*The Annals of Applied Probability*
1994, Vol. 4, No. 1, 200–225

# A PHASE TRANSITION FOR THE SCORE IN MATCHING RANDOM SEQUENCES ALLOWING DELETIONS[1]

BY RICHARD ARRATIA AND MICHAEL S. WATERMAN

*University of Southern California*

We consider a sequence matching problem involving the optimal alignment score for contiguous subsequences, rewarding matches and penalizing for deletions and mismatches. This score is used by biologists comparing pairs of DNA or protein sequences. We prove that for two sequences of length $n$, as $n \to \infty$, there is a phase transition between linear growth in $n$, when the penalty parameters are small, and logarithmic growth in $n$, when the penalties are large. The results are valid for independent sequences with iid or Markov letters. The crucial step in proving this is to derive a large deviation result for matching with deletions. The longest common subsequence problem of Chvátal and Sankoff is a special case of our setup. The proof of the large deviation result exploits the Azuma–Hoeffding lemma. The phase transition is also established for more general scoring schemes allowing general letter-to-letter alignment penalties and block deletion penalties. We give a general method for applying the bounded increments martingale method to Lipschitz functionals of Markov processes. The phase transition holds for matching Markov chains and for nonoverlapping repeats in a single sequence.

**1. Introduction.** DNA and protein molecules can be represented as strings of letters from a finite alphabet, such as a 4-letter alphabet for DNA sequences and a 20-letter alphabet for protein sequences. A major undertaking in biology is to determine genetic sequences from various organisms. Important portions of the sequences are preserved over evolutionary time, so that relationships among sequences are indicative of evolutionary—and hence functional—relationships. Computer algorithms have been devised to detect these relationships; see Waterman (1984, 1989) for reviews of the application of sequence comparison algorithms to biology. Sequence comparison has appeared in a number of fields, such as speech recognition, bird song studies and geological strata comparisons. The book of Sankoff and Kruskal (1983) presents many of these applications as well as more formal, computer science aspects of sequence comparison. In the computer science literature these problems are known as string comparisons or string matching. See, for example, the books by Capocelli (1990) and Apostolico, Crochemore, Galil and Manbar (1992).

Received August 1991; revised July 1993.
[1]Supported by grants from the NIH and NSF.
*AMS* 1991 *subject classifications*. 62E20, 62P10.
*Key words and phrases.* Sequence matching, longest common subsequence, large deviations, Azuma-Hoeffding, phase transition, percolation.

More recently the probability distributions of random variables relating to these comparisons have been under study. See Karlin, Ghandour, Ost, Tavaré and Korn (1983), Arratia and Waterman (1985a), Karlin and Ost (1987), Arratia, Gordon and Waterman (1990), and Neuhauser (1994) for some of this literature. In this paper we study the random variable that is the score of the sequence comparison algorithm thought to be the most rigorous and useful for biology.

The phenomenon of phase transitions was first described in Waterman, Gordon and Arratia (1987). The overall setup for phase transitions is the following. We have a random variable $M \equiv M_n \equiv M(\mu, \delta)$ which is the optimal score $S(I, J)$, over all possible choices of two contiguous regions, $I$ taken from $A_1 A_2 \cdots A_n$ and $J$ taken from $B_1 B_2 \cdots B_n$, where all the letters $A_i$ and $B_j$ are independent, identically distributed random variables from a finite or countable alphabet. The score $S$ is the maximum, over all possible "alignments" of $I$ and $J$, of the number of matches, minus the penalty parameter $\mu$ times the number of mismatches, minus the penalty parameter $\delta$ times the number of deleted letters. Algorithms for computing $M$ and finding the optimal matching segments and their alignments are discussed in Smith and Waterman (1981) and Waterman and Eggert (1987).

Write $S_t \equiv S(I_1 \cdots I_t, J_1 \cdots J_t)$ for the score, with penalty parameters $\mu$ and $\delta$, for two regions of length $t$. Subadditivity implies that $S_t/t$ has a constant limit, almost surely and in expectation; call this limit $a(\mu, \delta)$. When the penalty parameters are small, this limit is positive and the overall score $M_n$ grows linearly with $n$, with the optimal contiguous regions having length close to $n$. When the penalty parameters are large, so that the average score $a$ per unit length is negative, the optimal regions for $M$ represent large deviation behavior. They are regions with a positive score, while the average score is negative. *If* the probability of regions of positive score and length $t$ decays *exponentially fast* in $t$, then since there are at most $n^2$ locations for the pair $(I, J)$ where such a large deviation might occur (and there are at least $n/t - 1$ independent events involved here), the overall score $M_n$ will grow like $\log n$, whenever the penalty parameters are such that the average score is negative. In greater detail, consider the large deviation rate $r(q) \equiv r(q; \mu, \delta) = \lim -t^{-1} \log P(S_t \geq qt)$, which obviously satisfies $r(q) \geq 0$. If $a(\mu, \delta) < 0$, then we can prove the crucial property that $r(q) > 0$ for all $q \geq 0$. In this case, defining $b = b(\mu, \delta) = \max_{q \geq 0} q/r(q)$, we prove in Lemma 2, using an easy block argument like that in Arratia and Waterman (1985a) (corresponding to the case $\mu = \delta = \infty$) that

$$P\big(b(\mu, \delta) - \varepsilon < M_n/\log n < 2b(\mu, \delta) + \varepsilon\big) \to 1.$$

It is easy to show that the average score per letter, $a(\mu, \delta)$, has the property that the set of values $(\mu, \delta)$ for which $a = 0$ is a line in the parameter space, separating the negative region for $a$ from the positive. To show that this line is the location of a phase transition between logarithmic and linear growth for $M$, the crucial and difficult step is to show that large deviations for $S_t$ have probability which is exponentially small as $t \to \infty$.

Our first proof of the large deviation property for $S$ was based on the corresponding proof of a large deviation property in first-passage percolation, given by Grimmett and Kesten (1984) and given in a slightly modified version in Kesten (1986). This proof involves a "big-block, small-block" argument, which is considered routine by workers in percolation. The modifications required to adopt the percolation proof to sequence matching were not easy and consumed several pages in the original version of this paper. The first proof has now been replaced by a much easier argument using the Azuma–Hoeffding inequality. A copy of the first proof is available from the authors.

To see the percolation correspondence, view $-S_t$ as a minimum cost over paths in the plane, from $(0, 0)$ to $(t, t)$. In these paths, three kinds of steps are possible:

1. one unit to the right, from $(i, j)$ to $(i + 1, j)$, which costs $\delta$ and corresponds to deleting $A_{i+1}$;
2. one unit up, from $(i,j)$ to $(i, j + 1)$, which corresponds to deleting $B_{j+1}$, and costs $\delta$;
3. diagonally from $(i, j)$ to $(i + 1, j + 1)$, which corresponds to matching $A_{i+1}$ against $B_{j+1}$, and is paid 1 (i.e., costs $-1$) if the letters are equal and costs $\mu$ otherwise.

Essentially, the difference between this setup and first passage percolation is that here, the $t^2$ diagonal costs for a $t$-by-$t$ square are not independent—they are determined by only $2t$ random letters, while in first passage percolation all the $t^2$ edge costs are independent. That results from first passage percolation do not automatically extend to sequence matching is shown by the qualitative behavior of large deviations on the opposite side of average behavior. The probability that the best path of length $t$ is $\varepsilon t$ *worse* than average decays like $\exp(-c(\varepsilon)t)$ for matching $d \geq 2$ sequences, but like $\exp(-k(\varepsilon)t^d)$ for first passage percolation in $d$ dimensions.

Our second proof of the large derivation property, presented in Section 3, is based on the Azuma–Hoeffding inequality for martingales with bounded increments. We became aware of the applicability of this from lectures by M. Steele and B. Bollabás at a conference, the proceedings of which are summarized in Tavaré (1992). The proof using the Azuma–Hoeffding inequality is not only much simpler than the percolation-style proof, but it is also more powerful. In particular, using the Azuma–Hoeffding inequality, there are no extra restrictions needed in Section 4 on generalized scoring schemes, which allow longer blocks of deletions to receive a single penalty.

We summarize our main result as Theorem 1.

THEOREM 1. *For iid letters $A_1, A_2, \ldots$ and $B_1, B_2, \ldots$, the optimal alignment score $M_n = M(A_1 A_2 \cdots A_n, B_1 \cdots B_n)$, with penalty parameters $\mu$ per mismatch and $\delta$ per deletion, has a phase transition between linear growth with $n$ for small $\mu$ and $\delta$, and logarithmic growth with $n$ for large $\mu$ and $\delta$.*

*The coefficient of* $\log(n)$ *is given by* (7)–(9). *The coefficient of* $n$ *is given by* (6) *and* (14).

PROOF. The result follows immediately by combining Lemmas 1, 2 and 3. □

REMARK. In practice, two sequence of different lengths, say $m$ and $n$, are compared. Provided that $m, n \to \infty$, the result of Theorem 1 holds. More precisely, there is a phase transition between linear and logarithmic growth, in the sense that if $a(\mu, \delta) > 0$, then $M(A_1A_2 \cdots A_m, B_1B_2 \cdots B_n)$ grows at least as fast as $a(\mu, \delta)\min(m, n)$, and if $a(\mu, \delta) < 0$, then $M(A_1A_2 \cdots A_m, B_1B_2 \cdots B_n)$ grows at most as fast as $b(\mu, \delta)\log(mn)$ and at least as fast as $b(\mu, \delta)\log(\min(m,n))$. The proof we give for Lemma 2, corresponding to the special case $m = n$, easily extends to this more general case. For the case of the longest perfect matching, that is, $\mu = \delta = \infty$, it is possible to give the explicit coefficient of $\log(mn)$ when $m, n \to \infty$ with $\lim \log(m)/\log(n) = \theta \in (0, \infty)$. This coefficient is a continuous, nonanalytic function of $\theta$. See Arratia and Waterman (1985b).

This paper is organized as follows. In Section 2 we show how the phase transition result follows rigorously from the large deviation result. Section 3 gives the proof of the large deviation result. Section 4 extends the phase transition result to generalized alignment scoring schemes, Section 5 gives the extension to Markov chains, Section 6 gives the extension to nonoverlapping repeats in a single sequence and Section 7 presents some numerical results and illustrative examples.

In Waterman, Gordon and Arratia (1987) we pointed out a correspondence between sequence alignment, where similar letters are aligned, and helical structures where complementary bases (AT or GC) form base pairs. A generalized scoring scheme, for example, $s(A, T) = s(T, A) = 1.7$, and $s(G, C) = s(C, G) = 2.1$, $s(a, b) = -\mu$ otherwise, handles both the notion of complementary matching and the fact that the free energy of an AT base pair is weaker than that of a GC base pair. In addition, unpaired regions have more complex destabilization energy functions, often taken to be the logarithm of the length of the regions. For simple helical regions between distinct sequences, the theorems we prove in this paper apply. Related problems arise from the formation of DNA or RNA secondary structure where a single stranded molecule folds back on itself to form helical regions. The behavior of the free energy of the optimal structure does not follow from our theorems because of more complex dependence; we do conjecture that analogous results hold.

**2. Subadditive theory and large deviation theory imply phase transition.** We recall some definitions from the introduction. The variable $M_n$ is defined to be the optimal score $S(I, J)$ over all possible choices of two contiguous regions, $I$ taken from $A_1A_2 \cdots A_n$ and $J$ taken from $B_1B_2 \cdots B_n$. Formally

$$M_n = M(A_1 \cdots A_n, B_1 \cdots B_n) = \max_{I,J} S(I, J), \tag{1}$$

where $I = A_{g+1} \cdots A_{g+i}$ and $J = B_{h+1} \cdots B_{h+j}$ with $1 \le g+1 \le g+i \le n$ and $1 \le h+1 \le h+j \le n$. The alignment score $S(A_{g+1} \cdots A_{g+i}, B_{h+1} \cdots B_{h+j})$ is defined by

$$S(A_{g+1} \cdots A_{g+i}, B_{h+1} \cdots B_{h+j}) = \max\left\{ -\delta(i-l+j-l) + \sum_{k=1}^{l} s(A_{a(k)}, B_{b(k)}) \right\}, \tag{2}$$

where the maximum is over all alignments, given by increasing subsequences

$$\begin{aligned} g = a(0) < a(1) < a(2) < \cdots < a(l) < a(l+1) = g+i+1, \\ h = b(0) < b(1) < b(2) < \cdots < b(l) < b(l+1) = h+j+1. \end{aligned} \tag{3}$$

The scoring function for aligned pairs is

$$s(x,y) = \begin{cases} 1, & \text{if } x = y, \\ -\mu, & \text{if } x \ne y, \end{cases} \tag{4}$$

and the term $-\delta(i-l+j-l)$ is $-\delta$ times the number of letters deleted. Essentially, the difference between the scores $M$ and $S$ is that $M$ allows an initial and terminal segment to be deleted, at no cost, from each sequence.

Let

$$S_k = S(A_1 \cdots A_k, B_1 \cdots B_k)$$

and observe that

$$S_{k+l} \ge S_k + S(A_{k+1} \cdots A_{k+l}, B_{k+1} \cdots B_{k+l}), \tag{5}$$

where $S(A_{k+1} \cdots A_{k+l}, B_{k+1} \cdots B_{k+l})$ equals $S_l$ in distribution. Thus $ES_{k+l} \ge ES_k + ES_l$, so that subadditive theory implies the following limit exists, and equals the supremum

$$a(\mu, \delta) = \lim_{k \to \infty} \frac{ES_k}{k} = \sup_{k \ge 1} \frac{ES_k}{k}. \tag{6}$$

Observe that for $\mu = \infty$, $\delta = 0$, $S_n$ is the length of a longest common subsequence of $A_1 A_2 \cdots A_n$ and $B_1 B_2 \cdots B_n$, so that $a(\infty, 0) = c$, the Chvátal–Sankoff [Chvátal and Sankoff (1975)] constant. In the language of ergodic theory, $1 - c = \bar{f}$ is the $\bar{f}$ distance between $A_1 A_2 \cdots$ and $B_1 B_2 \cdots$. Steele (1986) proves an Efron–Stein inequality for functionals of iid variables and illustrates its use to obtain the bound $\operatorname{var}(S_n) \le nP(A_1 \ne B_1)$ for the variance of the length of a longest common subsequence. Steele's inequality applies for all $\mu, \delta \in [0, \infty]$, to yield $\operatorname{var}(S_n) \le nP(A_1 \ne B_1)(1 + \min(\mu, \delta))$.

In the remainder of the section, we assume Theorem 2 of Section 3 and prove our result on phase transitions. Lemma 1 states that $\{a(\mu, \delta) = 0\}$ defines a line in $[0, \infty]^2$. Lemma 2, using Theorem 2, shows that $M_n$ grows like $\log(n)$ when $a(\mu, \delta) < 0$, while Lemma 3 shows that $M_n$ grows like $n$ when $a(\mu, \delta) > 0$. Theorem 1 of the introduction is proved by the combination of these three lemmas.

LEMMA 1. *The set* $\{(\mu, \delta)\colon a(\mu, \delta) = 0\}$ *defines a line in the parameter space* $[0, \infty]^2$, *separating the negative and positive regions* $\{a < 0\}$ *and* $\{a > 0\}$.

PROOF. First, $a$ is obviously nonincreasing in each of its parameters, and we have the global inequality $a(\mu + \varepsilon, \delta + \varepsilon/2) \geq a(\mu, \delta) - \varepsilon$, because the corresponding inequality is satisfied by each possible alignment, and taking maxima, expectation and limit preserves this inequality. This shows that $a$ is continuous. In detail, with $Q = (\mu, \delta)$ and $Q' = (\mu', \delta')$ we have $|a(Q) - a(Q')| \leq \varepsilon \equiv |\mu - \mu'| + 2|\delta - \delta'|$, since with $R = (\mu_0, \delta_0) = (\mu \wedge \mu', \delta \wedge \delta')$ and $S = (\mu_0 + \varepsilon, \delta_0 + \varepsilon/2)$, monotonicity and the global inequality give $a(R) \geq a(Q) \geq a(S) \geq a(R) - \varepsilon$, and similarly for $a(Q')$. Second, although $a$ is not strictly monotone in each parameter everywhere in the parameter space, it is strictly monotone in the $(1, 1)$ direction, in a neighborhood of the line $a = 0$. To see this, let $\gamma \equiv \max(\mu, 2\delta)$ and observe that in alignments which score $g$ or less per pair of letters, the proportion $x$ of nonmatching pairs satisfies $-\gamma x + (1 - x) \leq g$, so that $x \geq (1 - g)/(\gamma + 1)$. For such alignments, increasing each of the penalty parameters by $\varepsilon > 0$ must decrease the score by at least $\varepsilon x$. It follows that $a(\mu + \varepsilon, \delta + \varepsilon) \leq a(\mu, \delta) - \varepsilon(1 - a(\mu, \delta))/(1 + \mu + 2\delta)$ for all $\varepsilon, \mu, \delta > 0$. The cases where $\mu = \infty$ or $\delta = \infty$ require a separate but similar argument. □

REMARK. The preceding lemma does not state that, for all $\varepsilon > 0$, $a(\mu + \varepsilon, \delta) < a(\mu, \delta)$ and $a(\mu, \delta + \varepsilon) < a(\mu, \delta)$. In fact, the first is clearly false when $2\delta < \mu$.

The following conjecture embodies rigorously the intuition that, except for cases with $2\delta \leq \mu$, any optimal alignment of typical $A_1 \cdots A_n$ and $B_1 \cdots B_n$, for large $n$, uses both deletions and mismatches a significant proportion of $n$ times. The analogous result for first passage percolation has been proved in van den Berg and Kesten (1993).

CONJECTURE 1. *For all* $\varepsilon, \mu, \delta > 0$, $a(\mu, \delta + \varepsilon) < a(\mu, \delta)$. *For all* $\varepsilon, \mu, \delta > 0$ *with* $2\delta > \mu$, $a(\mu + \varepsilon, \delta) < a(\mu, \delta)$.

Define

$$r(q) = \lim\left\{-\frac{1}{k}\log P(S_k \geq qk)\right\} = \inf\left\{-\frac{1}{k}\log P(S_k \geq qk)\right\}. \tag{7}$$

This limit exists and equals the infimum using the subadditive property

$$P\big(S_{j+k} \geq q(j + k)\big) \geq P(S_j \geq qj)P(S_k \geq qk).$$

The next section, using the Azuma-Hoeffding inequality, shows that if $a(\mu, \delta) < 0$ and $q \geq 0$, then $r(q) > 0$. This is a corollary of Theorem 2 of Section 3. Observe that subadditivity allows the possibility that $r(q) = 0$. Indeed this is the case when $a(\mu, \delta) > q$.

If $a(\mu, \delta) < 0$, then we can define

$$b \equiv b(\mu, \delta) = \max_{q \geq 0} \frac{q}{r(q)}. \tag{8}$$

Note that $b > 0$ since $r(1) = -\log P(A_1 = B_1) < \infty$.

In the following lemma the upper and lower bounds differ by a factor of 2, which arises from the need for independent blocks in the proof of the lower bound. One would need control over the correlation of the events in (13) below, to be able to improve the lower bound by a factor of 2.

CONJECTURE 2. *For all* $(\mu, \delta) \in [0, \infty]^2$, *the upper bound in Lemma* 2 *is sharp, in the sense that* $M_n/\log n$ *converges in probability to* $2b$, *whenever* $a(\mu, \delta) < 0$. [*This was proved for the special case* $\delta = \infty$, *allowing mismatches but no deletions, in Arratia and Waterman* (1989).]

The next two conjectures give a refinement of Conjecture 2. They express the belief that in the proof of the lower bound for Lemma 2, there is a unique optimal $q$, which governs the length of the optimal subregions. If Conjectures 2 and 3 were proved, the result in Conjecture 4 would follow immediately, via considerations like those in the proof of Lemma 2.

CONJECTURE 3. *For each* $(\mu, \delta) \in [0, \infty]^2$, *there is a unique value* $q = f(\mu, \delta) > 0$ *which achieves the value of* $b$, *that is,*

$$q/r(q) = b, \quad \text{if } q = f(\mu, \delta),$$
$$q/r(q) < b, \quad \text{if } q > 0, q \neq f(\mu, \delta).$$

CONJECTURE 4. *If* $a(\mu, \delta) < 0$, *then the lengths of the optimal regions grow like* $2\log(n)/r(q)$, *where* $q = f(\mu, \delta)$. *More precisely, with the notation from* (2), *so that* $i$ *and* $j$ *represent lengths of optimal regions* $I$ *and* $J$, *as* $n \to \infty$,

$$P\left\{S(I, J) = M_n \quad \textit{implies} \quad i, j \in \left[(2 - \varepsilon)\frac{\log(n)}{r(q)}, (2 + \varepsilon)\frac{\log(n)}{r(q)}\right]\right\} \to 1.$$

LEMMA 2. *For all* $(\mu, \delta) \in [0, \infty]^2$, *if* $a(\mu, \delta) < 0$, *then* $b(\mu, \delta)$ *is the coefficient of* $\log n$. *More precisely, as* $n \to \infty$,

$$P\left\{(1 - \varepsilon)b < \frac{M_n}{\log n} < (2 + \varepsilon)b\right\} \to 1. \tag{9}$$

PROOF. *Lower bound.* $P(M_n \geq (1 - \varepsilon)b \log n) \to 1$ *as* $n \to \infty$.

Given $\varepsilon > 0$, pick $\delta > 0$ small and $q > 0$ so that $q/r(q)$ approximates $b$ closely, so that

$$(1 - \varepsilon)b\left(\frac{r(q) + \delta}{q}\right) < 1 - \frac{\varepsilon}{2}.$$

Let $t = (1 - \varepsilon)b \log n$ and let $k = \lfloor t/q \rfloor$, so $k \sim c \log n$. For sufficiently large $n$, $k$ is sufficiently large that

$$-\frac{1}{k}\log P(S_k \geq qk) \leq r(q) + \delta,$$

so

$$\begin{aligned} P(S_k \geq qk) &\geq \exp(-k(r(q) + \delta)) \\ &\geq \exp\left(-t\left(\frac{r(q) + \delta}{q}\right)\right) \quad \left(\text{using } k \leq \frac{t}{q}\right) \\ &\geq \exp\left(-\left(1 - \frac{\varepsilon}{2}\right)\log n\right) = n^{-1+\varepsilon/2}. \end{aligned}$$

To conclude the proof of the lower bound, we will consider nonoverlapping blocks of length $k + 1$, so that we have about $n/k \sim n/(c \log n)$ independent chances to get a large score. Each chance has size at least $n^{-1+\varepsilon/2}$, so the expected number of successes goes to infinity like $n^{\varepsilon/2}/(c \log n)$. Formally, with $j = k + 1$ so that $t < qj$,

$$\begin{aligned} P(M_n < (1 - \varepsilon)b \log n) &= P(M_n < t) \\ &\leq P(M_n < qj) \\ &\leq P\left(\bigcap_{0 \leq i \leq \lfloor n/j \rfloor - 1} \{S(A_{ij+1} \cdots A_{ij+j}, B_{ij+1} \cdots B_{ij+j}) < qj\}\right) \\ &= P(S_j < qj)^{\lfloor n/j \rfloor} \\ &< (1 - n^{-1+\varepsilon/2})^{\lfloor n/j \rfloor} \quad \text{(for sufficiently large } n) \\ &\to 0. \end{aligned}$$

*Upper bound.* $P(M_n \geq (2 + \varepsilon)b \log n) \to 0$ as $n \to \infty$. Recall

$$r(q) = \lim -\frac{1}{k}\log P\left(\frac{S_k}{k} \geq q\right),$$

where $S_k = S(A_1 \cdots A_k, B_1 \cdots B_k)$. Now define $S_{ij} = S(A_1 \cdots A_i, B_1 \cdots B_j)$ and

$$r'(q) = \lim\left\{-\frac{1}{k}\log \max_{i+j=2k} P\frac{S_{ij}}{k} \geq q\right\}, \tag{10}$$

so that it is obvious that $r' \leq r$.

The next step is to show $r' = r$. We need to show $r' \geq r - \varepsilon$ for all $\varepsilon > 0$. Subadditivity implies

$$r(q) = \inf\left\{-\frac{1}{k}\log P\left(\frac{S_k}{k} \geq q\right)\right\}$$

and

$$r'(q) = \inf\left\{-\frac{1}{k}\log \max_{i+j=2k} P\left(\frac{S_{ij}}{k} \geq q\right)\right\}.$$

Pick $i, j, k = (i + j)/2$ such that

$$-\frac{1}{k}\log P\left(\frac{S_{ij}}{k} \geq q\right) < r' + \varepsilon.$$

Note that

$$P\left(\frac{S_{2k}}{2k} \geq q\right) \geq P\left(S_{ij} \geq qk, S(A_{i+1} \cdots A_{i+j}, B_{j+1} \cdots B_{j+i}) \geq qk\right)$$
$$= P\left(\frac{S_{ij}}{k} \geq q\right)^2.$$

This relies on the symmetry between $A$ and $B$, so that $S_{ij}$ and $S_{ji}$ have the same distribution. Hence

$$r(q) \leq -\frac{1}{2k}\log P\left(\frac{S_{2k}}{2k} \geq q\right) \leq -\frac{1}{2k}\log P\left(\frac{S_{ij}}{k} \geq q\right)^2 < r' + \varepsilon.$$

This establishes the claim that $r' = r$.

Since

$$r(q) = r'(q) = \inf_k \left\{-\frac{1}{k}\log \max_{i+j=2k} P\left(\frac{S_{ij}}{k} \geq q\right)\right\}, \tag{11}$$

we have, for all $i, j, k = (i + j)/2$ and for all $q$,

$$P(S_{ij} \geq qk) \leq e^{-kr(q)}. \tag{12}$$

Let $t = (2 + \varepsilon)b \log n$. The event $\{M \geq t\}$ is naturally expressed as a union of about $n^4$ events, by choosing the starting and ending points for the high-scoring regions. We break this up into a union containing on the order of $(n \log n)^2$ events that contribute substantially to the probability, and a second union for the remaining events. Let $C = 5/r(0)$, noting that $r(0) > 0$:

$$\{M_n \geq t\} \subseteq \left[\bigcup_{\substack{i_0, j_0 \in [1, n] \\ i, j \leq n;\ i+j \leq 2C \log n}} \left\{S(A_{i_0+1} \cdots A_{i_0+i}, B_{j_0+1} \cdots B_{j_0+j}) \geq t\right\}\right] \tag{13}$$
$$\cup \left[\bigcup_{\substack{i_0, j_0 \leq n \\ i, j \leq n;\ 2C \log n < i+j}} \left\{S(A_{i_0+1} \cdots A_{i_0+i}, B_{j_0+1} \cdots B_{j_0+j}) \geq 0\right\}\right].$$

In the first union, each event has probability at most

$$P(S_{ij} \geq t) = P(S_{ij} \geq qk) \leq e^{-kr(q)} = e^{-tr(q)/q} \leq n^{-(2+\epsilon)},$$

since with $k = (i + j)/2$, $t = qk$ we have

$$t\frac{r(q)}{q} = (2 + \epsilon)\left(\max \frac{c}{r(c)}\right)(\log n)\frac{r(q)}{q}$$
$$\geq (2 + \varepsilon)\log n.$$

Since this first union involves at most $n^2(2C \log n)^2$ events, the probability of the union satisfies

$$P\left( \bigcup_{\substack{i_0, j_0 \in [1,n] \\ i,j \le n;\, i+j \le 2C \log n}} \left\{ S(A_{i_0+1} \cdots A_{i_0+i}, B_{j_0+1} \cdots B_{j_0+j} \ge t \right\} \right)$$
$$\le n^2 (2C \log n)^2 n^{-(2+\varepsilon)}.$$

The second union in (13) is composed of at most $n^4$ events of the form $\{S_{ij} \ge 0\}$. Each of these events has probability satisfying

$$P(S_{ij} \ge 0) \le \exp[-kr(0)] \le \exp[-(C \log n) r(0)] \le \exp(-5 \log n),$$

since $k = (i+j)/2 > C \log n$ and $C = 5/r(0)$. Therefore,

$$P\left( \bigcup_{\substack{i_0, j_0 \in [1,n] \\ i,j \le n;\, 2C \log n < i+j}} \left\{ S(A_{i_0+1} \cdots A_{i_0+i}, B_{j_0+1} \cdots B_{j_0+1} \ge 0 \right\} \right)$$
$$\le n^4 \exp[-C \log nr(0)]$$
$$\le n^4 \exp(-5 \log n) = \frac{1}{n}.$$

This concludes the proof of Lemma 2, that $M_n$ grows like $\log(n)$. □

LEMMA 3. *If $a(\mu, \delta) > 0$, then $a(\mu, \delta)$ is the coefficient of $n$. More precisely, we have the following convergence in probability as $n \to \infty$:*

$$\frac{M_n}{n} \to_P a(\mu, \delta); \qquad \frac{S_n}{n} \to_P a(\mu, \delta). \tag{14}$$

PROOF. Since $M_n \ge S_n$, we need

$$P(M_n > (1+\varepsilon)na) \to 0 \tag{15}$$

and

$$P(S_n < (1-\varepsilon)na) \to 0. \tag{16}$$

The second half, (16), is just a corollary of (5) and subadditive ergodic theory, which implies that $S_n/n \to a$ almost surely.

To prove (15), set $t = (1+\varepsilon)na$.

Let

$$r = r((1+\varepsilon)a) = \lim \left\{ -\frac{1}{k} \log P\left( \frac{S_k}{k} \ge (1+\varepsilon)a \right) \right\}$$
$$= \inf \left\{ -\frac{1}{k} \max_{i+j=2k} \log P\left( \frac{S_{ij}}{k} \ge (1+\varepsilon)a \right) \right\},$$

where the last equality uses (11) from the proof of Lemma 2. Theorem 2 of Section 3 says $r > 0$. From the infimum above,

$$P\big(S_{ij} \geq a(1+\varepsilon)k\big) \leq e^{-rk},$$

for all $i, j, k = (i+j)/2$. For $i, j \leq n$, $k = (i+j)/2 \leq n$ so that

$$P\big(S_{ij} \geq t\big) = P\big(S_{ij} \geq (1+\varepsilon)na\big) \leq P\big(S_{ij} \geq ka(1+\varepsilon)\big) \leq e^{-rk}.$$

Now $\{S_{ij} \geq t\}$ requires $k = (i+j)/2 \geq t$ since each match scores $+1$, and hence

$$P\big(S_{ij} \geq t\big) \leq e^{-rk} \leq e^{-rt}. \tag{17}$$

Finally,

$$P(M_n \geq t) = P\Big(\bigcup_{i,j,k,l} \big\{S(A_{i+1} \cdots A_{i+k}, B_{j+1} \cdots B_{j+l}) \geq t\Big)$$

$$\leq n^4 e^{-rt} \to 0.$$

This completes the proof that $M_n$ grows like $n$ when $a > 0$. □

D. Haas observed that $M_n$ is also a Lipschitz functional of $n$ pairs of letters, so that the Azuma-Hoeffding inequality as used in Section 3 could be applied directly to $M_n$ in place of $S_n$. This shows at $P(|M_n - EM_n| \geq \epsilon n) \leq 2e^{-\epsilon^2 n/2c}$, which is summable in $n$. If we know that $EM_n/n \to a$, it would then follow that $M_n/n \to a$ both in probability and almost surely. However, it is not even obvious that $\lim EM_n/n$ exists; subadditivity is not directly applicable. The $n^4$ proof above shows that $M_n/n \to a$ in probability; and since $M_n/n$ is bounded it follows that $EM_n/n \to a$. Thus, using Azuma-Hoeffding on $M_n$ and $S_n$ proves an extension of Lemma 3: there is almost sure convergence in (14).

**3. Large deviations have exponentially small probability.** For the sake of proving a logarithmic versus linear phase transition, we are interested in applying the following theorem in situations where $a(\mu, \delta) < 0 \leq q$. Note, however, that Theorem 2 holds without restriction on the sign of $a(\mu, \delta)$.

THEOREM 2. *With* $A_1, A_2, \ldots$ *and* $B_1, B_2, \ldots$ *iid, for* $q > a(\mu, \delta) \equiv \lim(1/k)ES_k$,

$$\lim\left\{-\frac{1}{k}\log P(S_k \geq qk)\right\} > 0.$$

PROOF. We show a stronger result: that, without taking limits, we have $P(S_k \geq qk) \leq \exp(-k(q-a)^2/(2c^2))$ with $c = \min(2 + 4\delta, 2 + 2\mu)$.

We apply the Azuma–Hoeffding inequality, which is the following lemma.

LEMMA 4 (Azuma–Hoeffding). *Let* $X_i$ *be a martingale with* $X_0 = 0$ *such that, for some sequence* $c_i$, $i \geq 1$, *of positive constants,*

$$|X_{i-1} - X_i| \leq c_i.$$

*Then, for* $x > 0$,

$$P\Big(\sup_{i\le k} X_k \ge x\Big) \le \exp\left\{-\frac{x^2}{2}\Big/\sum_{i=1}^{k} c_i^2\right\}.$$

An outline of the proof of this lemma is given in Williams (1991).

We will apply this in a situation for which $c_i = c = \min(2 + 4\delta, 2 + 2\mu)$ and $X_k = S_k - ES_k$, so that

$$\begin{aligned} P(S_k \ge ES_k + \epsilon k) &= P(X_k \ge \epsilon k) \\ &\le P\Big(\sup_{i\le k} X_i \ge \epsilon k\Big) \\ &\le \exp\left(-\frac{1}{2}(\epsilon k)^2 \Big/ \sum_{i=1}^{k} c^2\right) \text{(Lemma 4)} \\ &= \exp\left(-\frac{\varepsilon^2 k}{2c^2}\right). \end{aligned}$$

We use this with $\varepsilon = q - a(\mu, \delta) = q - a > 0$. From subadditivity, $ES_k \le ka$, so $P(S_k \ge qk) \le P((S_k - ES_k) \ge (q - a)k)$. Combining this with Azuma-Hoeffding, taking logarithms and dividing by $-k$, we get

$$-\frac{1}{k}\log P(S_k \ge qk) \ge \frac{(q-a)^2}{2c^2} > 0.$$

In detail, our martingale is $X_i = E(Y|\mathcal{F}_i)$, where $Y = S_k - ES_k$ and $\mathcal{F}_i = \sigma(C_1, C_2, \ldots, C_i)$, where $C_i = (A_i, B_i)$ is the $i$th pair of letters. Since $S_k$ is $\mathcal{F}_k$-measurable, our martingale has $X_k = S_k - ES_k$, and since $\mathcal{F}_0$ is trivial, $X_0 = 0$.

To bound the martingale increments, we first give a deterministic bound

$$S - S' \le c = \min\{2 + 4\delta, 2 + 2\mu\}, \tag{18}$$

where $S = S(a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_k) = S(c_1, c_2, \ldots, c_k)$ is the score for $k$ pairs of letters, and $S' = S(c_1, \ldots, c_{i-1}, c_i', c_{i+1}, \ldots, c_k)$ is the score when the $i$th pair of letters is changed. The argument for (18) starts by fixing a particular optimal alignment for $S$. Now there are three cases, depending how the $i$th pair of letters was aligned. In the first and dominant case, both $a_i$ and $b_i$ were matched successfully, $a_i$ matching $b_j$ for $i \ne j$ and $b_i$ matching $a_l$ for $l \ne i$. No matter what letters $a_i'$ and $b_i'$ are, good alignments for $S'$ are given by (1) deleting the four letters $a_i', b_j, b_i', a_l$, which scores $S - 2 - 4\delta$, and by (2) scoring $a_i$ aligned to $b_j$ and $b_i$ aligned to $a_l$ as two mismatches, which scores $S - 2 - 2\mu$. Thus $S' \ge S - \min(2 + 4\delta, 2 + 2\mu)$.

It is elementary to go from a bound on changes in the deterministic scoring function, namely, $|S(c_1, \ldots, c_k) - S(c_1', \ldots, c_k')| \le c$ whenever $c_i = c_i'$ for all but one $i$, to a bound on the martingale increments, namely, $|X_i - X_{i-1}| \le c$. [See, e.g., Alon and Spencer (1992).] We give the proof below for completeness,

and to highlight where the assumption that $C_1, C_2, \ldots$ are independent gets used. In Section 5, we show how to extend this to the case where $C_1, C_2, \ldots$ are Markov. The martingale increment $X_i - X_{i-1}$ is given by a deterministic function $g$ applied to $i$ pairs of letters: $X_i - X_{i-1} = g(C_1 C_2 \cdots C_i)$, where $g$ satisfies

$$\begin{aligned} g(c_1, \ldots, c_i) &= \sum_{c_{i+1}, \ldots, c_k} P(C_{i+1} \cdots C_k = c_{i+1} \cdots c_k) S(c_1 \cdots c_i \cdots c_k) \\ &\quad - \sum_{c_i', c_{i+1}, \ldots, c_k} P(C_i C_{i+1} \cdots C_k = c_i' c_{i+1} \cdots c_k) S(c_1 \cdots c_i' \cdots c_k) \\ &= \sum_{c_i', c_{i+1}, \ldots, c_k} P(C_i C_{i+1} \cdots C_k = c_i' c_{i+1} \cdots c_k)(S - S'). \end{aligned}$$

Thus

$$|g(c_1, \ldots, c_i)| \leq \max(S - S').$$

**4. Generalized scoring and gapping.** Our results on phase transitions have been established for a simple alignment scoring function, where aligned letters $x$ and $y$ score

$$s(x, y) = \begin{cases} 1, & \text{if } x = y, \\ -\mu, & \text{if } x \neq y, \end{cases} \tag{19}$$

and deleted letters score $-\delta$ per letter. Biological sequence alignment often requires much more complex scoring schemes. Protein sequences are sequences over the 20-letter alphabet of amino acids. There are, then, 210 distinct unordered pairs of letters that each could receive values depending on an array of chemical and physical properties of the amino acids. The gap functions also can be more complex. A gap of length $k$ receives weight $w(k) = \delta k$ in Section 3, while in practice $w(k) = \alpha + \beta k$ is popular for biological and algorithmic reasons. Also algorithms for alignment with concave $w(k)$ have been studied [Waterman (1984), Miller and Myers (1988) and Eppstein, Galil, Giancarlo and Italiano (1989)]. In this section we sketch the extension of our theorems for these more general scoring schemes.

In general, a gap weight function $g(k) \geq 0$ might not be concave or even monotone. The goal is to delete $i$ adjacent letters at minimum cost: $w(i) = \min\{g(i_1) + g(i_2) + \cdots + g(i_l): i_1 + i_2 + \cdots + i_l = i\}$. Using the function $w$ in the alignment algorithm will give correct alignment scores and, at the cost of recalling the exact composition of each long "gap," correct alignments. Note that $w$ is necessarily subadditive: $w(k + l) \leq w(k) + w(l)$. Note, for example, with $g(0) = 0$, $g(1) = 1$, $g(2) = 0$, $g(k) = k$ for $k \geq 3$, we get $w(k) = 0$ if $k$ is even and $w(k) = 1$ if $k$ is odd, which is not concave or monotone, but is subadditive!

The alignment in (3), using $l$ pairs of letters, involves $l + 1$ gaps in each of the two sequences. The lengths of the $k$th gap are $a(k) - a(k - 1) - 1$ in the $A$ sequence, $b(k) - b(k - 1) - 1$ in the $B$ sequence, for $k = 1$ to $l + 1$. The score with penalty $w(i)$ for a gap of length $i$ [we define $w(0) = 0$] for the

alignment in (3) is

$$(20)\qquad \begin{aligned} &-\sum_{k=1}^{l+1} w(a(k)-a(k-1)-1) \\ &-\sum_{k=1}^{l+1} w(b(k)-b(k-1)-1)+\sum_{k=1}^{l} s(A_{a(k)},B_{b(k)}) \end{aligned}$$

In the special case $w(i)=\delta i$, this reduces to the alignment score whose maximum is taken in (2). The subadditive property in (5) used in Section 2 holds for our generalized scoring since $w(i+j)\le w(i)+w(j)$, for all $i,j\ge 0$.

THEOREM 1′. *Consider iid letters $A_1, A_2, \ldots$ and $B_1, B_2, \ldots$ and the optimal alignment score $M_n = M(A_1A_2\cdots A_n, B_1B_2\cdots B_n)$ with symmetric scoring $s(x,y)=s(y,x)$ and subadditive gap weights $w(k)$ used in* (1) *and* (20). *Let $a=\lim ES_k/k$ be the limiting score per letter. If $a>0$, then $M_n$ grows linearly with coefficient $a$, and if $a<0$, then $M_n$ grows logarithmically with coefficient in $[b,2b]$ in the sense of* (9).

PROOF. Lemmas 2 and 3 proceed as above. Note that in the argument just before (11), for $S_{ij}$ and $S_{ji}$ to have the same distribution requires that the scoring matrix be symmetric. In the argument leading up to (17), the event $\{S_{ij}\ge t\}$ requires $k=(i+j)/2\ge t/s^*$, where $s^*=\max s(a,b)>0$. (If $s^*\le 0$, we do not have $a>0$.) Now (17) becomes

$$P(S_{ij}\ge t)\le e^{-rt/s^*}.$$

Observe that Theorem 1′ is not exactly parallel to Theorem 1; we are no longer discussing a "phase transition line" separating the regions $a<0$ from $a>0$. In part, this is because we do not have a single notion of the space of parameters $\{s(\cdot,\cdot), w(\cdot,\cdot)\}$. Next we discuss two examples.

EXAMPLE 1. For an alphabet of size $d$, we consider $d(d+1)/2$ parameters $(s_{ij})$ with $s_{ij}=s_{ji}$, and two additional parameters for the gap penalty $g(k)=\alpha+\beta k$. In this parameter space of dimension $d(d+1)/2+2$, there is a surface of codimension 1 separating the regions $\{a>0\}$ and $\{a<0\}$. This can be seen by an argument like the proof of Lemma 1.

EXAMPLE 2. Let $w(k)=\delta\log(1+k)$, which has $\lim w(k)/k=0$, and hence $a\ge 0$, independent of the choice of $\delta$ and $s(\cdot,\cdot)$. Consider examples with $s(a,b)=\mathbf{1}(a=b)-\mu\mathbf{1}(a\ne b)$. We conjecture that, for sufficiently large $\delta$ and $\mu$, $P(S_t/t\ge 0)\to 0$ as $t\to\infty$. If this conjecture were proven, then for this class of examples, the region $\{a=0\}$ in the parameter space $[0,\infty]^2$ would have nonempty interior and positive area.

In order to prove Theorem 1′, we need an analogue of Theorem 2 for these general scoring schemes.

THEOREM 2′. *For the generalized scoring scheme given by* (20), *with* $A_1, A_2, \ldots$ *and* $B_1, B_2, \ldots$ *iid and with* $a = \lim ES_k/k$, *for* $q > a$

$$\lim\left\{-\frac{1}{k}\log P(S_k \geq qk)\right\} > 0.$$

PROOF. The proof of Theorem 2 requires only one modification. With

$$s^* = \max_{i,j} s(i,j),$$

$$s_* = \min_{i,j} s(i,j),$$

the upper bound $c$ on the martingale increments is

$$c = \max\{\min(2s^* + 4w(1), 2s^* - 2s_*), 0\}.$$

This comes from the deterministic bound on scores where the $i$th pair of letters is changed, which we show below:

$$S - S' \leq c.$$

If $a_i$ is matched to $b_j$ and $b_i$ is matched to $a_l$, scoring at most $2S^*$ for these matches, then after changes to $a_i', b_i'$, feasible alignments include matching $a_i'$ to $b_j$ and $b_i'$ to $a_l$, for a score of $2s_*$, or else deleting all four letters, for an additional gap penalty (using subadditivity of $w$) of at most $4w(1)$. Otherwise $a_i$ is matched to $b_i$, and $S - S' \leq \max\{0, \min(1s^* + 2w(1), s^* - s_*)\}$, corresponding to replacing one good match by a pair of deletions or a bad match. The maximum with 0 in the value of $c$ is needed in case $s^* + 2w(1) \leq 0$, in which case $S_k = -2w(k)$ is constant, regardless of the sequences; here $c = 0$ and the upper bound $P(S_k \geq qk) \leq \exp(-k(q - a)^2/0)$ is correct in the sense that $0 \leq \exp(-\infty)$. □

**5. Matching Markov chains.** In this section, we extend the results of Theorems 1 and 2 from the case of iid sequences to the case of Markov chains. This requires a small extension of the usual method of applying the Azuma–Hoeffding inequality.

*Azuma–Hoeffding for Lipschitz functions of Markov chains.* In the usual setup, $C_1, \ldots, C_n$ are independent (not necessarily identically distributed), $f$: $R^n \to R$ is a deterministic function, $\mathcal{F}_i = \sigma(C_1, \ldots, C_i)$, $S = f(C_1, \ldots, C_n)$, and the martingale is $X_i = E(S - ES|\mathcal{F}_i)$, with $X_0 = 0$, $X_n = S - ES$. We assume that $f$ is Lipschitz in the sense that changing a single coordinate of the input to $f$ changes the value of $f$ by at most $c$. It follows, using the independence of the coordinates $C_i$, that the martingale increments satisfy $|X_i - X_{i-1}| \leq c$. We presented this, in the special case of scoring, at the end of Section 3. For the case of Markov chains, we present a bounded martingale increments argument in a general setting, since it should be broadly applicable. Note that irreducible, aperiodic Markov chains with a finite state space always satisfy the uniformly bounded expected coupling time hypothesis of the following lemma.

LEMMA 5. *Assume* $f$: $\mathbb{N}^n \to R$ *is Lipschitz*$(c)$, *in the sense that if* $\mathbf{y}, \mathbf{y}' \in \mathbb{N}^n$ *differ in at most one coordinate, then* $|f(\mathbf{y}) - f(\mathbf{y}')| \le c$. *Assume* $Y_1, Y_2, \ldots$ *is a Markov chain with state space* $\mathbb{N}$ *which can be coupled, in expected time less than or equal to* $t$, *uniformly over initial states. Consider* $S = f(Y_1, \ldots, Y_n)$ *and* $\mathcal{F}_i = \sigma(Y_1, Y_2, \ldots, Y_i)$, *and suppose that* $E|S| < \infty$ *so that* $X_i = E(S - ES|\mathcal{F}_i)$ *defines a martingale. Then, for* $i = 1$ *to* $n$, *the martingale increments satisfy*

$$|X_i - X_{i-1}| \le ct.$$

PROOF. In detail, our coupling assumption is the following. For all $a, b \in \mathbb{N}$, it is possible to construct $Y_0 = a, Y_1, Y_2, \ldots$ and $Y_0' = b, Y_1', Y_2', \ldots$, each a realization of the Markov chain, so that with the coupling time $\tau$ defined by

$$\tau = \min\{j \ge 0 \colon Y_j = Y_j'\},$$

we have $Y_j = Y_j'$ for all $j \ge \tau$ and

$$E_{ab}(\tau) \le t < \infty.$$

The martingale $X_i$, on the event $\{Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}, Y_i = y_i = a\}$, has value

$$\begin{aligned} X_i &= h(y_1, y_2, \ldots, y_{i-1}, a) \\ &= \sum_{y_{i+1}, \ldots, y_n} f(y_1, \ldots, y_n) P(Y_{i+1} \cdots Y_n = y_{i+1} \cdots y_n | Y_i = a) \\ &= \sum_{y_{i+1}, \ldots, y_n} f(y_1, \ldots, y_n) P^a(Y_1 \cdots Y_{n-i} = y_{i+1} \cdots y_n). \end{aligned}$$

Observe that the previous martingale value is an average of these $X_i$:

$$X_{i-1} = \sum_b h(y_1, \ldots, y_{i-1}, b) P(Y_i = b | Y_{i-1} = y_{i-1}).$$

Thus the martingale increment on the event $\{Y_1 \cdots Y_i = y_1 \cdots y_{i-1}a\}$ is

$$\begin{aligned} X_i - X_{i-1} = \sum_b & P(Y_i = b | Y_{i-1} = y_{i-1}) \\ & \times (h(y_1, \ldots, y_{i-1}, a) - h(y_1, \ldots, y_{i-1}, b)). \end{aligned}$$

Our lemma follows if we can show, for all $y_1, \ldots, y_{i-1}, a, b$,

$$|h(y_1, \ldots, y_{i-1}, a) - h(y_1, \ldots, y_{i-1}, b)| \le ct.$$

With $Y_1, Y_2, \ldots$ and $Y_1', Y_2', \ldots$ copies of the Markov chain, starting from $a$ and $b$, we have

$$\begin{aligned} h(y_1, \ldots, y_{i-1}, a) &= E^a f(y_1, \ldots, y_{i-1}, a, Y_1, \ldots Y_{n-i}), \\ h(y_1, \ldots, y_{i-1}, b) &= E^b f(y_1, \ldots, y_{i-1}, b, Y_1', \ldots Y_{n-i}'). \end{aligned}$$

The function $f$ is applied above at two points of $\mathbb{N}^n$ which differ in $\min(n+1-i, \tau)$ coordinates, so that the Lipschitz property of $f$ implies

$$|f(y_1,\ldots,y_{i-1},a,Y_1,\ldots,Y_{n-i}) - f(y_1,\ldots,y_{i-1},b,Y'_1,\ldots,Y'_{n-1})| \le c\tau.$$

Taking expectations completes the proof. □

In our version of Theorem 1″, we assume for convenience that each sequence $A_1A_2\cdots$ and $B_1B_2\cdots$ is governed by an irreducible aperiodic Markov chain on a finite alphabet $\mathscr{A}$ (so that there exists $n$ such that for all $i, j \in \mathscr{A}$, $[P^n]_{ij} > 0$) and that the two sequences are independent. This makes the process $C_1C_2\cdots$ of pairs of letters, with $C_i = (A_i, B_i)$, an irreducible aperiodic Markov chain with state space $S = \mathscr{A}^2$.

*Subadditivity for the average score.* Subadditivity for the sequence of scores $S_k = S(C_1\cdots C_k)$ needs to be handled more carefully for Markov chains than for iid sequences. It is only by starting the chain $C_1C_2\cdots$ in *equilibrium* that the random variables $X_{mn} = S(C_{m+1}\cdots C_n)$, for $0 \le m \le n$, have the required stationary distribution. We thus require the assumption that both sequences $A_1A_2\cdots$ and $B_1B_2$ start in equilibrium in (6),

$$a(\mu,\delta) = \lim_{k\to\infty}\frac{ES_k}{k} = \sup_{k\ge 1}\frac{ES_k}{k},$$

the definition of $a(\mu,\delta)$.

Now, for the sake of Theorem 1″, there is no need to restrict the initial distributions of our sequences. To show that $\lim(ES_k/k)$ does not depend on the initial distribution, and for proving the extension of Lemma 3 to the Markov case with arbitrary initial distribution, we can use a coupling of $C_1C_2\cdots$ to another copy of the chain, $C'_1C'_2\cdots$, starting in equilibrium. Observe that, if $C_i = C'_i$ for all $i \ge \tau$, then $|S(C_1\cdots C_k) - S(C'_1\cdots C'_k)| \le 4\delta\tau$, by considering alignments which delete all letters before the coupling time $\tau$. Using $E\tau < \infty$, we see, for example, that $\lim(ES_k/k)$ does not depend on the initial distribution. We note also that Lemma 3 can be proved, even for the Markov case, entirely by the Azuma–Hoeffding inequality applied to $X_i = -E(S_k - ES_k|\mathscr{F}_i)$, and we do not need to rely on Kingman's subadditive ergodic theorem.

*Subadditivity for the exponential decay rate.* For Markov chains in place of iid sequences, the subadditivity needed to establish (7), that

$$r(q) = \lim\left\{-\frac{1}{k}\log P(S_k \ge qk)\right\} = \inf\left\{-\frac{1}{k}\log P(S_k \ge qk)\right\},$$

is slightly more delicate. It is necessary to keep track of the initial state $c_1 = (a_1, b_1)$ when analyzing probabilities involving the score $S_k = S(C_1,\ldots,C_k)$.

We sketch two arguments: a simple one given in this and the next paragraph, in the case $P_{ij} > 0$ for all $i, j \in \mathscr{A}$, and a more complicated argument for the case $P$ irreducible and aperiodic, but with some value $P_{ij} = 0$. We define

$$p(k, q) = \max_{c_1 \in \mathscr{A}^2} P^{c_1}(S_k \geq qk).$$

Defining this $p$ with minimum in place of maximum would make it easy to prove subadditivity, but then we would not get the required upper bound on probabilities. Consider

$$\lambda = \min\{P_{ij}: i, j \in \mathscr{A}\} > 0.$$

Then $\lambda^2$ is the corresponding minimum for our Markov chain on $S = \mathscr{A}^2$. Now, using the deterministic subadditive property of scoring only in the first line below,

$$\begin{aligned} P^{c_1}\big(S_{j+k} \geq q(j+k)\big) &\geq P^{c_1}\big(S_j \geq qj \text{ and } S(C_{j+1} \cdots C_{j+k}) \geq qk\big) \\ &= \sum_{c_{j+1}} P^{c_1}(S_j \geq qj \text{ and } C_{j+1} = c_{j+1}) P^{c_{j+1}}(S_k \geq qk) \\ &\geq \sum_{c_{j+1}} P^{c_1}\big(S_j \geq qj\big) \lambda^2 P^{c_{j+1}}(S_k \geq qk) \\ &\geq P^{c_1}\big(S_j \geq qj\big) \lambda^2 p(k, q). \end{aligned} \tag{21}$$

In the last step, we bound a sum from below by one of its terms. Taking the maximum over $c_1$ yields the following: for all $q$, for all $j, k \geq 1$,

$$p(j+k, q) \geq p(j, q) \lambda^2 p(k, q).$$

Thus, $-\log(\lambda^2 p(k, q))$ is subadditive in $k$, so that

$$r(q) = \inf\left\{-\frac{1}{k}\log\big(\lambda^2 p(k, q)\big)\right\}$$

satisfies

$$r(q) = \lim\left\{-\frac{1}{k}\log p(k, q)\right\}.$$

For the proof of the upper bound in Lemma 2, we also consider the large deviation rate $r'$ for scoring sequences of possibly different lengths. In the context of Markov chains, (10), the definition of $r'$, should be modified to include a maximum over initial states $c_1$, so that now

$$r'(q) \equiv \inf\left\{-\frac{1}{k}\log\left[\lambda^2 \max_{i+j=2k} \max_{c_1 \in S} P^{c_1}\big(S_{ij} \geq qk\big)\right]\right\}.$$

As before, one must prove that $r = r'$. It then follows that, for any initial distribution, for any $i, j$ with $i + j = 2k$,

$$P\big(S_{ij} \geq qk\big) \leq \lambda^{-2} e^{-k r(q)}.$$

With the extra factor of $\lambda^{-2}$, our proof of the upper bound in Lemma 2, based on the probability of a union being less than the sum of the probabilities, goes through as in the iid case.

To handle the case of $P$ irreducible and aperiodic on $\mathscr{A}$, but not strictly positive, we need a more complicated subadditivity argument. This is because $\lim\{-(1/k)\log P^c(S_k \geq qk)\}$ can vary with the initial state $c$, as we see from the example with $q = 1$ where $\{S_k \geq qk\}$ requires perfect matching, which has probability zero starting from a mismatch and has nonzero but exponentially decaying probability starting from a match. With various scoring schemes, there can be less trivial examples in which the event $\{S_k \geq qk\}$, of exponentially small positive probability, confines the chain to some part of the state space, so that the irreducible aperiodic property is not seen. In any case, we cannot bound $\liminf P^{c_1}(C_j = b | S_j \geq qj)$ away from zero, which was the essence of the argument in (21).

For the irreducible aperiodic case in general, we define

$$r(q) = \lim\left\{ -\frac{1}{n}\log \min_{c \in \mathscr{A}^2} P^c(S_n \geq qn) \right\}, \tag{22}$$

which exists by subadditivity. Noting that $r(\cdot)$ is a nondecreasing function, we define

$$r(q-) = \sup_{q' < q} r(q'),$$

so that $r(q-) \leq r(q)$. Next we argue that, for each $\varepsilon > 0$ and $q$, there exists $k_0$ finite so that for all $k \geq k_0$, for all initial $c \in \mathscr{A}$,

$$P^c(S_k \geq qk) \leq \exp[-(r(q-) - \varepsilon)k]. \tag{23}$$

Observing that $b = \max_{q>0} r(q)/q = \max_{q>0} r(q-)/q$, it is then straightforward to extend the proof of the upper bound in Lemma 2 to the case of irreducible, aperiodic Markov chains.

To prove (23), fix $m$ so that $\lambda = \min[P^m]_{ij} > 0$, and fix $\varepsilon > 0$ and $q$. Pick $q' < q$ so that $r(q') > r(q-) - \varepsilon/2$. Pick $k_0$ sufficiently large that $k_0 q - 2\delta m \geq (k_0 + m)q'$ and $\exp(-k_0\varepsilon/2) < \lambda^2$. Now let $k \geq k_0$ and consider blocks of length $l = k + m$. For a block of length $l$, if the $k$ pairs of letters score $qk$, then by considering an alignment in which the other $m$ pairs are deleted we see that the net score is at least $kq - 2\delta m \geq lq'$. We have

$$P^c(S_k \geq qk \text{ and } C_{l+1} = c) \geq \lambda^2 P^c(S_k \geq qk).$$

For $n = ld$ a multiple of $l$, for any $c_1$, $P^{c_1}(S_n \geq q'n) \geq (\lambda^2 P^c(S_k \geq qk))^d$, by insisting on $C_{m+1} = C_{m+l+1} = \cdots = C_{m+(d-1)l+1} = c$, deleting the first $m$ pairs of letters and scoring at least $qk$ from the remaining $k$ pairs in each successive block of $l$ pairs of letters. Thus, taking $l \to \infty$ and picking $c_1$ to achieve the minimum in the definition of $r$, (22), we have

$$\begin{aligned} e^{-nr(q')} &\approx P^{c_1}(S_n \geq q'n) \geq \left(\lambda^2 P^c(S_k \geq qk)\right)^d \\ &\geq \left(e^{-k\varepsilon/2} P^c(S_k \geq qk)\right)^{n/k} \end{aligned}$$

so that

$$e^{-r(q')} \geq e^{-\varepsilon/2}\big(P^c(S_k \geq qk)\big)^{1/k}.$$

Hence

$$P^c(S_k \geq qk) \leq \big(\exp[-r(q') + \varepsilon/2]\big)^k \leq \exp[-(r(q-) - \varepsilon)k],$$

which proves (23). The remaining details needed to extend the upper bound in Lemma 2 to this irreducible, aperiodic Markov case may be considered routine.

*Doeblin's method.* For the lower bound in Lemma 2, our argument required independent blocks. In the Markov case, Doeblin's method provides these. Here are some of the details. As before, we take $t = (1 - \varepsilon)b \log n$, $k = \lfloor t/q \rfloor$ and $j = k + 1$. Pick a pair of letters $c$ so that, for sufficiently large $k$,

$$P^c(S_j \geq qj) \geq \exp[-j(r(q) + \delta)] \geq n^{-1+\varepsilon/2}.$$

Consider the successive returns to $c$: $T(0) \equiv 0$, $T(i+1) = \min\{t > T(i): C_t = c\}$. The $i$th excursion is $C_{T(i)}C_{T(i)+1} \cdots C_{T(i+1)-1}$, and for $i = 1, 2, \ldots$, these excursions are iid (with an infinite state space for the excursions, even for finite $S$). Also, successive blocks of $j$ consecutive excursions are iid. Consider $\mu_c$, the equilibrium probability of $c$. The weak law of large numbers, $T(i)/i \to_P 1/\mu_c$, implies that the "good" event $G$, defined by

$$G = \{T_{lj} \leq n\}, \quad \text{where } l = n\mu_c/(2j),$$

has $P(G) \to 1$ as $n \to \infty$. Let $E_i$ be the event that the first $j$ pairs of letters in the $i$th block score less than $qj$:

$$E_i = \big\{S(C_{T(ij)+1} \cdots C_{T(ij)+j}) < qj\big\}.$$

From the iid property of excursions,

$$P\Big(\bigcap_{0 \leq i < l} E_i\Big) = P(E_1)^l = P^c(S_j \geq qj) \leq (1 - n^{-1+\varepsilon/2})^l \to 0$$

since $ln^{-1+\varepsilon/2} \to \infty$. Finally, $\{M_n < t\} \cap G \subset \bigcap_{0 \leq i < l} E_i$ so

$$P(M_n < t) \leq P(\cap E_i) + (1 - P(G)) \to 0.$$

These arguments have proven the following two theorems.

THEOREM 1″. *The results of Theorems* 1 *and* 1′ *remain valid in case the two sequences* $A_1A_2 \cdots$ *and* $B_1B_2 \cdots$ *are independent, each governed by the same irreducible aperiodic Markov chain on a finite alphabet, with arbitrary initial distributions for each sequence.*

THEOREM 2″. *The results of Theorems* 2 *and* 2′ *extend to the Markov situation in Theorem* 1″.

**6. Repeats in a sequence.** Biological sequences can evolve by duplicating intervals of sequence. These duplications can appear adjacent to or distant from the original interval of sequence. This motivates us to consider our theorems for approximate repeats within a sequence. We will take these repeats to be nonoverlapping.

First the definition of $M_n$ must be modified appropriately to $M_n'$:

$$M_n' = M'(A_1 A_2 \cdots A_n) = \max_{I,J} S(I, J)$$

where $I = A_{g+1} \cdots A_{g+i}$ and $J = A_{h+1} \cdots A_{h+j}$ and $1 \le g+1 \le g+i < h+1 \le h+j \le n$. Other definitions such as that for $r(q)$ remain the same, as the behavior of $S(I, J)$ is still the key for the proofs.

First we handle the logarithmic region. The lower bound is straightforward. For all $\varepsilon > 0$,

$$\begin{aligned} M'(A_1 \cdots A_n) &\ge M(A_1 \cdots A_{\lfloor n/2 \rfloor}, A_{\lfloor n/2 \rfloor + 1} \cdots A_n) \\ &\ge (1-\varepsilon) b \log(n/2) \end{aligned}$$

with probability tending to 1. Since $\log(n/2)$ is asymptotic to $\log n$, it follows that for all $\varepsilon > 0$,

$$\mathbb{P}(M'(A_1 \cdots A_n) > (1-\varepsilon) b \log n) \to 1.$$

For the upper bound, the unions in (13) are over fewer events, namely, $i_0 < i_0 + i \le j_0 < j_0 + j$ and $i + j \le 2C \log n$ (first union) or $2C \log n < i + j$ (second union). The first union still involves at most $n^2(2C \log n)^2$ events and the second union still involves at most $n^4$ events. Since all other bounds hold without change,

$$M'(A_1 \cdots A_n) \le (2+\varepsilon) b \log n$$

with probability tending to 1, and Lemma 2 holds with $M_n'$ replacing $M_n$.

When $a(\mu, \delta) > 0$, we have the following convergence in probability:

$$\frac{M_n'}{n} \to_P \frac{a(\mu, \delta)}{2}, \tag{24}$$

which is the counterpart of Lemma 3 but with an extra factor of $\frac{1}{2}$. To prove this, first note that

$$M_n' \ge S(A_1 \cdots A_{\lfloor n/2 \rfloor}, A_{\lfloor n/2 \rfloor + 1} \cdots A_n).$$

Therefore,

$$\liminf \frac{M_n'}{n} \ge \frac{a}{2},$$

and the lower bound is established. To check the upper bound note that the union following (17) is over $1 \le i < i + k \le j < j + l \le n$, which is fewer than $n^4$ events and here $(i + j)/2 \le n/2$.

This generalization to repeats is summarized in the next theorem.

THEOREM 1‴. *Consider iid letters* $A_1, A_2, \ldots$ *and the optimal alignment score* $M'_n = M'(A_1 A_2 \cdots A_n)$ *with symmetric scoring* $s(x, y) = s(y, x)$ *and subadditive gap weights* $w(k)$ *used in* (1) *and* (20). *Let* $a = \lim ES_k/k$ *be the limiting score per letter. If* $a > 0$, *then* $M'_n$ *grows linearly with coefficient* $a/2$, *and if* $a < 0$, *then* $M_n$ *grows logarithmically with coefficient in* $[b, 2b]$ *in the sense of* (9).

**7. Examples.** This paper was motivated by the application of sequence matching algorithms to the study of DNA sequences with a 4-letter alphabet and protein sequences with a 20-letter alphabet. In Fall 1993 there were about $150 \times 10^6$ letters of DNA sequence in the international databases, which were contained in about 100,000 sequence entries that average around 1000 letters each. The longest contiguous sequence of DNA is the complete sequence of a yeast chromosome, 315,357 letters in length. The Human Genome Project promises to accelerate the rate of DNA sequencing. Currently the databases increase in size by about 50% per year.

To determine relationships between sequences, the Smith-Waterman algorithm computes $M \equiv M_n \equiv M(\mu, \delta)$, the optimal score $S(I, J)$ over all possible contiguous regions where $I$ is from $A_1 A_2 \cdots A_n$ and $J$ is from $B_1 B_2 \cdots B_n$. As earlier in this paper, $\mu$ is the mismatch penalty and $\delta$ is the single letter deletion penalty. Our theorems show that there is a phase transition between linear growth in $n$, when the penalty parameters are small, and logarithmic growth in $n$, when the penalties are large. Not much is known theoretically about the location of the phase transition curve in $(\mu, \delta) \in [0, \infty]^2$. If $p = P(A_i = B_j)$, then the point $(p/(1 - p), \infty)$ lies on the curve since $\mu = p/(1 - p)$ solves $1 \cdot p - \mu(1 - p) = 0$. Similarly, if $c$ is the Chvátal–Sankoff constant, $(\infty, c/(2(1 - c)))$ lies on the curve. Noting that whenever $\mu \ge 2\delta$, $M(\mu, \delta) = M(2\delta, \delta)$, we have $(\mu, c/(2(1 - c)))$ on the curve when $\mu \ge (c/(1 - c))$.

To obtain more information about the shape of the phase transition curve, we studied

$$\{(\mu, \delta): S(A_1 A_2 \cdots A_{5000}, B_1 B_2 \cdots B_{5000}) = 0\} \tag{25}$$

for simulated DNA sequences of length 5000 and $p = P(A_i = B_j) = 1/4$. The study is motivated by the definition of the phase transition curve, $\{(\mu, \delta): \lim_{k \to \infty} ES_k/k = 0\}$. The simulated curve appears as Figure 1. To check the simulation, note that $p/(1 - p) = 1/3$ when $p = 1/4$. For a uniform four-letter alphabet, it is known that $c \in (0.45, 0.77)$ so that $c/(2(1 - c)) \in (0.41, 1.67)$. These values are consistent with the results of our simulation.

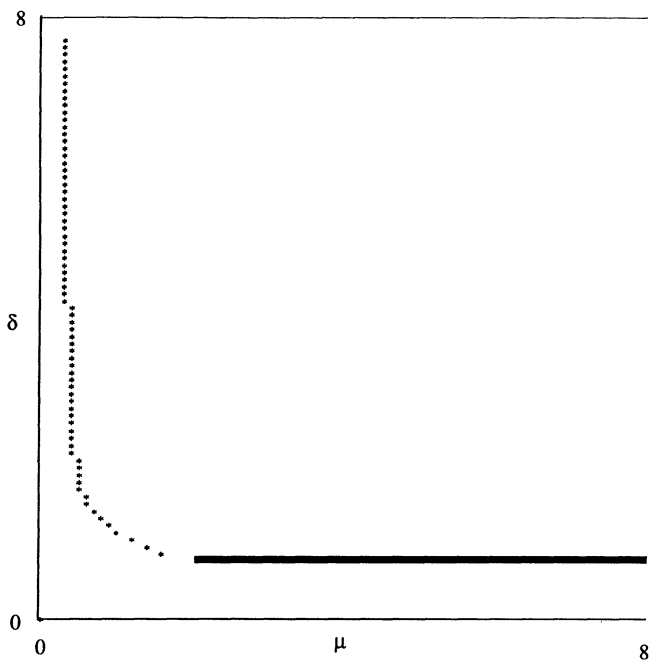


FIG. 1. *The approximate phase-transition curve for* $p = 1/4$; *phase-transition boundary for two length-400 DNA sequences.*

The phase transition curve in $[0,\infty]^2$ does not tell the whole story. It is instructive to look at a line through the parameter space and study the behavior of optimal alignments along the line. We chose the line $\mu = \delta$. For the simulated length-64 sequences

**A** = GTCTGACAAAGGCAACTCAAGGTAGGACTGGCGTCCAATAGCCAATCAGCATATCTTTTTTATC,
**B** = CGTCGGGTGTACTCAATCCTTGAGTTCGCTTAGCTATCTGGCCAGCCGCATCTCGAGACGTAAC,

there are nine regions of optimal alignments, that is, $[0,\infty] = \cup_{i=1}^{9} I_i$, where $I_1 = [0, 0.2]$, $I_2 = [0.2, 0.5]$, $I_3 = [0.5, 0.7]$, $I_4 = [0.7, 0.8]$, $I_5 = [0.8, 0.9]$, $I_6 = [0.9, 1.0]$, $I_7 = [1.0, 1.25]$, $I_8 = [1.25, 3.0]$ and $I_9 = [3, \infty]$. On the interior of each $I_i$, the set of optimal alignments is constant. It should be pointed out that optimal alignments are not always unique. For these sequences, letting $n_i$ = number of optimal alignments for interval $I_i$, $n_1 = 302{,}400$, $n_2 = 14{,}688$, $n_3 = 1632$, $n_4 = 1440$, $n_5 = 1440$, $n_6 = 6$, $n_7 = 6$, $n_8 = 1$ and $n_9 = 1$. In Figure 2 we have taken a representative alignment from each interval. Note that Figure 1 suggests the phase transition should occur around $(\mu, \delta) = (1, 1)$, the region in Figure 2 where the alignment lengths and scores are dramatically changing.


FIG. 2. *Selected alignments from each interval of optimal alignments.*

**Acknowledgments.** We are indebted to Rick Durrett, who pointed out the article of Grimmett and Kesten to the first author when he was in residence at the Institute of Mathematics and its Applications at the University of Minnesota. We are also grateful to Mark Eggert for computational assistance.

## REFERENCES

ALON, N and SPENCER, J. H. (1992). *The Probabilistic Method.* Wiley, New York.

APOSTOLICO, A., CROCHEMORE, M., GALIL, Z. and MANBAR, U., eds. (1992). *Combinatorial Pattern Matching. Lecture Notes in Computer Science* **644**. Springer, Berlin.

ARRATIA, R. A., GORDON, L. and WATERMAN, M.S. (1990). The Erdös–Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.* **18** 539–570.

ARRATIA, R. A. and WATERMAN, M.S. (1985a). An Erdös–Rényi law with shifts. *Adv. Math.* **55** 13–23.

ARRATIA, R. A. and WATERMAN, M.S. (1985b). Critical phenomena in sequence matching. *Ann. Probab.* **13** 1236–1249.

ARRATIA, R. A. and WATERMAN, M.S. (1989). The Erdös–Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.* **17** 1152–1169.

CAPOCELLI, R. M. (1990). *Sequences: Combinatorics, Compression, Security and Transmission* (R. M. Capocelli, ed.). Springer, New York.

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

EPPSTEIN, D., GALIL, Z., GIANCARLO, R. and ITALIANO, G. (1992). Sparse dynamic programming II: Convex and concave cost functions. *J. Amer. Comp. Mach.* **39** 547–567.

GRIMMETT, G. R. and KESTEN, H. (1984). First passage percolation, network flows and electrical resistances. *Z. Wahrsch. Verw. Gebiete* **66** 335–366.

HAMMERSLEY, J. M and WELSH, D. J. A. (1965). First-passage percolation, subadditive processes, stochastic networks and generalized renewal theory. In *Bernoulli, Bayes, Laplace Anniversary Volume* (J. Neyman and L. M. Le Cam, eds.). Springer, Berlin.

KARLIN, S., GHANDOUR, G., OST, F., TAVARÉ, S. and KORN, L. J. (1983). New approaches for computer analysis of nucleic acid sequences. *Proc. Nat. Acad. Sci. U.S.A.* **80** 5660–5664.

KARLIN, S. and OST, F. (1987). Counts of long aligned word matches among random letter sequences. *Adv. in Appl. Probab.* **19** 293–351.

KESTEN, H. (1986). Aspects of first-passage percolation. *Ecole d'Eté de Probabilités de Saint Flour XIV. Lecture Notes in Math.* **1180** 125–264. Springer, Berlin.

MILLER, W. and MYERS, E. W. (1988). Sequence comparison with concave weighting functions. *Bull. Math. Biol.* **50** 97–120.

NEUHAUSER, C. (1994). A Poisson approximation for sequence comparisons with insertions and deletions. *Ann. Statist.* To appear.

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

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

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

TAVARÉ, S. (1992). International conference on random mappings, partitions and permutations. *Adv. in Appl. Probab.* **24** 761–777.

VAN DEN BERG, J and KESTEN, H. (1993). Inequalities for the constant in first passage percolation. *Ann. Appl. Probab.* **3** 56–80.

WATERMAN, M. S. (1984). General methods of sequence comparison. *Bull. Math. Biol.* **46** 473–500.

WATERMAN, M. S., ed. (1989). *Mathematical Methods for DNA Sequences*. CRC Press, Boca Raton, FL.

WATERMAN, M. S. and EGGERT, M. (1987). A new algorithm for best subsequence alignments with application to tRNA–rRNA comparisons. *Journal of Molecular Biology* **197** 723–728.

WATERMAN, M. S., GORDON, L. and ARRATIA, R. (1987). Phase transitions in sequence matches and nucleic acid structure. *Proc. Nat. Acad. Sci. U.S.A.* **84** 1239–1243.

WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge Univ. Press.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
1052 W. 36TH PLACE, DRB 155
LOS ANGELES, CALIFORNIA 90089-1113

DEPARTMENTS OF MATHEMATICS
AND MOLECULAR BIOLOGY
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089-1113