The rate of the convergence of the mean score in random sequence comparison

J. Lember*, H. Matzinger, F. Torres†

November 18, 2010

Jüri Lember, Tartu University, Institute of Mathematical Statistics
Liivi 2-513 50409, Tartu, Estonia. *E-mail address:* jyril@ut.ee
Heinrich Matzinger, Georgia Tech, School of Mathematics
Atlanta, Georgia 30332-0160, U.S.A. *E-mail address:* matzing@math.gatech.edu
Felipe Torres, University of Bielefeld, Faculty of Mathematics
Postfach 100131 - 33501 Bielefeld, Germany. *E-mail address:* ftorres@math.uni-bielefeld.de

Abstract

We consider a general class of superadditive scores measuring the similarity of two independent sequences of n i.i.d. letters from a finite alphabet. Our object of interest is the mean score by letter l_n. By subadditivity l_n is nondecreasing and converges to a limit l. We give a simple method of bounding the difference $l - l_n$ and obtaining the rate of convergence. Our result generalizes the previous result of Alexander [1], where only the special case of the longest common subsequence was considered.

Keywords. *Random sequence comparison, longest common sequence, rate of convergence.*

AMS. 60K35, 41A25, 60C05

1 Introduction

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.

*Supported by the Estonian Science Foundation Grant nr. 7553 and the German Science Foundation (DFG) through CRC 701 at Bielefeld University

†Supported by the German Science Foundation (DFG) through CRC 701 at Bielefeld University

The problem of measuring the similarity of X and Y is central in many areas of applications including computational molecular biology [7, 14, 23, 24, 25] and computational linguistics [8, 18, 20, 21]. In this paper, 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 and we denote by F and A the largest possible score and the largest possible change of score by one variable, respectively. Formally (recall that S is symmetric)

$$F := \max_{a,b \in \mathbb{A}} S(a,b), \quad A := \max_{a,b,c \in \mathbb{A}} |S(a,b) - S(a,c)|.$$

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. $1 \leq \pi_1 < \pi_2 < ... < \pi_k \leq n$ and $1 \leq \mu_1 < \mu_2 < \ldots < \mu_k \leq n$. The integer k is the number of aligned letters and $n - k$ is the number of *gaps* in the alignment. 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

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

In our general scoring scheme δ can also be positive, although usually $\delta \leq 0$ penalizing the mismatch (in this case $-\delta$ is usually called the gap penalty). We naturally assume $\delta \leq F$.

The (optimal) score of X and Y is defined to be best score over all possible alignments, i.e.

$$L_n := L(X;Y) := \max_{(\pi,\mu)} U_{(\pi,\mu)}(X,Y).$$

The alignments achieving the maximum are called *optimal*. Such a similarity criterion is most commonly used in sequence comparison [3, 14, 24, 25, 26]. When $S(a,b) = 1$ for $a = b$ and $S(a,b) = 0$ for $a \neq b$, then for $\delta = 0$ the optimal score is equal to the length of the *longest common subsequence* (LCS) of X and Y.

It is well-known that the sequence EL_n, $n = 1, 2 \ldots$ is superadditive, i.e. $EL_{n+m} \geq EL_n + EL_m$ for all $n, m \geq 1$. Hence, by Fekete's lemma the ratios $l_n := \frac{EL_n}{n}$ are nondecreasing and converge to the limit

$$l := \lim_n l_n = \sup_n l_n.$$

In fact, from Kingman's subadditivity ergodic theorem, it follows that l is also the a.s. limit of $\frac{L_n}{n}$. The limit l (which for the LCS-case is called *Chvatal-Sankoff constant*) is not known exactly even for the simplest scoring scheme and the simplest

model for X and Y, so it is usually estimated by simulations. Using McDiarmid's inequality (see (3.6)) one can estimate l_n with prescribed accuracy; to obtain confidence intervals for l, the difference $l - l_n$ should be estimated. This is the aim of the present paper.
To our best knowledge, the difference $l - l_n$ has been theoretically studied only by Alexander in [1], though there exist many numeric results on the value of l_n or its distribution in various contexts [4, 6, 9, 11, 12, 15, 16, 17, 22]. Alexander proved that in the case of the LCS, for any $C > (2+\sqrt{2})$ there exists an integer $n_o(C)$ such that

$$l - l_n \leq C\sqrt{\frac{\log n}{n}}, \quad \text{provided } n > n_o(C). \tag{1.1}$$

The bound (1.1) is independent of the common law of X and Y, and the integer $n_o(C)$ can be exactly determined. Hence the bound (1.1) can be used for the calculation of explicit confidence intervals.

Our main result is the following:

Theorem 1.1 *Let $n \in \mathbb{N}$ be even. Then, with any $c > \sqrt{A}$,*

$$l - l_n \leq c\sqrt{\frac{2}{n-1}\Big(\frac{n+1}{n-1} + \ln(n-1)\Big)} + \frac{F}{n-1}. \tag{1.2}$$

Note that by the monotonicity of l_n, the assumption on n even actually is not restrictive. In fact, Alexander's main result (Prop. 2.4 in [1]) is also proven for n even. Theorem 1.1 and its proof generalize Alexander's result in many ways:

1. Theorem 1.1 applies for a general scoring scheme, not just for the LCS. This is due to the fact that our proof is based solely on McDiarmid's large deviation equality, whilst Alexander's proof, although using also McDiarmid's inequality, is mainly based on first passage percolation techniques. Despite the fact that the percolation approach applies in many other situations rather than sequence comparison (see [2]), it is not clear whether it can be efficiently applied to our general scoring scheme. For McDiarmid's inequality, however, it makes no difference what kind of scoring is used. This gives us reasons to believe that our proof is somehow "easier" than the one in [1].

2. The proof of Theorem 1.1 relates the rate of the convergence of l_n to the cardinality of the set of partitions $\mathcal{B}_{k,n}$ (see Lemma 3.1) so that finding the good rate boils down to the good estimation of $|\mathcal{B}_{k,n}|$. The bound (1.2) corresponds to a particular estimate of $|\mathcal{B}_{k,n}|$, any better estimate would give a sharper bound and, probably, also a faster rate. In a sense, the cardinality $|\mathcal{B}_{k,n}|$ could be interpreted as the complexity of the model and the relation between the rate of convergence and the complexity of the model is a well-known fact in statistics (see e.g. [5]).

3. When applied to the LCS, our bound (1.2) is sharper than (1.1). Indeed, for the case of LCS the constants A and F in (1.2) can be taken equal to one and the smaller constants make the difference. In other words, for the case of LCS

both results yield the rate $C\sqrt{\frac{\ln n}{n}}$, but the constant C is different ($C > 3.42$ in Alexander's result and $c > \sqrt{2}$ in ours).

We can easily compare (1.2) and (1.1) by comparing the decay of the two following functions:

$$\begin{aligned} R : \{1, \ldots, 10000\} &\rightarrow \mathbb{R}^+ \\ R(n) &= (3.42 + 0.1)\sqrt{\frac{\ln n}{n}} \qquad (1.3) \\ Q_F : \{1, \ldots, 10000\} \times \{0.1, \ldots, 2\} &\rightarrow \mathbb{R}^+ \\ Q_F(n, A) &= \sqrt{A}\,\sqrt{\frac{2}{n-1}\Big(\frac{n+1}{n-1} + \ln(n-1)\Big) + \frac{F}{n-1}} \qquad (1.4) \end{aligned}$$

In figure 1, we can see the improved bound (1.2) given by function (1.4) (changes of A are represented in colours, $F = 1$) over the bound by Alexander (1.1) given by function (1.3) (in black). Note that the dark blue curve corresponds to $A = 0.1$ whilst the light violet curve to $A = 2$ (namely, the colour gets lighter as A increases). The curve in green corresponds to the case $A = 1$ (ie, our bound for the LCS case).

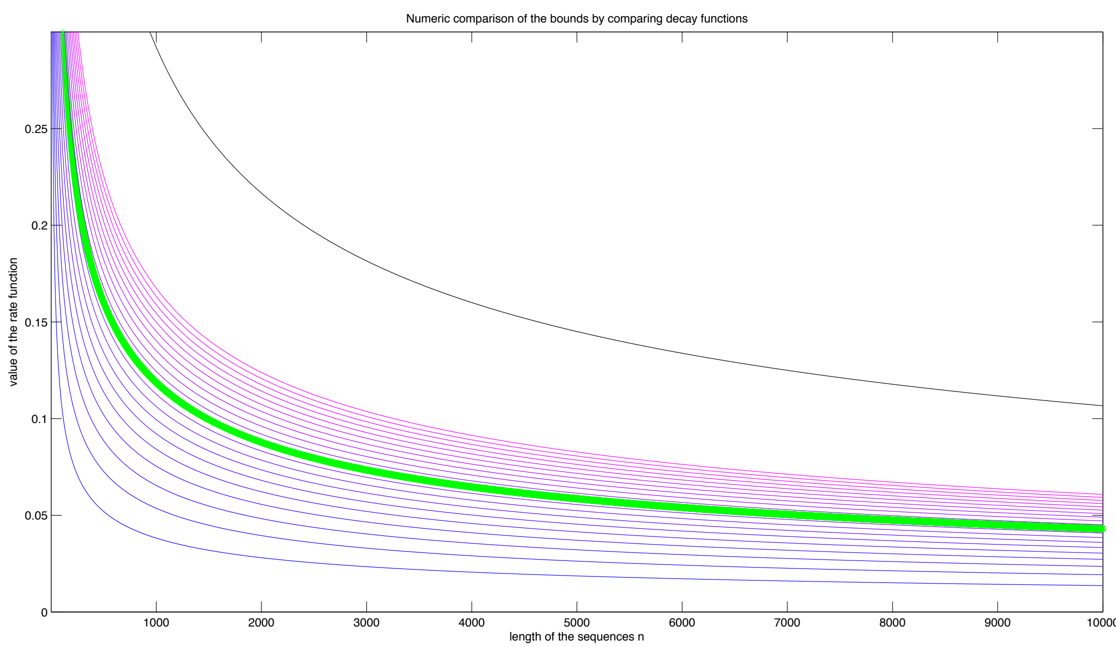

Figure 1: comparison of the bounds (1.2) and (1.1) through the functions (1.4) and (1.3), respectively.

2 Confidence bounds for l

Suppose that k samples of $X^i = X_1^i, \ldots, X_n^i$ and $Y^i = Y_1^i, \ldots, Y_n^i$, $i = 1, \ldots, N$ are generated. Let L_n^i be the score of the i-th sample. Thus $EL_n^i = n\, l_n$. By McDiarmid's inequality (see (3.5) below), for every $\rho > 0$

$$P\left(\frac{1}{kn}\sum_{i=1}^{k} L_n^i - l_n < -\rho\right) = P\left(\sum_{i=1}^{k} L_n^i - kn\, l_n < -kn\rho\right) \leq \exp\left[-\frac{\rho^2 kn}{A^2}\right]. \quad (2.1)$$

Let

$$\bar{L}_n := \frac{1}{kn}\sum_{i=1}^{k} L_n^i.$$

If n is even, by (1.2) and (1.4) we have that $l \leq l_n + Q_F(n, A)$ and then

$$P(\bar{L}_n + \rho + Q_F(n, A) \geq l) \geq P(\bar{L}_n + \rho \geq l_n) = P(\bar{L}_n - l_n \geq -\rho) \geq 1 - \exp\left[-\frac{\rho^2 kn}{A^2}\right]. \quad (2.2)$$

Now, given $\varepsilon > 0$, choose $\rho = \rho(n, \varepsilon)$ so that the right hand side in the last inequality is equal to $1 - \varepsilon$:

$$\rho(n, \varepsilon) = A\sqrt{\frac{\ln(1/\varepsilon)}{kn}}.$$

So, with probability $1 - \varepsilon$, we obtain one side confidence interval as follows:

$$l \leq \bar{L}_n + Q_F(n, A) + A\sqrt{\frac{\ln(1/\varepsilon)}{kn}}. \quad (2.3)$$

In statistical learning, the inequalities of type (2.3) are known as PAC inequality (probably almost correct inequalities). The two-sided confidence bounds are, with probability $1 - \varepsilon$, as follows:

$$\bar{L}_n - A\sqrt{\frac{\ln(2/\varepsilon)}{kn}} \leq l \leq \bar{L}_n + Q_F(n, A) + A\sqrt{\frac{\ln(2/\varepsilon)}{kn}}. \quad (2.4)$$

The bounds in (2.4) suggest to use the estimate

$$\hat{l}_n := \bar{L}_n + \frac{Q_F(n, A)}{2}$$

so that the confidence bounds for this estimate are

$$P\left(|\hat{l}_n - l| \leq A\sqrt{\frac{\ln(2/\varepsilon)}{kn}} + \frac{Q_F(n, A)}{2}\right) \geq 1 - \varepsilon. \quad (2.5)$$

Alexander [1] obtained, for $n = 100000$, $k = 2$ and $A = F = 1$ (for the LCS case), the following bounds:

$$P(|\hat{l}_n - l| \leq 0.0264) \geq 0.95. \quad (2.6)$$

By using (2.5) and (1.4) we obtain, for $n = 100000$, $k = 2$ and $A = F = 1$ (for the LCS case), the following bounds:

$$P(|\hat{l}_n - l| \leq 0.0122) \geq 0.95. \tag{2.7}$$

It is clear that (2.7) is sharper than (2.6). To our best knowledge, the best previous lower and upper bounds for l, in the LCS context for $\mathbb{A} = \{0, 1\}$, were due to Dancik [10], Dancik and Paterson [11, 22] (0.773911 and 0.837623, respectivley) and Lueker [19] (0.788071 and 0.826280, respectively).

Remark: The inequality (2.3) confirms the well-known fact that it is better to generate one sample of length kn rather than k samples of length n. Indeed, with one sample of length kn, the inequality (2.3) becames

$$l \leq \bar{L}_n + Q_F(kn, A) + A\sqrt{\frac{\ln(1/\varepsilon)}{kn}} \tag{2.8}$$

and since $Q_F(kn, A) < Q_F(n, A)$, the bounds get narrower.

3 Proof of the main result

3.1 The set of partitions $\mathcal{B}_{k,n}$

In this section, we shall consider the sequences X and Y with length kn where k, n are nonnegative integers. Let (π, μ) be an arbitrary alignment of X and Y. Let $\nu = (\nu_1, \ldots, \nu_{r+1})$ and $\tau = (\tau_1, \ldots, \tau_{r+1})$ be vectors satisfying

$$1 = \nu_1 \leq \nu_2 \leq \ldots \leq \nu_r \leq \nu_{r+1} = kn + 1, \quad 1 = \tau_1 \leq \tau_2 \leq \ldots \leq \tau_r \leq \tau_{r+1} = kn + 1. \tag{3.1}$$

We say that the pair (ν, τ) forms a r**-partition of the alignment** (π, μ) if for any $j = 1, \ldots, r$, the following conditions are simultaneously satisfied:
1) if, for some $i = 1, \ldots k$, it holds that $\nu_j \leq \pi_i < \nu_{j+1}$, then $\tau_j \leq \mu_i < \tau_{j+1}$;
2) if, for some $i = 1, \ldots k$, it holds that $\tau_j \leq \mu_i < \tau_{j+1}$, then $\nu_j \leq \pi_i < \nu_{j+1}$.

Thus (ν, τ) is a r-partition, if the sequences X and Y can be partitioned into r pieces

$$(X_1, \ldots, X_{\nu_2 - 1}), (X_{\nu_2}, \ldots, X_{\nu_3 - 1}), \ldots, (X_{\nu_r}, \ldots, X_{kn})$$
$$(Y_1, \ldots, Y_{\tau_2 - 1}), (Y_{\tau_2}, \ldots, Y_{\tau_3 - 1}), \ldots, (Y_{\tau_r}, \ldots, Y_{kn})$$

such that the alignment (π, μ) aligns a piece $(X_{\nu_j}, \ldots, X_{\nu_{j+1}-1})$ with the piece $(Y_{\tau_j}, \ldots, Y_{\tau_{j+1}-1})$, where $j = 1, \ldots r$. It is important to note that the pieces might be empty, i.e. it might be that $\nu_j = \nu_{j+1}$ (or $\tau_j = \tau_{j+1}$), meaning that $(\tau_j, \ldots, \tau_{j+1} - 1)$ cannot contain any elements of μ, otherwise the requirement 2) would be violated (or $(\mu_j, \ldots, \mu_{j+1} - 1)$ cannot contain any elements of τ, otherwise the requirement 1) would be violated). Hence, if for a partition a piece of X is empty, then the corresponding piece of Y cannot have any aligned letter.

The following observation shows that any alignment of X and Y can be partitioned into r pieces such that $k \le r \le \lceil \frac{2kn}{2n-1} \rceil$ and such that the sum of the lengths of aligned pairs in each partition is always at most $2n$. We believe that the idea of the proof as well as the meaning of the partition becomes transparent by an example.

Example. Let $n = 3, k = 4$. Let $\pi = (1, 5, 6, 9, 10, 12)$ and $\mu = (2, 3, 4, 6, 9, 10)$. The alignment (π, μ) can be represented as follows

X	-	1	2	3	4	5	6	7	8	-	9	-	-	10	11	12	-	-
Y	1	2	-	-	-	3	4	-	-	5	6	7	8	9	-	10	11	12

The table above indicates that X_1 is aligned with Y_2, X_5 is aligned with Y_3 and so on; the rest of the letters are unaligned, so we say that they are aligned with gaps. In the table, there are two types of columns: the columns with two figures (aligned pairs) and the columns with one figure (unaligned pairs). Let $u_i \in \{1, 2\}$ be the number of figures in the i-th column, and let $s_j = u_1 + \cdots + u_j$ be the corresponding cumulative sum. To get an r-partition proceed as follows: start from the beginning of the table (most left position) and find j such that $s_j = 2n$. Since the cumulative sum increases by one or two, such a j might not exist. In this case find j such that $s_j = 2n - 1$. In the present example $n = 3$, thus we are looking for j such that $s_j = 6$. Such a j is 5. The first five columns thus form the first part of the partition and there are exactly $2n = 6$ elements in the first part (those elements are X_1, X_2, X_3, X_4, Y_1 and Y_2). Now disregard the first five columns from the table and start the same procedure afresh. Then the second part is obtained and so on. In the following table the vertical lines indicate the different parts obtained by the aforementioned procedure: the first two parts have six elements, the third and fourth has five elements and the last part consists of one element:

X	- 1 2 3 4	5 6 7 8	- 9 - -	10 11 12 -	-
Y	1 2 - - -	3 4 - -	5 6 7 8	9 - 10 11	12

From the table, we read the corresponding pieces from the X-side: $(1, 4), (5, 8), (9, 9), (10, 12), \emptyset$ as well as the ones from the Y-side: $(1, 2), (3, 4), (5, 8), (9, 11), (12, 12)$. The corresponding vectors ν and τ are thus $\nu = (1, 5, 9, 10, 13, 13)$, $\tau = (1, 3, 5, 9, 12, 13)$. The number of parts in such a partition is clearly at least k (corresponding to the case that all pairs sum up to $2n$) and at most $\lceil \frac{2kn}{2n-1} \rceil$ (corresponding to the case that all pairs except the last one sum up to $2n-1$). In our example is $r = 5 = \lceil \frac{24}{5} \rceil$. Now, it is clear that the following claim holds.

Claim 3.1 *Let X, Y be sequences of length kn and let (π, μ) be an arbitrary alignment of X and Y. Then there exist an integer r such that $k \le r \le \lceil \frac{2kn}{2n-1} \rceil$ and an r-partition (ν, τ) of (π, μ) such that for every $j = 1, \ldots, r-1$, it holds*

$$(\nu_{j+1} - \nu_j) + (\tau_{j+1} - \tau_j) \in \{2n, 2n-1\} \quad \textit{and} \quad (\nu_{r+1} - \nu_r) + (\tau_{r+1} - \tau_r) \le 2n. \tag{3.2}$$

Let, for every r, $\mathcal{B}^r_{k,n}$ be the set of vectors $\nu = (\nu_1, \ldots, \nu_{r+1})$ and $\tau = (\tau_1, \ldots, \tau_{r+1})$ satisfying (3.1) and (3.2). Let

$$\mathcal{B}_{k,n} = \bigcup_{r=k}^{\lceil \frac{2kn}{2n-1} \rceil} \mathcal{B}^r_{k,n}.$$

We shall call the elements of $\mathcal{B}_{k,n}$ as the partitions. For every partition $(\nu, \tau) \in \mathcal{B}^r_{k,n}$, we define

$$L_{kn}(\nu, \tau) := \sum_{i=1}^{r} L(X_{\nu_j}, \ldots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \ldots, Y_{\tau_{j+1}-1}),$$

where $L(X_{\nu_j}, \ldots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \ldots, X_{\tau_{j+1}-1})$ is the optimal score between $X_{\nu_j}, \ldots, X_{\nu_{j+1}-1}$ and $Y_{\tau_j}, \ldots, Y_{\tau_{j+1}-1}$. The key observation is the following: if (π, μ) is optimal for X, Y and (ν, τ) is a r-partition of (π, μ), then $L_{kn} = L_{kn}(\nu, \tau)$. By Claim 3.1, every alignment, including the optimal one, has at least one partition from the set $\mathcal{B}_{k,n}$, hence it follows that

$$L_{kn} = \max_{(\nu,\tau) \in \mathcal{B}_{k,n}} L_{kn}(\nu, \tau). \tag{3.3}$$

Claim 3.2 *For every r-partition $(\nu, \tau) \in \mathcal{B}_{k,n}$,*

$$E\big(L_{kn}(\nu, \tau)\big) \leq \frac{r}{2} EL_{2n} \leq \frac{1}{2} \left\lceil \frac{2kn}{2n-1} \right\rceil EL_{2n}. \tag{3.4}$$

Proof. Let $(\nu, \tau) \in \mathcal{B}^r_{k,n}$ with $r \leq \lceil \frac{2nk}{2n-1} \rceil$. Let j be such that $(\nu_{j+1} - \nu_j) + (\tau_{j+1} - \tau_j) = 2n$. Thus, there exists an integer $u \in \{-n, \ldots, n\}$ such that $\nu_{j+1} - \nu_j = n - u$ and $\tau_{j+1} - \tau_j = n + u$. Since $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ are i.i.d., we have

$$E\big(L(X_{\nu_j}, \ldots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \ldots, Y_{\tau_{j+1}-1})\big) = E\big(L(X_1, \ldots, X_{n-u}; Y_1, \ldots, Y_{n+u})\big) =$$
$$E\big(L(X_{n-u+1}, \ldots, X_{2n}; Y_{n+u+1}, \ldots, Y_{2n})\big) \leq \frac{1}{2} E\big(L(X_1, \ldots, X_{2n}; Y_1, \ldots, Y_{2n})\big) = \frac{1}{2} EL_{2n}.$$

The last inequality follows from the superadditivity:

$$\begin{aligned} L(X_1, \ldots, X_{n-u}; Y_1, \ldots, Y_{n+u}) + L(X_{n-u+1}, \ldots, X_{2n}; Y_{n+u+1}, \ldots, Y_{2n}) \\ \leq L(X_1, \ldots, X_{2n}; Y_1, \ldots, Y_{2n}). \end{aligned}$$

If $(\nu_{j+1} - \nu_j) + (\tau_{j+1} - \tau_j) < 2n$, then by the same argument

$$E\big(L(X_{\nu_j}, \ldots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \ldots, Y_{\tau_{j+1}-1})\big) \leq E\big(L(X_1, \ldots, X_{n-u}; Y_1, \ldots, Y_{n+u})\big) \leq \frac{1}{2} EL_{2n}.$$

Hence the first inequality in (3.4) follows. The second inequality follows from the condition $r \leq \lceil \frac{2nk}{2n-1} \rceil$. ∎

3.2 The size of $\mathcal{B}_{k,n}$ and the rate of convergence

In the following we prove the main theoretical result that links the rate of the convergence to the rate at which the number of elements in $|\mathcal{B}_{k,n}|$ grows as k increases. Our proof is entirely based on McDiarmid's inequality, so let us recall it for the sake of completeness: 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 A. Then for any $\Delta > 0$,

$$P\Big(f(Z_1, \ldots, Z_{2m}) - Ef(Z_1, \ldots, Z_{2m}) > \Delta\Big) \leq \exp\left[-\frac{\Delta^2}{mA^2}\right]. \tag{3.5}$$

For the proof, we refer [13]. We apply (3.5) with L in the role of f to the independent (but not necessarily identically distributed) random variables $X_1, \ldots, X_m, Y_1, \ldots, Y_m$. It is easy but important to see that independently of the value of δ, changing one random variable changes the score at most by A so that in our case (3.5) is

$$P\Big(L_m - EL_m > \Delta\Big) \leq \exp\left[-\frac{\Delta^2}{mA^2}\right]. \tag{3.6}$$

Lemma 3.1 *Suppose that for any n and for k big enough*

$$|\mathcal{B}_{k,n}| \leq \exp[\big(\psi(n) + o(k)\big)kn], \tag{3.7}$$

where $\psi(n)$ does not depend on k. Let $u(n) > A\sqrt{\psi(n)}$. Then

$$l - l_{2n} \leq u(n) + \frac{l_{2n}}{2n-1} \leq u(n) + \frac{l}{2n-1} \leq u(n) + \frac{F}{2n-1}. \tag{3.8}$$

Proof. Let $(\nu, \tau) \in \mathcal{B}_{k,n}$. Recall (3.4). Thus, from (3.6), we get that for any $\rho > 0$,

$$P\Big(L_{kn}(\nu,\tau) - \frac{1}{2}\left\lceil\frac{2kn}{2n-1}\right\rceil EL_{2n} > \rho kn\Big) \leq P\Big(L_{kn}(\nu,\tau) - E\big(L_{kn}(\nu,\tau)\big)\rho kn\Big) \leq \exp\left[-\frac{\rho^2 kn}{A^2}\right]. \tag{3.9}$$

From (3.3) and (3.7) it now follows that, for big k

$$\begin{aligned} P\left(\frac{L_{kn}}{kn} - \frac{1}{k}\left\lceil\frac{2kn}{2n-1}\right\rceil l_{2n} > \rho\right) &\leq \sum_{(\nu,\tau)\in\mathcal{B}_{k,n}} P\Big(L_{kn}(\nu,\tau) - \frac{1}{2}\left\lceil\frac{2kn}{2n-1}\right\rceil EL_{2n} > \rho kn\Big) \\ &\leq |\mathcal{B}_{k,n}| \exp\left[-\frac{\rho^2 kn}{A^2}\right] \leq \exp\left[\Big(\psi(n) + o(k) - \Big(\frac{\rho}{A}\Big)^2\Big)kn\right]. \end{aligned}$$

We consider n fixed and let k go to infinity. If $u(n) > A\sqrt{\psi(n)}$, then there exists $K(n) < \infty$ so that for every $k > K(n)$,

$$\psi(n) + o(k) - \left(\frac{u(n)}{A}\right)^2 < \frac{1}{2}\Big(\psi(n) - \left(\frac{u(n)}{A}\right)^2\Big).$$

Hence, replacing in the inequalities above ρ with $u(n)$, we obtain for every $k > K(n)$,

$$P\left(\frac{L_{kn}}{kn} - \frac{1}{k}\left\lceil\frac{2kn}{2n-1}\right\rceil l_{2n} > u(n)\right) \leq \exp\left[\frac{1}{2}\left(\psi(n) - \left(\frac{u(n)}{A}\right)^2\right)nk\right] = \exp[-d_n k], \tag{3.10}$$

where

$$d_n := \left(\left(\frac{u(n)}{A}\right)^2 - \psi(n)\right)n > 0.$$

Now recall the assumption that $\delta \leq F$. Hence for any n and k, the random variable $\frac{L_{kn}}{kn}$ is bounded by F. From (3.10), it thus follows that for any k

$$E\Big(\frac{L_{kn}}{kn}\Big) = l_{kn} \leq \frac{1}{k}\left\lceil\frac{2kn}{2n-1}\right\rceil l_{2n} + u(n) + F\exp[-d_n k].$$

Since $l_{kn} \to l$ as $k \to \infty$ and

$$\frac{1}{k}\left\lceil\frac{2kn}{2n-1}\right\rceil \leq \frac{2n}{2n-1} + \frac{1}{k},$$

we obtain that for any n,

$$l \leq \left(\frac{2n}{2n-1}\right) l_{2n} + u(n) = l_{2n}\left(1 + \frac{1}{2n-1}\right) + u(n).$$

■

The proof of Theorem 1.1. From Lemma 3.1, it follows that to obtain a bound to $l - l_n$, a suitable estimator of $|\mathcal{B}_{k,n}|$ satisfying (3.7) should be found.
Let us estimate $|\mathcal{B}^r_{k,n}|$. The number of parts in the X side is bounded above by the number of combination with repetition from $nk+1$ by $r-1$. The repetitions allow empty parts. When the size of a part in X-side is m, then, except from the last part, the size of the corresponding part on Y side has two possibilities: $2n-1-m$ or $2n-m$. Hence to any r-partition of X-size corresponds at most $2^{r-1}2n$ options in Y side. In the following we use the fact that the number of combination with repetition from $nk+1$ by $r-1$ is $\binom{nk+r-1}{r-1}$ and for any non-negative integers $a > b$ it holds

$$\binom{a}{b} \leq \exp\left[h_e\left(\frac{b}{a}\right)a\right],$$

where $h_e(q) := -q\ln q - (1-q)\ln(1-q)$ is the binary entropy function. Since $r \leq \lceil\frac{2nk}{2n-1}\rceil$ implies that $r - 1 \leq \frac{2nk}{2n-1}$, we thus have for $n \geq 2$

$$\begin{aligned}
|\mathcal{B}^r_{k,n}| &\leq (2^{r-1}2n)\binom{nk+r-1}{r-1} \\
&\leq \exp\left[(r-1)(\ln 2) + \ln(2n) + h_e\left(\frac{r-1}{nk+r-1}\right)(nk+r-1)\right] \\
&\leq \exp\left[\left(\frac{\ln 4}{2n-1} + \frac{\ln(2n)}{nk} + h_e\left(\frac{r-1}{nk+r-1}\right)\left(1 + \frac{2}{2n-1}\right)\right)nk\right] \\
&\leq \exp\left[\left(\frac{\ln 4}{2n-1} + \frac{\ln(2n)}{nk} + h_e\left(\frac{2}{2n+1}\right)\left(\frac{2n+1}{2n-1}\right)\right)nk\right].
\end{aligned}$$

The last inequality follows from the inequalities

$$\frac{r-1}{nk+r-1} \leq \frac{\frac{2nk}{2n-1}}{nk+\frac{2nk}{2n-1}} = \frac{2}{2n+1}$$

so that if $n \geq 2$, then $\frac{2}{2n+1} \leq 0.5$ and

$$h_e\left(\frac{r-1}{nk+r-1}\right) \leq h_e\left(\frac{2}{2n+1}\right).$$

Hence

$$\begin{aligned}
|\mathcal{B}_{k,n}| &\leq \left(\frac{2nk}{2n-1} - k + 2\right) \exp\left[\left(\frac{\ln 4}{2n-1} + \frac{\ln(2n)}{nk} + h_e\left(\frac{2}{2n+1}\right)\left(\frac{2n+1}{2n-1}\right)\right) nk\right] \\
&= \left(\frac{k}{2n-1} + 2\right) \exp\left[\left(\frac{\ln 4}{2n-1} + \frac{\ln(2n)}{nk} + h_e\left(\frac{2}{2n+1}\right)\left(\frac{2n+1}{2n-1}\right)\right) nk\right] \\
&= \exp\left[\ln\left(\frac{k}{2n-1} + 2\right) + \left(\frac{\ln 4}{2n-1} + \frac{\ln(2n)}{nk} + h_e\left(\frac{2}{2n+1}\right)\left(\frac{2n+1}{2n-1}\right)\right) nk\right] \\
&= \exp\left[\left(\frac{\ln\left(\frac{k}{2n-1} + 2\right) + \ln(2n)}{nk} + \frac{\ln 4}{2n-1} + h_e\left(\frac{2}{2n+1}\right)\left(\frac{2n+1}{2n-1}\right)\right) nk\right]
\end{aligned}$$

$$\begin{aligned}
&= \exp\left[\left(o(k) + \frac{\ln 4}{2n-1} + h_e\left(\frac{2}{2n+1}\right)\left(\frac{2n+1}{2n-1}\right)\right) nk\right] \\
&\leq \exp\left[\left(o(k) + \frac{2}{2n-1}\left(\frac{2n+1}{2n-1} + \ln(2n-1)\right)\right) nk\right],
\end{aligned}$$

where the last inequality follows from the inequality

$$h_e\left(\frac{2}{2n+1}\right) \leq \frac{2}{2n+1}\left(\frac{2n+1}{2n-1} + \ln\left(\frac{2n-1}{2}\right)\right). \tag{3.11}$$

Hence (3.7) holds with

$$\psi(n) = \frac{2}{2n-1}\left(\frac{2n+1}{2n-1} + \ln(2n-1)\right).$$

The inequality (1.2) now follows from Lemma 3.1. ■

Acknowledgments

The authors would like to thank the support of the German Science Foundation (DFG) through the Collavorative Research Center 701 "Spectral Structures and Topological Methods in Mathematics" (CRC 701) at Bielefeld University and the support of Barbara Gentz with the research stay of J. Lember at the CRC 701. Additionally, F. Torres would like to thank the partial support of the International Graduate College "Stochastics and Real World Models" (IRTG 1132) at Bielefeld University.

References

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

[2] Kenneth S. Alexander. Approximation of subadditive functions and convergence rates in limiting-shape results. *Ann. Appl. Probab.*, 25(1):30–55, 1997.

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

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

[5] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[6] H. S. Booth, S.F. MacNamara, O.M. Nielsen, and S.R. Wilson. An iterative approach to determining the length of the longest common subsequence of two strings. *Methodology Comput. Appl. Probab.*, 6(4):401–421, 2004.

[7] N. Christianini and M. W. Hahn. *Introduction to Computational Genomics*. Cambridge University Press, 2007.

[8] Kar Wing Li Christopher C. Yang. Automatic construction of english/chinese parallel corpora. *Journal of the American Society for Information Science and Technology*, 54:730–742, 2003.

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

[10] Vlado Dancik. *Expected length of longest common subsequences*. PhD dissertation, Department of Computer Science, University of Warwick, 1994.

[11] Vlado Dancik and Mike Paterson. Upper bounds for the expected length of a longest common subsequence of two binary sequences. *Random Structures Algorithms*, 6(4):449–458, 1995.

[12] Joseph G. Deken. Some limit results for longest common subsequences. *Discrete Math.*, 26(1):17–31, 1979.

[13] Luc Devroe, Gabor Lugosi, and Laszlo Gyorfi. *A probabilistic theory of pattern recognition*. Springer-Verlag, New York, 1996.

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

[15] J. Fu and W. Lou. Distribution of the length of the longest common subsequence of two multi-state biological sequences. *Journal of Statistical Planning and Inference*, 138:3605–3615, 2008.

[16] Raphael Hauser, Heinrich Matzinger, and Clement Durringer. Approximation to the mean curve in the lcs-problem. *Stochastic Proc. and Appl.*, 118(4):629–648, 2008.

[17] Marcos A. Kiwi, Martin Loebl, and Jirí Matousek. Expected length of the longest common subsequence for large alphabets. *Advances in Mathematics*, 197(2):480–498, 2005.

[18] Chin-Yew Lin and Franz Josef 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*, page 605, 2004.

[19] George S. Lueker. Improved bounds on the average length of longest common subsequences. *J. ACM*, 56(3):1–38, 2009.

[20] 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*, 1995.

[21] I. Dan Melamed. Bitext maps and alignment via pattern recognition. *Comput. Linguist.*, 25(1):107–130, 1999.

[22] Mike Paterson and Vlado Dancik. Longest common subsequences. In *Mathematical foundations of computer science 1994 (Kosice, 1994)*, volume 841 of *Lecture Notes in Comput. Sci.*, pages 127–142. Springer, Berlin, 1994.

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

[24] Temple F. Smith and Michael S. Waterman. Identification of common molecular subsequences. *J. Mol. Bio.*, 147:195–197, 1981.

[25] Michael S. Waterman. *Introduction to Computational Biology.* Chapman & Hall, 1995.

[26] Michael S. Waterman and M. Vingron. Sequence comparison significance and Poisson approximation. *Statistical Science*, 9(3):367–381, 1994.