The Annals of Applied Probability
1994, Vol. 4, No. 4, 1074–1082

THE RATE OF CONVERGENCE OF THE MEAN LENGTH OF THE LONGEST COMMON SUBSEQUENCE[1]

BY KENNETH S. ALEXANDER

University of Southern California

Given two i.i.d. sequences of n letters from a finite alphabet, one can consider the length L_n of the longest sequence which is a subsequence of both the given sequences. It is known that EL_n grows like γn for some $\gamma \in [0,1]$. Here it is shown that $\gamma n \geq EL_n \geq \gamma n - C(n \log n)^{1/2}$ for an explicit numerical constant C which does not depend on the distribution of the letters. In simulations with $n = 100{,}000$, EL_n/n can be determined from k such trials with 95% confidence to within $0.0055/\sqrt{k}$, and the results here show that γ can then be determined with 95% confidence to within $0.0225 + 0.0055/\sqrt{k}$, for an arbitrary letter distribution.

1. Introduction. Given a finite alphabet A and two sequences $x_1, \dots, x_n$ and $y_1, \dots, y_n$ in A, there is said to be a common subsequence of length k if for some $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_k \leq n$, we have $x_{i_m} = y_{j_m}$ for all $1 \leq m \leq k$. We wish to consider the length L_n of the longest common subsequence (LCS) of two A-valued i.i.d. sequences $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ with a common law μ. This problem and its variants have been much studied in probability theory [6, 7, 19], computer science [1, 3, 14] and mathematical biology [11, 15, 16, 18]; see also the volume [17] for several articles. In mathematical biology, the alphabet $A = \{a, c, t, g\}$ of greatest interest consists of the four DNA bases, and one may want to test whether an observed common subsequence between two base sequences could be due to chance. The quantity $2(n - L_n)$ is the minimal number of insertions and deletions needed to change either sequence to the other one; in computer science this "edit distance" is used as a metric on strings.

It is easy to see that $\{EL_n,\ n \geq 1\}$ is a superadditive sequence, that is,

$$EL_{n+m} \geq EL_n + EL_m \quad \text{for all } n, m \geq 1. \tag{1.1}$$

It therefore follows from standard methods that EL_n/n has a limit and the convergence is from below, that is, there exists $\gamma = \gamma(\mu) \in [0,1]$ such that

$$\lim_n EL_n/n = \sup_n EL_n/n = \gamma. \tag{1.2}$$

Kingman's subadditive ergodic theorem [13] further implies that $L_n/n \to \gamma$ a.s. For fair coin tossing, where $A = \{H, T\}$ and $\mu(H) = \mu(T) = 1/2$, simula-

Received September 1992; revised December 1993.
[1]Research supported by NSF Grants DMS-90-06395 and DMS-92-06139.
AMS 1991 *subject classification*. 60C05.
Key words and phrases. Longest common subsequence, subadditivity, first-passage percolation.

tions and heuristics (see [19] and Section 3 below) suggest that γ is between 0.81 and 0.82.

What interests us here is the rate at which the convergence in (1.2) occurs. The following is our main result.

THEOREM 1.1. *There exists a constant C such that for every alphabet A, law μ and $n \geq 1$,*

$$\gamma n \geq EL_n \geq \gamma n - C(n \log n)^{1/2}. \tag{1.3}$$

For a given n_0 our calculations will give an explicit value of C valid for $n \geq n_0$. This C will be smaller for larger n_0 as lower-order terms become negligible. In fact, we will show in Section 2 that for any $C > 3.42$, (1.3) is valid for all sufficiently large n.

The bound in (1.3) is useful in conjunction with simulations, which can really only estimate EL_n, in estimating γ. Simulations with $n = 100{,}000$ will be discussed in Section 3, together with simulations which suggest that the $(n \log n)^{1/2}$ rate in Theorem 1.1 is nearly the best obtainable by the methods of this paper.

Our method is modeled after that of [2], where a rate of convergence problem for first-passage percolation was considered. The analog of L_n is the passage time from the origin to a point n units out on an axis. The applicability of the method is not surprising in view of the fact that the LCS problem can be reformulated as a dependent first-passage percolation problem, as noted in [1, 4, 14, 20].

2. Proof of the main result. In place of L_n it is more convenient to work with

$$U_n := 2(n - L_n).$$

If we think of the corresponding letters of the two maximal identical subsequences as being matched, then U_n represents the number of letters unused in this matching. More generally, for $1 \leq i \leq j + 1$ and $1 \leq m \leq n + 1$, we define $U([i, j], [m, n])$ to be the number of letters unused after matching the corresponding letters of a longest common subsequence of $X_i, \dots, X_j$ and $Y_m, \dots, Y_n$. When $j + 1 = i$ and/or $n + 1 = m$, we interpret the corresponding sequence here as being empty, so that, for example, $U([i, i - 1], [m, n]) = n - m + 1$ if $m \leq n + 1$. When $j + 1 < i$ and/or $n + 1 < m$, we use the convention that $U([i, j], [m, n]) = \infty$. We will abbreviate $U([1, j], [1, n])$ to $U(j, n)$. Define

$$V_n := \min_{-n \leq k \leq n} U(n + k, n - k).$$

These quantities appear naturally in the first-passage reformulation of LCS, so we will briefly describe that reformulation now. Consider the integer lattice in $[0, 2n] \times [0, 2n]$, with horizontal and vertical bonds between nearest-neighbor sites of the lattice (that is, pairs x, y with $|x - y| = 1$) and a diagonal bond from each $(i-1, j-1)$ to (i, j), $1 \le i \le 2n$ and $1 \le j \le 2n$. The passage time of each horizontal and vertical bond is defined to be 1, and the passage time of the diagonal bond from $(i-1, j-1)$ to (i, j) is 0 if $X_i = Y_j$, and ∞ otherwise. Then $U([i, j], [r, s])$ represents the minimal total passage time among all paths from $(i-1, r-1)$ to (j, s) for which each coordinate in nondecreasing. We will call such a path a nondecreasing path. Let l_n denote the diagonal from $(0, 2n)$ to $(2n, 0)$. Then V_n represents the minimal total passage time among all nondecreasing paths starting at $(0,0)$ and ending on l_n.

For $n \ge 1$ and $\beta > 0$, define the generating functions

$$g_n(\beta) := -\log\Bigl(\sum_{-n \le k \le n} E\exp(-\beta(U(n+k, n-k) - 2))\Bigr).$$

Heuristically one expects the sum in the definition of $g_n(\beta)$ to behave like its largest term, so that $g_n(\beta)$ behaves like $-\log E\exp(-\beta V_n)$. In fact, by Jensen's inequality we have

$$g_n(\beta) \le -\log(Ee^{-\beta(V_n - 2)}) \le \beta(EV_n - 2). \tag{2.1}$$

The key property of $g_n(\beta)$ is given in the following result.

PROPOSITION 2.1. *For each $\beta > 0$, the sequence $\{g_n(\beta): n \ge 1\}$ is superadditive, that is,*

$$g_{n+m}(\beta) \ge g_n(\beta) + g_m(\beta) \quad \text{for all } m, n \ge 0. \tag{2.2}$$

Consequently, for some constants $\nu_\beta \le 2(1-\gamma)$,

$$\lim_n g_n(\beta)/n = \sup_n g_n(\beta)/n = \beta\nu_\beta \quad \text{for each } \beta > 0. \tag{2.3}$$

Before proving Proposition 2.1 we note that together with (1.2) it tells us that, for each fixed n and β, we have

$$g_n(\beta)/\beta \le \nu_\beta n \le 2(1-\gamma)n \le EU_n. \tag{2.4}$$

Thus because EU_n is subadditive and $g_n(\beta)$ is superadditive, $g_n(\beta)/\beta$ and EU_n are on opposite sides of the limiting approximation $2(1-\gamma)n$. It follows that

$$\begin{aligned} EU_{2n} - 4(1-\gamma)n &\le EU_{2n} - 2g_n(\beta)/\beta \\ &= (EU_{2n} - 2EV_n) + 2(EV_n - g_n(\beta)/\beta). \end{aligned} \tag{2.5}$$

Note that all of this is valid even if β is chosen depending on n.

To prove Proposition 2.1 we will need the following result.

LEMMA 2.2. *For each $n, m \ge 0$ and $0 \le k \le n + m$,*

$$\begin{aligned} &U(n+m+k, n+m-k) + 2 \\ &\quad \ge \min\{U([1, n+j], [1, n-j]) + U([n+j+1, n+m+k], \\ &\qquad [n-j+1, n+m-k]): -n \le j \le n, k-m \le j \le k+m\}. \end{aligned}$$

PROOF. Let Γ be a nondecreasing path of minimal total passage time from $(0,0)$ to $(n+m+k, n+m-k)$. Then Γ intersects l_n in a unique point $(n+x, n-x)$, and for some integer j, $-n \le j < n$, either $x = j$ or $x = j + 1/2$. Since the path is nondecreasing, we have $-n \le x \le n$ and $k - m \le x \le k + m$. If $x = j$, then breaking Γ into two pieces at $(n+j, n-j)$ shows that

$$\begin{aligned} &U(n+m+k, n+m-k) \\ &\quad = U([1, n+j], [1, n-j]) \\ &\qquad + U([n+j+1, n+m+k], [n-j+1, n+m+k]). \end{aligned}$$

If $x = j + 1/2$, then replacing the bond from $(n+j, n-j-1)$ to $(n+j+1, n-j)$ in Γ with the bond from $(n+j, n-j-1)$ to $(n+j, n-j)$ and the bond from $(n+j, n-j)$ to $(n+j+1, n-j)$ adds 2 to the passage time. Breaking the altered Γ at $(n+j, n-j)$ then shows that

$$\begin{aligned} &U(n+m+k, n+m-k) + 2 \\ &\quad \ge U([1, n+j], [1, n-j]) \\ &\qquad + U([n+j+1, n+m+k], [n-j+1, n+m-k]). \end{aligned}$$

In both cases, the desired result follows. □

PROOF OF PROPOSITION 2.1. From Lemma 2.2, independence and translation invariance,

$$\begin{aligned} &\sum_{-(n+m)\le k\le n+m} E\exp(-\beta[U(n+m+k, n+m-k) - 2]) \\ &\quad \le \sum_{-(n+m)\le k\le n+m} E\exp(-\beta\min\{U([1, n+j], [1, n-j]) \\ &\qquad\qquad + U([n+j+1, n+m+k], [n-j+1, n+m-k]) - 4: \\ &\qquad\qquad\qquad -n \le j \le n, k-m \le j \le k+m\}) \\ &\quad \le \sum_{-(n+m)\le k\le n+m} \sum_{\substack{j:\, -n\le j\le n \\ k-m\le j\le k+m}} E\exp(-\beta\{U([1, n+j], [1, n-j]) \\ &\qquad\qquad + U([n+j+1, n+m+k], [n-j+1, n+m-k]) - 4\}) \\ &\quad = \sum_{-(n+m)\le k\le n+m} \sum_{\substack{j:\, -n\le j\le n \\ k-m\le j\le k+m}} E\exp(-\beta[U(n+j, n-j) - 2]) \\ &\qquad\qquad \times E\exp(-\beta[U(m+k-j, m-k+j) - 2]) \\ &\quad = \left(\sum_{j:\, -n\le j\le n} E\exp(-\beta[U(n+j, n-j) - 2])\right) \\ &\qquad \times\left(\sum_{r:\, -m\le r\le m} E\exp(\{\beta[U(m+r, m-r) - 2])\right). \end{aligned}$$

Taking logarithms then yields (2.2). Standard subadditivity arguments then give (2.3). The fact that $\nu_\beta \le 2(1-\gamma)$ follows from (2.1) and the fact that $V_n \le U_n$. □

The following lemma gives a special case of Azuma's inequality [5] and is essentially a martingale version of Theorem 2 of Hoeffding [10].

LEMMA 2.3. *Suppose* $f(x_1, \dots, x_n, y_1, \dots, y_n)$ *is a function on* A^{2n} *with the property that changing any one argument of* f *while holding the others fixed changes the value of* f *by at most* 2. *Then for* $Z := f(X_1, \dots, X_n, Y_1, \dots, Y_n)$ *and* $u \ge 0$,

$$P[Z - EZ \ge u] \le \exp(-u^2/4n).$$

In particular,

$$P[L_{2n} - EL_{2n} \ge u/2] = P[U_{2n} - EU_{2n} \le -u] \le \exp(-u^2/8n), \tag{2.6}$$

$$P[L_{2n} - EL_{2n} \le -u/2] = P[U_{2n} - EU_{2n} \ge u] \le \exp(-u^2/8n), \tag{2.7}$$

$$P[V_n - EV_n \le -u] \le \exp(-u^2/8n) \tag{2.8}$$

and

$$P[V_n - EV_n \ge u] \le \exp(-u^2/8n). \tag{2.9}$$

Theorem 1.1 is an immediate consequence of the next proposition, since monotonicity of EU_n handles odd indices n. Any fixed values of $\lambda > 1$ and $\theta > \sqrt{2}$ suffice for proving Theorem 1.1, so if $C > (2 + \sqrt{2})$, then (1.3) is valid for all sufficiently large n. For the explicit confidence intervals of Section 3, the more detailed requirements in (2.10) and (2.11) are important.

PROPOSITION 2.4. *Suppose* $n \ge 8$ *and* $\lambda, \theta > 0$ *satisfy*

$$\begin{aligned} \lambda^2 \ge 1 &+ 1/(2n \log 2n) + 2\lambda/(2n \log 2n)^{1/2} \\ &+ (\log 5.1\lambda)/\log 2n + (\log\log 2n)/(2 \log 2n) \end{aligned} \tag{2.10}$$

and

$$\theta^2 \ge 2 + (\log 4)/\log 2n. \tag{2.11}$$

Then

$$\begin{aligned} EU_{2n}/(2n) \le 2(1-\gamma) &+ 2(2\lambda + \theta)((\log 2n)/(2n))^{1/2} \\ &+ ((8 \log 2)/n)^{1/2} \end{aligned} \tag{2.12}$$

and

$$EL_{2n}/(2n) \ge \gamma - (2\lambda + \theta)((\log 2n)/(2n))^{1/2} - ((2 \log 2)/n)^{1/2}. \tag{2.13}$$

PROOF. We will use (2.5). Let

$$\beta_n := \lambda((\log 2n)/(2n))^{1/2}.$$

Let us first bound $EV_n - g_n(\beta_n)/\beta_n$. From integration by parts and Lemma 2.3,

$$
\begin{aligned}
&E\exp(-\beta_n V_n)\\
&\quad= \int_0^\infty \beta_n \exp(-\beta_n x) P[V_n \le x]\,dx\\
&\quad\le \exp(-\beta_n EV_n) + \int_0^{EV_n} \beta_n \exp(-\beta_n x)\exp\bigl(-(EV_n - x)^2/(8n)\bigr)\,dx\\
&\quad\le \exp(-\beta_n EV_n) + \beta_n (8\pi n)^{1/2} \exp\bigl(-\beta_n EV_n + 2n\beta_n^2\bigr).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\exp(-g_n(\beta_n)) &= \sum_{-n\le k\le n} E\exp(-\beta_n(U(n+k, n-k) - 2))\\
&\le (2n+1)E\exp(-\beta_n(V_n - 2))\\
&\le (2n+1)\{\exp(-\beta_n(EV_n - 2))\}\bigl[1+\beta_n(8\pi n)^{1/2}\exp(2n\beta_n^2)\bigr].
\end{aligned}
\tag{2.14}
$$

Observe that

$$
\begin{aligned}
&\log\bigl(1.011(8\pi n\beta_n^2)^{1/2}\exp(2n\beta_n^2)\bigr)\\
&\qquad\le \log(2.860\pi^{1/2}\lambda) + \log\bigl((n\beta_n^2)^{1/2}/\lambda\bigr) + 2n\beta_n^2.
\end{aligned}
\tag{2.15}
$$

Taking logs in (2.14), rearranging and using $n\beta_n^2 \ge 2\log 2$, (2.15) and (2.10), we obtain

$$
\begin{aligned}
&EV_n - g_n(\beta_n)/\beta_n\\
&\quad\le 2 + \beta_n^{-1}\log(2n+1) + \beta_n^{-1}\log\bigl(1 + (8\pi n\beta_n^2)^{1/2}\exp(2n\beta_n^2)\bigr)\\
&\quad\le 2 + \beta_n^{-1}(\log 2n + 1/(2n)) + \beta_n^{-1}\log\bigl(1.011(8\pi n\beta_n^2)^{1/2}\exp(2n\beta_n^2)\bigr)\\
&\quad\le \lambda^{-1}(2n\log 2n)^{1/2}\\
&\qquad\times\bigl[1 + 1/(2n\log 2n) + 2\lambda/(2n\log 2n)^{1/2}\\
&\qquad\quad + (\log 5.1\lambda)/\log 2n + (\log\log 2n)/2\log 2n + \lambda^2\bigr]\\
&\quad\le 2\lambda(2n\log 2n)^{1/2}.
\end{aligned}
\tag{2.16}
$$

Let us next bound $EU_{2n} - 2EV_n$. In terms of the first-passage formulation, we use what is essentially a reflection argument across the diagonal l_n. From (2.6) of Lemma 2.3, we have

$$
\begin{aligned}
1/2 &\le P\bigl[V_n \le EV_n + (8n\log 2)^{1/2}\bigr]\\
&\le \sum_{-n<j<n} P\bigl[U(n+j, n-j) \le EV_n + (8n\log 2)^{1/2}\bigr],
\end{aligned}
\tag{2.17}
$$

where $j = n$ and $-n$ need not be included in the sum because the minimum of $U(n+j, n-j)$ always occurs with $-n < j < n$. Therefore, there exists an index j with

$$P\left[U(n+j, n-j) \le EV_n + (8n\log 2)^{1/2}\right] \ge 1/(4n-2).$$

Since $U([n+j+1, 2n], [n-j+1, 2n])$ has the same distribution as $U(n+j, n-j) = U([1, n+j], [1, n-j])$ and is independent of it, we have

$$\begin{aligned} 1/(4n-2)^2 &\le P\Big[U([1, n+j], [1, n-j]) \le EV_n + (8n\log 2)^{1/2}, \\ &\qquad U([n+j+1, 2n], [n-j+1, 2n]) \\ &\qquad \le EV_n + (8n\log 2)^{1/2}\Big] \\ &\le P\left[U_{2n} \le 2\left(EV_n + (8n\log 2)^{1/2}\right)\right]. \end{aligned} \tag{2.18}$$

However, from (2.6) of Lemma 2.3 and (2.11),

$$P\left[U_{2n} \le EU_{2n} - 2\theta(2n\log 2n)^{1/2}\right] \le \exp(-\theta^2 \log 2n) \le 1/(4n)^2, \tag{2.19}$$

which with (2.18) shows

$$EU_{2n} \le 2EV_n + 2(8n\log 2)^{1/2} + 2\theta(2n\log 2n)^{1/2}.$$

With (2.5) and (2.16) this proves (2.12), which implies (2.13). □

Our method should be applicable to other problems which have a first-passage formulation. The main ingredients for which one must have analogs are, first, that there is enough independence that something like Lemma 2.3 holds, and second, that for a path Γ as in the proof of Lemma 2.2 which meets l_n at a particular point, the two segments into which Γ is split by that point are independent or at least are appropriately comparable to independent segments as in [2].

3. Confidence bounds and simulations. For $2n \geq 100{,}000$, (2.10) and (2.11) are satisfied with $\lambda = 1.123$ and $\theta = 1.457$. Therefore, from Proposition 2.4,

$$EL_{2n}/2n \le \gamma \le EL_{2n}/(2n) + 0.0450. \tag{3.1}$$

Given k independent observations of L_{2n}, let $\overline{L}_{2n}$ denote the sample mean of these observations. For optimal confidence bounds, we place our estimate of γ in the center of the interval suggested by (3.1), defining

$$\hat{\gamma}_{2n} := \overline{L}_{2n}/(2n) + 0.0225. \tag{3.2}$$

From Lemma 2.3, for $2n \ge 100{,}000$, $x \ge 0.0055/\sqrt{k}$ and $t := \gamma - EL_{2n}/2n$,

$$\begin{aligned} P&\left[|\hat{\gamma}_{2n} - \gamma| > x + 0.0225\right] \\ &\le P\left[\overline{L}_{2n}/(2n) - EL_{2n}/(2n) \ge x + t\right] \\ &\quad + P\left[\overline{L}_{2n}/(2n) - EL_{2n}/(2n) \le -(x + 0.0450 - t)\right] \\ &\le \exp\left(-2kn(x+t)^2\right)2 + \exp\left(-2kn(x + 0.0450 - t)^2\right) \\ &\le \exp(-2knx^2) + \exp\left(-2kn(x+0.0450)^2\right) \\ &\le 0.05. \end{aligned}$$

In particular, for $2n \geq 100{,}000$ and $k = 2$,

$$P\big[|\hat{\gamma}_{2n} - \gamma| \leq 0.0264\big] \geq 0.95. \tag{3.3}$$

Eggert and Waterman [9] simulated two trials of L_{2n} for fair coin tossing with $2n = 100{,}000$. The observed values were 81,223 and 81,146, yielding the estimate $\hat{\gamma}_{2n} = 0.8343$. Therefore, from (3.3),

$$0.8079 \leq \gamma \leq 0.8607. \tag{3.4}$$

with 95% confidence. By contrast, the best bounds known with certainty for fair coin tossing are $0.7615 \leq \gamma \leq 0.8376$ [7, 8]. By using this upper bound we can improve on (3.4) as follows. From Lemma 2.3 we have $P[\bar{L}_{2n}/2n - \delta \leq \gamma] \geq 1 - \exp(-2kn\delta^2)$. In particular, with $k = 2$ and $\delta = 0.0039$ this yields $0.8079 \leq \gamma \leq 0.8376$ with 95% confidence.

Additional simulations in [21] suggest that the variance of L_n is approximately proportional to n. If this is true, then Lemma 2.3 cannot be valid with any smaller power of n in the denominator of the exponent. This means that our method cannot yield a better power of n than the $1/2$ which appears in (1.3). Of course, the actual difference $\gamma n - EL_n$ may well be $o(n^{1/2})$, but quite different methods would apparently be needed to obtain such an improved result.

NOTE ADDED IN PROOF. An additional reference by P. Jaillet [(1992) *Math Oper. Res.* **17** 964–980), has a proof of Lemma 2.3 (ii) and gives two-sided bounds with rates as in (1.4) for several functionals including TSP and MST.

Acknowledgments. The author would like to thank J. M. Steele for helpful comments on an earlier version of this manuscript, and M. Eggert and M. Waterman for the simulations described in Section 3.

REFERENCES

[1] AHO, A. V. (1990). Algorithms for finding patterns in strings. In *Handbook of Theoretical Computer Science* (J. van Leeuwen, ed.) 256–300. North-Holland, Amsterdam.

[2] ALEXANDER, K. S. (1993). A note on some rates of convergence in first-passage percolation. *Ann. Appl. Probab.* **3** 81–90.

[3] APOSTOLICO, A. and GUERRA, C. (1987). The longest common subsequence problem revisited. *Algorithmica* **2** 315–332.

[4] 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.

[5] AZUMA, K. (1967). Weighted sums of certain dependent random variables. *Tohuku Math. J.* **19** 357–367.

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

[7] DANČÍK, V. and PATERSON, M. (1994). Upper bound for the expected length of a longest common subsequence of two binary sequences. In *STACS 94. Lecture Notes in Comput. Sci.* **775** 669–678. Springer, New York.

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

[9] EGGERT, M. and WATERMAN, M. S. (1992). Personal communication.

[10] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

[11] HUANG, X. and WATERMAN, M. S. (1992). Dynamic programming algorithms for restriction map comparison. *Comp. Appl. Bio. Sci.* **8** 511–520.

[12] KESTEN, H. (1993). On the speed of convergence in first-passage percolation. *Ann. Appl. Probab.* **3** 296–338.

[13] KINGMAN, J. F. C. (1968). The ergodic theory of subadditive stochastic processes. *J. Roy. Statist. Soc. Ser. B* **30** 499–510.

[14] MYERS, E. W. (1986). An $O(ND)$ difference algorithm and it variations. *Algorithmica* **1** 251–266.

[15] NEEDLEMAN, S. B. and WUNSCH, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48** 443–453.

[16] PEVZNER, P. A. and WATERMAN, M. S. (1993). Generalized sequence alignment and duality. *Adv. in Appl. Math.* **14** 139–171.

[17] SANKOFF, D. and KRUSKAL, J. B., eds. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, Reading, MA.

[18] SMITH, T. F. and WATERMAN, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **18** 38–46.

[19] STEELE, J. M. (1986). An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758.

[20] UKKONEN, E. (1985). Algorithms for approximate string matching. *Inform. Control* **64** 100–118.

[21] WATERMAN, M. S. (1994). Estimating statistical significance of sequence alignments. *Phil. Trans. Roy. Soc. London Ser. B* **344** 383–390.

DEPARTMENT OF MATHEMATICS DRB 155
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089-1113