Let L denote the maximal alignment score of the two texts X and Y, i.e.,

$$L = L(X, Y) := \max_{(\pi,\eta)} S(\pi, \eta),$$

where the maximum is taken over all alignments of the text X with the text Y.

Definition 2.3 *An* optimal alignment of the texts X and Y *is an alignment* (π, η) *that has maximal alignment score, i.e., such that*

$$L = S(\pi, \eta).$$

Naturally, there may exist several optimal alignments.

Let us look at our alignment of the word "think" with the word "denke". Assume that $s(\cdot, \cdot)$ is such that for identical letters the score is 1, whilst for similar letters the score is 0.5. Let the score for a pair of dissimilar aligned letters be -0.5, and the gap penalty be 0.5. We assume that "t" and "d" are similar to each other and that so are "e" and "i". The score of the alignment

t	h	i	n	k	$-$
d	-	e	n	k	e

is then $2(0.5)+2-2(0.5) = 2$. In this alignment there are two pairs of similar letters and two pair of identical letters which get aligned. One can check that this is also the alignment with maximum score. Hence, the alignment (π, η) where $(\pi(1), \pi(2), \pi(3), \pi(4)) = (1, 3, 4, 5)$ and $(\eta(1), \eta(2), \eta(3), \eta(4)) = (1, 2, 3, 4)$ is an optimal alignment of the word "think" with the word "denke".

We defined L to be the the maximal alignment score of the two texts X and Y. To indicate that the texts have length n, we may sometimes write L_n for L. Throughout this paper, we suppose that the following holds:

Assumption. *We are in the linear phase (see Arratia and Waterman [3]), and so*

$$\lim_{n\to\infty} \frac{L_n}{n} > 0. \tag{2.1}$$

Suppose that (π, η) is an alignment of length k. Let f be the continuous map obtained by linear interpolation from the discrete map:

$$\pi(i) \mapsto \eta(i).$$

We call f the *path associated with* (π, η). More precisely, f is the continuous map from $[\pi(1), \pi(k)]$ to $[\eta(1), \eta(k)]$, such that both of the following conditions hold:

- For all $i \in \{1, 2, \dots, k\}$, we have: $f(\pi(i)) = \eta(i)$.

- For all $i \in \{1, 2, \ldots, k-1\}$ and all $t \in [\pi(i), \pi(i+1)]$, we have:

$$f(t) = \eta(i) + (t - \pi(i))\frac{\eta(i+1) - \eta(i)}{\pi(i+1) - \pi(i)}.$$

Let $f : [a_1, b_1] \to \mathbb{R}^+$ and $g : [a_2, b_2] \to \mathbb{R}^+$ be two continuous maps. Let $x \in [a_1, b_1] \cap [a_2, b_2]$. We say that f and g *cross at the point* x iff $f(x) = g(x)$. We say that f and g *do not cross on an interval* I if there exists no point $x \in I \cap [a_1, b_1] \cap [a_2, b_2]$ such that $f(x) = g(x)$.

Let us introduce the following important notations: first, let I_1 denote the interval $I_1 := [0, (n+1)/2)$ and let $I_2 := ((n+1)/2, n]$.

Definition 2.4 *We say that an alignment (π', η') is close to optimal, if*

$$|S(\pi', \eta') - L| \leq 2 \max_{a,b \in \mathcal{A}} |s(a, b)| + 4q.$$

Let A be the event defined as follows:

$$\begin{aligned} A \;=\; & \{\text{there exist an optimal alignment } (\pi, \eta) \text{ of } X \text{ and } Y \text{ and} \\ & \text{a close-to-optimal alignment } (\pi', \eta') \text{ which do not cross} \\ & \text{on at least one of the two intervals } I_1,\ I_2\}. \end{aligned} \tag{2.2}$$

The next theorem is one of the two main results of this paper. It gives a lower bound for the probability that there exist simultaneously two close-to-optimal alignments which do not cross each other on a long stretch.

Theorem 2.1 *We have that*

$$P(A) \geq \frac{1}{1+n}. \tag{2.3}$$

Let

$$r_1 := \max_{(\pi,\eta)} \pi(1)$$

where the maximum is taken over all optimal alignments (π, η) of the texts X and Y. We will show in the next proposition that with high probability, the interval $[r_1, (n+1)/2]$ has length of linear order in n. If $|r_1 - n/2|$ is not of linear order in n, then the statement "there exist two close-to-optimal alignments which do not cross on $[0, n/2]$" does not convey a lot of information: they do not cross simply because they are not defined on most of that interval.

Note that $|r_1 - n/2|$ is of linear order only when we are in the linear phase, that is when we have (2.1). This is why throughout this paper we assume that (2.1) holds.

Proposition 2.1 *There exist a constant $c_{LD} > 0$ not depending on n, (but depending on the scoring function and the gap penalty), such that*

$$P(r_1 \geq 0.4n) \leq 2n^{-c_{LD} \ln n}. \tag{2.4}$$

Although Theorem 2.1 gives a lower bound for the probability of macroscopic non-uniqueness, we do not believe that the inequality (2.3) is the best possible. In the rest of this section we will try to strengthen this inequality, assuming a fact about the so-called *mean curve* that we believe is true.

Let us define the mean curve. Let $p \in [-1, 1]$. We define $L_n(p)$ to be the optimal alignment score when we align two independent i.i.d. sequences with unequal lengths $n(1+p)$ and $n(1-p)$ (to simplify notations we assume that np and $n(1-p)$ are integers). Hence:

$$L_n(p) := L(X_1 X_2 \ldots X_{n(1+p)}; Y_1 Y_2 \ldots Y_{n(1-p)}).$$

(Recall that, for two sequences V and W, we denoted by $L(V; W)$ the optimal alignment score of V and W.) By subadditivity, the limit

$$\gamma(p) := \lim_{n \to \infty} \frac{E[L_n(p)]}{n}$$

exists. The curve $p \mapsto \gamma(p)$ is called the *mean curve* and is concave and symmetric around the origin. Concavity follows from a subadditivity argument. Not a lot is known, however, about differentiability and curvature properties of the mean curve, although simulations seem to indicate non-zero curvature everywhere.

In FPP language (recall that optimal alignment is a FPP problem), the mean curve corresponds to the asymptotic shape (rescaled limit) of the wet region. There are known special examples for which that shape contains segments of straight lines (what we actually need is that the mean curve does not contain a segment of straight line at $p = 0$). It is not assumed that this is typical, although nothing on this is known rigorously [8]. Hence, the exact geometry of the mean curve is an extremely hard question. For general first passage and last passage percolation this has been open for decades. (Recall that optimal alignment is a last passage percolation with correlated weights.) Note that in [9] and [12], Howard and Newman as well as Newman and Piza also use an assumption similar to ours (of Theorem 2.2) to establish an upper bound for transversal fluctuation.

We established Theorem 2.1 using that Z and $\tilde{Z}$ take values in $[0, n]$. If the values taken by Z and $\tilde{Z}$ lie with high enough probability in an interval of smaller order, then the lower bound (2.3) can be improved. This is the idea behind the next theorem:

Theorem 2.2 *Assume that $p \mapsto \gamma(p)$ has continuous second derivative in an open neighborhood of $p = 0$ and $\gamma(p)'' < 0$ (positive curvature at $p = 0$). Then,*

$$P(A) \geq \frac{1}{2n^{0.75} \ln n}, \tag{2.5}$$

for all n large enough.

We believe that one should be able to prove that the polynomial lower bounds (2.3) and (2.5) imply also a similar bound for two macroscopically different optimal alignments. (In this paper the result is for one optimal and one close to optimal alignment, which is somewhat weaker.) We plan to investigate this issue in a forthcoming paper.

Remark 2.1 *Using the polynomial lower bound for the probability of macroscopically different close to optimal alignments, there is a* heuristic argument *which suggests that the path of the optimal alignments is typically non-unique on stretches of size $n^{0.5+\epsilon}$, where $\epsilon > 0$. To see this, divide the intervals $[1, n]$ into $n^{0.5+\epsilon}$ pieces of length $n^{0.5-\epsilon}$ each. According to Theorem 2.1, the probability that the optimal path in non-unique on one of the stretches of length $n^{0.5-\epsilon}$ is larger than $n^{-0.5+\epsilon}$. Since there are $n^{0.5+\epsilon}$ of them, the expected number of stretches of length $n^{0.5-\epsilon}$ where the optimal alignment is non-unique is at least $n^{2\epsilon}$. In other words, typically, the optimal alignment is non-unique on a stretch of polynomial length in polynomially many places.*

This is in stark contrast to what we expect when X and Y are not independent but were obtained by random mutations from a common ancestor (computational biology). Then, it is believed that in many cases the optimal alignment should be non-unique on stretches of length of at most logarithmic order in n. This very different behaviour should be useful for distinguishing if two sequences are related or not.

3 Transversal fluctuation and the probability of macroscopic non-uniqueness

Let Z be the smallest value in the point $(n+1)/2$ taken by the path associated with an optimal alignment of the texts X and Y. More precisely:

$$Z := \min_f f((n+1)/2),$$

where the minimum is taken over all paths f associated with an optimal alignment (π, η) of X and Y.

Note that the fluctuation of Z in First and Last Passage Percolation language is called *transversal fluctuation*.

For a random variable W we define the interquartile distance by

$$q_w = F_W^{-1}(3/4) - F_W^{-1}(1/4),$$

where $F_W^{-1}(.)$ designates the inverse distribution function of W.

We are now ready to state the second main result of this paper. This result says that if the probability that there exists two close-to-optimal alignments different on large stretches is small, than the fluctuation of Z is large.

Recall that A is the event that there exist simultaneously two close-to-optimal alignments different on a large stretch. The event A was defined just before Theorem 2.1.

Theorem 3.1 *Let $\alpha \in [0,1]$ and $C > 0$. Assume that*

$$P(A) \leq Cn^{-\alpha}, \tag{3.1}$$

then

$$q_Z \geq (2(Cn^{-\alpha} + n^{-1}))^{-1} - 4, \tag{3.2}$$

where q_Z is the interquartile distance of the random variable Z.

We saw in Theorem 2.1 that $P(A) \geq n^{-1}$. Hence we are only interested in the case when $\alpha < 1$. For $\alpha \in (0,1]$, the order of the expression in the right-hand side of (3.2) is n^{α}. In other words, we get a lower bound on the transversal fluctuation of order equal to the inverse of the probability $P(A)$.

The result of Theorem 3.1 can also be described in the following way: a *small quenched fluctuation implies a large annealed fluctuation*. To see this, assume that we hold X and Y fixed and select one optimal path at random.

The "fluctuation" between one optimal path and the other for X and Y fixed, can be interpreted as quenched fluctuation. If the different optimal paths do not differ a lot macroscopically then this fluctuation is small. We think of the annealed fluctuation as of how much the optimal paths changes when we redraw X and Y. The fluctuation of Z is a good measure of these fluctuations.

4 Proofs

4.1 Main idea: a measure-preserving transformation

The main technique used in this paper is introducing a measure preserving map (transformation) $\sim$. This map is defined as follows: we transform the text Y by removing the first letter of Y and placing it at the end. The text obtained in this way is denoted by $\tilde{Y}$. More precisely:

- For i with $0 \leq i < n$, we define $\tilde{Y}_i := Y_{i+1}$.
- We define $\tilde{Y}_n := Y_1$.

Obviously, the transformation $\sim$ does not change the distribution, since we assumed that the texts are i.i.d.

We denote by $\tilde{S}(\pi, \eta)$ the score obtained when we use the alignment (π, η) to align the texts X and $\tilde{Y}$:

$$\tilde{S}(\pi, \eta) := \sum_{i=1}^{k} s(X_{\pi(i)}, \tilde{Y}_{\eta(i)}) - 2q(n-k).$$

Let $\tilde{L}$ denote the optimal alignment score of the texts X and $\tilde{Y}$:

$$\tilde{L} := \max_{(\pi,\eta)} \tilde{S}(\pi, \eta)$$

where the maximum is taken over all alignments of the text X with the text $\tilde{Y}$.

Again, let us look at the numerical example where X is the word "think" and Y is equal to "denke". Take the scoring scheme as described in the numerical example before. In this case, we find that the transformed text $\tilde{Y}$

is equal to "enked". We obtain this by removing the "d" at the beginning of "denke" and moving it to the end of the word.

The optimal alignment of X and $\tilde{Y}$ is:

t	h	i	n	k	_	_
_	_	e	n	k	e	d

In this alignment there is one pair of similar letters aligned, and two identical letters aligned. Four letters are aligned with gaps. The score of the above alignment is hence equal to $0.5 + 2 - 4(0.5) = 0.5$. This is the maximum possible alignment score and hence $\tilde{L} = 0.5$.

Note that between Y and $\tilde{Y}$ the only difference is the first letter of Y and the last letter of $\tilde{Y}$. The above alignment of X with $\tilde{Y}$ can be obtained from the alignment of X with Y presented previously. For this we simply align all the letters of $\tilde{Y}$, except Y_1, with the same letter as before. The last letter of $\tilde{Y}$ gets aligned with a gap. In this way every alignment of X with Y induces an alignment of X with $\tilde{Y}$. We denote by $(\tilde{\pi}, \tilde{\eta})$ the alignment of X with $\tilde{Y}$ induced by the alignment (π, η) of X with Y.

The difference in score of the two alignments is at most the maximum possible score for a pair of letters, plus twice the gap penalty.

Let (π, η) be an alignment of length k (recall that we think of (π, η) as an alignment of X with Y). As mentioned in the previous paragraph, we write $(\tilde{\pi}, \tilde{\eta})$ for the alignment of the text X with the text $\tilde{Y}$, *induced by the alignment* (π, η). By this we mean that, except for the letter Y_1, the two alignments align the same pair of letters. More precisely: Let $1^* := 2$ if $\eta(1) = 1$ and $1^* := 1$ otherwise. We define $\tilde{\pi}$ to be the increasing sequence:

$$\pi(1^*) < \pi(1^* + 1) < \ldots < \pi(k).$$

Let $\tilde{\eta}$ be the increasing sequence

$$\eta(1^*) - 1 < \eta(1^* + 1) - 1 < \ldots < \eta(k) - 1.$$

It is straightforward to note that the length of alignment $(\tilde{\pi}, \tilde{\eta})$ is $k - 1$ if $\eta(1) = 1$ and k if $\eta(1) > 1$. Similarly, we define the alignment $\hat{\pi}, \hat{\eta}$ to be the alignment of length $k - 1$ defined by the equation

$$(\hat{\pi}(i), \hat{\eta}(i)) := (\pi(i), \eta(i) + 1),$$

which is to hold for every $i \in [1, k-1]$. If we think of (π, η) as an alignment of X with $\tilde{Y}$, then $(\hat{\pi}, \hat{\eta})$ aligns the texts X and Y in such a way that, roughly, the same letters get aligned for both alignments. The only exception is the last aligned pair of letters aligned by (π, η).

The next lemma states what we already saw in the last numerical example: the difference in score between an alignment and its induced alignment is less than the maximal score for a pair of letters, plus twice the gap penalty.

Lemma 4.1 *Let (π, η) be an alignment. We have that*

$$|S(\pi, \eta) - \tilde{S}(\tilde{\pi}, \tilde{\eta})| \leq q^* \tag{4.1}$$

and

$$|\tilde{S}(\pi, \eta) - S(\hat{\pi}, \hat{\eta})| \leq q^*. \tag{4.2}$$

whilst

$$|L - \tilde{L}| \leq q^*, \tag{4.3}$$

where

$$q^* := \max_{a,b \in \mathcal{A}} |s(a, b)| + 2q \tag{4.4}$$

(compare with Definition 2.4).

Proof. The alignment $(\tilde{\pi}, \tilde{\eta})$ contains one pair of aligned letters less than (π, η). The loss incurred for that is at most $\max_{a,b \in \mathcal{A}} |s(a, b)|$. If two letters are matched with gaps the additional penalty occurring is $2q$. Hence,

$$S(\tilde{\pi}, \tilde{\eta}) \geq S(\pi, \eta) - q^*,$$

whilst, by definition, we have

$$S(\pi, \eta) \geq S(\tilde{\pi}, \tilde{\eta}).$$

Therefore, (4.1) follows. Similarly we prove (4.2).

For every alignment (π, η) we have that $(\tilde{\pi}, \tilde{\eta})$ is well defined. Hence, we know that for every alignment (π, η) of X and Y, there is an alignment of X and $\tilde{Y}$ with a score closer than q^* from the score $S(\pi, \eta)$. The converse is also true. Hence (4.3). □

Recall that Z was defined as the smallest value in the point $(n + 1)/2$ taken by a path associated with an optimal alignment of the texts X and Y, i.e.,

$$Z := \min_f f((n + 1)/2),$$

where the minimum is taken over all paths f associated with optimal alignments of the two texts. Similarly, we define $\tilde{Z}$:

$$\tilde{Z} := \min_f f((n+1)/2),$$

where the minimum is taken over the set of all paths associated with an optimal alignment (π, η) of X and $\tilde{Y}$ (we could take any other point instead of $(n+1)/2$, which is at linear distance from 1 and n, to make the same argument).

Let us come back to our numerical example where $X =$ *think* and $Y =$ *denke*, whilst $\tilde{Y} =$ *enked*. The optimal alignment of X and Y is:

e					
k					$\bullet$
n				$\bullet$	
e			$\bullet$		
d	$\bullet$				
	t	h	i	n	k

The optimal alignment of X with $\tilde{Y}$ is

d					
e					
k					$\bullet$
n				$\bullet$	
e			$\bullet$		
	t	h	i	n	k

In this case we have that $n = 5$ and $(n+1)/2 = 3$. In each case, for both Y and $\tilde{Y}$, the optimal alignment is unique. It follows that Z and $\tilde{Z}$ are equal to the value at 3 of the paths associated with the respective optimal alignments. We find $Z = 2$ and $\tilde{Z} = 1$. Note that, in this case, $\tilde{Z} = Z - 1$.

Look at the $\bullet$'s in the two diagrams above. As mentioned, they represent pairs of aligned letters. Between the first alignment and the second most points are moved downwards by one unit. The only exception is the first $\bullet$. As we will argue later, this is a typical situation: the transformation $\sim$ has in most cases the effect of moving down the map associated with the optimal alignment except possibly at its beginning and end.

Note that (X, Y) has the same distribution as $(X, \tilde{Y})$. Hence, Z and $\tilde{Z}$ have the same distribution. It follows that

$$E[Z] = E[\tilde{Z}]. \tag{4.5}$$

This implies that the equation

$$\tilde{Z} = Z - 1$$

can not hold with a too large probability. This is one of the main ideas of this paper.

We saw that in many cases, $\tilde{Z} = Z - 1$. Hence, in many cases $\tilde{Z} > Z - 1$ does not hold. The next lemma gives a necessary condition for the inequality $\tilde{Z} > Z-1$ to hold. Roughly speaking, this condition is that there exist a close-to-optimal alignment whose path does not cross the optimal alignment on a large interval. Recall the notations $I_1 = [0, (n+1)/2)$ and $I_2 = ((n+1)/2, n]$.

Lemma 4.2 *If*

$$\tilde{Z} \neq Z - 1, \tag{4.6}$$

then there exist:

- *an optimal alignment* (π, η) *of* X *and* Y, *and*
- *another alignment* (π', η'),

such that the following two conditions are satisfied:

1. *We have that the alignment* (π', η') *is close to optimal (recall Definition 2.4 and (4.4)):*
$$|S(\pi', \eta') - L| \leq 2q^*. \tag{4.7}$$
2. *If* f *designates the path associated with* (π, η) *and* h' *designates the path associated with* (π', η'), *then* f *and* h' *do not cross on at least one of the intervals* I_1, I_2 *(i.e., the event* A *of (2.2) occurs).*

In other words, inequality (4.6) implies that there exist two close-to-optimal alignments which are macroscopically different on the scale n.

Proof. We use the following notation. Let f be a continuous strictly increasing map with convex domain in $[1, n]$ and image space $[1, n]$. We write $S[f]$ for the score obtained by aligning the texts X and Y along f. More precisely:

$$S[f] := \left(\sum_i s(X_i, Y_{f(i)}) \right) - r \cdot q, \tag{4.8}$$

where the sum is taken over those i's in the domain of f for which $f(i)$ is an integer and r is equal to the number of non-matched letters. Note that r is equal to $2n$ minus twice the number of terms in the sum (4.8).

Similarly, let $\tilde{S}[f]$ denote the score obtained by aligning X with $\tilde{Y}$ along f. More precisely,

$$\tilde{S}[f] := \left(\sum_i s(X_i, \tilde{Y}_{f(i)}) \right) - r \cdot q, \tag{4.9}$$

where again the sum is taken over those i's in the domain of f for which $f(i)$ is an integer and r is equal to the number of non-matched letters.

Let $1 \leq c < d \leq n$. We write $S_c^d(f)$ for the score obtained by aligning X with Y but calculated only on the interval $[c, d]$. Hence, $S_c^d(f)$ is equal to (4.8), where the sum is taken over all $i \in [c, d]$ such that $f(i)$ is an integer, and r is the sum of

$$\Big(\text{the number of } i\text{'s in } [c,d] \text{ such that } f(i) \text{ is not an integer}\Big)$$

and

$$\Big(\text{the number of } j\text{'s in } [f(c), f(d)] \text{ such that } f^{-1}(j) \text{ is not an integer}\Big).$$

We write $\tilde{S}_c^d(f)$ for the score obtained by aligning X with $\tilde{Y}$ but calculated only on the interval $[c, d]$.

For a map f we denote by $f_{|[a,b]}$ the restriction of f to $[a, b]$.

Let (π, η) denote an optimal alignment of X and Y. Let f be the path associated with (π, η). We assumed that (π, η) minimizes $f(n/2)$ among all optimal paths. Hence, we have that $f(n/2) = Z$.

Let g denote the path associated with $(\tilde{\pi}, \tilde{\eta})$. Let h denote the path associated with an optimal alignment of X with $\tilde{Y}$ which is minimal in the point $n/2$. Hence, $h(n/2) = \tilde{Z}$. Let h' be the path which is defined on the interval $h^{-1}([0, n-1])$ by the equation $h'(x) := h(x) + 1$.

Now comes an important point: *assume that every optimal path and every close to optimal path cut each other on I_1 and on I_2.* In other words, we assume that any optimal alignment (π, η) of X and Y and every alignment (π', η') satisfying (4.7), cross each other on I_1 and on I_2.

Then: Since f is the path of an optimal alignment of X and Y, since furthermore h is an optimal alignment of X and $\tilde{Y}$, Lemma 4.1 implies that

$$|S[f] - \tilde{S}[h]| \leq q^*. \tag{4.10}$$

By an argument similar to the one used in Lemma 4.1, one obtains

$$|S[h'] - \tilde{S}[h]| \leq q^*. \tag{4.11}$$

Combining (4.10) and (4.11), we obtain that

$$|S[f] - S[h']| \leq 2q^*. \tag{4.12}$$

Because of our assumption, f and h' cut each other on I_1 and on I_2. Hence, g and h also cut each other on I_1 and on I_2.

Denote by $a = (a_1, a_2)$ the point where g and h cut each other to the left of $n/2$. Hence, $a_1 < n/2$ and $h(a_1) = g(a_1) = a_2$. Denote by $b = (b_1, b_2)$ the point where g and h cut each other to the right of $n/2$. We have that $n/2 < b_1$ and $h(b_1) = g(b_1) = b_2$. Between a and b, we can replace the path h by g. In this way we obtain a new path (increasing continuous function) which is equal to g on $[a_1, b_1]$ and is equal to h outside $[a_1, b_1]$. This yields an admissible alignment. Since h is the path of an optimal alignment of X and $\tilde{Y}$ we find

$$\tilde{S}_{a_1}^{b_1}[g] \leq \tilde{S}_{a_1}^{b_1}[h]. \tag{4.13}$$

By definition of $(\tilde{\pi}, \tilde{\eta})$, we have that $f(a_1) = g(a_1) + 1$ and $f(b_1) = g(b_1) + 1$. We can thus replace the path f on the interval $[a_1, b_1]$ by h'. In this manner, we get an admissible path which is equal to h' on $[a_1, b_1]$ and is equal to f outside $[a_1, b_1]$. Since f is an optimal alignment of X with Y, we get

$$S_{a_1}^{b_1}[f] \geq S_{a_1}^{b_1}[h'],$$

which is equivalent to

$$\tilde{S}_{a_1}^{b_1}[g] \geq \tilde{S}_{a_1}^{b_1}[h]. \tag{4.14}$$

Together, (4.13) and (4.14) imply

$$\tilde{S}_{a_1}^{b_1}[g] = \tilde{S}_{a_1}^{b_1}[h]. \tag{4.15}$$

From (4.15) it follows that on $[a_1, b_1]$ we can replace the path h by g and still get an optimal alignment of X and $\tilde{Y}$. Thus if we take the alignment which is equal to g on $[a_1, b_1]$ and is equal to h outside $[a_1, b_1]$, this gives an optimal alignment of X with $\tilde{Y}$. Hence, by definition of $\tilde{Z}$ we obtain that at the point $n/2$ that alignment does not go below $\tilde{Z}$:

$$g(n/2) \geq \tilde{Z}.$$

From the last inequality above and the facts that $g(n/2) = f(n/2) - 1$ and $f(n/2) = Z$, we find

$$Z - 1 \geq \tilde{Z}. \tag{4.16}$$

We can now use a similar argument and, in the path f, replace the part on $[a_1, b_1]$ by $h + 1$. This yields an alignment which is equal to $h + 1$ on $[a_1, b_1]$ and is equal to f outside $[a_1, b_1]$. With the same line of argument as before, we find that this new alignment is an optimal alignment for X and Y. Hence its value at $n/2$ cannot be below Z. This gives

$$Z \leq h(n/2) + 1$$

and hence

$$Z - 1 \leq \tilde{Z}. \tag{4.17}$$

Together, (4.16) and (4.17) imply

$$Z - 1 = \tilde{Z}. \tag{4.18}$$

We have just proven that if every optimal and close to optimal alignment of X and Y cut each other on I_1 and on I_2, then (4.18) follows. This implies that when

$$Z - 1 \neq \tilde{Z}$$

holds, then there exists a close to optimal alignment and an optimal alignment which do not cut each other on either I_1 or I_2. $\square$

4.2 Proof of Theorem 2.1

Proof. Let W be the random variable

$$W := \tilde{Z} - Z + 1.$$

Since $\tilde{Z}$ and Z have same distribution, we find that

$$E[W] = 1. \tag{4.19}$$

Note that, since Z and $\tilde{Z}$ take values in $[0, n]$, we have that

$$P(W \in [-n + 1, n + 1]) = 0.$$

Let p_i be equal to the probability $p_i := P(W = i)$. We have

$$\begin{aligned} 1 &= E[W] \\ &= \sum_{i\in[-n+1,n+1]} ip_i \\ &\leq \sum_{i\in[1,n+1]} ip_i \\ &\leq \sum_{i\in[1,n+1]} (n+1)p_i \\ &\leq (n+1)\cdot P(W>0). \end{aligned}$$

The last inequality implies that

$$P(W > 0) \geq \frac{1}{n+1}. \tag{4.20}$$

Now, $W > 0$ is equivalent to $\tilde{Z} > Z - 1$. But, we saw in Lemma 4.2, that if $\tilde{Z} > Z - 1$, then A holds. Hence, $W > 0$ implies the event A, so that (4.20) implies that

$$P(A) \geq \frac{1}{n+1}.$$

□

4.3 Proof of Proposition 2.1

Proof. In order to simplify the notations, we assume that $0.25n$ is an integer.

Note that $r_1 > 0.4n$ is equivalent to

$$L_n = L(X_{0.4n}X_{0.25n+1}\ldots X_n; Y) \tag{4.21}$$

Let L^* be defined by

$$L^* := L(X_{0.4n}X_{0.25n+1}\ldots X_n; Y).$$

We have that

$$E[L^*] = E[L(X_1X_2\ldots X_{0.6n}; Y) = E[L_{0.8n}(0.25)]. \tag{4.22}$$

Furthermore,

$$L^* = 0.8n\gamma(0.25) + (E[L^*] - 0.8n\gamma(0.25)) + (L^* - E[L^*])\,,$$

and with (4.22) we find

$$L^* = 0.8n\gamma(0.25) + (E[L_{0.8n}(0.25)] - 0.8n\gamma(0.25)) + (L^* - E[L^*])\,. \quad (4.23)$$

By definition

$$\lim_{n\to\infty} E[L_{0.8n}(0.25)]/(0.8n) = \gamma(0.25). \quad (4.24)$$

The speed of convergence in the limit (4.24) is faster than $\frac{\ln n}{\sqrt{n}}$, as proved by Alexander [2]. Hence, for n large enough:

$$|E[L_{0.8n}(0.25)] - \gamma(0.25)| \leq \sqrt{n}\ln n \quad (4.25)$$

We know that the map $p \mapsto \gamma(p)$ is concave and symmetric in $p = 0$. It follows that

$$\gamma(0) \geq \gamma(0.25). \quad (4.26)$$

Let F^n be the event that

$$|L^* - E[L^*]| \leq \sqrt{n}\ln n.$$

When F^n holds, with the help of (4.23), (4.25) and (4.26), we obtain that

$$L^* \leq 0.8n\gamma(0) + 2\sqrt{n}\ln n. \quad (4.27)$$

Similarly, by the speed of convergence result [2] we obtain that for all n large enough

$$|E[L_n] - n\gamma(0)| \leq \sqrt{n}\ln n. \quad (4.28)$$

Let B^n be the event that

$$|L_n - E[L_n]| \leq \sqrt{n}\ln n.$$

When B^n occurs, we find with the help of (4.28) that

$$L_n \geq n\gamma(0) - 2\sqrt{n}\ln n. \quad (4.29)$$

We made the assumption that we are in the linear phase, i.e., that

$$\gamma(0) = \lim_{n\to\infty} E[L_n]/n > 0.$$

From $\gamma(0) > 0$, it follows that for all n large enough: the equations (4.27) and (4.29) jointly imply $L_n > L^*$. But when $L_n > L^*$ holds, then $r_1 < 0.4n$. Hence, F^n and B^n imply $r_1 < 0.4n$. Thus,

$$(F^n \cap B^n) \subset \{r_1 < 0.4n\},$$

from which it follows that

$$P(r_1 \geq 0.4n) \leq P(F^{nc}) + P(B^{nc}), \tag{4.30}$$

where F^{nc}, respectively, B^{nc} denotes the complement of F^n, respectively, B^n. By a large deviation result of Arratia and Waterman [3] we have that there exist a constant $c_{LD} > 0$ not depending on n, such that

$$\max\{P(F^{nc}), P(B^{nc})\} \leq n^{-c_{LD} \ln n}$$

for all n. The last inequality together with (4.30) implies (2.4). □

4.4 Proof of Theorem 2.2

First, the question is when is Z typically taking values in an interval of smaller order than n. We know of degenerate situations when this is not true. On the other hand, if we assume the mean curve to have non zero curvature at the origin, one can prove that Z typically takes values in an interval of size order $n^{0.75}$.

Lemma 4.3 *Assume that $p \mapsto \gamma(p)$ has continuous second derivative in a open neighborhood of $p = 0$ and $\gamma(p)'' < 0$. Then, we find that for all n large enough*

$$P(\ Z \notin [n - n^{0.75} \ln n, n + n^{0.75} \ln n]\ ) \leq 2n^{2 - c \ln n}, \tag{4.31}$$

where $c > 0$ is a constant not depending on n and (i, j).

Proof. Let $L(i, j)$ denote the optimal score obtained by aligning $X_1 X_2 \dots X_i$ with $Y_1 Y_2 \dots Y_j$:

$$L(i, j) := L(X_1 \dots X_i; Y_1 \dots Y_j).$$

Let

$$\bar{L}(i, j) := L(X_{i+1} \dots X_n; Y_{j+1} \dots Y_n).$$

Let $a > 0$ be an integer with $a \leq n/2$. We have that

$$\begin{aligned} L(n/2, n/2-a) &= 0.5(n-a)\gamma(p_a) \\ &\quad + E[L(n/2, n/2-a)] - 0.5(n-a)\gamma(p_a) \\ &\quad + L(n/2, n/2-a) - E[L(n/2, n/2-a)], \end{aligned} \tag{4.32}$$

where

$$p_a = \frac{a}{0.25(n-a)}.$$

According to our notation, we have that

$$L(n/2, n/2-a) = L_{0.5(n-a)}(p_a).$$

Using the fact that the convergence of $L_n(p)$ is faster than $\frac{\ln n}{\sqrt{n}}$, we obtain

$$\mid E[L(n/2, n/2-a)] \;-\; 0.5(n-a)\gamma(p_a) \mid \;\leq\; \sqrt{n}\ln n \tag{4.33}$$

for n large enough (the inequality (4.33) holds true because it is possible to find a uniform bound for all $a \in [0, n]$).

Let $F(i,j)$ be the event that

$$|L(i,j) - E[L(i,j)| \leq \sqrt{n}\ln n.$$

Similarly, we define $\bar{F}(i,j)$ to be the event that

$$|\bar{L}(i,j) - E[\bar{L}(i,j)| \leq \sqrt{n}\ln n.$$

Let F^n_{tot} be the event:

$$F^n_{\text{tot}} := \bigcap_{i,j\in[0,n]} (F(i,j) \cap \bar{F}(i,j)).$$

Assume that

$$a \geq n^{0.75} \cdot \ln n. \tag{4.34}$$

The map $p \mapsto \gamma(p)$ is convex and symmetric around the origin. We assumed that it has strictly positive curvature at $p = 0$. Hence, $\gamma(0)' = 0$ and there exists an open neighbourhood $\mathcal{O}$ such that $0 \in \mathcal{O}$ and for all $p \in \mathcal{O}$ we have

$$\gamma(0) \geq \gamma(p) + \kappa \cdot p^2, \tag{4.35}$$

where $\kappa > 0$ is a constant not depending on $p \in \mathcal{O}$. Let p_0 designate the value of p_a when a is taken equal to $n^{0.75} \cdot \ln n$. Assuming that (4.34) holds, and that n is large enough so that $p_0 \in \mathcal{O}$, we find that

$$\gamma(0) \geq \gamma(p_0) + \kappa \cdot {p_0}^2. \tag{4.36}$$

Since the map $p \mapsto \gamma(p)$ is concave and symmetric around the origin, it follows that on $[0,1]$ it must be decreasing. Hence, when (4.34) holds, we have that $\gamma(p_0) \geq \gamma(p_a)$, so that inequality (4.36) becomes

$$\gamma(0) \geq \gamma(p_a) + \kappa \cdot {p_0}^2. \tag{4.37}$$

Hence, when the event F^n_{tot} holds, we find using (4.32), (4.33) and (4.37) that

$$L(n/2, n/2 - a) \leq 0.5(n-a)\gamma(0) - 0.5(n-a)\kappa \cdot {p_0}^2 + 2\sqrt{n}\ln n. \tag{4.38}$$

Let

$$\bar{p}_a = \frac{-a}{0.25(n+a)}.$$

Let $\bar{p}_0$ denote the value of $\bar{p}_a$ when a is equal to $n^{0.75}\ln n$. According to our notation, we have that

$$E[\bar{L}(n/2, n/2 - a)] = E[L_{0.5(n+a)}(\bar{p}_a)].$$

Using the same rate of convergence as we did in (4.33), we obtain the inequality

$$|E[\bar{L}(n/2, n/2 - a)] - 0.5(n+a)\gamma(\bar{p}_a)| \leq \sqrt{n}\ln n. \tag{4.39}$$

With a similar argument as was used for (4.38), we obtain that if the event F^n_{tot} holds together with (4.34), then

$$\bar{L}(n/2, n/2 + a) \leq 0.5(n+a)\gamma(0) - 0.5(n+a)\kappa(*)^2 + 2\sqrt{n}\ln n. \tag{4.40}$$

The inequalities (4.38) and (4.40) together imply:

$$\begin{aligned} & L(n/2, n/2 - a) + \bar{L}(n/2, n/2 + a) \\ \leq \quad & n\gamma(0) - 0.5\kappa((n-a)(p_0)^2 + (n+a)(*)^2 + 4\sqrt{n}\ln n, \end{aligned} \tag{4.41}$$

and hence

$$L(n/2, n/2 - a) + \bar{L}(n/2, n/2 + a) \leq n\gamma(0) - 0.5\kappa(n+a)(*)^2 + 4\sqrt{n}\ln n. \tag{4.42}$$

Note that

$$(*)^2 = (\ln n)^2 \cdot n^{-0.5} \frac{1}{0.625(1+n^{-0.25}\ln n)^2}. \tag{4.43}$$

Also, when n is large enough, we get

$$1 \leq \frac{1}{0.625(1+n^{-0.25}\ln n)^2}. \tag{4.44}$$

Since $a > 0$, we have that $n+a \geq n$. Combining this with the formulas (4.42), (4.43) and (4.44), we obtain

$$L(n/2, n/2-a)+\bar{L}(n/2, n/2+a) \leq n\gamma(0) - 0.5\kappa(\ln n)^2\sqrt{n}+4\sqrt{n}\ln n. \tag{4.45}$$

Similarly, when F^n_{tot} holds, we find that

$$L_n \geq n\gamma(0) - 2\sqrt{n}\ln n. \tag{4.46}$$

Together, (4.45) and (4.46) imply

$$L_n - \big(L(n/2, n/2-a)+\bar{L}(n/2, n/2+a)\big) \geq 0.5\kappa(\ln n)^2\sqrt{n} - 6\sqrt{n}\ln n. \tag{4.47}$$

For n large enough, we have

$$0 < 0.5\kappa(\ln n)^2\sqrt{n} - 6\sqrt{n}\ln n,$$

so that

$$L_n > L(n/2, n/2-a) + \bar{L}(n/2, n/2+a). \tag{4.48}$$

But if (4.48) holds, then there is no path f of an optimal alignment such that $f(n/2) \in [n-a, n-a+1)$. Hence, when (4.47) holds for all $a \geq n^{0.75}\ln n$, then there is no path f of an optimal alignment such that $f(n/2) \in [0, n-n^{0.75}\ln n]$ and hence $Z \geq n - n^{0.75}\ln n$. Summarizing, we have proved that for all n large enough

$$F^n_{\text{tot}} \subset \{Z \geq n - n^{0.75}\ln n\}.$$

By symmetry we obtain also

$$F^n_{\text{tot}} \subset \{Z \leq n + n^{0.75}\ln n\}.$$

The two last inclusions finally imply that

$$F^n_{\text{tot}} \subset \{Z \in [n - n^{0.75}\ln n, n + n^{0.75}\ln n]\},$$

so

$$P(Z \notin [n - n^{0.75}\ln n, n + n^{0.75}\ln n]) \leq P(F^{nc}_{\text{tot}}). \tag{4.49}$$

Note also that

$$P(F^{nc}_{\text{tot}}) \leq \sum_{i,j\in[0,n]} (\ P(F^c(i,j)) + P(\bar{F}^c(i,j))\ ). \tag{4.50}$$

By the large deviation result of Arratia and Waterman [3], we find that

$$\max\{P(F^c(i,j)), P(\bar{F}^c(i,j))\} \leq n^{-c\ln n}, \tag{4.51}$$

where $c > 0$ is a constant not depending on n. Using (4.51) in (4.50), we obtain

$$P(F^{nc}_{\text{tot}}) \leq 2n^{2-c\ln n}. \tag{4.52}$$

Inequalities (4.49) and (4.52) together imply (4.31). □

Now, we are able to prove Theorem 2.2.

Proof of Theorem 2.2. Let W be the random variable

$$W := \tilde{Z} - Z + 1.$$

Since $\tilde{Z}$ and Z have the same distribution, we find that

$$E[W] = 1. \tag{4.53}$$

Note that W takes values in $[-n+1, n+1]$ and that due to (4.31):

$$P(W \in [-2n^{0.75}\ln n, 2n^{0.75}\ln n]) \geq 1 - 4n^{2-c\ln n}. \tag{4.54}$$

Let I denote the interval $I := [1, 2n^{0.75}\ln n]$. Let p_i be equal to the probability $p_i := P(W = i)$. We have

$$\begin{aligned}
1 &= E[W] \\
&= \sum_{i\in[-n+1,n+1]} ip_i \\
&\leq \sum_{i>0} ip_i \\
&\leq \sum_{i\in I}(2n^{0.75}\ln n)p_i + (n+1)\cdot P(W > 2n^{0.75}\ln n).
\end{aligned}$$

From the last inequality and with the help of (4.54), we find:

$$1 \leq P(W > 0)(2n^{0.75} \ln n) + 4(n+1)n^{2-c\ln n}. \tag{4.55}$$

Note that

$$\lim_{n \to \infty} 4(n+1)n^{2-c\ln n} = 0,$$

so that for n large enough we have

$$4(n+1)n^{2-c\ln n} \leq \frac{1}{2}. \tag{4.56}$$

Using (4.55) and (4.56), we find that

$$1 \leq P(W > 0)(2n^{0.75} \ln n) + \frac{1}{2}, \tag{4.57}$$

from which it follows that

$$\frac{1}{4n^{0.75} \ln n} \leq P(W > 0). \tag{4.58}$$

But $W > 0$ is equivalent to $\tilde{Z} > Z - 1$. If $\tilde{Z} > Z - 1$, then the event A occurs, as was proven in Lemma 4.2. Hence

$$\{\tilde{Z} > Z - 1\} \subset A,$$

so that

$$P(A) \geq P(\tilde{Z} > Z - 1) = P(W > 0) \geq \frac{1}{4n^{0.75} \ln n}.$$

□

4.5 Proof of Theorem 3.1

We saw that Z and $\tilde{Z}$ have the same distribution, but yet in many cases $\tilde{Z} = Z - 1$. There is another important consequence which follows from this seeming contradiction. For two variables V and W with the same distribution, the only possibility that the equation $V = W - 1$ holds with high probability is when the fluctuation of V is large.

Let us give two examples to illustrate this. First let the distribution $\mathcal{L}(V)$ be mono-atomic. Then, if $\mathcal{L}(V) = \mathcal{L}(W)$, we find that

$$P(V = W - 1) = 0.$$

On the other hand, if V and W have both uniform (discrete or continuous) distributions in the interval $[0, n]$, then it is possible to couple V and W in such a way that

$$P(V = W - 1) \geq 1 - \frac{1}{n}.$$

The next lemma show that a high probability for $V = W - 1$ to hold, implies a large fluctuation of V when the two variables have same distribution.

Lemma 4.4 *Assume that V and W are two random variables with identical distribution such that*

$$P(V = W - 1) \geq 1 - Cn^{-\alpha}, \tag{4.59}$$

and such that

$$P(V \in [0, n]) = P(W \in [0, n]) = 1.$$

Then, the interquartile distance q_w satisfies

$$q_w > (2(Cn^{-\alpha} + n^{-1}))^{-1} - 4. \tag{4.60}$$

Proof. We begin by recalling a simple fact about discrete probability distributions. For two distributions μ and ν on a finite (or countable) set B, define the total variation distance by

$$\|\mu - \nu\| = \frac{1}{2} \sum_{z \in B} |\mu(z) - \nu(z)|.$$

It is well-known that

$$\|\mu - \nu\| = \inf P(U_1 \neq U_2), \tag{4.61}$$

where the infimum is taken over all possible couplings of variables U_1 (having distribution μ) and U_2 (with distribution ν). Now, denote $p_i = P(V \in (i-1, i])$, for $i = 1, \ldots, n$, and let $p_0 = P(V = 0)$. From (4.59) and (4.61) it is straightforward to obtain that

$$\sum_{i=1}^{n} |p_i - p_{i-1}| \leq Cn^{-\alpha}. \tag{4.62}$$

Let $k_0 = \arg\min p_k$, $k_1 = \arg\max p_k$; without restriction of generality we now suppose that $k_0 < k_1$. Clearly, $p_{k_0} < n^{-1}$, so we obtain from (4.62) that

$$p_{k_1} = p_{k_1} - p_{k_0} + p_{k_0} \le p_{k_0} + \sum_{i=k_0+1}^{k_1} |p_i - p_{i-1}| \le Cn^{-\alpha} + n^{-1},$$

which means that

$$p_k \le Cn^{-\alpha} + n^{-1} \text{ for all } k = 0, \dots, n. \tag{4.63}$$

Now, let Q^1 and Q^3 be the first and the third quartiles of the random variable V (and W). Suppose that $Q^1 \in (m_0 - 1, m_0]$, $Q^3 \in (m_1, m_1 + 1]$ for some $m_0 \le m_1$. By (4.63), we have that $P(V \in [Q^1, Q^3]) \ge \frac{1}{2} - 2(Cn^{-\alpha} + n^{-1})$, so

$$P(V \in (m_0, m_1]) \ge \frac{1}{2} - 4(Cn^{-\alpha} + n^{-1}). \tag{4.64}$$

On the other hand, again by (4.63),

$$P(V \in (m_0, m_1]) \le (m_1 - m_0)(Cn^{-\alpha} + n^{-1}). \tag{4.65}$$

Thus, by (4.64) and (4.65),

$$q_w \ge m_1 - m_0 \ge \frac{1}{2(Cn^{-\alpha} + n^{-1})} - 4.$$

□

Now, we are ready to finish the proof of Theorem 3.1.

Proof of Theorem 3.1. We apply Lemma 4.4. For this we take the variables Z and $\tilde{Z}$ as the variables V and W of Lemma 4.4. Note that Z and $\tilde{Z}$ have the same distribution. Furthermore, we showed in Lemma 4.2 that, if A does not hold, then $\tilde{Z} = Z - 1$. Hence, inequality (3.1) implies that

$$P(\tilde{Z} = Z - 1) \ge 1 - Cn^{-\alpha}.$$

The last inequality above is the condition (4.59) of Lemma 4.4. Thus, all the conditions of Lemma 4.4 hold, and we obtain (3.2). □

References

[1] David Aldous and Persi Diaconis. Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem. *Bull. Amer. Math. Soc. (N.S.)*, 36(4):413–432, 1999.

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

[3] Richard Arratia and Michael S. Waterman. A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, 4(1):200–225, 1994.

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

[5] Václáv Chvatal and David Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probability*, 12:306–315, 1975.

[6] Peter Clote and Rolf Backofen. *Computational molecular biology*. Wiley Series in Mathematical and Computational Biology. John Wiley & Sons Ltd., Chichester, 2000. An introduction.

[7] Raphael Hauser and Heinrich Matzinger. Local uniqueness of alignments with af fixed proportion of gaps. Submitted, 2005.

[8] C. Douglas Howard. Models of first-passage percolation. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 125–173. Springer, Berlin, 2004.

[9] C.D. Howard and C.M. Newman. Geodesics and spanning trees for euclidian first-passage percolation. *Ann. Probab.*, 29:577–623, 2001.

[10] Kurt Johansson. Transversal fluctuations for increasing subsequences on the plane. *Probab. Theory Related Fields*, 116(4):445–456, 2000.

[11] Jyri Lember and Heinrich Matzinger. Variance of the LCS for 0 and 1 with different frequencies. Submitted, 2005.

[12] C.M. Newman and M.S.T. Piza. Divergence of shape fluctuations in two dimensions. *Ann. Probab.*, 23:977–1005, 1995.

[13] Pavel A. Pevzner. *Computational molecular biology.* Computational Molecular Biology. MIT Press, Cambridge, MA, 2000. An algorithmic approach, A Bradford Book.

[14] Michael J. Steele. An Efron-Stein inequality for non-symmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

[15] Michael S. Waterman. Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 344:383–390, 1994.