Proportion of Gaps and Fluctuations of the Optimal Score in Random Sequence Comparison

Jüri Lember, Heinrich Matzinger, and Felipe Torres

Abstract We study the asymptotic properties of optimal alignments when aligning two independent i.i.d. sequences over finite alphabet. Such kind of alignment is an important tool in many fields of applications including computational molecular biology. We are particularly interested in the (asymptotic) proportion of gaps of the optimal alignment. We show that when the limit of the average optimal score per letter (rescaled score) is considered as a function of the gap penalty, then given a gap penalty, the proportion of the gaps converges to the derivative of the limit score at that particular penalty. Such an approach, where the gap penalty is allowed to vary, has not been explored before. As an application, we solve the long open problem of the fluctuation of the optimal alignment in the case when the gap penalty is sufficiently large. In particular, we prove that for all scoring functions without a certain symmetry, as long as the gap penalty is large enough, the fluctuations of the optimal alignment score are of order square root of the length of the strings. This order was conjectured by Waterman [Phil. Trans. R. Soc. Lond. B 344(1):383–390, 1994] but disproves the conjecture of Chvatal and Sankoff in [J. Appl. Probab. 12:306–315, 1975].

Keywords Fluctuations • Longest common sequence • McDiarmid's inequality • Random sequence comparison • Waterman conjecture

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

H. Matzinger (✉)
Georgia Tech, School of Mathematics, Atlanta, GA 30332-0160, USA
e-mail: matzing@math.gatech.edu

F. Torres
Münster University, Institute for Mathematical Statistics, Einsteinstraße 62, 48149 - Münster, Germany
e-mail: ftorrestapia@math.uni-muenster.de

P. Eichelsbacher et al. (eds.), *Limit Theorems in Probability, Statistics and Number Theory*, Springer Proceedings in Mathematics & Statistics 42,
DOI 10.1007/978-3-642-36068-8_10, © Springer-Verlag Berlin Heidelberg 2013

AMS. 60K35, 41A25, 60C05

1 Introduction

1.1 Preliminaries

Throughout this paper $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are two independent sequences of i.i.d. random variables drawn from a finite alphabet $\mathbb{A}$ and having the same distribution. Since we mostly study the finite strings of length n, let $X = (X_1, X_2, \ldots X_n)$ and let $Y = (Y_1, Y_2, \ldots Y_n)$ be the corresponding n-dimensional random vectors. We shall usually refer to X and Y as random sequences.

The problem of measuring the similarity of X and Y is central in many areas of applications including computational molecular biology [4, 7, 18, 20, 24] and computational linguistics [13, 16, 17, 25]. In this paper we adopt the same notation as in [11], namely we consider a general scoring scheme, where $S : \mathbb{A} \times \mathbb{A} \to \mathbb{R}^+$ is a *pairwise scoring function* that assigns a score to each couple of letters from $\mathbb{A}$. We assume S to be symmetric, non-constant and we denote by F and E the largest and the second largest possible score, respectively. Formally (recall that S is symmetric and non-constant)

$$F := \max_{(a,b)\in\mathbb{A}\times\mathbb{A}} S(a,b), \quad E := \max_{(a,b):S(a,b)\neq F} S(a,b).$$

An *alignment* is a pair (π, μ) where $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ are two increasing sequences of natural numbers, i.e. $1 \leq \pi_1 < \pi_2 < \ldots < \pi_k \leq n$ and $1 \leq \mu_1 < \mu_2 < \ldots < \mu_k \leq n$. The integer k is the number of aligned letters, $n - k$ is the number of *gaps* in the alignment and the number

$$q(\pi, \mu) := \frac{n-k}{n} \in [0, 1]$$

is the *proportion of gaps of the alignment* (π, μ). The *average score of aligned letters* is defined by

$$t(\pi, \mu) := \frac{1}{k} \sum_{i=1}^{k} S(X_{\pi_i}, Y_{\mu_i}).$$

Note that our definition of gap slightly differs from the one that is commonly used in the sequence alignment literature, where a gap consists of maximal number of consecutive *indels* (insertion and deletion) in one side. Our gap actually corresponds to a pair of indels, one in X-side and another in Y-side. Since we consider the sequences of equal length, to every indel in X-side corresponds an indel in Y-side, so considering them pairwise is justified. In other words, the number of gaps in our sense is the number of indels in one sequence. We also consider a *gap price* δ.

Given the pairwise scoring function S and the gap price δ, the score of the alignment (π, μ) when aligning X and Y is defined by

$$U^{\delta}_{(\pi,\mu)}(X, Y) := \sum_{i=1}^{k} S(X_{\pi_i}, Y_{\mu_i}) + \delta(n - k)$$

which can be written down as the convex combination

$$U^{\delta}_{(\pi,\mu)}(X, Y) = n\,(t(\pi, \mu)(1 - q(\pi, \mu)) + \delta q(\pi, \mu))\,. \tag{1}$$

In our general scoring scheme δ can also be positive, although usually $\delta \leq 0$ penalizing the mismatch. For negative δ, the quantity $-\delta$ is usually called the *gap penalty*. We naturally assume $\delta \leq F$. The optimal alignment score of X and Y is defined to be

$$L_n(\delta) := \max_{(\pi,\mu)} U^{\delta}_{(\pi,\mu)}(X, Y),$$

where the maximum above is taken over all possible alignments. The alignments achieving the maximum are called *optimal*. For every $\delta \in \mathbb{R}$, let us denote

$$B_n(\delta) := \frac{L_n(\delta)}{n}. \tag{2}$$

Note that to every alignment (π, μ) corresponds an unique pair $(t(\pi, \mu), q(\pi, \mu))$, but different alignments can have the same $t(\pi, \mu)$ and $q(\pi, \mu)$, thus from (1) we get that

$$B_n(\delta) = \max_{(\pi,\mu)} \big(t(\pi, \mu)(1 - q(\pi, \mu)) + \delta q(\pi, \mu)\big) = \max_{(t,q)} \big(t(1 - q) + \delta q\big), \tag{3}$$

where in the right hand side the maximum is taken over all possible pairs (t, q) corresponding to an alignment of X and Y. In the following, we identify alignments with pairs (t, q), so a pair (t, q) always corresponds to an alignment (π, μ) of X and Y. Let $\mathcal{O}_n(\delta)$ denote the set of optimal pairs, i.e. $(t, q) \in \mathcal{O}_n(\delta)$ if and only if $t(1 - q) + \delta q = B_n(\delta)$. Note that the set $\mathcal{O}_n(\delta)$ is not necessarily a singleton. Let us denote

$$\underline{q}_n(\delta) := \min\{q : (t, q) \in \mathcal{O}_n(\delta)\}$$
$$\overline{q}_n(\delta) := \max\{q : (t, q) \in \mathcal{O}_n(\delta)\}.$$

By Kingman's subadditive ergodic theorem, for any δ there exists a constant $b(\delta)$ so that

$$B_n(\delta) \to b(\delta), \quad \text{a.s..} \tag{4}$$

1.2 The Organization of the Paper and Main Results

In this paper, we use a novel approach regarding the quantities of interest like the proportion of gaps, the rescaled score B_n, etc., as functions of δ. In Sect. 2, we derive some elementary but important properties of $B_n(\delta)$ and we explore the relation between the proportion of gaps of any optimal alignment and the derivatives of $B_n(\delta)$. In particular, we show (Claim 2.2) that for any n and δ,

$$B_n'(\delta_+) = \overline{q}_n(\delta), \quad B_n'(\delta_-) = \underline{q}_n(\delta). \tag{5}$$

In a sense these equalities, which almost trivially follow from the elementary calculus, are the core for the rest of the analysis.

In Sect. 3, we show that when the limit score function b is differentiable at δ, then a.s. $\overline{q}_n(\delta)$ and $\underline{q}_n(\delta)$ both converge to $b'(\delta)$ (by using expression (5)) so that $b'(\delta)$ can be interpreted as the *asymptotic proportion of gaps*. The section ends with an example showing that if b is not differentiable at δ, then the extremal proportions $\overline{q}_n(\delta)$ and $\underline{q}_n(\delta)$ can still a.s. converge to the corresponding one-side derivatives, namely $\underline{q}_n(\delta) \to b'(\delta_+)$ a.s. and $\overline{q}_n(\delta) \to b'(\delta_-)$ a.s.

Section 4 deals with large deviations bounds for the (optimal) proportion of gaps. The main result of this section is Theorem 4.1, which states that for every $\varepsilon > 0$ there exists a $c > 0$ such that for every n big enough the following large deviation inequality holds

$$P\big(b'(\delta_-) - \varepsilon \le \underline{q}_n(\delta) \le \overline{q}_n(\delta) \le b'(\delta_+) + \varepsilon\big) \ge 1 - 4\exp[-c(\varepsilon)n].$$

Combining this last inequality with the result on the speed of convergence proven in [11], we obtain the confidence intervals for the in general unknown quantities $b'(\delta_+)$ and $b'(\delta_-)$ in terms of $B_n(\delta)$ (the inequalities (27) and (27), respectively).

In Sect. 5 we obtain results on the fluctuations of the score of optimal alignments, namely we show that under some asymmetry assumption on the score function there exists a $c > 0$ so that for n large enough $\mathrm{Var}[L_n(\delta)] \ge cn$ provided that the gap penalty $-\delta$ is big enough (Theorem 5.2). This result implies that $\mathrm{Var}[L_n(\delta)] = \Theta(n)$, because as shown by Steele in [21], there exists another constant C such that $\mathrm{Var}[L_n(\delta)] \le Cn$. Our proof is based on the existence of the asymptotic proportion of gaps and, therefore, differs from the previous proofs in the literature.

Finally, Sect. 6 is devoted to the problem of determining the sufficiently large gap penalty δ_o so that the conditions of Theorem 5.2 are fulfilled. We show that when knowing the asymptotic upper bound $\overline{t}(\delta)$ of the average score of aligned letters, then δ_o can be easily found (Claim 6.1). Theorem 6.1 shows how the upper bound $\overline{t}(\delta)$ can be found. The proof of Theorem 6.1 uses similar ideas that the ones used in the proof of Theorem 5.2. The section ends with a practical example (Sect. 6.2).

It is important to notice that we could not find in the literature complete results on the fluctuations of the score in random sequences comparison. Though, a particular model for comparison of random sequences has had an interesting development in the past 4 decades: the longest common subsequence problem (abbreviated by *LCS problem*). In our setting, the LCS problem corresponds to choose $S(x, y) = 1$ if $x = y$ and $S(x, y) = -\infty$ if $x \neq y$. Already in 1975, Chvatal and Sankoff [5] conjectured that the fluctuations of the length of the LCS is of order $o(n^{2/3})$. But in 1994, Waterman [23] conjectured that those fluctuations should be of order $\Theta(n)$. This last order had been proven by Matzinger et al. [2, 8–10] in a series of relatively recent papers treating extreme models with low entropy. In 2009, the Ph.D. thesis of Torres [14, 15, 22] brought an improvement, proving that the length of the LCS of sequences built by i.i.d. blocks has also fluctuations of order $\Theta(n)$, turning it to be the first time Waterman's conjecture was proven for a model with relatively high entropy. Unfortunately, the block-model of Torres does not have enough ergodicity as to extend the result to the still open original Waterman's conjecture. We believe that the results on the fluctuations of the score of optimal alignments showed in the present paper are an important source of new evidence that Waterman's conjecture might be true, even in more general models of sequence comparison than the LCS problem, provided the score function does not have a certain symmetry.

Note that the LCS problem can be reformulated as a last passage percolation problem with correlated weights [1]. For several last passage percolation models, the order of the fluctuations has been proven to be power $2/3$ of the order of the expectation. But as the previous models and simulations have showed (for simulations, see e.g. [3]), this order seem to be different as the order of the fluctuations of the score in optimal alignments.

2 Basic Properties of B_n

We start by deriving some elementary properties of the function $\delta \mapsto B_n(\delta)$:

Claim 2.1. *For every X and Y, the function $\delta \mapsto B_n(\delta)$ is non-decreasing, piecewise linear and convex.*

Proof. The non-decreasing and piecewise linear properties follow from the definition. For the convexity, with $\lambda \in (0, 1)$ let $\delta = \lambda\delta_1 + (1-\lambda)\delta_2$ and $(t, q) \in \mathcal{O}_n(\delta)$. Note that the pair (t, q) is not necessarily optimal for the proportions δ_1 and δ_2, so that from (3) it follows

$$\begin{aligned} B_n(\lambda\delta_1 + (1-\lambda)\delta_2) &= t(1-q) + (\lambda\delta_1 + (1-\lambda)\delta_2)q \\ &= \lambda\big(t(1-q) + \delta_1 q\big) + (1-\lambda)\big(t(1-q) + \delta_2 q\big) \\ &\leq \lambda B_n(\delta_1) + (1-\lambda)B_n(\delta_2). \end{aligned}$$

□

Claim 2.2. *For any $\delta \in \mathbb{R}$ we have*

$$B_n'(\delta_-) := \lim_{s \searrow 0} \frac{B_n(\delta - s) - B_n(\delta)}{s} = \underline{q}_n(\delta)$$

$$B_n'(\delta_+) := \lim_{s \searrow 0} \frac{B_n(\delta + s) - B_n(\delta)}{s} = \overline{q}_n(\delta).$$

Thus, $\mathcal{O}_n(\delta)$ is singleton if and only if $B_n(\delta)$ is differentiable at δ.

Proof. Fix $\delta \in \mathbb{R}$ and $s > 0$. Let $(t, q) \in \mathcal{O}_n(\delta)$, thus

$$B_n(\delta + s) \geq t(1 - q) + q(\delta + s) = B_n(\delta) + qs$$

$$B_n(\delta - s) \geq t(1 - q) + q(\delta - s) = B_n(\delta) - qs.$$

Hence,

$$\frac{B_n(\delta) - B_n(\delta - s)}{s} \leq q \leq \frac{B_n(\delta + s) - B_n(\delta)}{s}.$$

The inequalities above hold for any optimal (t, q) and for any s, so letting $s \searrow 0$ we have

$$B_n'(\delta_-) \leq \underline{q}_n(\delta) \leq \overline{q}_n(\delta) \leq B_n'(\delta_+). \tag{6}$$

Thus, if B_n is differentiable at δ, then $\underline{q}_n(\delta) = \overline{q}_n(\delta)$ meaning that $\mathcal{O}_n(\delta)$ is a singleton, say $\mathcal{O}_n(\delta) = (t_n(\delta), q_n(\delta))$. To prove that $B_n'(\delta_+) = \overline{q}_n(\delta)$, it is enough to show that there exists a pair $(t, q) \in \mathcal{O}_n(\delta)$ such that $B_n'(\delta_+) = q$. Indeed, since B_n is piecewise linear, for every $\varepsilon > 0$ small enough B_n is differentiable at $\delta + \varepsilon$ and the derivative equals to $B_n'(\delta_+)$. Hence, for every $\varepsilon > 0$ small enough $q := q_n(\delta + \varepsilon) = B_n'(\delta_+)$. Let $t := t_n(\delta + \varepsilon)$. Thus, for every $\varepsilon > 0$ small enough there exists a pair $(t, q) \in \mathcal{O}_n(\delta + \varepsilon)$ such that $q = B_n'(\delta_+)$. This means $t(1 - q) + q\delta + q\epsilon = B_n(\delta + \varepsilon)$. Since B_n is continuous, we see that $\lim_{\varepsilon \to 0+} B_n(\delta + \varepsilon) = B_n(\delta) = t(1 - q) + q\delta$, i.e. $(t, q) \in \mathcal{O}_n(\delta)$. With similar arguments one can show that $\underline{q}_n(\delta) = B_n'(\delta_-)$. □

Function $B_n(\delta)$ for large δ. With fairly simple analysis, it is possible to determine $B_n(\delta)$ for large δ. Recall the definition of F. Clearly, when $\delta > F$, the optimal alignment only consists of gaps, namely $\delta \geq F \Rightarrow B_n(\delta) = \delta$. If we decrease the value of δ, say $\delta \in (E, F)$, the optimal alignment tries to align as many pairs of letters which score F as possible, thus minimizing the number of gaps. Formally, such optimal alignment can be obtained by defining a new score function

$$S_1(a, b) := \begin{cases} F & \text{if } S(a, b) = F \\ 0 & \text{if } S(a, b) < F \end{cases}$$

Let $B_n^1(\delta)$ be the corresponding expression (2) for the score function S_1. If (t_n^1, q_n^1) is such that $B_n^1(0) = t_n^1(1 - q_n^1) + 0 \cdot q_n^1$, then $t_n^1 = F$ and $1 - q_n^1$ is the maximal proportion of pairs that score F. Thus (t_n^1, q_n^1) is unique and, therefore, B_n^1 is differentiable at 0. For the original B_n, if $\delta \in [E, F]$, then we have

$$B_n(\delta) = F(1 - q_n^1) + \delta q_n^1 = B_n^1(0) + \delta q_n^1,$$

from where we have

$$B_n(F) = F = B_n^1(0) + F q_n^1. \tag{7}$$

If δ is slightly smaller than E, then the candidate alignments to be optimal alignments are obtained by aligning only those pair of letters that score E or F; amongst such alignments an optimal one will be the one having minimal number of gaps. Formally, we consider the score function

$$S_2(a,b) = \begin{cases} F \text{ if } S(a,b) = F \\ E \text{ if } S(a,b) = E \\ 0 \quad \text{otherwise} \end{cases}$$

Let $B_n^2(\delta)$ be the corresponding expression (2) for the score function S_2. Let (t_n^2, q_n^2) be such that $B_n^2(0) = t_n^2(1 - q_n^2) + 0 \cdot q_n^2$ with the additional property that $q_n^2 \leq q$ for any other optimal pair (t, q) for $B_n^2(0)$. By Claim 2.2, $q_n^2 = (B_n^2)'(0_-)$. Hence, if δ is slightly smaller than E, then $B_n(\delta) = t_n^2(1 - q_n^2) + \delta q_n^2 = B_n^2(0) + \delta q_n^2$.

Hence, we can write down

$$B_n(\delta) = \begin{cases} \delta & \text{if } \delta \geq F \\ F(1 - q_n^1) + \delta q_n^1 & \text{if } E \leq \delta \leq F \\ t_n^2(1 - q_n^2) + \delta q_n^2 & \text{if } E - \varepsilon \leq \delta \leq E \end{cases} \tag{8}$$

for a small $\varepsilon > 0$ which depends on X, Y. Indeed, if δ is much smaller than E but still above the value of the next score, then the optimal alignment (t, q) might align less F-valued letters for in order to achieve less gaps. In other words, the optimal alignment (t, q) can be such that $t(1 - q) < B_n^2(0)$. But it is not so for $\delta = E$ and due to the piecewise linearity of B_n, the $\varepsilon > 0$ described above exists.

By Claim 2.2, for any n we have that

$$\underline{q}_n(F) = q_n^1, \quad \overline{q}_n(F) = 1, \qquad \underline{q}_n(E) = q_n^2, \quad \overline{q}_n(E) = q_n^1. \tag{9}$$

Finally, note that by taking $\delta = E$, we obtain that

$$B_n(E) = B_n^2(0) + E q_n^2 \tag{10}$$

and

$$B_n^1(0) - B_n^2(0) = E(q_n^2 - q_n^1). \tag{11}$$

3 The Asymptotic Proportion of Gaps

¿From the convergence in (4), we see that the limit function $b(\cdot)$ inherits properties from $B_n(\cdot)$. More precisely, the (random) function $B_n(\cdot)$ is convex and non-decreasing, so the same holds for $b(\cdot)$. Moreover, due to the monotonicity, the convergence in (4) is uniform on δ, i.e.

$$\sup_{\delta \in \mathbb{R}} |B_n(\delta) - b(\delta)| \to 0 \quad \text{a.s. as } n \to \infty. \tag{12}$$

But we need to be a bit more careful in deriving properties of the derivative b' from B'_n, since the uniform convergence of convex functions implies the convergence of one side derivatives at x only when the limit function is differentiable at x. Indeed, let f_n and f be convex functions that converge pointwise, i.e. $f_n(x) \to f(x)$ as $n \to \infty$, for every x. Then, in general [19] it holds

$$f'(x_-) := \lim_{s \searrow 0} \lim_{n \to \infty} \frac{f_n(x-s) - f_n(x)}{s} \leq \liminf_{n \to \infty} \lim_{s \searrow 0} \frac{f_n(x-s) - f_n(x)}{s}$$
$$\leq \limsup_{n \to \infty} \lim_{s \searrow 0} \frac{f_n(x+s) - f_n(x)}{s} \leq \lim_{s \searrow 0} \lim_{n \to \infty} \frac{f_n(x+s) - f_n(x)}{s} = f'(x_+),$$

and these inequalities can be strict. In our case these inequalities are

$$b'(\delta_-) \leq \liminf_n \underline{q}_n(\delta) \leq \limsup_n \overline{q}_n(\delta) \leq b'(\delta_+), \quad \text{a.s..} \tag{13}$$

Lemma 3.1. *Let b be differentiable at δ. Then*

$$\underline{q}_n(\delta) \to b'(\delta) \quad \textit{and} \quad \overline{q}_n(\delta) \to b'(\delta) \quad \textit{a.s. as } n \to \infty. \tag{14}$$

Remark 3.1. An interesting question is the following: If b is not differentiable at δ, there exist $\underline{q}, \overline{q} \in (0,1)$ with $\underline{q} \geq b'(\delta_-)$ and $\overline{q} \leq b'(\delta_+)$ such that

$$\underline{q}_n(\delta) \to \underline{q} \quad \text{and} \quad \overline{q}_n(\delta) \to \overline{q} \quad \text{a.s. as } n \to \infty? \tag{15}$$

Numerical simulations of the difference $\overline{q}_n - \underline{q}_n$ as $n \to \infty$ do not conclusively show convergence nor boundedness, so perhaps such $\underline{q}, \overline{q}$ do not exist.

Thus, if b is differentiable at δ, the random proportion of gaps of optimal alignments tends to an unique number $q(\delta) := b'(\delta)$ that we can interpreted as the **asymptotic proportion of gaps** at δ. If the function b is not differentiable at δ, then it is not known whether the maximal or minimal proportion of gaps converge. However, as we shall now see this might be the case.

Asymptotic proportion of gaps for large δ. In general, it seems hard to determine where b is not differentiable and the asymptotic proportion of gaps does not exist.

However, based on the elementary properties of B_n and b, we can say something about the differentiability of b for large δ's. Recall B_n^1 and B_n^2, let b^1 and b^2 be the corresponding limits. The following claim shows that the proportions $\overline{q}_n(\delta)$ and $\underline{q}_n(\delta)$ might converge even if b is not differentiable at δ.

Claim 3.1. *The following convergences hold as $n \to \infty$:*

1. $\overline{q}_n(F) \to 1 = b'(F_+)$, *a.s.*
2. $\underline{q}_n(F) \to \frac{F-b^1(0)}{F} = b'(F_-)$, *a.s.*
3. $\overline{q}_n(E) \to \frac{F-b^1(0)}{F} = b'(E_+)$, *a.s.*
4. $\underline{q}_n(E) \to \frac{b(E)-b^2(0)}{E} = \frac{b^1(0)-b^2(0)}{E} + \frac{F-b^1(0)}{F} \geq b'(E_-)$, *a.s..*

If $b^2(0) > b^1(0) > 0$, then b is not differentiable at $\delta = E$ and $\delta = F$.

Proof. We are going to use the fact that the convergence (12) does not depend on the score function, so there exist constants $b^1(0)$ and $b^2(0)$ such that $B_n^1(0) \to b^1(0)$ and $B_n^2(0) \to b^2(0)$ as $n \to \infty$, a.s.. Hence, from (7)

$$q_n^1 \to \frac{F - b^1(0)}{F}, \quad \text{a.s..}$$

From (10), it follows that

$$q_n^2 \to \frac{b(E) - b^2(0)}{E} = \frac{b^1(0) - b^2(0)}{E} + \frac{F - b^1(0)}{F} \geq b'(E_-), \quad \text{a.s.,} \tag{16}$$

where the equality comes from (11) and the last inequality comes from (13). So that from (9) the convergences (1)–(4) now follow.

Let us now compare the limits with corresponding derivatives. From (8), we obtain

$$b(\delta) = \begin{cases} \delta & \text{if } \delta \geq F \\ b^1(0) + \delta\frac{F-b^1(0)}{F} & \text{if } E \leq \delta \leq F \end{cases} \tag{17}$$

Hence, $b'(F_+) = 1$, $b'(F_-) = b'(E_+) = \frac{F-b^1(0)}{F}$. If $b_1(0) > 0$, then $b'(F_+) > b'(F_-)$ so that b is not differentiable at F. When $b^2(0) > b^1(0)$, then

$$b'(E_-) \leq \frac{b^1(0) - b^2(0)}{E} + \frac{F - b^1(0)}{F} < \frac{F - b^1(0)}{F} = b'(E_+),$$

so that b is not differentiable at E. □

We conclude with an important example (see [14, 15, 22]) showing that the case $b^2(0) > b^1(0) > 0$ is realistic.

Example 3.1. Let $m > 0$ be an integer, $\mathbb{A} = \{1, \ldots, m\}$ and $S(a, b) = a \wedge b$. Then $E = m - 1$ and $F = m$. Let every letter in $\mathbb{A}$ having a positive probability. Since $S(a, b) = m$ iff $a = b = m$, obviously $b^1(0) = mP(X_i = m)$ so that

$$b'(F_-) = b'(E_+) = \frac{m - b_1(0)}{m} = 1 - P(X_1 = m) < 1 = b'(F_+).$$

Since $B_n^2(0)$ is bigger than the score of the alignment obtained by aligning as many m-s as possible, thus $B_n^1(0)$, and aligning so many $m-1$'s as possible without disturbing already existing alignment of m's, clearly $b^2(0) > b^1(0)$.

4 Large Deviations

In this section, given $\delta \in \mathbb{R}$, we derive large deviations principle for $B'_n(\delta_+)$ resp. $B'_n(\delta_-)$ by using McDiarmid's inequality. From there, we also derive confidence bounds for $b'(\delta_+)$ resp. $b'(\delta_-)$. Recall that S is symmetric. Let

$$A := \max_{x,y,z \in \mathbb{A}} |S(x,y) - S(x,z)|. \tag{18}$$

For the sake of completeness, let us recall McDiarmid's inequality:

Let $Z_1, \ldots, Z_{2m}$ be independent random variables and $f(Z_1, \ldots, Z_{2m})$ be a function so that changing one variable changes the value at most $K > 0$. Then for any $\sigma > 0$ we have

$$P\Big(f(Z_1, \ldots, Z_{2m}) - Ef(Z_1, \ldots, Z_{2m}) > \sigma\Big) \le \exp\left[-\frac{\sigma^2}{mK^2}\right]. \tag{19}$$

For the proof, we refer to [6]. Another inequality which will be useful later is the so called Höffding's inequality, which is the consequence of McDiarmid's inequality when $f(Z_1, \ldots, Z_m) = \sum_{i=1}^m Z_i$, i.e. for any $\varepsilon > 0$ we have

$$P\left(\frac{1}{m}\sum_{i=1}^m Z_i - EZ_1 > \varepsilon\right) = P\left(\sum_{i=1}^m Z_i - E\big(\sum_{i=1}^m Z_i\big) > \varepsilon m\right)$$
$$\le \exp\left[-\frac{(\varepsilon m)^2}{K^2 \frac{m}{2}}\right] = \exp\left[-\frac{2\varepsilon^2}{K^2} m\right]. \tag{20}$$

In our case, for any $\delta \in \mathbb{R}$ changing the value of one of the $2n$ random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ changes the value of $L_n(\delta)$ at most A, hence for every $\varepsilon > 0$ inequality (19) is translated into

$$P\big(L_n(\delta) - EL_n(\delta) \ge \varepsilon n\big) \le \exp\left[-\frac{\varepsilon^2}{A^2} n\right]$$
$$P\big(L_n(\delta) - EL_n(\delta) \le -\varepsilon n\big) \le \exp\left[-\frac{\varepsilon^2}{A^2} n\right]. \tag{21}$$

Let us define $b_n(\delta) := EB_n(\delta)$. For every $\delta \in \mathbb{R}$, by dominated convergence we have $b_n(\delta) \to b(\delta)$ and by monotonicity the convergence is uniform, i.e.

$$\sup_{\delta\in\mathbb{R}} |b_n(\delta) - b(\delta)| \to 0 \qquad \text{as } n \to \infty.$$

Theorem 4.1. *Let $\delta \in \mathbb{R}$. Then, for every $\varepsilon > 0$ there exists $N(\varepsilon) < \infty$ and a constant $c(\varepsilon) > 0$ such that*

$$P\big(b'(\delta_-) - \varepsilon \le \underline{q}_n(\delta) \le \overline{q}_n(\delta) \le b'(\delta_+) + \varepsilon\big) \ge 1 - 4\exp[-c(\varepsilon)n] \tag{22}$$

for every $n > N(\varepsilon)$.

Proof. Given $\delta \in \mathbb{R}$ and $\varepsilon > 0$, we are looking for bounds on $P\big(B_n'(\delta_+) - b'(\delta_+) > \varepsilon\big)$. For any $s > 0$ and any function $\varphi : \mathbb{R} \to \mathbb{R}$ let us define

$$\Delta\varphi := \varphi(\delta + s) - \varphi(\delta). \tag{23}$$

Now, choose a small $1 > s > 0$ depending on ε such that

$$|\frac{\Delta b}{s} - b'(\delta_+)| \le \frac{\varepsilon}{4}$$

and take n large enough (also depending on ε) such that

$$|\Delta b_n - \Delta b| \le s\frac{\varepsilon}{4}.$$

Thus for those s and n chosen as before we have

$$\begin{aligned}\frac{\Delta B_n}{s} - b'(\delta_+) &= \left(\frac{\Delta B_n}{s} - \frac{\Delta b_n}{s}\right) + \left(\frac{\Delta b_n}{s} - \frac{\Delta b}{s}\right) + \left(\frac{\Delta b}{s} - b'(\delta_+)\right) \\ &\le \left(\frac{\Delta B_n}{s} - \frac{\Delta b_n}{s}\right) + \frac{\varepsilon}{2}.\end{aligned} \tag{24}$$

From (21), it follows

$$\begin{aligned} P\left(B_n(\delta) - b_n(\delta) \le -s\frac{\varepsilon}{4}\right) &\le \exp\left[-\frac{\varepsilon^2 s^2}{16A^2}n\right] = \exp[-c_1(\varepsilon)n] \\ P\left(B_n(\delta + s) - b_n(\delta + s) \ge s\frac{\varepsilon}{4}\right) &\le \exp[-c_1(\varepsilon)n] \end{aligned} \tag{25}$$

where $c_1(\varepsilon) := \varepsilon^2 s^2/(16A^2)$ is a positive constant depending on ε (recall that our s depends on ε). Hence

$$P\left(\frac{\Delta B_n}{s}-\frac{\Delta b_n}{s}\geq\frac{\varepsilon}{2}\right)\leq P\left(B_n(\delta)-b_n(\delta)\leq -s\frac{\varepsilon}{4}\right)+P\left(B_n(\delta+s)-b_n(\delta+s)\geq s\frac{\varepsilon}{4}\right)$$
$$\leq 2\exp[-c_1(\varepsilon)n]. \tag{26}$$

Since B_n is convex, it holds that $B_n'(\delta_+)\leq\Delta B_n/s$ so that for ε and s chosen as before (24) and (26) yield

$$P\left(B_n'(\delta+)-b'(\delta+)\geq\varepsilon\right)\leq P\left(\frac{\Delta B_n}{s}-b'(\delta+)\geq\varepsilon\right)\leq 2\exp[-c_1(\varepsilon)n]. \tag{27}$$

By similar arguments, there exists a positive constant $c_2(\varepsilon)>0$ so that

$$P\left(B_n'(\delta_-)-b'(\delta_-)\leq\varepsilon\right)\leq 2\exp[-c_2(\varepsilon)n]. \tag{28}$$

for every n big enough. Finally, by taking $c:=\min\{c_1,c_2\}$, the inequality (6) implies the inequality (22). □

Note that if b is differentiable at δ, the inequality (22) is satisfied for $q(\delta)$ instead of $b'(\delta_-)$ or $b'(\delta_+)$, namely

Corollary 4.1. *Let b be differentiable at δ. Then, for every $\varepsilon>0$ there exists $N(\varepsilon)<\infty$ and a constant $c(\varepsilon)>0$ such that*

$$P\left(q(\delta)-\varepsilon\leq\underline{q}_n(\delta)\leq\overline{q}_n(\delta)\leq q(\delta)+\varepsilon\right)\geq 1-4\exp[-c(\varepsilon)n] \tag{29}$$

for every $n>N(\varepsilon)$, where $q(\delta)$ is the unique asymptotic proportion of gaps.

We now derive confidence bounds for $b'(\delta_+)$ resp. $b'(\delta_-)$. Recall the definition $b_n(\delta)=EB_n(\delta)=EL_n(\delta)/n$ and the notation (23). From [11] we have

$$b_n(\delta)\leq b(\delta)\leq b_n(\delta)+v(n)$$

for $n\in\mathbb{N}$ even, where

$$v(n):=A\sqrt{\frac{2}{n-1}\Big(\frac{n+1}{n-1}+\ln(n-1)\Big)}+\frac{F}{n-1},$$

so it follows

$$\Delta b-v(n)\leq\Delta b_n\leq\Delta b+v(n).$$

Suppose that k samples of $X^i=X_1^i,\dots,X_n^i$ and $Y^i=Y_1^i,\dots,Y_n^i$, $i=1,\dots,k$ are generated. Let $L_n^i(\delta)$ be the score of the i-th sample. Let

$$\bar{B}_n(\delta):=\frac{1}{kn}\sum_{i=1}^{n}L_n^i(\delta).$$

From (25) we have

$$\begin{aligned} P\big(\Delta\bar{B}_n - \Delta b_n < -c\big) &= P\big(\bar{B}_n(\delta+s) - b_n(\delta+s) + b_n(\delta) - \bar{B}_n(\delta) < -c\big) \\ &\leq P\left(\bar{B}_n(\delta+s) - b_n(\delta+s) < -\frac{c}{2}\right) + P\left(b_n(\delta) - \bar{B}_n(\delta) < -\frac{c}{2}\right) \\ &\leq 2\exp\left[-\frac{c^2k}{4A^2}n\right]. \end{aligned}$$

By convexity $sb'(\delta+) \leq \Delta b$ for every $s > 0$, so from the last inequality it follows

$$\begin{aligned} P\big(\Delta\bar{B}_n + c + v(n) \geq sb'(\delta+)\big) &\geq P\big(\Delta\bar{B}_n + c + v(n) \geq \Delta b\big) = P\big(\Delta\bar{B}_n + c \geq \Delta b - v(n)\big) \\ &\geq P\big(\Delta\bar{B}_n + c \geq \Delta b_n\big) = P\big(\Delta\bar{B}_n - \Delta b_n \geq -c\big) \\ &\geq 1 - 2\exp\left[-\frac{c^2k}{4A^2}n\right], \end{aligned}$$

from where we obtain that with probability $1-\varepsilon$

$$b'(\delta+) \leq \frac{1}{s}\left(\bar{B}_n(\delta+s) - \bar{B}_n(\delta) + 2A\sqrt{\frac{\ln(2/\varepsilon)}{kn}} + v(n)\right). \tag{30}$$

Since (30) holds for every $s > 0$, we have that with probability $1-\varepsilon$

$$b'(\delta+) \leq \min_{s>0}\frac{1}{s}\left(\bar{B}_n(\delta+s) - \bar{B}_n(\delta) + 2A\sqrt{\frac{\ln(2/\varepsilon)}{kn}} + v(n)\right). \tag{31}$$

Similarly, we have that with probability $1-\varepsilon$

$$b'(\delta-) \geq \max_{s>0}\frac{1}{s}\left(\bar{B}_n(\delta-s) - \bar{B}_n(\delta) - 2A\sqrt{\frac{\ln(2/\varepsilon)}{kn}} - v(n)\right). \tag{32}$$

5 Fluctuations of the Score in Optimal Alignments

In this section we prove $\mathrm{Var}[L_n(\delta)] = \Theta(n)$. The $\Theta(n)$ notation means that there exist two constants $0 < c < C < \infty$ such that $cn \leq \mathrm{Var}[L_n(\delta)] \leq Cn$ for n large enough. The upper bound follows from an Efron-Stein type of inequality proved by Steele in [21], so we aim to provide conditions on the scoring function that guarantee the existence of the lower bound. In this section we show that when $\delta < 0$ and $|\delta|$ is large enough in the sense of Assumption 5.1 (see below), then there exists $c > 0$ so that $\mathrm{Var}(L_n(\delta)) > cn$ for n large enough. In comparison with previous results, here we solve—for the first time—the problem of the fluctuations of the score in optimal alignments for rather realistic high entropy models.

5.1 Order of the Variance

All above mentioned fluctuations results are based in the following strategy: the inequality $\mathrm{Var}[L_n(\delta)] \geq cn$ is satisfied as soon as we are able to establish that changing at random one symbol in the sequences has a biased effect on the optimal alignment score. In details, we choose two letters $a, b \in \mathbb{A}$ and fix a realization of $X = X_1 \dots X_n$ and $Y = Y_1 \dots Y_n$. Then among all the a's in X and Y we choose one at random (with equal probability). That chosen letter a is replaced by a letter b. The new sequences thus obtained are denoted by $\tilde{X}$ and $\tilde{Y}$. The optimal score for the strings $\tilde{X}$ and $\tilde{Y}$ is denoted by

$$\tilde{L}_n(\delta) := \max_{(\pi,\mu)} U^{\delta}_{(\pi,\mu)}(\tilde{X}, \tilde{Y}).$$

The following important theorem postulates the mentioned strategy. In full generality, it is proven in [12], for special case of two colors and S corresponding to LCS, see Sect. 3 in [10]; for a special case of $S(a,b) = a \wedge b$ and $\mathbb{A} = \{m-1, m, m+1\}$, see Theorem 2.1 in [14].

Theorem 5.1. *Assume that there exist* $\varepsilon > 0$, $d > 0$ *and* $n_0 < \infty$ *such that*

$$P\left(E[\tilde{L}_n(\delta) - L_n(\delta) | X, Y] \geq \varepsilon \right) \geq 1 - e^{-dn} \tag{33}$$

for all $n > n_0$. *Then, there exists a constant* $c > 0$ *not depending on* n *such that* $\mathrm{Var}[L_n(\delta)] \geq cn$ *for every* n *large enough.*

Now, our aim is to show that if δ is small enough and the scoring function satisfies some asymmetry property, then there exist letters $a, b \in \mathbb{A}$ so that the condition (33) is fulfilled. Typically, to satisfy the assumptions, δ should be negative so that the main result holds if the gap penalty $|\delta|$ is large enough. Let us introduce our asymmetry assumption on the scoring function:

Assumption 5.1. *Suppose there exist letters* $a, b \in \mathbb{A}$ *such that*

$$\sum_{c \in \mathbb{A}} P(X_1 = c)\big(S(b,c) - S(a,c)\big) > 0. \tag{34}$$

Remark 5.1. For the alphabet $\mathbb{A} = \{a, b\}$, condition (34) says

$$\big(S(b,a) - S(a,a)\big)P(X_1 = a) + \big(S(b,b) - S(b,a)\big)P(X_1 = b) > 0.$$

Since S is symmetric and one could exchange a and b, the condition (34) actually means

$$\big(S(b,a) - S(a,a)\big)P(X_1 = a) + \big(S(b,b) - S(b,a)\big)P(X_1 = b) \neq 0.$$

When $S(b,b) = S(a,a)$, then Assumption 5.1 is satisfied if and only if $P(X_1 = a) \neq P(X_1 = b)$. For, example when $S(b,b) = S(a,a) > S(b,a)$ (recall that S is assumed to be symmetric and non-constant), then (34) holds if $P(X_1 = a) \neq P(X_1 = b)$.

In the present paper, the main result on the fluctuations of the score in optimal alignments can be formulated as following:

Theorem 5.2. *Suppose Assumption 5.1 holds. Then, there exist constants δ_0 and $c > 0$ not depending on n such that*

$$\mathrm{Var}[L_n(\delta)] \geq cn$$

for all $\delta \leq \delta_0$ and for n large enough.

Before proving the above-stated theorem, we need a preliminary lemma. Suppose Assumption 5.1 holds, then take $a, b \in \mathbb{A}$ satisfying (34) and define the functions $\zeta^x : \mathbb{A} \times \mathbb{A} \mapsto \mathbb{R}$ and $\zeta^y : \mathbb{A} \times \mathbb{A} \mapsto \mathbb{R}$ in the following way:

$$\zeta^x(x,y) = \begin{cases} S(b,y) - S(a,y) & \text{if } x = a \\ 0 & \text{otherwise} \end{cases}$$

$$\zeta^y(x,y) = \begin{cases} S(x,b) - S(x,a) & \text{if } y = a \\ 0 & \text{otherwise.} \end{cases}$$

Note that $S(x,y) = S(y,x)$ implies $\zeta^y(x,y) = \zeta^x(y,x)$. We now define the random variable Z by

$$Z := \zeta^x(X_1,Y_1) + \zeta^y(X_1,Y_1) = \zeta^x(X_1,Y_1) + \zeta^x(Y_1,X_1). \tag{35}$$

Note that (5.1) ensures that Z has strictly positive expectation:

$$\begin{aligned}\rho := EZ &= E\big(\zeta^x(X_1,Y_1) + \zeta^y(X_1,Y_1)\big) = 2E\zeta^x(X_1,Y_1)\\ &= 2E[\zeta^x(a,Y_1)|X_1 = a]P(X_1 = a) = 2\sum_{c\in\mathbb{A}} \zeta^x(a,c)P(Y_1 = c)P(X_1 = a)\\ &= 2P(X_1 = a)\sum_{c\in\mathbb{A}}(S(b,c) - S(a,c))P(X_1 = c) > 0.\end{aligned}$$

Let Λ^* be the Legendre-Fenchel transform of the logarithmic moment generating function of $-Z$, namely

$$\Lambda^*(c) = \sup_{t\in\mathbb{R}} (ct - \ln E[\exp(-Zt)]) \qquad \forall\, c \in \mathbb{R}.$$

It is known that the supremum above can be taken over non-negative t's and, for any $c > E(-Z) = -\rho$, it holds $\Lambda^*(c) > 0$. Since $\rho > 0$, we have for $c = 0$

$$\Lambda^*(0) = -\inf_{t\in\mathbb{R}} \ln E[\exp(-tZ)] = -\inf_{t\geq 0} \ln E[\exp(-tZ)] = -\ln \inf_{t\geq 0} E[\exp(-tZ)] > 0.$$

Let $Z_1, \ldots, Z_k$ be i.i.d. random variables distributed as $-Z$, then for any $c > -\rho$ the following large deviation bound holds

$$P\left(\sum_{i=1}^{k} Z_i > ck\right) \leq \exp[-\Lambda^*(c)k]. \tag{36}$$

Finally, denote $h(q)$ the binary entropy function $h(q) := -q\ln q - (1-q)\ln(1-q)$ and note that the inequality

$$2h(q) < \Lambda^*(0)(1-q)$$

holds when $q > 0$ is small enough, since Λ^* and h are both continuous and $\Lambda^*(0) > 0$.

In what follows, let for any $q \in (0,1)$, $\mathcal{A}^n(q)$ be the set of all alignments with no more than qn gaps, i.e.

$$\mathcal{A}^n(q) := \{(\pi, \mu) : q(\pi, \mu) \leq q\}.$$

We are interested in the event that the sequences X and Y are such that for every alignment (π, μ) with no more than qn gaps we have a biased effect of the random change of at least $\varepsilon > 0$. Let $D_q^n(\varepsilon)$ denote that event i.e.

$$D_q^n(\varepsilon) := \bigcap_{(\pi,\mu)\in\mathcal{A}^n(q)} D_{(\pi,\mu)}^n(\varepsilon) \tag{37}$$

where

$$D_{(\pi,\mu)}^n(\varepsilon) := \left\{ E\left[\sum_{i=1}^{k} \left(S(\tilde{X}_{\pi_i}, \tilde{Y}_{\mu_i}) - S(X_{\pi_i}, Y_{\pi_i})\right) \big| X, Y\right] \geq \varepsilon \right\}.$$

Now, we are ready to state the key lemma.

Lemma 5.1. *Suppose Assumption 5.1 is fulfilled and take $a, b \in \mathbb{A}$ satisfying (34). Let $q > 0$ small enough such that*

$$2h(q) < \Lambda^*(0)(1-q). \tag{38}$$

Then, there exist $\varepsilon > 0$, $\alpha > 0$ and $n_2 < \infty$, all depending on q, such that

$$P\left((D_q^n(\varepsilon))^c\right) \leq \exp[-\alpha n] \tag{39}$$

for every $n > n_2$.

Proof. Let $x = x_1, \dots, x_n$ and respectively $y = y_1, \dots, y_n$ be fixed realizations of X and Y, respectively. Let n_a be the number of a's in both sequences. Let $\pi = (\pi_1, \pi_2, \dots, \pi_k)$ and $\mu = (\mu_1, \mu_2, \dots, \mu_k)$ be a fixed alignment of X and Y. Recall that $\tilde{X}$ and $\tilde{Y}$ are obtained by choosing at random one a among all the a's in x and y. Hence, such an a is chosen with probability $1/n_a$. Our further analysis is based on the following observation:

$$E\left[\sum_{i=1}^{k}\big(S(\tilde{X}_{\pi_i}, \tilde{Y}_{\mu_i}) - S(X_{\pi_i}, Y_{\pi_i})\big)\Big| X = x, Y = y\right]$$
$$= \frac{1}{n_a}\sum_{i=1}^{k}\big(\zeta^x(x_{\pi_i}, y_{\mu_i}) + \zeta^y(x_{\pi_i}, y_{\mu_i})\big).$$

Thus, it holds

$$P\big((D^n_{(\pi,\mu)}(\varepsilon))^c\big) = P\left(E\left[\sum_{i=1}^{k}\big(S(\tilde{X}_{\pi_i}, \tilde{Y}_{\mu_i}) - S(X_{\pi_i}, Y_{\pi_i})\big)\Big| X, Y\right] < \varepsilon\right)$$
$$= P\left(\sum_{i=1}^{k} Z_i < N_a\varepsilon\right), \tag{40}$$

where N_a is the (random) number of a's in X and Y and the random variables $Z_1, \dots, Z_k$ are defined as follows:

$$Z_i := \zeta^x(X_{\pi_i}, Y_{\mu_i}) + \zeta^y(X_{\pi_i}, Y_{\mu_i}) = \zeta^x(X_{\pi_i}, Y_{\mu_i}) + \zeta^x(Y_{\mu_i}, X_{\pi_i})$$

for $i = 1, \dots, k$. Let us mention again that the random variables Z_i depend on fixed alignment (π, μ) (which is omitted in the notation) and, since $X_1, \dots, X_n, Y_1, \dots, Y_n$ are i.i.d., so are the random variables $Z_1, \dots, Z_k$. Clearly, Z_i is distributed as Z defined in (35). Suppose now that the fixed alignment (π, μ) has the proportion of gaps less or equal than q, i.e. $(\pi, \mu) \in \mathcal{A}^n(q)$. Then $\frac{k}{n} \geq 1 - q$ and, since obviously $N_a \leq 2n$, we have

$$\left\{\sum_{i=1}^{k} Z_i < \varepsilon N_a\right\} \subseteq \left\{\sum_{i=1}^{k} Z_i < \varepsilon 2n\right\} = \left\{\sum_{i=1}^{k} Z_i < k2\varepsilon\frac{n}{k}\right\} \subseteq \left\{\sum_{i=1}^{k} Z_i < k\frac{2\varepsilon}{(1-q)}\right\}. \tag{41}$$

Fix now q satisfying (38). Since Λ^* is continuous, there exists ε, depending on the chosen q so that the following two conditions are simultaneously satisfied

$$\frac{-2\varepsilon}{1-q} > -\rho \quad \text{and} \quad -2\alpha := 2h(q) - \Lambda^*\left(-\frac{2\varepsilon}{1-q}\right)(1-q) < 0. \tag{42}$$

Using the large deviations bound (36) with $c = \frac{-2\varepsilon}{1-q}$ and the fact that $\frac{k}{n} \geq 1-q$, we have

$$P\left(-\sum_{i=1}^{k} Z_i > -k\frac{2\varepsilon}{(1-q)}\right) \leq \exp\left[-\Lambda^*\left(\frac{-2\varepsilon}{1-q}\right)k\right]$$
$$\leq \exp\left[-\Lambda^*\left(-\frac{2\varepsilon}{1-q}\right)(1-q)n\right]. \qquad (43)$$

By (37), (40), (41) and (43), we obtain

$$P\big((D_q^n(\varepsilon))^c\big) \leq |\mathcal{A}^n(q)| \exp\left[-\Lambda^*\left(-\frac{2\varepsilon}{1-q}\right)(1-q)n\right]. \qquad (44)$$

In order to bound $|\mathcal{A}^n(q)|$, note that the number of different alignment with exactly $(n-k)$ gaps is bounded above by $\binom{n}{n-k}^2$ so that for $q \leq 0.5$ we have

$$|\mathcal{A}^n(q)| \leq \sum_{i \leq qn} \binom{n}{i}^2 \leq \sum_{i \leq qn} \binom{n}{qn}^2 \leq qn\binom{n}{qn}^2 \leq \exp[2h(q)n + \ln(qn)], \quad (45)$$

where $h(q)$ is the binary entropy function. In the second inequality the relation $q \leq 0.5$ was used, while the last inequality is based on the well-known relation $\binom{n}{\gamma n} \leq \exp[h(\gamma)n]$, for any $\gamma \in (0,1)$. Thus, from (42), (44) and (45) we have

$$P\big((D_q^n(\varepsilon))^c\big) \leq \exp\left[\left(2h(q) - \Lambda^*\left(-\frac{2\varepsilon}{1-q}\right)(1-q) + \frac{\ln(qn)}{n}\right)n\right]$$
$$= \exp\Big[-\Big(2\alpha - \frac{\ln(qn)}{n}\Big)n\Big]. \qquad (46)$$

This implies that there exists n_2 big enough (recall $nq \geq 1$) such that (39) holds. □

Proof of Theorem 5.2. Let $\mathcal{O}(X,Y)$ denote the set of all optimal alignments of (X,Y), i.e.

$$(\pi,\mu) \in \mathcal{O}(X,Y) \Leftrightarrow L_n(\delta) = U^{\delta}_{(\pi,\mu)}(X,Y) = \sum_{i=1}^{k} S(X_{\pi_i}, Y_{\mu_i}) + \delta q(\pi,\mu)n.$$

Note that the difference $\tilde{L}_n(\delta) - L_n(\delta)$ is bounded from below by

$$\tilde{L}_n(\delta) - L_n(\delta) \geq U^{\delta}_{(\pi,\mu)}(\tilde{X},\tilde{Y}) - U^{\delta}_{(\pi,\mu)}(X,Y) = \sum_{i=1}^{k}\big(S(\tilde{X}_{\pi_i},\tilde{Y}_{\mu_i}) - S(X_{\pi_i},Y_{\mu_i})\big).$$

Thus, for every $\varepsilon > 0$ we have

$$\Big\{\exists(\pi,\mu) \in \mathcal{O}(X,Y) : E\left[\sum_{i=1}^{k}\big(S(\tilde{X}_{\pi_i},\tilde{Y}_{\mu_i}) - S(X_{\pi_i},Y_{\mu_i})\big)\Big| X,Y\right] \geq \varepsilon\Big\}$$
$$\subseteq \Big\{E[\tilde{L}_n(\delta) - L_n(\delta)|X,Y] \geq \varepsilon\Big\}. \quad (47)$$

Recall that the event $D^n_q(\varepsilon)$ means that every alignment (π,μ) with no more than qn gaps has a biased effect of the random change at least ε. Now, it is clear that the right side of (47) holds if $D^n_q(\varepsilon)$ holds and there exists an optimal alignment contains no more than qn gaps, i.e. we have the inclusion

$$\big\{\mathcal{O}(X,Y) \subseteq \mathcal{A}^n(q)\big\} \cap D^n_q(\varepsilon) \subseteq \Big\{E[\tilde{L}_n(\delta) - L_n(\delta)|X,Y] \geq \epsilon\Big\}. \quad (48)$$

Recall that $b(\delta)$ is convex and increasing, $b'(\delta) = 1$ if δ is big enough and $b'(\delta) = 0$ if δ is small enough. Hence, for every $q \geq 0$ there exists δ so that $b'(\delta+) < q$. Let δ be such and denote $\varepsilon_1 := q - b'(\delta+)$. Then by Theorem 4.1, there exist $c(\varepsilon_1)$ and $n_1(\varepsilon_1)$ such that

$$P\big(\overline{q}_n(\delta) \leq q\big) \geq 1 - 2\exp[c(\varepsilon_1)n] \quad (49)$$

for every $n > n_1$. Therefore we have

$$P\big(\mathcal{O}(X,Y) \subseteq \mathcal{A}^n(q)\big) \geq 1 - 2\exp[c(\varepsilon_1)n] \quad (50)$$

for $n > n_1$. From Lemma 5.1, it follows that if $q > 0$ is small enough to satisfy (38), then there exist $\varepsilon > 0$, $\alpha > 0$ and $n_2 < \infty$, all depending on q so that

$$P\big((D^n_q(\varepsilon))^c\big) \leq \exp[-\alpha n] \quad (51)$$

for every $n > n_2$. To finalize the proof, let us take q satisfying (38) and δ_0 be such that $b'(\delta_{0+}) < q$. Then, there exist $\varepsilon > 0$, $\alpha > 0$ and $n_0 := \max\{n_2, n_1\}$ so that (50) and (51) hold. Thus, from (48) we have

$$P\Big(E[\tilde{L}_n(\delta) - L_n(\delta)|X,Y] \geq \epsilon\Big) \geq 1 - 2\exp[c(\varepsilon_1)n] - \exp[-\alpha n]$$

for every $n > n_0$. Hence, the assumptions of Theorem 5.1 are satisfied. □

An alternative to Lemma 5.1. Recall that δ_0 in Theorem 5.2 was chosen to be such that $b'(\delta_{0+}) < q$, where q satisfies assumptions of Lemma 5.1, namely (38). This assumption comes from the large deviations bound (43). Although, asymptotically it is a sharp inequality, the rate-function Λ^* might not always be easy to compute. Clearly, the statement of Lemma 5.1 holds true for any other type of large deviations inequality giving the same exponential decay. An alternative would be to use Höffding's inequality (20) to get a version of Lemma 5.1 which does not rely on the computation of Λ^*. The Höffding's inequality gives smaller q, and, therefore, larger δ_0.

Lemma 5.2. *Suppose Assumption 5.1 is fulfilled and take a, b satisfying (34). Let $q > 0$ small enough such that*

$$h(q) < \frac{(1-q)\rho^2}{9A^2}. \tag{52}$$

Then there exist $\varepsilon > 0$, $\alpha > 0$ and $n_2 < \infty$, all depending on q, such that

$$P\big((D_q^n(\varepsilon))^c\big) \le \exp[-\alpha n] \tag{53}$$

for every $n > n_2$.

Proof. Recall the definition of A from (18). Let $q > 0$ be small enough satisfying (52). Then, there exists $\varepsilon > 0$ small enough such that both conditions simultaneously hold:

(1) $2\varepsilon < (1-q)\rho$, which means that $\sigma := \rho - \frac{2\varepsilon}{(1-q)} > 0$;
(2)
$$h(q) - \frac{((1-q)\rho - 2\varepsilon)^2}{9A^2(1-q)} =: -\alpha(\varepsilon) < 0.$$

Hence, there exists $n_2 < \infty$ such that

$$h(q) - \frac{((1-q)\rho - 2\varepsilon)^2}{9A^2(1-q)} + \frac{\ln(qn)}{2n} \le -\frac{\alpha}{2} \tag{54}$$

for every $n > n_2$. Recall that $k \ge (1-q)n$. To apply Höffding's inequality, we need to bound the random variable Z. Recall the definition of Z from (35). From the definition, $\zeta^x(x,y)$ and $\zeta^y(x,y)$ are simultaneously non-zero if and only if $x = y = a$, this means that the difference between the maximum and minimum value of Z_i is at most $3A$. For instance, if $S(b,a) < S(a,a)$ then $-2A \le 2(S(b,a) - S(a,a)) \le Z_i \le \max_{c \neq a}(S(b,c) - S(a,c)) \le A$. Then, by using (20), the large deviations bound (43) can be written down as (recall **(1)**)

$$P\left(-\sum_{i=1}^{k} Z_i > -k\frac{2\varepsilon}{(1-q)}\right) = P\left(\frac{1}{k}\sum_{i=1}^{k}(-Z_i) + \rho > \rho - \frac{2\varepsilon}{(1-q)}\right) \le \exp\left[-\frac{2\sigma^2}{(3A)^2}k\right]$$
$$\le \exp\left[-\frac{2\sigma^2(1-q)}{9A^2}n\right] = \exp\left[-\frac{2((1-q)\rho-2\varepsilon)^2}{9A^2(1-q)}n\right].$$

Finally, the inequality (46) can be now written down as

$$P\left(D_q^{nc}(\varepsilon)\right) \le \exp\left[2\Big(h(q) - \frac{((1-q)\rho - 2\varepsilon)^2}{9A^2(1-q)} + \frac{\ln(qn)}{2n}\Big)n\right] = \exp[-\alpha n], \tag{55}$$

where the result is proven by using (54). □

6 Determining δ

In the last section, we discuss how to determine δ_0 in Theorem 5.2. Recall, once again, the proof of that theorem: δ_0 is so small that $b'(\delta_{0+}) < q$, where q is small enough to satisfy condition (38). This condition depends on Λ^*, but knowing the distribution of X_1, Λ^* can be found. When Λ^* is unknown, then condition (38) can be substituted by (more restrictive) condition (52). The latter does not depend on Λ^* and can be also used when the distribution of X_1 is unknown. Hence finding q is not a problem. The problem, however, is to determine the function b or its derivatives. In Sect. 4, we found confidence upper bound for $b'(\delta_+)$ (31). That bound is random and holds only with certain probability. In the following, we investigate deterministic ways to estimate $b'(\delta+)$.

Let $(t_n, q_n) \in \mathcal{O}_n(\delta)$ be an optimal pair: $B_n = t_n(1-q_n) + \delta q_n$. Clearly, when (t'_n, q'_n) is another optimal pair and $q'_n > q_n$, then $t'_n > t_n$. Hence $(\bar{t}_n, \bar{q}_n) \in \mathcal{O}_n(\delta)$, where $\bar{t}_n = \max\{t : (t,q) \in \mathcal{O}_n(\delta)\}$. For every $q \in (0,1)$, let $\bar{t}(q)$ be an asymptotic upper bound for $\bar{t}_n$ in the sense that if $b'(\delta+) < q$ (i.e. $\limsup_n \bar{q}_n(\delta) < q$ almost surely) then

$$P(\text{eventually } \bar{t}_n \leq \bar{t}(q)) = 1.$$

Thus, $q \mapsto \bar{t}(q)$ is non-decreasing. In what follows, let $\underline{b}$ be the lower bound of $b(\delta)$ for every δ. Since the asymptotic proportion of gaps goes to zero as $\delta \to -\infty$, $\underline{b}$ can be taken as the limit of gapless alignments. This limit is obviously $ES(X_1, Y_1) =: \gamma$. If the distribution of X_1 is unknown, then $\underline{b}$ can be any lower bound for γ.

Let now $q_o \in (0,1)$ be fixed. We aim to find $\delta_o := \delta(q_o) \geq 0$ such that $b'(\delta_+) < q_o$ for every δ satisfying $-\delta > \delta_o$. The following claim shows that δ_o can be computed as follows:

$$\delta_o = \sup_{q \geq q_o} \frac{\bar{t}(q)(1-q) - \underline{b}}{q}. \tag{56}$$

Claim 6.1. *Let $\delta < 0$ be such that $-\delta > \delta_o$, where δ_o is as in (56). Then $b'(\delta_+) \leq q_o$.*

Proof. Take $\delta \leq 0$ so small that $-\delta \geq \delta_o$. Without loss of generality, we can assume that b is differentiable at δ. Thus b is differentiable at δ implies that $\bar{q}_n \to b'(\delta) > 0$ a.s. Let now $\varepsilon := |\delta| - \delta_o$ and let $q' > b'(\delta)$ be such that

$$\Big|\frac{\bar{t}(q')(1-q') - \underline{b}}{q'} - \frac{\bar{t}(q')(1-b'(\delta)) - \underline{b}}{b'(\delta)}\Big| < \varepsilon.$$

Suppose $b'(\delta) \geq q_o$. Then, $q' > q_o$ and by definition of δ_o

$$\delta(q_o) \geq \frac{\bar{t}(q')(1-q') - \gamma}{q'}$$

so that

$$|\delta| = \delta_o + \varepsilon > \frac{\bar{t}\big(q'\big)(1 - b'(\delta)) - \underline{b}}{b'(\delta)} \quad \Leftrightarrow \quad \bar{t}\big(q'\big)(1 - b'(\delta)) - |\delta| b'(\delta) < \underline{b}. \tag{57}$$

Since $b'(\delta) < q'$, then eventually $\bar{t}_n \leq \bar{t}(q')$ a.s. By the convergence $\bar{q}_n \to b'(\delta)$ from the r.h.s. of (57), follows then that eventually

$$B_n = \bar{t}_n(1 - \bar{q}_n) - |\delta|\bar{q}_n < \underline{b}$$

almost surely. We have a contradiction with the almost surely convergence $B_n \to b(\delta) \geq \underline{b}$. □

Remark 6.1. If $\bar{t}(q) \equiv \bar{t}$ is constant, then (56) is

$$\delta(q_o) := \frac{\bar{t}(1 - q_o) - \underline{b}}{q_o}. \tag{58}$$

6.1 Finding $\bar{t}(q)$

For applying (56) the crucial step is to find $\bar{t}$. Since the maximum value of the scoring function is F, a trivial bound is $\bar{t}(q) \equiv F$ and δ_o can be found from (58). However, using the same ideas as in the proof of Theorem 5.2, we could obtain a realistic bound for $\bar{t}(q)$ as follows. In the following theorem, let Λ^* be Legendre-Fenchel transform of $\Lambda(t) := \ln E\exp[tZ]$, where $Z := S(X, Y)$. Clearly Z is a nonnegative random variable $Z \leq F$ and EZ was denoted by γ.

Theorem 6.1. *Let $q_1 \in (0, 1)$ and let $\bar{t}(q_1)$ satisfy one of the following conditions*

$$\frac{2h(q_1 \wedge 0.5)}{1 - q_1} = \Lambda^*(\bar{t}(q_1)). \tag{59}$$

or

$$\bar{t}(q_1) = F\sqrt{\frac{h(q_1 \wedge 0.5)}{1 - q_1}} + \gamma. \tag{60}$$

Then for every δ such that $b'(\delta_+) < q_1$, the following holds

$$P\big(\text{eventually } \bar{t}_n(\delta) \leq \bar{t}(q_1)\big) = 1.$$

Proof. Let $q_1 \in (0, 1)$ be fixed. Let δ be such that $b'(\delta_+) < q_1$. Note that we can find q such that $b'(\delta_+) < q$ and the following conditions both hold

$$\frac{2h(q \wedge 0.5)}{1-q} < \Lambda^*(\bar{t}(q_1)). \tag{61}$$

and

$$\bar{t}(q_1) > F\sqrt{\frac{h(q \wedge 0.5)}{1-q}} + \gamma. \tag{62}$$

Note that (62) implies

$$h(q \wedge 0.5) < \frac{(\bar{t}(q_1) - \gamma)^2}{F^2}(1-q). \tag{63}$$

Let $(\pi_1, \ldots, \pi_k)$ and $(\mu_1, \ldots, \mu_k)$ be a fixed alignment, and let $Z_1, \ldots, Z_k$ be i.i.d. random variables, where $Z_i = S(X_{\pi_i}, Y_{\mu_i})$. Clearly Z_i is distributed as Z defined above. If the alignment (π, μ) is optimal, then

$$t_n = \frac{1}{k}\sum_{i=1}^{k} Z_i.$$

Recall that $\gamma = EZ_i$. Since $\Lambda^*(\gamma) = 0$, the conditions (59) and (60) both guarantee $\bar{t} > \gamma$. Let us define

$$D^n_q(\bar{t}(q_1)) := \bigcap_{(\pi,\mu)\in\mathcal{A}^n(q)} \left\{ \frac{1}{k}\sum_{i=1}^{k} Z_i \leq \bar{t}(q_1) \right\}.$$

The event $D^n_q(\bar{t}(q_1))$ states that the average score of aligned letters is smaller than $\bar{t}(q_1)$ for every alignment with proportion of gaps at most q. If all optimal alignments are so, then also $\bar{t}_n(\delta) \leq \bar{t}(q_1)$, namely

$$\{\mathcal{O}(X,Y) \in \mathcal{A}^n(q)\} \cap D^n_q(\bar{t}) \subseteq \{\bar{t}_n(\delta) \leq \bar{t}(q_1)\}.$$

In order to bound $P(D^n_q(\bar{t}(q_1)))$, we proceed as in Lemma 5.1. Using the large deviations bound

$$P\left(\sum_{i=1}^{k} Z_i > \bar{t}(q_1)k\right) \leq \exp[-\Lambda^*(\bar{t}(q_1))k] \leq \exp[-\Lambda^*(\bar{t}(q_1))(1-q)n] \tag{64}$$

we obtain the following estimate

$$P\left((D^n_q(\bar{t}))^c\right) \leq |\mathcal{A}^n(q)| \exp[-\Lambda^*(\bar{t}(q_1))(1-q)n] \tag{65}$$

For $q \leq 0.5$, we estimate $|\mathcal{A}^n(q)|$ as in Lemma 5.1 by

$$|\mathcal{A}^n(q)| \leq \exp\left[\left(2h(q) + \frac{\ln(qn)}{n}\right)n\right].$$

For $q \in (0.5, 1)$ note that

$$|\mathcal{A}^n(q)| \leq \sum_{i \leq qn} \binom{n}{i}^2 < qn\binom{n}{\frac{1}{2}n}^2 \leq \exp\left[\left(2h(0.5) + \frac{\ln(qn)}{n}\right)n\right].$$

Hence

$$P\big((D^n_q(\bar{t}))^c\big) \leq \exp\left[\left(2h(q \wedge 0.5) + \frac{\ln(qn)}{n} - \Lambda^*(\bar{t}(q_1))(1-q)\right)n\right]. \quad (66)$$

Since (61) holds, then just like in the proof of Theorem 5.2, there exists $\alpha > 0$ and n_o, both depending on $\bar{t}(q_1)$, so that

$$P\big(\bar{t}_n(\delta) > \bar{t}(q_1)\big) \leq \exp[-\alpha n], \quad \forall n > n_o. \quad (67)$$

Thus,by Borel-Cantelli we have

$$P\big(\text{eventually } \bar{t}_n(\delta) \leq \bar{t}(q_1)\big) = 1.$$

With Höffding's inequality the bounds (64) and (66) are

$$P\left(\sum_{i=1}^{k} Z_i > \bar{t}(q_1)k\right) \leq \exp\left[-\frac{2(\bar{t}(q_1)-\gamma)^2}{F^2}k\right] \leq \exp\left[-\frac{2(\bar{t}(q_1)-\gamma)^2}{F^2}(1-q)n\right]$$

$$P\big((D^n_q(\bar{t}))^c\big) \leq \exp\left[2\big(h(q \wedge 0.5) + \frac{\ln(qn)}{2n} - \frac{(\bar{t}(q_1)-\gamma)^2}{F^2}(1-q)\big)2n\right].$$

respectively, and the existence of $\alpha > 0$ and n_o comes from (63). □

6.2 Example

Consider a two letter alphabet $\mathbb{A} = \{a, b\}$ with probabilities $P(X_i = b) = P(Y_i = b) = 0.7$, $P(X_i = a) = P(Y_i = a) = 0.3$. Let the scoring function S assign 1 to identical letter pairs and 0 to unequal letters. Then the letters a, b satisfy (5.1). The random variable Z as in (35) is distributed as follows:

$$P(Z = -2) = (0.3)^2 = 0.09, \quad P(Z = 0) = (0.7)^2 = 0.49, \quad P(Z = 1) = 2 \cdot 0.3 \cdot 0.7 = 0.42.$$

Hence $EZ = \rho = 0.24$ and

$$E\exp[-tZ] = (0.09)\exp[2t] + 0.49 + 0.42\exp[-t].$$

This function achieves its minimum at a t^* that is the solution of the equation

$$(2 \cdot 0.09)\exp[2t] = \exp[-t]0.42$$

that is

$$t^* = \frac{1}{3}\ln\frac{42}{18} \approx 0.28.$$

Then

$$-\Lambda^*(0) = \ln\left[(3 \cdot 0.09)\exp[2t^*] + 0.49\right] = -0.03564$$

so that q satisfies (38) if and only if $q < q_o := 0.00255$, because q_o is a solution of the equation

$$h(q_o) = \frac{\Lambda^*(0)}{2}(1 - q_o) \quad \Leftrightarrow \quad h(q_o) = 0.01782(1 - q_o).$$

Let us see, how much q_o changes when we assume the stronger condition (52). Clearly $A = 1$, so to satisfy (52), the proportion of gaps should satisfy the inequality $q < q_o := 0.000784674$, because q_o is the solution of the inequality $9h(q_o) = (1 - q_o)(0.24)^2$ that is

$$h(q_o) = 0.0064(1 - q_o).$$

Determining $\boldsymbol{\delta_0}$. Let us find δ_0 so that $b'(\delta_{0+}) \leq q_0 = 0.00255$. In this example $F = 1$, $\gamma = (0.3)^2 + (0.7)^2 = 0.58$. Taking $\bar{t} = 1$, from (58) with $q_o = 0.00255$, we get

$$\delta_o := \frac{(1 - q_o) - 0.58}{q_o} = \frac{(1 - 0.00255) - 0.58}{0.00255} < 164.$$

The inequality (60) is

$$\bar{t}(q) = \sqrt{\frac{h(q \wedge 0.5)}{1 - q}} + \gamma.$$

Thus, from (56), we get

$$\delta_o = \sup_{q \geq q_o} \frac{\bar{t}(q)(1 - q) - \gamma}{q} = \sup_{q \geq q_o} \frac{(\sqrt{\frac{h(q \wedge 0.5)}{1-q}} + \gamma)(1 - q) - \gamma}{q}$$
$$= \sup_{q \geq q_o} \frac{\sqrt{h(q \wedge 0.5)(1 - q)}}{q} - \gamma.$$

Since

$$q \mapsto \frac{\sqrt{h(q \wedge 0.5)(1-q)}}{q}$$

is decreasing, we get a much better bound

$$\delta_o = \frac{\sqrt{h(q_o)(1-q_o)}}{q_o} - \gamma = \frac{\sqrt{h(0.00255)(1-0.00255)}}{0.00255} - 0.58 < 52.$$

The random variable $Z := S(X_1, Y_1)$ has Bernoulli distribution with parameter γ, so it is well known that

$$\Lambda^*(t) = t \ln\left(\frac{t}{\gamma}\right) + (1-t)\ln\left(\frac{1-t}{1-\gamma}\right),$$

provided $t > \gamma$. Therefore, the maximum value of $\Lambda^*(t)$ is achieved for $t = 1$ and that is the solution of (59) for $q_1 = 0.0698$. Hence, (59) has solution $\bar{t}(q_1)$ for every $q_1 \in [0, 0.0698]$ and in the range $[0, 0.0698]$ the function

$$q \mapsto \frac{\bar{t}(q)(1-q)-\gamma}{q}$$

is decreasing. This means that δ_o can be taken as

$$\delta_o = \frac{\bar{t}(q_o)(1-q_o)-\gamma}{q_o}.$$

Since, $\bar{t}(0.00255) = 0.709053$, we thus get

$$\delta_o = \frac{\bar{t}(0.00255)(1-0.00255)-0.58}{0.00255} = \frac{0.709053(1-0.00255)-0.58}{0.00255} = 49.9 < 50.$$

Hence in this example, the bound (59) gives only a slight improvement over the bound (60). The reason is that, for Bernoulli random variables with parameter close to 0.5, the Höffding's inequality is almost as good as the one given by the large deviations principle.

Acknowledgements The author "Jüri Lember" was supported by the Estonian Science Foundation Grant nr. 9288 and targeted financing project SF0180015s12. The research work of "Felipe Torres" was supported by the DFG through the SFB 878 at University of Münster.

References

1. K.S. Alexander, The rate of convergence of the mean length of the longest common subsequence. Ann. Appl. Probab. **4**(4), 1074–1082 (1994)
2. F. Bonetto, H. Matzinger, Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets. Latin Am. J. Probab. Math. **2**, 195–216 (2006)

3. J. Boutet de Monvel, Extensive simulations for longest common subsequences. Eur. Phys. J. B **7**, 293–308 (1999)
4. N. Christianini, M.W. Hahn, *Introduction to Computational Genomics* (Cambridge University Press, Cambridge, 2007)
5. V. Chvatal, D. Sankoff, Longest common subsequences of two random sequences. J. Appl. Probab. **12**, 306–315 (1975)
6. L. Devroye, G. Lugosi, L. Gyorfi, *A Probabilistic Theory of Pattern Recognition* (Springer, New York, 1996)
7. R. Durbin, S. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, 1998)
8. C. Durringer, J. Lember, H. Matzinger, Deviation from the mean in sequence comparison with a periodic sequence. ALEA **3**, 1–29 (2007)
9. C. Houdre, H. Matzinger, Fluctuations of the optimal alignment score with and asymmetric scoring function. [arXiv:math/0702036]
10. J. Lember, H. Matzinger, Standard deviation of the longest common subsequence. Ann. Probab. **37**(3), 1192–1235 (2009)
11. J. Lember, H. Matzinger, F. Torres, The rate of the convergence of the mean score in random sequence comparison. Ann. Appl. Probab. **22**(3), 1046–1058 (2012)
12. J. Lember, H. Matzinger, F. Torres, General approach to the fluctuations problem in random sequence comparison. arXiv:1211.5072 (Submitted, 2012).
13. C.-Y. Lin, F.J. Och, Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain (Association for Computational Linguistics, Stroudsburg 2004), p. 605
14. H. Matzinger, F. Torres, Fluctuation of the longest common subsequence for sequences of independent blocks. [arvix math.PR/1001.1273v3]
15. H. Matzinger, F. Torres, Random modification effect in the size of the fluctuation of the LCS of two sequences of i.i.d. blocks. [arXiv math.PR/1011.2679v2]
16. I.D. Melamed, Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons, in *Proceedings of the Third Workshop on Very Large Corpora* (Massachusetts Institute of Technology, Cambridge, 1995), pp. 184–198. http://books.google.ee/books?id=CHswHQAACAAJ
17. I.D. Melamed, Bitext maps and alignment via pattern recognition. Comput. Linguist. **25**(1), 107–130 (1999)
18. P. Pevzner, *Computational Molecular Biology*. An algorithmic approach, A Bradford Book (MIT, Cambridge, 2000)
19. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)
20. T.F. Smith, M.S. Waterman, Identification of common molecular subsequences. J. Mol. Bio. **147**, 195–197 (1981)
21. M.J. Steele, An Efron-Stein inequality for non-symmetric statistics. Ann. Stat. **14**, 753–758 (1986)
22. F. Torres, On the probabilistic longest common subsequence problem for sequences of independent blocks. Ph.D. thesis, Bielefeld University, 2009. Online at http://bieson.ub.uni-bielefeld.de/volltexte/2009/1473/
23. M.S. Waterman, Estimating statistical significance of sequence alignments. Phil. Trans. R. Soc. Lond. B **344**(1), 383–390 (1994)
24. M.S. Waterman, *Introduction to Computational Biology* (Chapman & Hall, London, 1995)
25. K. Wing Li, C.C. Yang, Automatic construction of english/chinese parallel corpora. J. Am. Soc. Inform. Sci. Tech. **54**, 730–742 (2003)