Adv. Appl. Prob. **38,** 827–852 (2006)
Printed in Northern Ireland
© *Applied Probability Trust* 2006

LARGE DEVIATIONS-BASED UPPER BOUNDS ON THE EXPECTED RELATIVE LENGTH OF LONGEST COMMON SUBSEQUENCES

RAPHAEL HAUSER,* *University of Oxford*
SERVET MARTÍNEZ,** *Universidad de Chile*
HEINRICH MATZINGER,*** *Universität Bielefeld and Georgia Institute of Technology*

Abstract

Consider the random variable L_n defined as the length of a longest common subsequence of two random strings of length n and whose random characters are independent and identically distributed over a finite alphabet. Chvátal and Sankoff showed that the limit $\gamma = \lim_{n\to\infty} \mathrm{E}[L_n]/n$ is well defined. The exact value of this constant is not known, but various methods for the computation of upper and lower bounds have been discussed in the literature. Even so, high-precision bounds are hard to come by. In this paper we discuss how large deviation theory can be used to derive a consistent sequence of upper bounds, $(q_m)_{m\in\mathbb{N}}$, on γ, and how Monte Carlo simulation can be used in theory to compute estimates, $\hat{q}_m$, of the q_m such that, for given $\Xi > 0$ and $\Lambda \in (0, 1)$, we have $\mathrm{P}[\gamma < \hat{q} < \gamma + \Xi] \geq \Lambda$. In other words, with high probability the result is an upper bound that approximates γ to high precision. We establish $\mathcal{O}((1-\Lambda)^{-1}\,\Xi^{-(4+\varepsilon)})$ as a theoretical upper bound on the complexity of computing $\hat{q}_m$ to the given level of accuracy and confidence. Finally, we discuss a practical heuristic based on our theoretical approach and discuss its empirical behavior.

Keywords: Longest common subsequence problem; Chvátal–Sankoff constant; upper bound; large deviation theory; Monte Carlo simulation

2000 Mathematics Subject Classification: Primary 05A16; 62F10
Secondary 92E10

1. Introduction

In pattern matching, speech recognition, DNA and protein analysis, and various other domains of application, scoring methods to decide the degree of similarity between two finite strings play a central role; see, e.g. [31], [29], [21], [25], [12], [11], and [3]. The most widely used family of scoring functions consists of those defined as the maximum alignment score over a set of admissible alignments, where each alignment score is computed as the sum of scores of individually aligned characters. Insertions and deletions are typically allowed in the admissible alignments, subject to penalization. In most applications the actual solution of interest is the set of alignments that are (almost) score maximizing.

Received 14 February 2006.

* Postal address: Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom. Email address: hauser@comlab.ox.ac.uk
** Postal address: CMM-DIM-CNRS 2071, Universidad de Chile, Casilla 170-3 Correo 3, Santiago, Chile.
Email address: smartine@dim.uchile.cl
*** Postal address: Fakultät für Mathematik, Universität Bielefeld, D-33501 Bielefeld, Germany.
Email address: matzing@mathematik.uni-bielefeld.de

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

For example, if every pair of correctly aligned characters counts as a unit score, incorrectly aligned characters are penalized by (the addition to the score of the negative amount) $-\delta_{\neq}$ and the insertion of a gap, $\sqcup$, is allowed subject to penalization by $-\delta_{\sqcup}$, then $\begin{smallmatrix} \mathrm{b} & \mathrm{r} & \mathrm{o} & \mathrm{t} \\ \mathrm{b} & \mathrm{a} & \sqcup & \mathrm{t} \end{smallmatrix}$ is an admissible alignment of the strings 'brot' and 'bat' and the corresponding alignment score is $1-\delta_{\neq}-\delta_{\sqcup}+1$. The similarity score $S_{\delta_{\neq},\delta_{\sqcup}}(\text{'brot'};\text{'bat'})$ is obtained by computing the maximum score over all admissible alignments of the two strings. A special case of this scoring method is obtained by setting $\delta_{\neq}=\infty$ and $\delta_{\sqcup}=0$. In this case the similarity score of two strings x and y becomes the length, $L(x;\,y)$, of a *longest common subsequence (LCS)* of the two strings, where a common subsequence is any string that can be obtained from both x and y by deleting some of their entries.

Sequence alignment algorithms typically produce multiple near-optimal alignments, so that practitioners are faced with the problem of deciding which of these are significant to their specific application. In some cases it may not be possible to do this manually. Owing to statistical fluctuations, a short sequence x will have high-scoring alignments in any sufficiently long random sequence Y. It is thus natural to study the statistical behavior of optimal alignment scores of random sequences; see, e.g. [30].

Let $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$ be independent, identically distributed random variables with distribution ξ over a finite alphabet $\mathbb{A}$, and let

$$S_{\delta_{\neq},\delta_{\sqcup},n}:=S_{\delta_{\neq},\delta_{\sqcup}}(X_1\cdots X_n;\,Y_1\cdots Y_n)$$

be the similarity score of the random strings $X_1\cdots X_n$ and $Y_1\cdots Y_n$ under the scoring method defined above. We write $L_n:=S_{\infty,0,n}$ for the random length of a longest common subsequence of these two random strings. Using a subadditivity argument, we can show that the limit

$$\gamma_{\delta_{\neq},\delta_{\sqcup}}:=\lim_{n\to\infty}\frac{\mathrm{E}[S_{\delta_{\neq},\delta_{\sqcup},n}]}{n} \tag{1.1}$$

is well defined. Of course, the value of $\gamma_{\delta_{\neq},\delta_{\sqcup}}$ also depends on the distribution ξ, which is omitted from the notation for simplicity. The existence of $\gamma_{\delta_{\neq},\delta_{\sqcup}}$ was first established by Chvátal and Sankoff [13] in the special case $\gamma:=\gamma_{\infty,0}=\lim_{n\to\infty}\mathrm{E}[L_n]/n$.

Arratia and Waterman [5] identified a phase transition phenomenon in the above-described context by showing that there exists a region of pairs $(\delta_{\neq},\,\delta_{\sqcup})\in\mathbb{R}^2_+$ where $\mathrm{E}[S_{\delta_{\neq},\delta_{\sqcup},n}]$ grows linearly in n and another region where it grows logarithmically. The case $(\delta_{\neq},\,\delta_{\sqcup})=(\infty,0)$ is in the linear phase and the Chvátal–Sankoff constant, γ, captures the asymptotic growth rate, whose exact value remains unknown. Let $|\cdot|$ denote set cardinality. Steele [26] conjectured that

$$\gamma=\frac{2}{1+\sqrt{|\mathbb{A}|}} \tag{1.2}$$

when ξ is the uniform distribution over the alphabet $\mathbb{A}$. However, our numerical results in Section 5 suggest that the expression on the right-hand side of (1.2) may be too large. Chvátal and Sankoff [13] conjectured that

$$\lim_{|\mathbb{A}|\to\infty}\frac{2}{\gamma\sqrt{|\mathbb{A}|}}=1.$$

In other words, they conjectured that the Steele conjecture holds asymptotically when the size of the alphabet grows to infinity. This was recently proven to be true in [19].

Although the constant γ is not known exactly, several methods have been discussed in the literature to compute lower and upper bounds; see [13], [16], [2], [14], [24], and [9]. The

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

subadditivity property used to show the existence of γ can be readily used to generate lower bounds via Monte Carlo simulation, though some understanding of the fluctuations of L_n about its mean is necessary to establish confidence intervals; see more on this below. Other techniques to generate lower bounds are based on Markov chains and common subsequence machines; see [14], [24], and [9]. The conceptually more difficult problem of generating upper bounds can be approached via combinatorial arguments and by representing longer sequences as concatenations of shorter pieces, which leads to 'over-counting'. The methods of [14] and [24] are of this kind, as is the method discussed in the present paper. Other approaches, e.g. that of [9], are based on information theoretic ideas and the notion of Kolmogorov complexity. A third approach, used in [2], relies on adding a guaranteed level of approximation to the lower bound $\mathrm{E}[L_n]/n$. We note that the above-mentioned algorithms comprise both deterministic and randomized algorithms.

The circle of questions pertaining to the understanding of optimal alignments of random sequences is sometimes referred to as the *LCS problem* in the longest common subsequence context. Another of its important aspects concerns the convergence speed of (1.1) and the order of the fluctuations that occur in Monte Carlo simulations. Waterman [30] conjectured that the fluctuations of $S_{\delta_{\neq},\delta_{\sqcup},n}$ about its mean are of order $\mathcal{O}(\sqrt{n})$. In the special case of L_n, Arratia and Waterman [5] derived a law of large deviations for fluctuations on scales larger than $\sqrt{n}$, but the exact order of the fluctuations is unknown. In fact, it is not even known if the order is larger than a power of n. Using first passage percolation methods, Alexander [2] established that $\mathrm{E}[L_n]/n$ converges at a rate $\mathcal{O}(\sqrt{\log(n)/n})$.

We conclude this section by pointing out some other questions related to the LCS problem and the state of knowledge about them. An important problem that was open for decades concerns the longest *increasing* subsequence of random permutations. It is suspected that insights into the longest increasing subsequence problem can be used to study the LCS problem; see [10] and [1]. Another problem related to the LCS problem is that of comparing sequences X and Y by looking for longest common words; there are generalizations of this problem where the word does not need to appear in exactly the same form in the two sequences. The distributions that appear in this context have been studied in [6], [7], and [23]. A crucial role is played by the Chen–Stein method for the Poisson approximation. In [7] and [4] some light was shed on the relation between the Erdős–Rényi law for random coin tossing and the above-mentioned problem. The authors of [7] and [4] also developed an extreme value theory for this problem.

1.1. Overview

Our paper consists of three distinct parts that contribute in different ways to the understanding of the LCS problem.

The contribution of Section 3 is of a theoretical nature. In writing long pairs of strings as concatenations of shorter pairs with LCS length m, fundamental links emerge between upper bounds on the Chvátal–Sankoff constant, γ, on the one hand, and the large deviations of a naturally defined measure $\nu^{[m]}$ on $\mathbb{N}$, on the other. If this measure were known exactly, then the computation of upper bounds, q_m, on γ would reduce to the optimization problem

$$q_m = \inf\left\{q \in [0,1]\colon \text{ there exists a } t > 0 \text{ such that } \sum_{k\in\mathbb{N}} \mathrm{e}^{t(2m/q-k)}\nu^{[m]}(k) < 1\right\}. \tag{1.3}$$

Furthermore, it is true that $\lim_{m\to\infty} q_m = \gamma$. This yields a conceptual algorithm for the computation of a consistent sequence of upper bounds on γ.

The importance of this result is in pointing out that large deviation theory yields a mechanism for augmenting information gained from shorter pairs of strings to gain asymptotic information

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

(as $n \to \infty$). Thus, we are estimating a subadditive limit constant using large deviations of a finite-m measure. Since this paper was written, this mechanism (which seems to be new in the form presented) has turned out to be a surprisingly useful tool in analyzing the asymptotic behavior of L_n; see [17], where it was used to bound the proportion of random sequences where optimal alignments are locally unique, or [22], where it was used to derive bounds on the order of the fluctuations of L_n for certain models of random sequences. Alexander [2] also derived bounds on the order of the fluctuations using subadditivity and large deviations, but his approach relies on first passage percolation rather than a finite-m measure.

The contribution of Section 4 is a theoretical analysis of the practical algorithm obtained from (1.3) when the measure $\nu^{[m]}$ is replaced by a simulated approximation $\hat{\nu}^{[m]}$. Since this approximate measure is the outcome of a random experiment, the 'bounds' $\hat{q}_m$ obtained in this framework are random variables that can actually take values below γ, but with quantifiably small probability. We call such bounds *probabilistic upper bounds*. For any given confidence level Λ and absolute approximation error Ξ, we establish worst-case upper bounds on the size of m and the number of simulations, ℓ_0, needed to guarantee that $\hat{q}_m$ satisfies the following two properties with probability greater than Λ:

(i) $\hat{q}_m$ is indeed an upper bound on γ;

(ii) $\hat{q}_m$ overestimates the true value of γ by less than Ξ.

We establish that $\mathcal{O}((1-\Lambda)^{-1}\Xi^{-(4-\varepsilon)})$ is a conservative upper bound on the total complexity of computing $\hat{q}_m$ with these properties, where ε is a small, arbitrary constant. It turns out that the approximation error, $\Xi = \mathcal{O}(m^{-(1-\varepsilon)/2})$, depends only on the parameter m, and that the confidence level, $\Lambda = 1 - \mathcal{O}(\ell_0^{-1})$, depends only on the number of simulations. The bounds we prove are very conservative and do not reflect the true convergence speed of the algorithm. Nevertheless, they are important in establishing that the estimators $\hat{q}_m$ approximate the true value of γ to arbitrary precision and confidence level in reasonable time.

Finally, Section 5 concerns a heuristic practical version of our conceptual algorithm. We cannot claim mathematical rigor for the results presented there, because they are based on the assumption that certain unbiased estimators of the variance of a random variable $W(t, q)$ are sufficiently symmetrically distributed. However, we do discuss the empirical behavior of these estimators and argue that, when they are properly designed, the assumption is plausible. Numerical results are reported for random sequences with various alphabet sizes.

The previously known best bounds by a randomized algorithm with confidence bounds were obtained in [2]. The method there is based on estimating $\mathrm{E}[L_n]/n$ for large n and using a result which shows that $\mathrm{E}[L_n]/n \geq \gamma - C(n\log(n))^{1/2}$ for some $C > 0$. With this method one has to choose $n = 100\,000$ to achieve an accuracy of $\Xi = 0.045$ in the case of random sequences obtained by fair coin tossing. Since the best algorithm for computing the LCS of two sequences of length n takes $\mathcal{O}(n^2)$ 'work', it becomes impractical to compute γ to much higher accuracy. In contrast, to simulate the measure $\hat{\nu}^{[m]}$ in our method with $m = 1000$, it suffices to compute the LCS of multiple pairs of sequences of lengths up to $n \approx 2000$, and the accuracy achieved with these values seems to be approximately $\Xi \approx 0.005$ (in terms of the approximation error). We also note that in [9] tighter bounds were obtained than in [2], but without confidence bounds.

2. Some useful notation and a key inequality

Let $\mathbb{A}$ be a finite alphabet and $\mathbb{A}^* = \bigcup_{n\in\mathbb{N}} \mathbb{A}^n$ be the set of finite words. Recall that we denote by $|\mathbb{A}|$ the cardinality of $\mathbb{A}$, that is, the number of symbols in the alphabet. For $x \in \mathbb{A}^*$,

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

denote by $|x|$ its length, that is, the number of letters in x. Trivially, $|xy| = |x| + |y|$ for every pair $(x, y) \in \mathbb{A}^* \times \mathbb{A}^*$, where xy denotes the concatenation of x and y, that is, the string consisting of the letters of x followed by the letters of y.

Let Π^n be the class of increasing subsequences of the integer interval $[1, \dots, n]$. We denote the cardinality of any $\pi \in \Pi^n$ by $|\pi|$, and its components by $\pi(i)$, $i \in [0, |\pi|]$. For $x \in \mathbb{A}^*$ and $\pi \in \Pi^{|x|}$, we use the notation x_π for the string $x_{\pi(1)} \cdots x_{\pi(|\pi|)}$. The main object of study in this paper is the quantity

$$L(x; y) = \max\{k\colon \text{ there exist } \pi \in \Pi^{|x|} \text{ and } \sigma \in \Pi^{|y|} \text{ with } |\pi| = k = |\sigma| \text{ and } x_\pi = y_\sigma\},$$

that is, the length of a longest common subsequence of x and y.

Let $\mathbb{A}^{\mathbb{N}}$ be the set of infinite sequences of elements of $\mathbb{A}$. For the analysis, it is convenient to use the set of elementary events $\Omega = \mathbb{A}^{\mathbb{N}} \times \mathbb{A}^{\mathbb{N}}$ endowed with the canonical product σ-algebra. We will also sometimes identify Ω with $(\mathbb{A} \times \mathbb{A})^{\mathbb{N}}$, and we denote the points of Ω by $\omega = (x, y)$, where $x = (x_n)_{\mathbb{N}}$ and $y = (y_n)_{\mathbb{N}}$. We use the following notation for the canonical projections defined on Ω: $X(\omega) = x$, $X_i(\omega) = x_i$, $Y(\omega) = y$, and $Y_j(\omega) = y_j$. We endow Ω with the probability measure $\mathrm{P} = \xi^{\mathbb{N}} \times \xi^{\mathbb{N}}$, where ξ is a probability distribution on the finite alphabet $\mathbb{A}$ with $\xi(a) > 0$ for all $a \in \mathbb{A}$. In other words, all entries in X and Y are independent, identically distributed random variables with values in $\mathbb{A}$ and distribution ξ.

Remark 2.1. It is interesting to note that some of the results presented in this paper extend to the situation where P is an ergodic shift-invariant measure on Ω. For example, the proof of relation (2.1) below remains valid unchanged, and the relation that we will present in (2.3) extends to the more general model via Birkhoff's ergodic theorem. We restrict the exposition to the simpler setup because it represents the model of interest in the vast majority of applications.

We write $x[i, j]$ for the string, $x_i \cdots x_j$, formed by the letters between the ith and jth coordinates of x (inclusive), and we adopt a similar notation when x is random. Any pair of strings $(x, y) \in \mathbb{A}^* \times \mathbb{A}^*$ defines a measurable set as follows:

$$[[x, y]] = \{\omega \in \Omega\colon X[1, |x|](\omega) = x,\ Y[1, |y|](\omega) = y\}.$$

Extending this notation, we write $[[S]] = \bigcup_{(x,y)\in S}[[x, y]]$ for all $S \subseteq \mathbb{A}^* \times \mathbb{A}^*$.

Let $\{L^i_j\colon i, j \in \mathbb{N}\}$ be the family of random variables

$$L^i_j\colon \Omega \to \mathbb{N}, \qquad \omega \mapsto \begin{cases} L(X[i, j](\omega); Y[i, j](\omega)) & \text{if } i \le j, \\ 0 & \text{otherwise.} \end{cases}$$

For ease of notation, we write L_j for L^1_j. The family $\{L^i_j\}$ satisfies the hypotheses of Kingman's subadditive ergodic theorem, which implies that

$$\lim_{n\to\infty} \frac{L_n}{n} \overset{\text{a.s.}}{=} \sup_{n\ge 1} \frac{\mathrm{E}[L_n]}{n} = \lim_{n\to\infty} \frac{\mathrm{E}[L_n]}{n} =: \gamma \tag{2.1}$$

for some real number γ; see, e.g. [20]. (By '$\overset{\text{a.s.}}{=}$' we denote almost-sure equality.) The limit γ, trivially seen to be lying in the interval $(0, 1)$, is, recall, the Chvátal–Sankoff constant associated with the law P. Note that it follows from (2.1) that, for any $q < \gamma$, we have $\lim_{n\to\infty} \mathrm{P}[L_n \ge qn] = 1$. Therefore, for all $q \in (0, 1)$,

$$\lim_{n\to\infty} \mathrm{P}[L_n \ge qn] < 1 \quad \Longrightarrow \quad q \ge \gamma. \tag{2.2}$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

We write $S_1^n(q) := \{(x, y) \in \mathbb{A}^n \times \mathbb{A}^n \colon L(x; y) \geq qn\}$. Note that, with the notation introduced above,

$$\{\omega \in \Omega \colon L_n(\omega) \geq qn\} = [[S_1^n(q)]].$$

This notation will be useful in the proof of Lemma 2.1.

In the analysis in the subsequent sections we will some times need to know a lower bound on γ. For this purpose we use the following elementary relation, where $\mathbf{1}$ denotes the indicator function:

$$\gamma \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{L_n}{n} > \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k = Y_k\}} \overset{\text{a.s.}}{=} \sum_{a\in\mathbb{A}} \xi([[a]])^2 \geq \frac{1}{|\mathbb{A}|}. \tag{2.3}$$

Definition 2.1. For any string $x \in \mathbb{A}^*$ of length $|x| \geq 1$, let $x^- := x_1 \cdots x_{|x|-1}$ be the string obtained by removing the last of its letters. For $m \in \mathbb{N}$, we say that a pair $(x, y) \in \mathbb{A}^* \times \mathbb{A}^*$ is an *m-match* if

$$L(x; y) = m, \qquad L(x^-; y) = m - 1, \qquad L(x; y^-) = m - 1.$$

We write $\mathcal{M}^m$ for the set of m-matches in $\mathbb{A}^* \times \mathbb{A}^*$.

It follows immediately from Definition 2.1 that

$$(x, y) \in \mathcal{M}^m \quad \Longrightarrow \quad \min\{|x|, |y|\} \geq m, \tag{2.4}$$

$$(X[1, i], Y[1, j]) \in \mathcal{M}^m, \quad k \neq j \quad \Longrightarrow \quad (X[1, i], Y[1, k]) \notin \mathcal{M}^m, \tag{2.5}$$

$$(X[1, i], Y[1, j]), (X[1, k], Y[1, l]) \in \mathcal{M}^m, \quad k > i \quad \Longrightarrow \quad l < j. \tag{2.6}$$

The following families of random variables will play an important role in all parts of this paper:

$$\begin{aligned} L_{i,j} &:= L(X[1, i]; Y[1, j]), \\ Z_{i,j}^{[m]} &:= \mathbf{1}_{\mathcal{M}^m}(X[1, i], Y[1, j]), \\ Z_k^{[m]} &:= \sum_{\{(i,j) \colon i+j=k\}} Z_{i,j}. \end{aligned}$$

We will use the simplified notation $Z_{i,j}^{[m]} \equiv Z_{i,j}$ and $Z_k^{[m]} \equiv Z_k$ whenever we treat m as a fixed parameter. It follows immediately from (2.4) and (2.5) that

$$0 \leq Z_k \leq (k + 1 - 2m)_+, \tag{2.7}$$

that is, $Z_k \equiv 0$ for $k < 2m$.

Definition 2.2. The following measure on $\mathbb{N}$ plays a fundamental role in our analysis. For $k \in \mathbb{N}$, let

$$\nu^{[m]}(k) \equiv \nu(k) := \mathrm{E}[Z_k],$$

and let this measure be extended to subsets of $\mathbb{N}$ by σ-additivity. (The equivalent notation holds in the same situation as above.)

In Lemma 3.3 we will prove that $\nu(\mathbb{N}) \leq |\mathbb{A}|m$. Furthermore, the following trivial identity is sometimes useful:

$$\nu(k) = \sum_{\{(i,j) \colon i+j=k\}} \mathrm{P}[L_{i,j} = m,\ L_{i-1,j} = m - 1,\ L_{i,j-1} = m - 1]. \tag{2.8}$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

We are now ready to prove one of the key inequalities behind our approach. For any $z \in \mathbb{R}$, let $\lfloor z \rfloor$ and $\lceil z \rceil$ denote the integers obtained by rounding z down and, respectively, up.

Lemma 2.1. *Let $m \in \mathbb{N}$ and $q \in [0,1]$, and let $\nu^{*\lfloor qn/m \rfloor}$ be the $\lfloor qn/m \rfloor$-fold convolution of the measure ν with itself. Then*

$$\mathrm{P}[L_n \geq qn] \leq \nu^{*\lfloor qn/m \rfloor}([0, 2n]).$$

Proof. The family of strings $S_2^n(q) := \bigcup_{\{(i,j)\colon i+j=2n\}} \{(x,y) \in \mathbb{A}^i \times \mathbb{A}^j \colon L(x;y) \geq qn\}$ contains the set $S_1^n(q)$ defined above. Let us furthermore define

$$S_3^{n,m}(q) := \biggl\{(x^1 \cdots x^{\lfloor qn/m \rfloor} r^1, y^1 \cdots y^{\lfloor qn/m \rfloor} r^2)\colon \\ (x^i, y^i) \in \mathcal{M}^m \text{ for all } i \in [1, \lfloor qn/m \rfloor],\ r^1, r^2 \in \mathbb{A}^*, \\ \sum_{i=1}^{\lfloor qn/m \rfloor} |x^i| + \sum_{i=1}^{\lfloor qn/m \rfloor} |y^i| + |r^1| + |r^2| = 2n\biggr\}.$$

We claim that $S_2^n(q) \subseteq S_3^{n,m}(q)$. In fact, for any pair $(x,y) \in S_2^n(q)$, there exist two strictly increasing maps, $\pi\colon [1, \lceil qn \rceil] \to [1, |x|]$ and $\sigma\colon [1, \lceil qn \rceil] \to [1, |y|]$, such that $x_\pi = y_\sigma$, and it is possible to choose π and σ to be minimal in the sense that, for each pair $(\hat{\pi}, \hat{\sigma}) \in \Pi^{|x|} \times \Pi^{|y|}$ that satisfies $|\hat{\pi}| = |\hat{\sigma}| = \lceil qn \rceil$, $x_{\hat{\pi}} = y_{\hat{\sigma}}$, and

$$\hat{\pi}(i) \leq \pi(i), \quad \hat{\sigma}(i) \leq \sigma(i), \qquad i \in [1, \lceil qn \rceil],$$

we have $\hat{\pi} = \pi$ and $\hat{\sigma} = \sigma$. It is easy to see that, when π and σ are minimal in this sense,

$$(x^i, y^i) := (x_{\pi(m(i-1))+1} \cdots x_{\pi(mi)}, y_{\sigma(m(i-1))+1} \cdots y_{\sigma(mi)})$$

is an m-match for all $i \in [1, \lfloor qn/m \rfloor]$. Therefore, $(x^1 \cdots x^{\lfloor qn/m \rfloor} r^1, y^1 \cdots y^{\lfloor qn/m \rfloor} r^2) \in S_3^n(q)$, where $r^1 := x_{\pi(\lfloor qn/m \rfloor)+1} \cdots x_{|x|}$ and $r^2 := y_{\sigma(\lfloor qn/m \rfloor)+1} \cdots y_{|y|}$. This shows that $S_2^n(q) \subseteq S_3^{n,m}(q)$, as claimed above.

It is now useful to introduce the index set

$$\mathcal{I}(q,n,m) = \biggl\{\kappa = (\kappa_1, \ldots, \kappa_{\lfloor qn/m \rfloor}) \in \mathbb{N}^{\lfloor qn/m \rfloor}\colon \sum_{i=1}^{\lfloor qn/m \rfloor} \kappa_i \leq 2n\biggr\}.$$

For $\kappa \in \mathcal{I}(q,n,m)$, we define

$$S_3^{n,m}(q,\kappa) := \{(x^1 \cdots x^{\lfloor qn/m \rfloor} r^1, y^1 \cdots y^{\lfloor qn/m \rfloor} r^2) \in S_3^{n,m}(q)\colon \\ |x^i| + |y^i| = \kappa_i \text{ for all } i \in [1, \lfloor qn/m \rfloor]\},$$

so we can write $S_3^{n,m}(q) = \bigcup_{\kappa \in \mathcal{I}(q,n,m)} S_3^{n,m}(q,\kappa)$. It follows that

$$\begin{aligned} \mathrm{P}\Bigl[[[S_3^{n,m}(q)]]\Bigr] &\leq \sum_{\kappa \in \mathcal{I}(q,n,m)} \mathrm{P}\Bigl[[[S_3^{n,m}(q,\kappa)]]\Bigr] \\ &\leq \sum_{\kappa \in \mathcal{I}(q,n,m)} \sum_{S_3^n(q,\kappa)} \mathrm{P}\Bigl[[[(x^1 \cdots x^{\lfloor qn/m \rfloor} r^1, y^1 \cdots y^{\lfloor qn/m \rfloor}, r^2)]]\Bigr] \\ &\leq \sum_{\kappa \in \mathcal{I}(q,n,m)} \sum_{\{(x^i,y^i) \in \mathcal{M}^m\colon |x^i|+|y^i|=\kappa_i\ \forall i \in [1,\lfloor qn/m \rfloor]\}} \prod_{i=1}^{\lfloor qn/m \rfloor} \mathrm{P}\Bigl[[[x^i, y^i]]\Bigr]. \end{aligned}$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Since

$$\nu^{*\lfloor qn/m\rfloor}([0,2n]) = \sum_{\kappa\in\mathcal{I}(q,n,m)} \prod_{i=1}^{\lfloor qn/m\rfloor} \nu(\kappa_i) \tag{2.9}$$

and

$$\prod_{k=1}^{\lfloor qn/m\rfloor} \nu(\kappa_k) = \sum_{\{(x^i,y^i)\in\mathcal{M}^m :\ |x^i|+|y^i|=\kappa_i\ \forall i\in[1,\lfloor qn/m\rfloor]\}} \prod_{i=1}^{\lfloor qn/m\rfloor} \mathrm{P}\Big[[[a^k,b^k]]\Big],$$

we conclude that

$$\mathrm{P}\Big[[[S_3^{n,m}(q)]]\Big] \le \nu^{*\lfloor qn/m\rfloor}([0,2n]). \tag{2.10}$$

The claim of the lemma now follows from the relation

$$\{\omega\in\Omega\colon L_n(\omega)\ge qn\} = [[S_1^n(q)]] \subseteq [[S_2^n(q)]] \subseteq [[S_3^{n,m}(q)]].$$

3. A large deviations-based upper bound on γ

In this section we apply large deviation techniques to find the exponential rate of the bound (2.10). Since ν is not a probability measure in general, we derive the relevant results from first principles. Using measure-theoretic notation, we have

$$\begin{aligned}\bigg(\int_{\mathbb{N}} \mathrm{e}^{t(2m/q-x)}\,\mathrm{d}\nu(x)\bigg)^{\lfloor qn/m\rfloor} &= \sum_{\kappa\in\mathbb{N}^{\lfloor qn/m\rfloor}} \exp\bigg(t\sum_{i=1}^{\lfloor qn/m\rfloor}\bigg(\frac{2m}{q}-\kappa_i\bigg)\bigg)\prod_{i=1}^{\lfloor qn/m\rfloor}\nu(\kappa_i)\\ &\ge \sum_{\kappa\in\mathcal{I}(q,n,m)} \exp\bigg(t\sum_{i=1}^{\lfloor qn/m\rfloor}\bigg(\frac{2m}{q}-\kappa_i\bigg)\bigg)\prod_{i=1}^{\lfloor qn/m\rfloor}\nu(\kappa_i)\\ &\ge \mathrm{e}^{-2mt/q}\nu^{*\lfloor qn/m\rfloor}([0,2n]),\end{aligned}$$

where the last inequality holds since every $\kappa\in\mathcal{I}(q,n,m)$ satisfies the relation

$$\sum_{i=1}^{\lfloor qn/m\rfloor}\bigg(\frac{2m}{q}-\kappa_i\bigg) \ge -\frac{2m}{q}.$$

Equation (2.9) therefore implies that

$$\nu^{*\lfloor qn/m\rfloor}([0,2n]) \le \bigg(\int_{\mathbb{N}} \mathrm{e}^{t(2m/q-x)}\,\mathrm{d}\nu(x)\bigg)^{\lfloor qn/m\rfloor}\mathrm{e}^{2mt/q}. \tag{3.1}$$

This leads to the following theorem, providing the main tool for the construction of our upper bounds on γ.

Theorem 3.1. *Let $q\in[0,1]$. If there exists a $t>0$ such that $\sum_{k\in\mathbb{N}}\mathrm{e}^{t(2m/q-k)}\nu(k)<1$, then $\gamma<q$.*

Proof. If the stipulated condition holds then, for all sufficiently large n, the right-hand side of (3.1) is less than 1. The result then follows from Lemma 2.1 and (2.2).

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Definition 3.1.

$$q_m := \inf\left\{q \in [0,1]\colon \text{ there exists a } t > 0 \text{ such that } \sum_{k\in\mathbb{N}} \mathrm{e}^{t(2m/q-k)}\nu(k) < 1\right\}. \qquad (3.2)$$

This quantity will be studied in the remainder of this section.

Theorem 3.1 revealed that $\gamma \le q_m$ for all $m \in \mathbb{N}$. We now set out to prove that $(q_m)_{\mathbb{N}}$ is a consistent sequence of upper bounds on γ in the sense that $\lim_{m\to\infty} q_m = \gamma$; see Theorem 3.2. The analysis that leads to this result also serves as a basis for understanding the practical Monte Carlo methods of Section 4. We start by recalling the following large deviation inequality.

Lemma 3.1. (Azuma–Hoeffding.) *Let $t \in \mathbb{N}$, let $\mathcal{F} = \bigcup_{s\in\mathbb{N}_0} \mathcal{F}_s$ be a filtration, and let $V_0, V_1, \dots, V_t$ be an $\mathcal{F}$-adapted martingale such that $V_0 = 0$. Let $a > 0$ and $\Delta > 0$, and let us assume that, for all $s \in [0, t-1]$, we have $|V_t - V_{t+1}| \le a$ almost surely. Then the following inequality holds:*

$$\mathrm{P}[V_t \ge \Delta t] \le \mathrm{e}^{-t\Delta^2/(2a^2)}.$$

Lemma 3.1 is due to Azuma [8] and Hoeffding [18]. A modern proof can be found in, e.g. [28, Section 11.1.4]. We now use Lemma 3.1 to show that $L_{i,j}$ decays exponentially.

Lemma 3.2. *For all $\Delta \ge 0$, we have*

$$\mathrm{P}\left[L_{i,j} \ge \frac{i+j}{2}(\gamma+\Delta)\right] \le \mathrm{e}^{-(i+j)\Delta^2/8}.$$

Proof. If σ_X and σ_Y denote the respective left-shift operators on the X and Y components of (X, Y), then $L_{i+j,j+i} \ge L_{i,j} + L_{j,i} \circ (\sigma_X^i, \sigma_Y^j)$. Since $L_{i,j}$ and $L_{j,i} \circ (\sigma_X^i, \sigma_Y^j)$ are identically distributed, this implies that $\mathrm{E}[L_{i+j,j+i}] \ge 2\,\mathrm{E}[L_{i,j}]$. Since subadditivity furthermore implies that $\mathrm{E}[L_{i+j,j+i}] \le \gamma(i+j)$, we obtain $\mathrm{E}[L_{i,j}] \le \gamma(i+j)/2$ and, hence,

$$\mathrm{P}\left[L_{i,j} \ge \frac{i+j}{2}(\gamma+\Delta)\right] \le \mathrm{P}\left[L_{i,j} \ge \mathrm{E}[L_{i,j}] + \frac{(i+j)}{2}\Delta\right]. \qquad (3.3)$$

Let us next consider a fixed path $\Gamma\colon [0, i+j] \to \mathbb{Z}^2$ that leads from $\Gamma(0) = (0,0)$ to $\Gamma(i+j) = (i,j)$ by moving one unit in the positive direction of either coordinate in each step. Let $r(k)$ and $s(k)$ be defined by $\Gamma(k) = (r(k), s(k))$, let $\mathcal{F}_0 := \{\mathbb{R}, \varnothing\}$ be the trivial σ-algebra on $\mathbb{R}$, and let

$$\mathcal{F}_k := \sigma(X_u, Y_v\colon u = 1, \dots, r(k),\ v = 1, \dots, s(k)), \qquad k = 1, \dots, i+j.$$

Here and elsewhere the notation extends in a natural way to the case where an index set is empty. For example, if $r(k) = 0$ then $\mathcal{F}_k = \sigma(Y_1, \dots, Y_{s(k)})$. For all $k \in [0, i+j]$, let us define $V_k := \mathrm{E}[L_{i,j} - \mathrm{E}[L_{i,j}] \mid \mathcal{F}_k]$. Then the sequence $V_0, V_1, \dots, V_{i+j}$ is a martingale that satisfies the conditions of Lemma 3.1 with $a = 1$. By applying the lemma, we obtain the inequality

$$\mathrm{P}\left[L_{i,j} - \mathrm{E}[L_{i,j}] \ge (i+j)\frac{\Delta}{2}\right] \le \mathrm{e}^{-(i+j)\Delta^2/8}.$$

Combined with (3.3), this yields the result.

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Remark 3.1. By applying the Azuma–Hoeffding lemma to the martingale $(-V_0, \dots, -V_{i+j})$, where V_k is as defined in the proof of Lemma 3.2, we find the similar inequality

$$\mathrm{P}\bigg[L_{i,j} - \mathrm{E}[L_{i,j}] \le -\frac{(i+j)}{2}\Delta\bigg] \le \mathrm{e}^{-(i+j)\Delta^2/8}. \tag{3.4}$$

As a consequence of Lemma 3.2, we can now bound $\nu(k)$ for small k.

Corollary 3.1. *For $k \le 2m/\gamma$ and $\Delta'_k = 2m/k - \gamma$, we have*

$$\nu(k) \le 2m|\mathbb{A}|\mathrm{e}^{-(\Delta'_k)^2 k/8}.$$

Proof. For $k \le 2m/\gamma$ we have $\Delta'_k \ge 0$. By applying (2.8) and Lemma 3.2, we obtain

$$\begin{aligned}\nu(k) &\le \sum_{\{(i,j)\colon i+j=k\}} \mathrm{P}[L_{i,j} \ge m] \\ &= \sum_{\{(i,j)\colon i+j=k\}} \mathrm{P}\bigg[L_{i,j} \ge \frac{i+j}{2}(\gamma + \Delta')\bigg] \\ &\le \frac{2m}{\gamma}\mathrm{e}^{-(\Delta'_k)^2 k/8}.\end{aligned}$$

The claim now follows from (2.3).

Next we show that ν is a finite measure.

Lemma 3.3. *For every $m \in \mathbb{N}$, $\sum_{k\ge1} \nu(k) \le |\mathbb{A}|m$.*

Proof. Let $m \in \mathbb{N}$ be fixed, so that the measure ν is well defined. The relations $\mathcal{Z}_0 \equiv 0$ and $\mathcal{Z}_m := \min\{k \ge 0\colon Z^{[m]}_{m,k} = 1\}$ define an increasing sequence, $(\mathcal{Z}_m)_{\mathbb{N}}$, of random variables, where $Z^{[m]}_{i,j}$ is as defined in Section 2. In particular, we have $\mathcal{Z}_1 = \min\{k \ge 1\colon Y_k = X_1\}$ and

$$\mathrm{P}[\mathcal{Z}_1 = k] = \sum_{a\in\mathbb{A}} \xi(a)(1-\xi(a))^{k-1}\xi(a),$$

whence $\mathrm{E}[\mathcal{Z}_1] = |\mathbb{A}|$. Furthermore, $\mathcal{Z}_{k+1} - \mathcal{Z}_k$ is independent of $\mathcal{Z}_l$, $l < k$, and distributed identically to $\mathcal{Z}_1$, which implies that $\mathrm{E}[\mathcal{Z}_m] = m\,\mathrm{E}[\mathcal{Z}_1] = m|\mathbb{A}|$. It now follows from (2.4), (2.5), and (2.6) that

$$\begin{aligned}\sum_{k\ge1}\nu(k) = \mathrm{E}\bigg[\sum_{i,j>0} Z_{i,j}\bigg] &\le \mathrm{E}[\mathcal{Z}_m - m] \\ &\le m(|\mathbb{A}| - 1).\end{aligned} \tag{3.5}$$

This lemma allows us to extend Corollary 3.1 and bound $\nu(k)$ for all k.

Corollary 3.2. *For all $a \in \mathbb{A}$, let $\eta(a) := (1-\xi(a))^{1/m}$. Then, for all $k \in \mathbb{N}$, we have*

$$\nu(k) \le \Big(\max_{a\in\mathbb{A}} \eta(a)\Big)^{k-2} \sum_{a\in\mathbb{A}} m\xi(a)\frac{k-(k-1)\eta(a)}{(1-\eta(a))^2}. \tag{3.6}$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Proof. We continue to use the notation in the proof of Lemma 3.3 and the facts derived there. For $k < m$ we have $\nu(k) = 0$ and (3.6) is trivially true. For $k \geq m$, (2.6) shows that $Z_k^{[m]} > 0$ implies $k \leq \mathcal{Z}_m$. Therefore,

$$\begin{aligned}
\nu(k) = \mathrm{E}[Z_k^{[m]}] &= \sum_{r=1}^{\infty} \mathrm{P}[Z_k^{[m]} \geq r] \\
&= \sum_{s=m}^{\infty} \sum_{r=1}^{s} \mathrm{P}[Z_k^{[m]} \geq r \mid \mathcal{Z}_m = s]\, \mathrm{P}[\mathcal{Z}_m = s] \\
&\leq \sum_{s=k}^{\infty} s\, \mathrm{P}\biggl[\sum_{i=1}^{\infty} (\mathcal{Z}_i - \mathcal{Z}_{i-1}) \geq s\biggr] \\
&\leq \sum_{s=k}^{\infty} s \sum_{i=1}^{m} \mathrm{P}\biggl[\mathcal{Z}_i - \mathcal{Z}_{i-1} > \frac{s-1}{m}\biggr] \\
&= \sum_{s=k}^{\infty} sm \sum_{a \in \mathbb{A}} \xi(a)(1 - \xi(a))^{\lfloor (s-1)/m \rfloor} \\
&\leq \sum_{a \in \mathbb{A}} \frac{m\xi(a)}{\eta(a)} \sum_{s=k}^{\infty} s\eta(a)^{s-1} \\
&= \sum_{a \in \mathbb{A}} \eta(a)^{k-2} m\xi(a) \frac{k - (k-1)\eta(a)}{(1 - \eta(a))^2}.
\end{aligned}$$

Our next result is instrumental in proving the consistency of the bounds q_m.

Lemma 3.4. *Let* $\Delta > 0$ *be such that* $q := \Delta + \gamma \leq 1$. *Then, for any value of* t *such that* $0 < t \leq \Delta/(8|\mathbb{A}|^2)$ *and any value of* m *such that* $m \geq \max\{q^2/(\Delta(1-q^2)), q/(2-\Delta q)\}$, *we have*

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} \nu(k) \leq (m|\mathbb{A}| + 4m^2|\mathbb{A}|^2)\mathrm{e}^{-t\Delta m}. \tag{3.7}$$

Proof. It follows from the hypotheses that $1/\gamma = \Delta/(\gamma q) + 1/q$. Thus, $1/\gamma \geq \Delta + 1/q$ and

$$a := \frac{2m}{q} + m\Delta < \frac{2m}{\gamma} < 2m|\mathbb{A}|, \tag{3.8}$$

where the last inequality follows from (2.3). Since $2m/q - k \leq 2m/q - a = -\Delta m$ for all $k \geq a$, we have

$$\sum_{k \geq a} \mathrm{e}^{t(2m/q-k)} \nu(k) \leq \mathrm{e}^{-t\Delta m} \sum_{k \geq a} \nu(k) \leq m|\mathbb{A}|\mathrm{e}^{-t\Delta m}, \tag{3.9}$$

where the second inequality follows from Lemma 3.3. Next, note that (2.4) implies $\nu(k) = 0$ for all $k < 2m$. Together with (3.8) and Corollary 3.1, this implies that

$$\begin{aligned}
\sum_{k<a} \mathrm{e}^{t(2m/q-k)} \nu(k) &\leq 2m|\mathbb{A}| \sum_{k=2m}^{\lceil a-1 \rceil} \mathrm{e}^{t(2m/q-k)} \mathrm{e}^{-(\Delta'{}_k)^2 k/8} \\
&\leq 2m|\mathbb{A}| \sum_{k=2m}^{a-1} \mathrm{e}^{t(2m/q-k)} \mathrm{e}^{-(\Delta'{}_k)^2 m/4},
\end{aligned} \tag{3.10}$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

where, recall, $\Delta'_k = 2m/k - \gamma$. Introducing the new variable $\bar{k} := a - k$, and using the fact that

$$\lceil a \rceil - k - m\Delta = \lceil 2m/q + m\Delta \rceil - m\Delta - k \geq 2m/q - k,$$

we find that

$$\sum_{k=2m}^{\lceil a-1 \rceil} \mathrm{e}^{t(2m/q-k)} \mathrm{e}^{-(\Delta'_k)^2 m/4} \leq \mathrm{e}^{-tm\Delta} \sum_{\bar{k}=1}^{\lceil a \rceil - 2m} \mathrm{e}^{t\bar{k}} \mathrm{e}^{-(\Delta''_{\bar{k}})^2 m/4}, \tag{3.11}$$

where

$$\Delta''_{\bar{k}} := \frac{2m}{\lceil a \rceil - \bar{k}} - \gamma = \frac{2m}{\lceil a \rceil} \frac{1}{1 - \bar{k}/\lceil a \rceil} - \gamma \geq \frac{2m}{\lceil a \rceil} - \gamma + \frac{2m\bar{k}}{\lceil a \rceil^2}.$$

Because of the hypothesis on m, we have

$$\begin{aligned} \frac{2m}{\lceil a \rceil} - \gamma \geq \frac{2m}{a+1} - \gamma &= \frac{q}{1 + (q/2)(\Delta + 1/m)} - \gamma \\ &\geq q\left(1 - \frac{q}{2}\left(\Delta + \frac{1}{m}\right)\right) - \gamma = \Delta - \frac{q^2}{2}\left(\Delta + \frac{1}{m}\right) \\ &\geq \frac{\Delta}{2}, \end{aligned}$$

and combined with (3.8) this yields

$$(\Delta''_{\bar{k}})^2 \geq \left(\frac{\Delta}{2} + \frac{\bar{k}}{2m|\mathbb{A}|^2}\right)^2 > \frac{\Delta\bar{k}}{2m|\mathbb{A}|^2}. \tag{3.12}$$

Substituting (3.12) into (3.11), we obtain

$$\begin{aligned} \sum_{k=2m}^{\lceil a-1 \rceil} \mathrm{e}^{t(2m/q-k)} \mathrm{e}^{-(\Delta'_k)^2 m/4} &\overset{(3.12)}{\leq} \mathrm{e}^{-tm\Delta} \sum_{\hat{k}=1}^{\lceil a \rceil - 2m} \mathrm{e}^{t\bar{k} - \Delta\bar{k}/(8|\mathbb{A}|^2)} \leq \mathrm{e}^{-tm\Delta}(\lceil a \rceil - 2m) \\ &\overset{(3.8)}{<} 2m|\mathbb{A}|\mathrm{e}^{-tm\Delta}, \end{aligned}$$

where the second inequality is a consequence of the hypothesis on t. The result now follows from (3.10) and (3.9).

We are finally ready to establish the consistency of the sequence $(q_m)_{\mathbb{N}}$, defined in (3.2).

Theorem 3.2. $\lim_{m\to\infty} q_m = \gamma$.

Proof. Because of Theorem 3.1 we already know that $q_m \geq \gamma$ for all $m \in \mathbb{N}$. The result will thus be shown if we can establish that

$$\limsup_{m\to\infty} q_m \leq \gamma. \tag{3.13}$$

For a fixed $\varepsilon > 0$, let us choose Δ and t as functions of m, as follows: $\Delta := m^{-1/(2+\varepsilon)}$ and $t := \Delta/(8|\mathbb{A}|^2)$. Then, for all sufficiently large m, the conditions of Lemma 3.4 are satisfied. Moreover, we have

$$\mathrm{e}^{-t\Delta m} = \exp\left(-\frac{m^{\varepsilon/(2+\varepsilon)}}{8|\mathbb{A}|^2}\right),$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

meaning that, again for sufficiently large m, $(m|\mathbb{A}| + 4m^2|\mathbb{A}|^2)\mathrm{e}^{-t\Delta m} < 1$. Lemma 3.4 thus implies that there exists an $m_0 \in \mathbb{N}$ such that, for all $m \geq m_0$,

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)}\nu(k) < 1$$

and, thus, by (3.2), $q_m \leq \gamma + \Delta = \gamma + m^{-1/(2+\varepsilon)}$, showing that (3.13) is indeed true.

4. Monte Carlo simulation

Recall that Theorem 3.1 revealed that the inequalities $0 \leq q \leq 1$, $t > 0$, and

$$\sum_{k\in\mathbb{N}} \mathrm{e}^{t(2m/q-k)}\nu(k) < 1 \tag{4.1}$$

imply that q is an upper bound on the Chvátal–Sankoff constant, γ. This led to a conceptual algorithm to compute a sequence of consistent upper bounds, $(q_m)_{\mathbb{N}}$. In order to use this tool in a practical algorithm, the measures $\nu^{[m]}$ have to be approximated via Monte Carlo simulation. In this section we derive theoretical bounds on the convergence rates of this approach.

We continue to use the notation of Sections 2 and 3, and we introduce the random variables

$$W(t,q) := \sum_{k>0} \mathrm{e}^{t(2m/q-k)} Z_k, \tag{4.2}$$

such that $\mathrm{E}[W(t,q)] = \sum_{k>0} \mathrm{e}^{t(2m/q-k)}\nu(k)$ is the expression of interest in (4.1). Equations (2.4)–(2.6) imply that the series in (4.2) contains only finitely many terms for every $\omega \in \Omega$. As in previous sections, we drop the dependence on m from the notation, for simplicity.

Let $\{X_i^\ell\colon i,\ell \in \mathbb{N}\}$ and $\{Y_j^\ell\colon j,\ell \in \mathbb{N}\}$ be two sets of independent random variables with common distribution ξ on $\mathbb{A}$, and let

$$\begin{aligned} Z_{i,j}^\ell &:= \mathbf{1}_{\mathcal{M}^m}(X^\ell[1,i], Y^\ell[1,j]),\\ Z_k^\ell &:= \sum_{i+j=k} Z_{i,j}^\ell,\\ W^\ell(t,q) &:= \sum_{k>0} \mathrm{e}^{t(2m/q-k)} Z_k^\ell. \end{aligned}$$

The random variable Z_k^ℓ counts the number of m-matches of length k that occur starting from the first letters in the ℓth pair of random sequences (X^ℓ, Y^ℓ). With these conventions,

$$\hat{\nu}_k := \frac{1}{\ell_0}\sum_{\ell=1}^{\ell_0} Z_k^\ell \tag{4.3}$$

is an unbiased estimator of $\nu(k)$ for all $\ell_0 \in \mathbb{N}$ and

$$\frac{1}{\ell_0}\sum_{\ell=1}^{\ell_0} W^\ell = \sum_{k>0} \mathrm{e}^{t(2m/q-k)}\hat{\nu}_k$$

is an unbiased estimator of the left-hand side of (4.1).

The main result of this section is the following theorem, which gives us a tool with which to determine the value of the parameter m and the number, ℓ_0, of simulations necessary to obtain an estimator, $\hat{q}_m$, of γ to within a specific precision and at a given confidence level.

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Theorem 4.1. *Let $\alpha, \delta \in (0, 1)$ and $\ell_0 \in \mathbb{N}$ be fixed, and for all $m \in \mathbb{N}$ let*

$$\begin{aligned} \Delta_m &:= m^{-\alpha/2}, \\ t_m &:= \frac{\Delta_m}{16|\mathbb{A}|^2}, \\ \hat{q}_m &:= \Delta_m + \inf\Bigg\{q > 0\colon \sum_{k=1}^{\infty} \mathrm{e}^{t_m(2m/q-k)}\hat{\nu}_k < 1\Bigg\}. \end{aligned} \tag{4.4}$$

Then there exists a number, $m_0 \equiv m_0(\alpha, \xi, \delta)$, such that, for all $m \geq m_0$,

$$\mathrm{P}[\gamma \leq \hat{q}_m \leq \gamma + 2\Delta_m] \geq 1 - \mathrm{e}^{-\ell_0(1-\delta)^2/2} - 2/\ell_0. \tag{4.5}$$

Furthermore, we have

$$m_0 < \max\Bigg\{\mathcal{O}\bigg(\bigg(\frac{1-\gamma}{2}\bigg)^{-2/\alpha}\bigg), \mathcal{O}\bigg(\bigg(|\mathbb{A}|^4 \log\bigg(\frac{2}{\delta}\bigg)\bigg)^{(1+\varepsilon)/(1-\alpha)}\bigg[1 - \log\Big(\min_{a\in\mathbb{A}} \xi(a)\Big)\bigg]\bigg)\Bigg\},$$

where $\varepsilon > 0$ is a small, arbitrary number.

In (4.5) it is interesting to note that, for fixed α, δ, and ξ, the approximation error, $2\Delta_m$, is a function of m only; hence, the only way to approximate the true value of γ to higher precision is to increase m. On the other hand, the confidence level is solely a function of the number, ℓ_0, of Monte Carlo simulations.

We prepare the proof of Theorem 4.1 via three preliminary results. The first lemma shows that when m is large enough, an m-match of moderate length occurs with high probability.

Lemma 4.1. *Let α and Δ_m be as defined in Theorem 4.1 and consider the event*

$$B := \{\textit{there exist } i \textit{ and } j \textit{ such that } Z_{i,j} = 1 \textit{ and } i + j \leq 2\lceil m/\gamma + m\Delta_m/2\rceil\}. \tag{4.6}$$

Then there exists a number, $m_1 \equiv m_1(\alpha, \xi)$, such that, for all $m \geq m_1$,

$$\mathrm{P}[B] \geq 1 - \exp\bigg(-\frac{m^{1-\alpha}|\mathbb{A}|^{-4}}{256}\bigg).$$

Proof. Alexander [2] proved the existence of a constant, $C > 0$, such that $0 \leq \gamma - \mathrm{E}[L_n]/n \leq C\sqrt{\log(n)/n}$ for all $n \geq 1$, independently of ξ and $\mathbb{A}$. A more quantitative version of this result is obtained as follows: by choosing $\lambda = 2$ and $\theta = 3$ in Proposition 2.4 of [2], a relaxation of Equation (2.13) of [2] shows that

$$0 < \gamma - \frac{\mathrm{E}[L_n]}{n} < 7\sqrt{\frac{\log n}{n}} \quad \text{for all } n \geq 16. \tag{4.7}$$

Let $k' = m/\gamma + m\Delta_m/2$ and $n' = \lceil k' \rceil$. Then, having $m \geq 16$ implies that

$$n' \geq k' > m \geq 16. \tag{4.8}$$

Moreover, if $m \geq m_2(\alpha, \xi) := \inf\{y > 0\colon \log x < x^{1-\alpha}/(2 \times 56^2|\mathbb{A}|^3) \text{ for all } x \geq y\}$, then (2.3) implies that

$$\log m < \frac{m^{1-\alpha}\gamma^3}{2 \times 56^2}. \tag{4.9}$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Finally, if $m \geq m_3(\alpha, |\mathbb{A}|) := (2 \times 56^2 |\mathbb{A}|^3 \log(|\mathbb{A}| + \frac{1}{2}))^{1/(1-\alpha)}$, then (2.3) implies that

$$\log\left(\frac{1}{\gamma} + \frac{1}{2}\right) < \frac{m^{1-\alpha}\gamma^3}{2 \times 56^2}. \tag{4.10}$$

For $m \geq m_4(\alpha, |\mathbb{A}|) := \max\{m_2(\alpha, |\mathbb{A}|), m_3(\alpha, |\mathbb{A}|)\}$, we thus have

$$\begin{aligned} \log k' &= \log(m/\gamma + m\Delta_m/2) \\ &= \log m + \log\bigl(1/\gamma + \tfrac{1}{2}\bigr) \\ &\overset{(4.9),(4.10)}{<} \frac{m^{1-\alpha}\gamma^3}{56^2} \\ &< \frac{m^{-\alpha}\gamma^4}{56^2}\left(\frac{m}{\gamma} + \frac{m \times m^{-\alpha/2}}{2}\right) \\ &= \left(\frac{\Delta_m \gamma^2}{8}\right)^2 \frac{k'}{7^2}. \end{aligned} \tag{4.11}$$

Equations (4.7), (4.8), and (4.11) now show that, for $m \geq m_5(\alpha, \xi) := \max\{16, m_4\}$, we have

$$0 < \gamma - \frac{\mathrm{E}[L_{n'}]}{n'} < 7\sqrt{\frac{\log n'}{n'}} \leq 7\sqrt{\frac{\log k'}{k'}} < \frac{\Delta_m \gamma^2}{8}. \tag{4.12}$$

With the notation $\gamma_{n'} := \mathrm{E}[L_{n'}]/n'$, (4.12) and (2.3) imply that

$$\gamma - \gamma_{n'} - \frac{\Delta_m \gamma^2}{4} \leq -\frac{\Delta_m \gamma^2}{8} \leq -\frac{\Delta_m |\mathbb{A}|^{-2}}{8}. \tag{4.13}$$

Now note that if the event $D := \{L_{n'} \geq m\}$ occurs, then an m-match of total length less than or equal to $2n'$ must have appeared within $(X[1, n'], Y[1, n'])$. Hence, $D \subseteq B$ and it follows that

$$\mathrm{P}[\Omega \setminus B] \leq \mathrm{P}[\Omega \setminus D] = \mathrm{P}\left[\frac{L_{n'}}{n'} - \gamma_{n'} < \frac{m}{n'} - \gamma_{n'}\right]. \tag{4.14}$$

Let us now assume that $m \geq 2^{2/\alpha}$, whence $\Delta_m \gamma/2 < m^{-\alpha/2}/2 < \frac{1}{4}$, and observe that for $x \in [0, \frac{1}{4}]$ we have $1/(1+x) \leq 1 - x/2$. Applying this inequality to $x = \Delta_m \gamma/2$, we find that

$$\frac{m}{n'} - \gamma_{n'} \leq \frac{m}{k'} - \gamma_{n'} = \frac{\gamma}{1 + \Delta_m \gamma/2} - \gamma_{n'} \leq \gamma - \gamma_{n'} - \frac{\Delta_m \gamma^2}{4}.$$

Substituting this into (4.14) yields

$$\mathrm{P}[\Omega \setminus B] \leq \mathrm{P}\left[\frac{L_{n'}}{n'} - \gamma_{n'} \leq \gamma - \gamma_{n'} - \frac{\Delta_m \gamma^2}{4}\right]. \tag{4.15}$$

Since (4.13) and (4.15) hold for $m \geq m_1(\alpha, \xi) := \max\{m_5, 2^{2/\alpha}\}$, we obtain

$$\begin{aligned} \mathrm{P}[\Omega \setminus B] &\leq \mathrm{P}\left[\frac{L_{n'}}{n'} - \frac{\mathrm{E}[L_{n'}]}{n'} \leq -\frac{\Delta_m |\mathbb{A}|^{-2}}{8}\right] \\ &\leq \exp\left(-n' \frac{\Delta_m^2 |\mathbb{A}|^{-4}}{256}\right) \qquad (4.16) \\ &\leq \exp\left(-\frac{m^{1-\alpha}|\mathbb{A}|^{-4}}{256}\right), \qquad (4.17) \end{aligned}$$

where (4.16) follows from (3.4) and (4.17) from $n' > m$ and $\Delta_m = m^{-\alpha/2}$.

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Our second lemma shows that if m is sufficiently large then the probability of finding an estimator value $\hat{q}_m$ significantly below γ is exponentially small in the number of Monte Carlo simulations.

Lemma 4.2. *Let α, δ, Δ_m, and t_m be as defined in Theorem 4.1, and $\hat{\nu}_k$ as defined in (4.3). Then there exists a number, $m_6 \equiv m_6(\alpha, \xi, \delta)$, such that, for all $m \geq m_6$, $t \geq t_m$, and $q \in (0, \gamma - \Delta_m)$,*

$$\mathrm{P}\Bigg[\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} \hat{\nu}_k < 1\Bigg] \leq \mathrm{e}^{-\ell_0 (1-\delta)^2/2}.$$

Proof. Since $q, \gamma < 1$, the assumption $|\gamma - q| \geq \Delta_m$ implies that $|2m/\gamma - 2m/q| \geq 2m\Delta_m$. This shows that if $q \leq \gamma - \Delta_m$ and $k \leq 2\lceil m/\gamma + m\Delta_m/2\rceil$, then $2m/q - k \geq m\Delta_m - 2$. It follows that

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} Z_k \geq \sum_{k=1}^{2\lceil m/\gamma + m\Delta_m/2\rceil} \mathrm{e}^{t(2m/q-k)} Z_k \geq \mathrm{e}^{t(\Delta_m m-2)} \mathbf{1}_B, \tag{4.18}$$

where $\mathbf{1}_B$ denotes the indicator function of the event B, defined in (4.6). By definition, we have

$$\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} \hat{\nu}_k = \frac{1}{\ell_0} \sum_{\ell=1}^{\ell_0} \sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} Z_k^{\ell}.$$

It therefore follows from (4.18) that

$$\mathrm{P}\Bigg[\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} \hat{\nu}_k < 1\Bigg] \leq \mathrm{P}\Bigg[\frac{1}{\ell_0} \sum_{\ell=1}^{\ell_0} \mathrm{e}^{t(\Delta_m m-2)} \mathbf{1}_B^{\ell} < 1\Bigg], \tag{4.19}$$

where $(\mathbf{1}_B^{\ell})_{\ell \in \mathbb{N}}$ denotes a sequence of independent, identically distributed copies of $\mathbf{1}_B$. Now, for all

$$m \geq m_7(\alpha, \xi, \delta) := \max\{2^{2/\alpha}, 1 + 16|\mathbb{A}|^2 \log(2/\delta)\}^{1/(1-\alpha)}$$

and $t \geq t_m$, we have

$$\mathrm{e}^{t(m\Delta_m - 2)} \geq \exp\bigg(\frac{m^{1-\alpha} - 2m^{-\alpha/2}}{16|\mathbb{A}|^2}\bigg) > \exp\bigg(\frac{m^{1-\alpha} - 1}{16|\mathbb{A}|^2}\bigg) \geq \frac{2}{\delta}. \tag{4.20}$$

Furthermore, it follows from Lemma 4.1 that, for

$$m \geq m_8(\alpha, \xi, \delta) := \max\{m_1, (256|\mathbb{A}|^4 \log(2/\delta))^{1/(1-\alpha)}\},$$

we have

$$\mathrm{E}[\mathbf{1}_B^{\ell}] = \mathrm{P}[B] \geq 1 - \exp\bigg(-\frac{m^{1-\alpha}|\mathbb{A}|^{-4}}{256}\bigg) \geq 1 - \frac{\delta}{2}. \tag{4.21}$$

Equations (4.20) and (4.21) show that, for $m \geq m_6 := \max\{m_7, m_8\}$ and $t \geq t_m$, we have

$$\begin{aligned}\mathrm{P}\Bigg[\frac{1}{\ell_0} \sum_{\ell=1}^{\ell_0} \mathrm{e}^{t(m\Delta_m-2)} \mathbf{1}_B^{l} < 1\Bigg] &\leq \mathrm{P}\Bigg[\frac{1}{\ell_0} \sum_{\ell=1}^{\ell_0} \mathbf{1}_B^{\ell} < \frac{\delta}{2}\Bigg] \\ &\leq \mathrm{P}\Bigg[\frac{1}{\ell_0} \sum_{l=1}^{\ell_0} (\mathbf{1}_B^{l} - \mathrm{E}[\mathbf{1}_B^{l}]) \leq -(1-\delta)\Bigg].\end{aligned} \tag{4.22}$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Applying Lemma 3.1 with $a = 1$ to the martingale defined by $V_0 \equiv 0$, $\mathcal{F}_0 = \{\varnothing, \mathbb{R}\}$, and

$$V_k = \sum_{\ell=1}^{k} (\mathrm{E}[\mathbf{1}_B^\ell] - \mathbf{1}_B^\ell), \qquad \mathcal{F}_k = \sigma(V_1, \dots, V_k), \qquad k = 1, \dots, \ell_0,$$

we find that

$$\mathrm{P}\bigg[\frac{1}{\ell_0} \sum_{\ell=1}^{\ell_0} (\mathrm{E}[\mathbf{1}_B^\ell] - \mathbf{1}_B^\ell) \geq 1 - \delta\bigg] \leq \mathrm{e}^{-\ell_0 (1-\delta)^2/2}.$$

Together with (4.19) and (4.22), this proves the claim.

The third lemma will allow us to bound $\operatorname{var}(W)$ in the proof of Theorem 4.1.

Lemma 4.3. *Let α, Δ_m, and t_m be as defined in Theorem 4.1. Then, for all $m \geq m_9(\alpha, \xi) := (1-\gamma)^{-2/\alpha}$, $q \in [\gamma + \Delta_m, 1]$, and $t \in (0, t_m]$, we have*

$$\mathrm{E}\bigg[\bigg(\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} Z_k\bigg)^2\bigg] \leq \bigg(34m^4 |\mathbb{A}|(|\mathbb{A}|+1)^3 + \frac{2m|\mathbb{A}|}{\min_{a\in\mathbb{A}} \xi(a)}\bigg) \mathrm{e}^{-2tm\Delta_m}.$$

Proof. The condition

$$m \geq m_9(\alpha, \xi) := (1-\gamma)^{-2/\alpha}$$

is only necessary to guarantee that $[\gamma + \Delta_m, 1] \neq \varnothing$. Let $k_m := 2m/q + m$. Since $q > \gamma \geq |\mathbb{A}|^{-1}$, by (2.3), we have

$$k_m < 2m(|\mathbb{A}| + 1). \tag{4.23}$$

We will use the splitting

$$\mathrm{E}\bigg[\bigg(\sum_{k=1}^{\infty} \mathrm{e}^{t(2m/q-k)} Z_k\bigg)^2\bigg] \leq 2\,\mathrm{E}\bigg[\bigg(\sum_{k\leq k_m} \mathrm{e}^{t(2m/q-k)} Z_k\bigg)^2\bigg] + 2\,\mathrm{E}\bigg[\bigg(\sum_{k>k_m} \mathrm{e}^{t(2m/q-k)} Z_k\bigg)^2\bigg] \tag{4.24}$$

and bound each term on the right-hand side separately. For $k > k_m$, we have $2m/q - k < -m < -\Delta_m m$ and, hence,

$$\sum_{k>k_m} \mathrm{e}^{t(2m/q-k)} Z_k \leq \mathrm{e}^{-t\Delta_m m} \sum_{k>0} Z_k \overset{(3.5)}{\leq} \mathrm{e}^{-t\Delta_m m} \sum_{k=1}^{m} (\mathcal{Z}_k - \mathcal{Z}_{k-1}),$$

where the random variables $\mathcal{Z}_k$, $k = 0, \dots, m$, are as defined in the proof of Lemma 3.3. It follows from that proof that the random variables $\mathcal{Z}_k - \mathcal{Z}_{k-1}$ are independent and identically distributed with moment generating function

$$\Phi(s) = \sum_{a\in\mathbb{A}} \frac{\xi^2(a)s}{1 - s(1-\xi(a))}.$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Since $\mathcal{Z}_0 \equiv 0$, this implies that, on the one hand,

$$\begin{aligned}\mathrm{E}\Bigg[\bigg(\sum_{k>k_m} \mathrm{e}^{t(2m/q-k)} Z_k\bigg)^2\Bigg] &\le \mathrm{e}^{-2tm\Delta_m}\,\mathrm{E}\Bigg[\bigg(\sum_{k=1}^{m}(\mathcal{Z}_k - \mathcal{Z}_{k-1})\bigg)^2\Bigg] \\ &= \mathrm{e}^{-2t\Delta_m m}\bigg(m\,\mathrm{E}[\mathcal{Z}_1^2] + 2\binom{m}{2}\mathrm{E}[\mathcal{Z}_1]^2\bigg) \\ &= \mathrm{e}^{-tm\Delta_m}(m\Phi''(1) + m\Phi'(1) + m(m-1)\Phi'(1)^2) \\ &= \mathrm{e}^{-2tm\Delta_m}\bigg(2m\sum_{a\in\mathbb{A}}\frac{1}{\xi(a)} - m|\mathbb{A}| + m(m-1)|\mathbb{A}|^2\bigg). \end{aligned} \tag{4.25}$$

On the other hand,

$$\begin{aligned}\bigg(\sum_{k\le k_m} \mathrm{e}^{t(2m/q-k)} Z_k\bigg)^2 &\le \bigg(\sum_{k\le k_m} Z_k\bigg)\bigg(\sum_{k\le k_m} \mathrm{e}^{2t(2m/q-k)} Z_k\bigg) \\ &\le k_m^2 \sum_{k>0} \mathrm{e}^{2t(2m/q-k)} Z_k, \end{aligned} \tag{4.26}$$

where the first inequality follows from the Cauchy–Schwartz inequality and the second inequality follows from (2.7). Since $2t \in (0, 2t_m]$ and $\Delta = \Delta_m$ satisfy the conditions of Lemma 3.4, and since $q \le 1$, equations (3.7), (4.23), and (4.26) imply that

$$\mathrm{E}\Bigg[\bigg(\sum_{k\le k_m} \mathrm{e}^{t(2m/q-k)} Z_k\bigg)^2\Bigg] \le 4m^3|\mathbb{A}|(|\mathbb{A}|+1)^2(1+4|\mathbb{A}|m)\mathrm{e}^{-2tm\Delta_m}. \tag{4.27}$$

Using (4.24), (4.25), and (4.27),we readily obtain the result.

We are finally ready to give a proof of Theorem 4.1.

Proof of Theorem 4.1. Consider the events

$$E_{m,1} := \bigg\{\sum_{k=1}^{\infty} \mathrm{e}^{t_m(2m/q-k)}\hat{\nu}_k \ge 1 \text{ for all } q \in (0, \gamma - \Delta_m)\bigg\},$$

$$E_{m,2} := \bigg\{\sum_{k=1}^{\infty} \mathrm{e}^{t_m(2m/(\gamma+\Delta_m)-k)}\hat{\nu}_k < 1\bigg\}.$$

Equation (4.4) shows that $E_{m,1} \subseteq \{\gamma \le \hat{q}_m\}$ and $E_{m,2} \subseteq \{\hat{q}_m \le \gamma + 2\Delta_m\}$, which implies that

$$1 - \mathrm{P}[\gamma \le \hat{q}_m \le \gamma + 2\Delta_m] \le \mathrm{P}[\Omega \setminus E_{m,1}] + \mathrm{P}[\Omega \setminus E_{m,2}]. \tag{4.28}$$

However, Lemma 4.2 shows that, for all $m \ge m_6$,

$$\mathrm{P}[\Omega \setminus E_{m,1}] \le \mathrm{e}^{-\ell_0(1-\delta)^2/2}. \tag{4.29}$$

To bound the second term in (4.28), let $W(t, q)$ be as defined in (4.2). Then

$$\mathrm{E}[W(t_m, \gamma + \Delta_m)] = \sum_{k=1}^{\infty} \mathrm{e}^{t_m(2m/(\gamma+\Delta_m)-k)}\nu(k).$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

By Chebyshev's inequality,

$$\begin{aligned}
\mathrm{P}\bigg[\bigg|\sum_{k=1}^{\infty}\mathrm{e}^{t_m(2m/(\gamma+\Delta_m)-k)}\hat{\nu}_k-\mathrm{E}[W(t_m,\gamma+\Delta_m)]\bigg|\geq\frac{1}{2}\bigg]
&=\mathrm{P}\bigg[\bigg|\frac{1}{\ell_0}\sum_{\ell=1}^{\ell_0}W^{\ell}(t_m,\gamma+\Delta_m)-\mathrm{E}[W(t_m,\gamma+\Delta_m)]\bigg|\geq\frac{1}{2}\bigg]\\
&\leq\frac{4\,\mathrm{E}[W(t_m,\gamma+\Delta_m)^2]}{\ell_0}. \qquad (4.30)
\end{aligned}$$

However, for all $m\geq m_9$, $t=t_m$ and $\Delta=\Delta_m$ satisfy the conditions of Lemma 3.4, meaning that, for all $m\geq\max\{m_9,m_{10}\}$ with

$$m_{10}(\alpha,\xi):=\inf\{y>0\colon (x|\mathbb{A}|+4x^2|\mathbb{A}|^2)\mathrm{e}^{-x^{1-\alpha}/(16|\mathbb{A}|^2)}\leq\tfrac{1}{2}\text{ for all }x\geq y\},$$

we have

$$\mathrm{E}[W(t_m,\gamma+\Delta_m)]\leq\tfrac{1}{2}. \qquad (4.31)$$

Likewise, Lemma 4.3 shows that, for all $m\geq\max\{m_9,m_{11}\}$ with

$$\begin{aligned}
m_{11}&\equiv m_{11}(\alpha,\xi)\\
&:=\inf\bigg\{y>0\colon\bigg(34x^4|\mathbb{A}|(|\mathbb{A}|+1)^3+\frac{2x|\mathbb{A}|}{\min_{a\in\mathbb{A}}\xi(a)}\bigg)\mathrm{e}^{-x^{1-\alpha}/(8|\mathbb{A}|^2)}\leq\frac{1}{2}\text{ for all }x\geq y\bigg\},
\end{aligned}$$

we have

$$\mathrm{E}[W(t_m,\gamma+\Delta_m)^2]\leq\tfrac{1}{2}. \qquad (4.32)$$

Therefore,

$$\begin{aligned}
\mathrm{P}[\Omega\setminus E_{m,2}]&\overset{(4.31)}{\leq}\mathrm{P}\bigg[\bigg|\sum_{k=1}^{\infty}\mathrm{e}^{t_m(2m/(\gamma+\Delta_m)-k)}\hat{\nu}_k-\mathrm{E}[W(t_m,\gamma+\Delta_m)]\bigg|\geq\frac{1}{2}\bigg]\\
&\overset{(4.30),(4.32)}{\leq}\frac{2}{\ell_0}. \qquad (4.33)
\end{aligned}$$

The inequalities (4.28), (4.29), and (4.33) show that the theorem holds for

$$m_0(\alpha,\xi,\delta)=\max\{m_6,m_9,m_{10},m_{11}\}.$$

The claim concerning the order of m_0 as a function of α, ξ, and δ is easy to check directly.

We use the remainder of this section to give a brief discussion of the complexity of the algorithm that is implicitly defined in the statement of Theorem 4.1: given a confidence level $\Lambda\in(0,1)$ and an approximation error $\Xi\in(0,1)$, we wish to simulate an estimate, $\hat{q}_m$, such that

$$\mathrm{P}[\gamma\leq\hat{q}_m\leq\gamma+\Xi]\geq\Lambda. \qquad (4.34)$$

Corollary 4.1. *Let* $\alpha\in(0,1)$. *In order to compute an estimate,* $\hat{q}_m$, *that satisfies (4.34), it suffices to choose* $m=\mathcal{O}(\Xi^{-2/\alpha})$ *and to average over* $\ell_0=\mathcal{O}(1/(1-\Lambda))$ *Monte Carlo runs.*

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Proof. Let us choose $m \geq \max\{(2/\Xi)^{2/\alpha}, m_0\}$, $\ell_0 \geq 4/(1-\Lambda)$, and

$$\delta \leq 1 - \sqrt{(2/\ell_0)\log(2/\ell_0)}.$$

Then $2\Delta_m \leq \Xi$ and

$$\mathrm{e}^{-\ell_0(1-\delta)^2/2} \leq \ell_0/2.$$

Furthermore, Theorem 4.1 implies that

$$\mathrm{P}[\gamma \leq \hat{q}_m \leq \gamma + \Xi] \geq \mathrm{P}[\gamma \leq \hat{q}_m \leq \gamma + 2\Delta_m] \geq 1 - \mathrm{e}^{-\ell_0(1-\delta)^2/2} - \frac{2}{\ell_0} \geq 1 - \frac{4}{\ell_0} \geq \Lambda.$$

We remark that the value of α that minimizes the required size of m depends on Ξ but is bounded away from both 0 and 1. Let us now determine an estimate of the expected number of elementary computer operations required to compute $\hat{q}_m$ when m and ℓ_0 are chosen as in the proof of Corollary 4.1. For simplicity, we use a computational model that assumes a unit cost per real-number operation. There is no a-priori upper bound on the work required to generate the ℓ_0 independent, identically distributed pairs of sequences (X^ℓ, Y^ℓ). However, the ℓth pair has to be generated only up to the finite, *random* length that contains the full set of m-matches that start at the initial positions within X^ℓ and Y^ℓ. Lemma 3.3 implies that the expected number of terms in (X^ℓ, Y^ℓ) is smaller than $m(|\mathbb{A}|+1)$. Furthermore, Corollary 3.2 shows that it is exponentially rare in m that more than $\mathcal{O}(m)$ terms need to be generated. Computing the (finite) set of all m-matches contained in a pair (X^ℓ, Y^ℓ) thus requires the evaluation of a tableau of size $\mathcal{O}(m) \times \mathcal{O}(m)$ in the dynamic programming algorithm of [27]. Since each entry requires the same amount of work to evaluate, it takes $\mathcal{O}(m^2)$ work to evaluate the tableau and extract all the m-matches contained in the pair (X^ℓ, Y^ℓ). There are ℓ_0 of these computations, so the total work is $\mathcal{O}(\ell_0 m^2)$. Computing $\hat{q}_m$ from these data takes only $\mathcal{O}(\ell_0 m)$ extra time, so the total complexity for simulating $\hat{q}_m$ is seen to be $\mathcal{O}(\ell_0 m^2)$.

Corollary 4.1 now implies that computing an upper bound on γ to within an approximation error of Ξ and at the confidence level Λ takes an amount of work

$$\mathcal{O}((1-\Lambda)^{-1}\Xi^{-4/\alpha}). \tag{4.35}$$

We remark that the complexity estimate (4.35) is an upper bound derived on the basis of conservative estimates. The practical complexity seems to be considerably lower, as we will see in the next section.

5. Monte Carlo simulation in practice

The complexity bound (4.35) is interesting mainly from a theoretical perspective, as it is valid only for very large values of m. However, our theoretical analysis was conservative and relied on having $\hat{q}_m \geq \gamma + \Delta_m$. In order to use smaller values of m we need to take a statistical approach. Let there be a given confidence level, $\Lambda \in (0,1)$. As a first step we will define a function $\hat{v}(t,q)$ which is larger than $\mathrm{var}(W(t,q))$ with high probability, for any fixed pair $(t,q) \in \mathbb{R}^2_+$. We will argue that the estimate

$$\mathrm{P}[\hat{v}(t,q) > \mathrm{var}(W(t,q))] \geq 1 - \frac{1-\Lambda}{2} \tag{5.1}$$

is heuristically conservative, and use this bound in subsequent computations. There is strong empirical evidence that the bound (5.1) holds, but we do not claim that the results of this

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

section are rigorous mathematical conclusions; instead, the method described here should be considered as a heuristic.

For all $k \in \mathbb{N}$ and for fixed $r, s, u \in \mathbb{N}$, let $\{Z_{i,k}^{h,j} : 1 \leq h \leq r,\ 1 \leq j \leq s,\ 1 \leq i \leq u\}$ be independent, identically distributed copies of Z_k independent of the variables Z_k^{ℓ} used in the definition of $\hat{v}_k$, and for all $(t, q) \in \mathbb{R}_+^2$ let

$$\begin{aligned} W_i^{h,j}(t, q) &:= \sum_{k>0} \mathrm{e}^{t(2m/q-k)} Z_{i,k}^{h,j}, \\ U_i^h(t, q) &:= \frac{1}{s} \sum_{j=1}^{s} W_i^{h,j}(t, q), \\ \bar{U}_i(t, q) &:= \frac{1}{r} \sum_{h=1}^{r} U_i^h(t, q), \\ \hat{v}_i(t, q) &:= \frac{1}{r-1} \sum_{h=1}^{r} (U_i^h(t, q) - \bar{U}_i(t, q))^2. \end{aligned}$$

Then $\hat{v}_i(t, q)$ is an unbiased estimator of $\operatorname{var}(W(t, q))$. In order to render (5.1) a plausible assumption, it is important to choose s to be not too small: the larger s is chosen to be, the more symmetric the distribution of the estimators $\hat{v}_i(t, q)$ becomes around their mean, $\operatorname{var}(W(t, q))$. This phenomenon occurs because the distribution tails of the random variables $U_i^h(t, q)$ decay faster for larger s, and averages of $(U_i^h(t, q) - \bar{U}_i(t, q))^2$ thus converge (weakly) to a Gaussian variable much faster. It is important to realize that, although it follows from the results of previous sections that the distribution tails of $U_i^h(t, q)$ decay exponentially for all choices of s, this happens only for large values of $U_i^h(t, q)$; thus, empirically (and for numerical purposes), the tail decay is algebraic. This effect is illustrated in Figure 1, where the results of 1000 independent simulations of $W(t, q)$ are displayed for the case where ξ is the uniform distribution over the binary alphabet, $m = 100$, $t = 0.3$, and $q = 0.825$. This data was used to compute 50 samples of $\hat{v}_i(t, q)$, first for $r = 1000$ and $s = 1$, then for $r = 200$ and $s = 5$, and finally for $r = 10$ and $s = 100$, i.e. the same data was used but the averaging, as determined by the value of s, is different in each case. Note that the tail decay of $U_i^h(t, q)$ becomes steeper and the histogram of the samples $\hat{v}_i(t, q)$ becomes more symmetrical as the value of s increases.

Since the distribution of $\hat{v}_i(t, q)$ is not perfectly symmetrical for any choice of s, enough data need to be simulated in numerical experiments to make symmetry a reasonable heuristic assumption. The criterion for symmetry used in our experiments was that the generated samples of $\hat{v}_i(t, q)$ not reject the null hypothesis of a Gaussian distribution when a Lilliefors test on the 5% level was applied to 10 sample points. In the method described below we will use the values of (t, q) as the optimizers of an optimization problem: for a certain function $\Psi(t, q)$, we set $t(q) = \arg\min_{t \geq 0} \Psi(t, q)$ and then compute $\hat{q} = \min\{q > 0 \colon \Psi(t(q), q) \leq 1\}$. We thus need to design our numerical algorithm so that $\hat{v}_i(t(\hat{q}), \hat{q})$ are distributed symmetrically enough in the sense above described. We remark that when this criterion was satisfied, for $q > \hat{q}$ the sample variances $\hat{v}_i(t(q), q)$ were observed to be even more symmetrically distributed, in the sense that a Lilliefors test for normality did not reject the null hypothesis even with a higher p-value.

Let $\hat{v}(t, q) := \hat{v}_{[9]}(t, q)$ be the ninth order statistic of $\hat{v}_1(t, q), \dots, \hat{v}_{10}(t, q)$. We found strong empirical evidence that, under the above set of conditions, the latter variables are

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

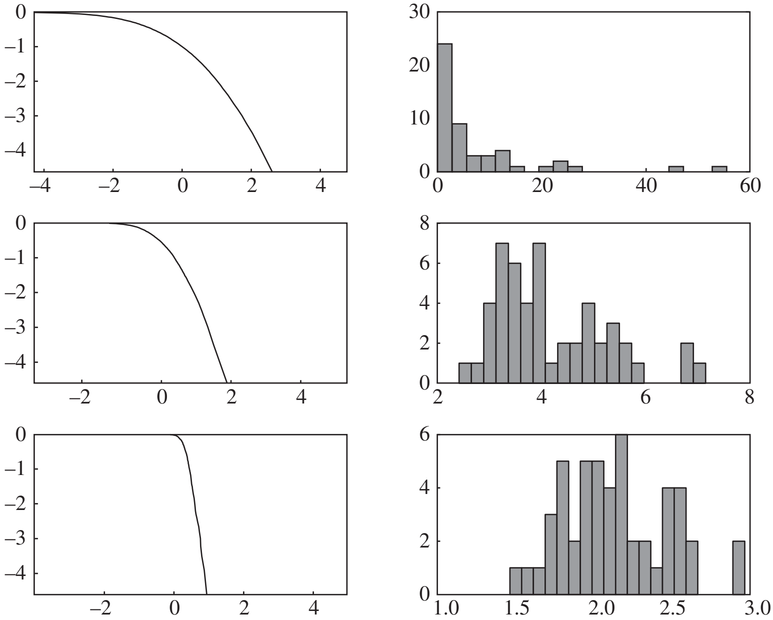

FIGURE 1: Plots of $\log \mathrm{P}[U_i^h(t,q) > \mathrm{e}^y]$ against y (*left*) and the histogram of $\hat{v}_i(t,q)$ (*right*) for parameter values $r = 1000$, $s = 1$ (*top*); $r = 200$, $s = 5$ (*middle*); and $r = 10$, $s = 100$ (*bottom*). The distribution of $\hat{v}_i(t,q)$ becomes more symmetrical as s increases.

sufficiently symmetrically distributed around their mean to render

$$\mathrm{P}[\hat{v}_i(t,q) \geq \mathrm{var}(W(t,q))] \geq 0.45 \tag{5.2}$$

a conservative estimate. It is now easy to see that (5.2) implies (5.1) for $\Lambda = 0.9$:

$$\mathrm{P}[\mathrm{var}(W(t,q)) \geq \hat{v}(t,q)] \leq \sum_{j=9}^{10} 0.55^j \times 0.45^{10-j}\binom{10}{j} = 0.034 < \frac{1-\Lambda}{2}. \tag{5.3}$$

We will henceforth make the heuristic assumption that *(5.3) (or equivalently, (5.1)) holds.*

Lemma 5.1. *The function*

$$\Psi\colon (t,q) \mapsto \sum_{k>0} \mathrm{e}^{t(2m/q-k)}\hat{v}_k + \sqrt{\frac{2\hat{v}(t,q)}{\ell_0(1-\Lambda)}}$$

is decreasing in q.

Proof. It suffices to show that the two summands are decreasing in q individually. For the first summand, we have

$$\frac{\partial}{\partial q}\sum_{k>0} \mathrm{e}^{t(2m/q-k)}\hat{v}_k = -\frac{2mt}{q^2}\sum_{k>0} \mathrm{e}^{t(2m/q-k)}\hat{v}_k < 0.$$

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

For the second summand, it suffices to prove that $\hat{v}_i(t,q)$ is decreasing in q, since the order statistics and the square root are both monotone increasing. Thus,

$$\begin{aligned}\frac{\partial}{\partial q}\hat{v}_i(t,q) &= \frac{1}{r-1}\sum_{h=1}^{r} 2(U_i^h(t,q)-\bar{U}_i(t,q))\Big(-\frac{2mt}{q^2}U_i^h(t,q)+\frac{2mt}{q^2}\bar{U}_i(t,q)\Big)\\ &= -\frac{4mt}{q^2}\hat{v}_i(t,q)\\ &< 0,\end{aligned}$$

which establishes the claim.

Next, let $q \in \mathbb{R}_+$ be fixed and let us design a test on the null hypothesis H_0, that $q < \gamma$, and the alternative hypothesis H_a that $\gamma \le q$. Let us assume that H_0 holds. Then, by virtue of Theorem 3.1 and (4.2), we have

$$\mathrm{E}\Bigg[\frac{1}{\ell_0}\sum_{\ell=1}^{\ell_0} W^\ell(t,q)\Bigg] = \mathrm{E}[W(t,q)] > 1 \tag{5.4}$$

for all $t \in \mathbb{R}_+$. By conditioning on the event $\{\hat{v}(t,q) \le \operatorname{var}(W(t,q))\}$ and its complement, we obtain

$$\begin{aligned}\mathrm{P}[\Psi(t,q)\le 1] &\le 1\times \mathrm{P}[\hat{v}(t,q)\le \operatorname{var}(W(t,q))] \qquad (5.5)\\ &\quad + \mathrm{P}\Bigg[\sum_{k>0}\mathrm{e}^{t(2m/q-k)}\hat{v}_k < 1-\sqrt{\frac{2\operatorname{var}(W(t,q))}{\ell_0(1-\Lambda)}}\Bigg]\times 1\\ &\overset{(5.1),(5.4)}{\le} \frac{1-\Lambda}{2} + \mathrm{P}\Bigg[\Bigg|\frac{1}{\ell_0}\sum_{\ell=1}^{\ell_0}W^\ell(t,q)-\mathrm{E}[W(t,q)]\Bigg| > \sqrt{\frac{2\operatorname{var}(W(t,q))}{\ell_0(1-\Lambda)}}\Bigg]\\ &\overset{\text{Chebyshev}}{\le} 1-\Lambda. \qquad (5.6)\end{aligned}$$

In particular, this holds for $t \equiv t(q) := \arg\min_{t\ge 0}\Psi(t,q)$. Therefore, if we design our test to reject H_0 when the criterion $\Psi(t(q),q) \le 1$ is satisfied (or, equivalently, if there exists a $t \in \mathbb{R}_+$ such that $\Psi(t,q) \le 1$), then the probability of a type-I error, i.e. of erroneously rejecting H_0, is $\mathrm{P}[\Psi(t(q),q) \le 1] \le 1-\Lambda$. We will be interested in the situation where H_0 is rejected, so the probability of a type-II error is irrelevant.

A Λ-confidence region is now associated with our test via the usual mechanism. Consider the random variable

$$\hat{q} := \min\{q\colon \text{ there exists a } t \in \mathbb{R}_+ \text{ such that } \Psi(t,q)\le 1\}, \tag{5.7}$$

where we follow the usual convention by setting $\hat{q} = \infty$ when the feasible set is empty. Then, from Lemma 5.1, H_0 is rejected if and only if $q \ge \hat{q}$. The probability that γ lies outside the confidence interval $[0,\hat{q}]$ is $\mathrm{P}[\hat{q} < \gamma] = 1-\Lambda$. This is easily seen, as

$$\mathrm{P}[\hat{q} < \gamma-\varepsilon] \le \mathrm{P}[\text{there exists a } t\in\mathbb{R}_+ \text{ such that } \Psi(t,\gamma-\varepsilon)\le 1] \overset{(5.6)}{\le} 1-\Lambda$$

holds for all $\varepsilon > 0$. It follows that $[0,\hat{q}]$ is a Λ-confidence interval for γ.

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

Table 1: New upper bounds $\hat{q}$ at the 90% confidence level, computed with $m = 1000$.

$\lvert\mathbb{A}\rvert$	BL	DL	AL	Est	$\hat{q}_{1000}$	DU	BU	AU	P	s	ℓ_0
2	0.7580	0.7739	0.8079	0.8118	0.8182	0.8376	0.8602	0.8607	0.0675	400	8 000
3	0.6338	0.6154	—	0.7172	0.7235	0.7658	0.7865	—	>0.2	400	12 000
4	0.5528	0.5455	—	0.6537	0.6601	0.7082	0.7297	—	>0.2	200	8 000

Table 2: New upper bounds $\hat{q}_m$ at the 98% confidence level, computed with $m = 100, 1000, 5000$.

$\lvert\mathbb{A}\rvert$	BL	DL	Est	$\hat{q}_{100}$	$\hat{q}_{1000}$	$\hat{q}_{5000}$	DU	BU
2	0.7580	0.7739	0.8118	0.8330	0.8203	0.8154	0.8376	0.8602
3	0.6338	0.6154	0.7172	0.7383	0.7266	—	0.7658	0.7865
4	0.5528	0.5455	0.6537	0.6777	0.6631	—	0.7082	0.7297
5	0.5095	0.5062	0.6069	0.6279	0.6152	—	0.6644	0.6861
6	0.4670	0.4717	0.5701	0.5928	0.5771	—	0.6293	0.6510
7	—	0.4450	0.5399	0.5596	0.5459	—	0.6002	0.6217
8	—	0.4224	0.5146	0.5352	0.5215	—	0.5754	0.5968
9	—	0.4032	0.4931	0.5098	0.4990	—	0.5539	0.5751
10	—	0.3866	0.4741	0.4951	0.4795	—	0.5349	0.5560

Corollary 5.1. *Subject to condition (5.3), the solution $\hat{q}$ to the optimization problem (5.7) constitutes an upper bound on γ with probability at least Λ.*

In our experiments we verified (5.3) empirically and solved (5.7) numerically. We implemented the above-described method in MATLAB® 6.1 and ran the experiments on a Sun Blade™ 100 Workstation for the uniform distributions over alphabets of size $|\mathbb{A}| = 2, 3, 4$. In all three experiments we chose $m = 1000$ and $\Lambda = 0.9$. Each of the experiments reported in Table 1 took a few days to complete, but there remains considerable room for code optimization. The value of $\hat{q}$ did not change significantly after a few hundred simulations, but more simulations were needed to obtain sufficiently symmetric variance estimators, as discussed above. The p-value, P, of the Lilliefors test and the number, s, of independent copies used in the computation of $\hat{v}_i(t, q)$ are listed in the tenth and eleventh columns, respectively. For comparison, we also list the best deterministic lower (DL) and upper (DU) bounds for these examples, as derived by Dančik and Paterson in [14] and [24], as well as the best known probabilistic lower (AL) and upper (AU) bounds at the 95% confidence level, obtained by Alexander [2] on the basis of two simulations of $\mathrm{E}[L_{50\,000}]$. Finally, we list the probabilistic lower (BL) and upper (BU) bounds of [9], although we do not know at which confidence level they apply, and their value (Est) estimated on the basis of ten simulations of $\mathrm{E}[L_{100\,000}]$, which is to be seen as a probabilistic lower bound without confidence guarantee.

Using a variant of our method, Decouvelaere [15] obtained further numerical results in his Masters thesis. He chose ξ to be the uniform distribution over alphabets of cardinality $|\mathbb{A}| = 2, \dots, 10$ and computed upper bounds $\hat{q}_m$ for $m = 100$ and $m = 1000$ at the confidence level $\Lambda = 98\%$. Furthermore, for $|\mathbb{A}| = 2$ he computed an upper bound for $m = 5000$ at the same confidence level. His results are reported in Table 2. Simulations with $m = 100$ are considerably faster than those with $m = 1000$, taking less than an hour to complete.

Steele [26] conjectured that, in the case where ξ is the uniform distribution over $|\mathbb{A}|$, the true value of γ is $2/(1 + \sqrt{|\mathbb{A}|})$. Since our new upper bounds are consistently smaller than the

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

conjectured values, and since our bounds have been obtained to a reasonable level of confidence, this indicates that the Steele conjecture may be wrong.

6. Conclusions

Our paper highlights the central role played by the theory of large deviations in revealing information about the Chvátal–Sankoff constant, γ. Exploiting large deviations, we designed a theoretical Monte Carlo algorithm to simulate upper bounds on γ to arbitrary precision, Ξ, and at an arbitrarily high confidence level, Λ. We presented a full complexity analysis, which shows that the amount of work required by the algorithm is polynomial in Ξ^{-1} and $(1-\Lambda)^{-1}$. A practical heuristic of our method turns out to be the first randomized algorithm that consistently generates bounds that are tighter than the deterministic bounds of Dančik and Paterson in [14] and [24].

Acknowledgements

We thank the anonymous referees for constructive comments that led to substantial improvements of the original manuscript. The first author's research was supported through grant GR/S34472 from the Physical Sciences Research Council of the UK and through grant NAL/00720/G of the Nuffield Foundation. The second author's research is supported by Nucleus Millennium Information and Randomness P04-069-F. The third author is grateful for partial support from Nucleus Millennium Information and Randomness P04-069-F for his visit to CMM-DIM in Santiago.

References

[1] Aldous, D. and Diaconis, P. (1999). Longest increasing subsequences: from patience sorting to the Baik–Deift–Johansson theorem. *Bull. Amer. Math. Soc.* **36,** 413–432.

[2] Alexander, K. S. (1994). The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Prob.* **4,** 1074–1082.

[3] Apostolico, A., Crochemore, M., Galil, Z. and Manber, U. (eds) (1993). *Combinatorial Pattern Matching* (Lecture Notes Comput. Sci. **684**). Springer, Berlin.

[4] Arratia, R. and Waterman, M. S. (1989). The Erdős–Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Prob.* **17,** 1152–1169.

[5] Arratia, R. and Waterman, M. S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Prob.* **4,** 200–225.

[6] Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen–Stein method. *Ann. Prob.* **17,** 9–25.

[7] Arratia, R., Gordon, L. and Waterman, M. S. (1990). The Erdős–Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.* **18,** 539–570.

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

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

[10] Baik, J., Deift, P. and Johansson, K. (1999). On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.* **12,** 1119–1178.

[11] Capocelli, R. M. (ed.) (1990). *Sequences*. Springer, New York.

[12] Capocelli, R., De Santis, A. and Vaccaro, U. (eds) (1993). *Sequences. II*. Springer, New York.

[13] Chvátal, V. and Sankoff, D. (1975). Longest common subsequences of two random sequences. *J. Appl. Prob.* **12,** 306–315.

[14] 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.

[15] Decouvelaere, Q. (2003). Upper bounds for the LCS problem. Masters Thesis, Computing Laboratory, University of Oxford.

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

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press

[17] Hauser, R. and Matzinger, H. (2005). Local uniqueness of alignments with a fixed proportion of gaps. Res. Rep. NA-05/08, Numerical Analysis Group, Computing Laboratory, University of Oxford. Available at http://web.comlab.ox.ac.uk/oucl/publications/natr/na-05-08.html.

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

[19] Kiwi, M., Loebl, M. and Matoušek, J. (2004). Expected length of the longest common subsequence for large alphabets. In *LATIN 2004: Theoretical Informatics* (Lecture Notes Comput. Sci. **2976**), Springer, Berlin, pp. 302–311.

[20] Krengel, U. (1985). *Ergodic Theorems*. De Gruyter, Berlin.

[21] Kruskal, J. B. (1983). An overview of sequence comparison: time warps, string edits, and macromolecules. *SIAM Rev.* **25,** 201–237.

[22] Lember, J. and Matzinger, H. (2005). Fluctuation of the LCS-score when letters are not equiprobable. Preprint.

[23] Neuhauser, C. (1994). A Poisson approximation for sequence comparisons with insertions and deletions. *Ann. Statist.* **22,** 1603–1629.

[24] Paterson, M. and Dančík, V. (1994). Longest common subsequences. In *Mathematical Foundations of Computer Science* (Lecture Notes Comput. Sci. **841**), Springer, Berlin, pp. 127–142.

[25] 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.

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

[27] Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *J. Assoc. Comput. Mach.* **21,** 168–173.

[28] Waterman, M. (1995). *Introduction to Computational Biology*. Chapman and Hall, London.

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

[30] Waterman, M. S. (1994). Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. London B* **344,** 383–390.

[31] Waterman, M. S. and Vingron, M. (1994). Sequence comparison significance and Poisson approximation. *Statist. Sci.* **9,** 367–381.

https://doi.org/10.1239/aap/1158685004 Published online by Cambridge University Press