The Annals of Applied Probability
2012, Vol. 22, No. 3, 1046–1058
DOI: 10.1214/11-AAP778

THE RATE OF THE CONVERGENCE OF THE MEAN SCORE IN RANDOM SEQUENCE COMPARISON

BY JÜRI LEMBER[1], HEINRICH MATZINGER[2] AND FELIPE TORRES[2]

Tartu University, Georgia Tech and Bielefeld University

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 [*Ann. Appl. Probab.* **4** (1994) 1074–1082], where only the special case of the longest common subsequence was considered.

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.

The problem of measuring the similarity of X and Y is central in many areas of applications including computational molecular biology [8, 13, 22–24] and computational linguistics [17, 19, 20, 26]. 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)

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

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, that is, $1 \le \pi_1 < \pi_2 < \cdots < \pi_k \le n$ and $1 \le \mu_1 < \mu_2 < \cdots < \mu_k \le n$. 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

Received November 2010; revised March 2011.

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

[2]Supported by the German Science Foundation (DFG) through CRC 701 at Bielefeld University.

MSC2010 subject classifications. 60K35, 41A25, 60C05.

Key words and phrases. Random sequence comparison, longest common sequence, rate of convergence.

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 $\delta \le 0$ penalizing the mismatch (in this case $-\delta$ is usually called the gap penalty). We naturally assume $\delta \le F$.

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

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

The alignments achieving the maximum are called *optimal*. Such a similarity criterion is most commonly used in sequence comparison [3, 13, 23–25]. When $S(a, b) = 1$ for $a = b$ and $S(a, b) = 0$ for $a \ne b$, 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, that is, $EL_{n+m} \ge EL_n + EL_m$ for all $n, m \ge 1$. 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 subadditive 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 [1], though there exist many numeric results on the value of l_n or its distribution in various contexts [4, 6, 7, 10, 11, 14–16, 21]. Alexander proved that in the case of the LCS, for any $C > (2 + \sqrt{2})$ there exists an integer $n_o(C)$ such that

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

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*

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

Note that by the monotonicity of l_n, the assumption on n even actually is not restrictive. In fact, Alexander's main result ([1], Proposition 2.4) 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, while 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 $\sqrt{2}$ in ours).

For simplicity in the writing that follows, let us define

$$\begin{aligned} Q_F &: \{1, 2, 3, \ldots\} \times \mathbb{R}^+ \to \mathbb{R}^+, \\ Q_F(n, A) &:= A\sqrt{\frac{2}{n-1}\left(\frac{n+1}{n-1} + \ln(n-1)\right)} + \frac{F}{n-1}. \end{aligned} \tag{1.3}$$

2. Confidence bounds for l. 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 be the score of the ith sample. Thus $EL_n^i = nl_n$. By McDiarmid's inequality [see (3.5) below], for every $\rho > 0$

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

Let

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

If n is even, by (1.2) and (1.3) we have that $l \le l_n + Q_F(n, A)$ and then

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

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-sided confidence interval as

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

The two-sided confidence bounds are, with probability $1 - \varepsilon$,

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

The bounds in (2.4) suggest using the estimate

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

so that the confidence bounds for this estimate are

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

Alexander [1] obtained, for $n = 100{,}000$, $k = 2$ and $A = F = 1$ (for the LCS case), the following bounds:

$$P(|\hat{l}_n - l| \le 0.0264) \ge 0.95. \tag{2.6}$$

By using (2.5) and (1.3) we obtain, for $n = 100{,}000$, $k = 2$ and $A = F = 1$ (for the LCS case), the following bounds:

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

It is clear that (2.7) is sharper than (2.6). To the best of our knowledge, the best previous confidence intervals for l, in the LCS context for $\mathbb{A} = \{0, 1\}$, were due to Dancik [9], Dancik and Paterson [10, 21] given by $[0.773911, 0.837623]$ and Lueker [18] given by $[0.788071, 0.826280]$.

REMARK 2.1. 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, inequality (2.3) becomes

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

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

$$\begin{aligned} &1 = \nu_1 \le \nu_2 \le \cdots \le \nu_r \le \nu_{r+1} = kn + 1, \\ &1 = \tau_1 \le \tau_2 \le \cdots \le \tau_r \le \tau_{r+1} = kn + 1. \end{aligned} \tag{3.1}$$

We say that the pair (ν, τ) forms an *r-partition of the alignment* (π, μ) if for every $j = 1, \ldots, r$, the following conditions are simultaneously satisfied:

(1) if, for some $i = 1, \ldots, k$, it holds that $\nu_j \le \pi_i < \nu_{j+1}$, then $\tau_j \le \mu_i < \tau_{j+1}$;

(2) if, for some $i = 1, \ldots, k$, it holds that $\tau_j \le \mu_i < \tau_{j+1}$, then $\nu_j \le \pi_i < \nu_{j+1}$.

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

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

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, that is, 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 in each partition there are always at most $2n$ aligned pairs. We believe that the idea of the proof as well as the meaning of the partition becomes transparent by an example.

EXAMPLE 3.1. 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

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 ith 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 (left most 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 have 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: $(1, 4), (5, 8), (9, 9), (10, 12), \varnothing$ 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 $j = 1, \dots, r-1$, we have*

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

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

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

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

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

where $L(X_{\nu_j}, \dots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \dots, X_{\tau_{j+1}-1})$ is the optimal score between $X_{\nu_j}, \dots, X_{\nu_{j+1}-1}$ and $Y_{\tau_j}, \dots, Y_{\tau_{j+1}-1}$. The key observation is the following: if (π, μ) is optimal for X, Y and (ν, τ) is an 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 $(\nu, \tau) \in \mathcal{B}_{k,n}$,*

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

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

$$\begin{aligned} &E(L(X_{\nu_j}, \dots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \dots, Y_{\tau_{j+1}-1})) \\ &\quad = E(L(X_1, \dots, X_{n-u}; Y_1, \dots, Y_{n+u})), \\ &E(L(X_{n-u+1}, \dots, X_{2n}; Y_{n+u+1}, \dots, Y_{2n})) \\ &\quad \le \tfrac{1}{2} E(L(X_1, \dots, X_{2n}; Y_1, \dots, Y_{2n})) \\ &\quad = \tfrac{1}{2} E L_{2n}. \end{aligned}$$

The last inequality follows from superadditivity,

$$L(X_1,\dots,X_{n-u};Y_1,\dots,Y_{n+u})+L(X_{n-u+1},\dots,X_{2n};Y_{n+u+1},\dots,Y_{2n})$$
$$\le L(X_1,\dots,X_{2n};Y_1,\dots,Y_{2n}).$$

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

$$E(L(X_{\nu_j},\dots,X_{\nu_{j+1}-1};Y_{\tau_j},\dots,Y_{\tau_{j+1}-1}))\le E(L(X_1,\dots,X_{n-u};Y_1,\dots,Y_{n+u}))$$
$$\le \tfrac{1}{2}EL_{2n}.$$

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

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 $Z_1,\dots,Z_{2m}$ be independent random variables and $f(Z_1,\dots,Z_{2m})$ be a function so that changing one variable changes the value at most A. Then for any $\Delta>0$,

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

For the proof, we refer to [12]. We apply (3.5) with L in the role of f to the independent (but not necessarily identically distributed) random variables $X_1,\dots,X_m,Y_1,\dots,Y_m$. 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

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

LEMMA 3.1. *Suppose that for every n and k*

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

where $\psi(n)$ does not depend on k and for every n we have that $a_{n,k}\to0$ as $k\to\infty$. Let $u(n)>A\sqrt{\psi(n)}$. Then

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

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

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

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

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

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 $k > K(n)$,

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

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

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

where

$$d_n := \frac{1}{2}\left(\left(\frac{u(n)}{A}\right)^2 - \psi(n)\right)n > 0.$$

Now recall the assumption that $\delta \le F$. 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 $k > K(n)$

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

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

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

we obtain that for any n,

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

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 nonnegative integers $a > b$ we have

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

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

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

The last inequality follows from the inequalities

$$\frac{r-1}{nk+r-1} \le \frac{2nk/(2n-1)}{nk+2nk/(2n-1)} = \frac{2}{2n+1}$$

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

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

Hence, with

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

$$
\begin{aligned}
&= \exp\Bigg[\ln\Big(\frac{k}{2n-1}+2\Big) \\
&\qquad + \Big(\frac{\ln 4}{2n-1} + \frac{\ln(2n)}{nk} + h_e\Big(\frac{2}{2n+1}\Big)\Big(\frac{2n+1}{2n-1}\Big)\Big)nk\Bigg] \\
&= \exp\Bigg[\Big(\frac{\ln(k/(2n-1)+2)+\ln(2n)}{nk} \\
&\qquad + \frac{\ln 4}{2n-1} + h_e\Big(\frac{2}{2n+1}\Big)\Big(\frac{2n+1}{2n-1}\Big)\Big)nk\Bigg] \\
&= \exp\Bigg[\Big(a_{n,k} + \frac{\ln 4}{2n-1} + h_e\Big(\frac{2}{2n+1}\Big)\Big(\frac{2n+1}{2n-1}\Big)\Big)nk\Bigg] \\
&\le \exp\Bigg[\Big(a_{n,k} + \frac{2}{2n-1}\Big(\frac{2n+1}{2n-1} + \ln(2n-1)\Big)\Big)nk\Bigg],
\end{aligned}
$$

where the last inequality follows from the inequality

$$
h_e\Big(\frac{2}{2n+1}\Big) \le \frac{2}{2n+1}\Big(\frac{2n+1}{2n-1} + \ln\Big(\frac{2n-1}{2}\Big)\Big). \tag{3.11}
$$

Hence, (3.7) holds with

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

Inequality (1.2) now follows from Lemma 3.1. □

REMARK 3.1. As it was already discussed in [1], the obtained rate might not be optimal. For example, it is reasonable to believe that the factor $\ln n$ could be removed. We have already mentioned that in order to get a faster rate with our method, one has to obtain a better upper bound for $|\mathcal{B}_{k,n}|$. Let us briefly examine the proof of Theorem 1.1 from this point of view: inequality (3.11) cannot be significantly improved. From the well-known fact that for any $\alpha \in (0,1)$ and $m \to \infty$,

$$
\ln\binom{m}{\alpha m} \text{ has the same order of magnitude that } m \cdot h_e(\alpha),
$$

it follows that the bound

$$
\ln\binom{nk+r-1}{r-1} \le h_e\Big(\frac{2}{2n+1}\Big)(nk+r-1)
$$

is fairly sharp as well. On the other hand, since the partitions in $\mathcal{B}_{k,n}$ should satisfy (3.2), not all possible $\binom{nk+r-1}{r-1}$ combinations with repetitions correspond to a valid partition in X-side. Our method is purely combinatorial, so it does not take into account the distribution and the entropy of the random vectors $(X_1, \dots, X_{kn})$

and $(Y_1, \ldots, Y_{kn})$. The advantage of the purely combinatorial approach is its simplicity and generality, but when a better rate is aimed, it could be helpful to note that (1) many partitions in $\mathcal{B}_{k,n}$ have negligible probability and (2) $|\mathcal{B}_{k,n}|$ could get smaller when only looking at partitions corresponding to the optimal alignments. Hence, discarding untypical and not optimal partitions might drastically reduce $|\mathcal{B}_{k,n}|$ and, therefore, our method could give us a better rate. This is the subject of further research.

Acknowledgments. The authors would like to thank the German Science Foundation (DFG) for support through the Collaborative Research Center 701 "Spectral Structures and Topological Methods in Mathematics" (CRC 701) at Bielefeld University and Barbara Gentz for support 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] ALEXANDER, K. S. (1994). The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.* **4** 1074–1082. MR1304773

[2] ALEXANDER, K. S. (1997). Approximation of subadditive functions and convergence rates in limiting-shape results. *Ann. Probab.* **25** 30–55. MR1428498

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

[4] BAEZA-YATES, R. A., GAVALDÀ, R., NAVARRO, G. and SCHEIHING, R. (1999). Bounding the expected length of longest common subsequences and forests. *Theory Comput. Syst.* **32** 435–452. MR1693383

[5] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

[6] BOOTH, H. S., MACNAMARA, S. F., NIELSEN, O. M. and WILSON, S. R. (2004). An iterative approach to determining the length of the longest common subsequence of two strings. *Methodol. Comput. Appl. Probab.* **6** 401–421. MR2108560

[7] CHVATAL, V. and SANKOFF, D. (1975). Longest common subsequences of two random sequences. *J. Appl. Probab.* **12** 306–315. MR0405531

[8] CRISTIANINI, N. and HAHN, M. W. (2007). *Introduction to Computational Genomics*. Cambridge Univ. Press, Cambridge. MR2427944

[9] DANCIK, V. (1994). Expected length of longest common subsequences. Ph.D. dissertation, Dept. Computer Science, Univ. Warwick.

[10] DANČÍK, V. and PATERSON, M. (1995). Upper bounds for the expected length of a longest common subsequence of two binary sequences. *Random Structures Algorithms* **6** 449–458. MR1368846

[11] DEKEN, J. G. (1979). Some limit results for longest common subsequences. *Discrete Math.* **26** 17–31. MR0535080

[12] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093

[13] DURBIN, R., EDDY, S., KROGH, A. and MITCHISON, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, Cambridge.

[14] DURRINGER, C., HAUSER, R. and MATZINGER, H. (2008). Approximation to the mean curve in the LCS problem. *Stochastic Process. Appl.* **118** 629–648. MR2394846

[15] FU, J. C. and LOU, W. Y. W. (2008). Distribution of the length of the longest common subsequence of two multi-state biological sequences. *J. Statist. Plann. Inference* **138** 3605–3615. MR2450100

[16] KIWI, M., LOEBL, M. and MATOUŠEK, J. (2005). Expected length of the longest common subsequence for large alphabets. *Adv. Math.* **197** 480–498. MR2173842

[17] LIN, C. Y. and OCH, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL'*04: *Proceedings of the* 42*nd Annual Meeting on Association for Computational Linguistics* 605. Association for Computational Linguistics, Stroudsburg, PA.

[18] LUEKER, G. S. (2009). Improved bounds on the average length of longest common subsequences. *J. ACM* **56** Art. 17, 38. MR2536132

[19] MELAMED, I. D. (1995). Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*. MIT, Boston.

[20] MELAMED, I. D. (1999). Bitext maps and alignment via pattern recognition. *Comput. Linguist.* 107–130.

[21] PATERSON, M. and DANČÍK, V. (1994). Longest common subsequences. In *Mathematical Foundations of Computer Science* 1994 (*KoŠice*, 1994). *Lecture Notes in Computer Science* **841** 127–142. Springer, Berlin. MR1319827

[22] PEVZNER, P. A. (2000). *Computational Molecular Biology*: *An Algorithmic Approach*, *a Bradford Book*. MIT Press, Cambridge, MA. MR1790966

[23] SMITH, T. F. and WATERMAN, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Bio.* **147** 195–197.

[24] WATERMAN, M. S. (1995). *Introduction to Computational Biology*. Chapman & Hall, New York.

[25] WATERMAN, M. S. and VINGRON, M. (1994). Sequence comparison significance and Poisson approximation. *Statist. Sci.* **9** 367–381. MR1325433

[26] YANG, C. C. and LI, K. W. (2003). Automatic construction of english/chinese parallel corpora. *Journal of the American Society for Information Science and Technology* **54** 730–742.

J. LEMBER
INSTITUTE OF MATHEMATICAL STATISTICS
TARTU UNIVERSITY
LIIVI 2-513 50409, TARTU
ESTONIA
E-MAIL: jyril@ut.ee

H. MATZINGER
SCHOOL OF MATHEMATICS
GEORGIA TECH
ATLANTA, GEORGIA 30332-0160
USA
E-MAIL: matzing@math.gatech.edu

F. TORRES
FACULTY OF MATHEMATICS
UNIVERSITY OF BIELEFELD
POSTFACH 100131–33501 BIELEFELD
GERMANY
E-MAIL: ftorres@math.uni-bielefeld.de