High precision simulations of the longest common subsequence problem

R. Bundschuh[a]

Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319, USA

Abstract. The longest common subsequence problem is a long studied prototype of pattern matching problems. In spite of the effort dedicated to it, the numerical value of its central quantity, the Chvátal-Sankoff constant, is not yet known. Numerical estimations of this constant are very difficult due to finite size effects. We propose a numerical method to estimate the Chvátal-Sankoff constant which combines the advantages of an analytically known functional form of the finite size effects with an efficient multi-spin coding scheme. This method yields very high precision estimates of the Chvátal-Sankoff constant. Our results correct earlier estimates for small alphabet size while they are consistent with (albeit more precise than) earlier results for larger alphabet size.

1 Introduction

Pattern matching problems have been of central interest in such different areas of science as image processing, speech recognition, time series analysis, and biological sequence comparison for a large number of years. In many practical applications of pattern matching, it is important to be able to characterize the amount and/or strength of patterns found in *random data* in order to be able to distinguish meaningful patterns from random ones [1]. This requires a *quantitative theory* of pattern matching procedures which does not only characterize universal characteristics like scaling laws but which also provides specific numerical values for non-universal quantities.

One of the simplest pattern matching problems is the longest common subsequence (LCS) problem. It is a special case of the sequence comparison problem, which is a very important tool in modern molecular biology [2]. As a prototype for pattern matching it has enjoyed the attention of mathematicians and computer scientists for a long time [3–13]. One of the central quantities in the LCS problem is the average fraction of matches in the longest common subsequence of two very long random strings also known as the *Chvátal-Sankoff constant a_c* [3]. However, in spite of the relative simplicity of the LCS problem and the long time it has been studied, the numerical value of the Chvátal-Sankoff constant is still unknown. There has

been a long standing conjecture

$$a_c \stackrel{?}{=} \frac{2}{\sqrt{c}+1} \qquad (1)$$

by Arratia [14] for its value based on numerical simulations where c is the size of the alphabet from which the sequences are chosen at random. This conjecture has been recently shown to hold for a *first-passage percolation version* of the LCS problem using the analogy [15] to surface growth as described by the Kardar-Parisi-Zhang (KPZ [17]) equation [16] and a cavity method [12,13]. In the language of surface growth the Chvátal-Sankoff constant takes the interpretation of the average growth velocity of the surface or of the ensemble averaged free energy per length in the directed polymer picture which is equivalent to the KPZ equation.

For the original Chvátal-Sankoff constant (*i.e.*, not the first-passage percolation version) a variety of *upper and lower bounds* have been rigorously proven in the mathematical literature [3,4,7,8,10,11] which are all consistent with Arratia's conjecture. Short of a rigorous result for the value of the Chvátal-Sankoff constant, *extensive numerical studies* are the only way to shed some light on the validity of Arratia's conjecture. The difficulty with numerical studies is that the Chvátal-Sankoff constant is defined as an asymptotic quantity in the limit that the sequences the longest common subsequence of which is computed are infinite. Simulations at finite sequence length always observe *finite size effects* the size and even functional form of which is not known. In fact, two recent numerical

---

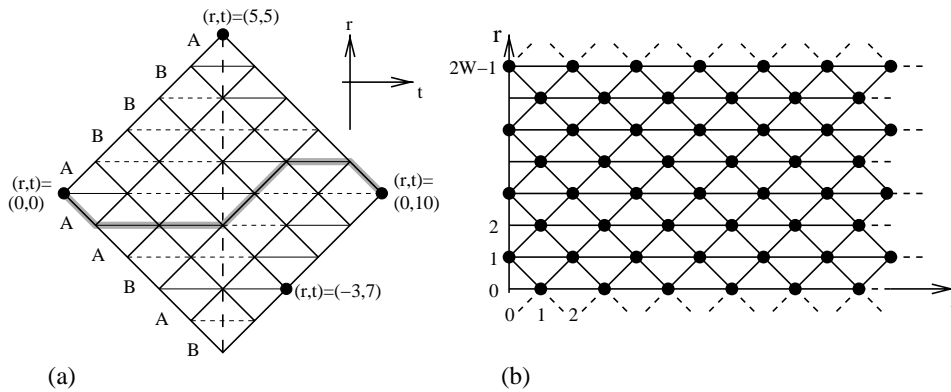[a] e-mail: `rbund@mps.ohio-state.edu`

Fig. 1. Lattice representation of the longest common subsequence problem. (a) shows how the two sequences ABBBA and AABAB are written at the edges of the lattice. The horizontal bonds of the lattice correspond to matches with a weight one (solid lines) or to mismatches with a weight zero (dashed lines). The diagonal bonds correspond to omitted letters. Each common subsequence corresponds to a path from the leftmost to the rightmost point of the lattice. The length of the common subsequence is the sum over all weights of the bonds which the path passed. The path shown as an example has a weight of three and corresponds to the maximal common subsequence ABA. The figure also shows how we use the coordinates r and t in order to denote lattice points and how the diamond shaped lattice can be split into two triangular lattices (long dashed line). (b) shows the rectangular lattice which we use instead of the lattice of (a) in order to be able to handle the effects of the finite lattice size.

studies [11–13] assume different functional forms of the finite size effect. Thus, they come to *different conclusions* about the value of the Chvátal-Sankoff constant although both studies agree on the fact that the Chvátal-Sankoff constant *differs* from its conjectured value equation (1).

In this communication, we will remedy this dilemma in two different ways. First, we will use a slightly different geometry in which we perform our simulations. The advantage of this geometry is that the form of the finite size behavior is *theoretically known* from the analogy to surface growth [15] and does not have to be inferred from the numerical data. We will prove this finite size behavior for the first-passage percolation case explicitly. While this result reproduces the known value of the first-passage percolation version of the Chvátal-Sankoff constant, it additionally provides the *complete finite-size correction* in terms of Legendre polynomials and an *explicit formula* for the prefactor of the leading finite size correction term. We will use this to validate our numerical procedure. Additionally, we will reformulate the LCS problem in a way that allows for a *multi-spin coding scheme*, in which 64 lattice sites are updated in parallel by simple bitwise logical operations. This extremely efficient code allows us to average over a large number of systems with several million letters each. This yields the value of the Chvátal-Sankoff constant with a precision of one part in 10^6.

2 Longest common subsequence problem

The LCS problem is defined as follows: Given two sequences $\boldsymbol{x} = x_1 \ldots x_N$ and $\boldsymbol{y} = y_1 \ldots y_N$ of letters x_i and y_j independently chosen with equal probabilities from an alphabet of size c, find their longest common subsequence. A common subsequence of length ℓ is given by a set of

(i_k, j_k)'s with $1 \leq i_1 < \ldots < i_\ell \leq N$ and $1 \leq j_1 < \ldots < j_\ell \leq N$ such that $x_{i_k} = y_{j_k}$ for all $k \in \{1, \ldots, \ell\}$. We denote the length of the longest common subsequence of two given sequences by $L(N)$. It is a random variable since it depends on the specific choice of the sequences. Its expectation value averaged over the ensemble of all sequence pairs is $\langle L(N) \rangle$. Chvátal and Sankoff noted [3] that this expectation value is asymptotically proportional to the length of the sequences, *i.e.*, $\langle L(N) \rangle \approx a_c N$. The prefactor a_c is now called the Chvátal-Sankoff constant and depends only on the size c of the alphabet. To find high precision estimates for this value for different alphabet sizes c is the purpose of this publication.

Numerically, the length of the longest common subsequence of two given sequences can be calculated very efficiently using a transfer matrix algorithm. This transfer matrix algorithm calculates the auxiliary quantity $\ell_{i,j}$ which is defined to be the length of the longest common subsequence of the substrings $x_1 \ldots x_i$ and $y_1 \ldots y_j$. The recursion for $\ell_{i,j}$ is

$$\ell_{i,j} = \max\{\ell_{i-1,j-1} + \eta_{i,j}, \ell_{i-1,j}, \ell_{i,j-1}\} \qquad (2)$$

where

$$\eta_{i,j} = \begin{cases} 1 & \text{if } x_i = y_j \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

This recursion is completed by the initial conditions

$$\ell_{i,0} = \ell_{0,j} = 0. \qquad (4)$$

The length of the longest common subsequence of the original strings $\boldsymbol{x}$ and $\boldsymbol{y}$ is then $L(N) = \ell_{N,N}$. The recursion equation can be visualized by assigning the $\ell_{i,j}$ to the nodes of the lattice shown in Figure 1a [18,19]. The horizontal bonds then correspond to the $\eta_{i,j}$ and the diagonal

bonds correspond to omitting a letter from one of the sequences. The recursive algorithm calculates the path from the origin to the end point of this lattice with the largest weight as given by the sum over the bond weights through which the path passes. This is clearly a discrete model of a directed polymer (the path) in a random medium [15] (the random matches and mismatches $\eta_{i,j}$). To make this more explicit, we will use "spatial" and "temporal" coordinates r and t instead of the indices i and j to refer to lattice points as shown in Figure 1a. In this coordinate system the dynamics given by the recursive algorithm reads

$$\ell(r, t+1) = \max \begin{Bmatrix} \ell(r, t-1) + \eta(r, t) \\ \ell(r-1, t) \\ \ell(r+1, t) \end{Bmatrix} . \qquad (5)$$

The total length $L(N) = \ell(0, 2N)$ of the longest common subsequence corresponds to the total free energy of the directed polymer. This is clearly an extensive quantity and the proportionality constant, the free energy per length, is the Chvátal-Sankoff constant. It can in principle be calculated numerically by running the algorithm (5) for a large number of sequences and averaging the free energies obtained. As already mentioned the difficulty with this procedure is that the free energy has terms which are subleading in the length N of the sequences. Thus, the ratio $\langle L(N) \rangle / N$ only converges towards the Chvátal-Sankoff constant a_c as the sequences become very long. For high precision measurements, the rate of this convergence has to be taken into account for a proper extrapolation towards infinite sequence lengths. This is relatively hard for the diamond shaped lattice, since there is no theoretical understanding of its finite size effects. The only reliable result is a *bound* [19] $\langle L(N) \rangle / N \leq a_c + O(N^{-1/2} \log N)$ on the finite size dependence. Dančík [11] assumed that this bound represents the exact functional form of the finite size correction and used it to fit his numerical data generated from four pairs of sequences of length up to one million to arrive at a value of $a_{c=2} = 0.81225 \pm 0.00025$ for a two letter alphabet. More recently, Boutet de Monvel tried to extract the finite size behavior directly from the numerical data without any theoretical foundation and concluded $\langle L(N) \rangle / N \approx a_c + O(N^{-1/2} / \log N)$ which lead him to a value[1] of $a_{c=2} = 0.81233 \pm 0.00005$ [12,13].

3 LCS on a rectangular lattice

In light of the confusion on the handling of finite size effects, we will consider a rectangular lattice of a given width W as the one shown in Figure 1b. In order to minimize the finite size effects, we apply periodic boundary conditions in the spatial (vertical) direction. For any fixed W

---

[1] Boutet de Monvel does not give an actual statistical error of his estimates. Instead he gives three different estimates for the Chvátal-Sankoff constant at each alphabet size. These three values stem from three different ways to treat the finite size effects. The statistical error cited here reflects the spread of these three values.

the dynamics given by equation (5) will after some startup phase become stationary which allows us to measure

$$a_c(W) \equiv \lim_{t \to \infty} \frac{1}{t} \langle \ell(W-1, 2t) \rangle \qquad (6)$$

$$= \lim_{t \to \infty} \frac{1}{t} \frac{1}{2W} \sum_{k=0}^{W-1} \langle \ell(2k+1, 2t) + \ell(2k, 2t-1) \rangle$$

with nearly arbitrary precision if we use long enough sequences (note also, that in this geometry we have to perform only WN instead of N^2 elementary steps as the one given in equation (5) on the diamond shaped lattice which allows for much larger sequence lengths N). This geometry directly corresponds to the finite size geometry usually imposed in studies of directed polymers in a random medium and equivalent KPZ surface growth problems [20]. From these studies it is well known [21], that

$$a_c(W) = a_c(\infty) - \frac{b_c}{W} + \text{higher order terms} \qquad (7)$$

with some constant b_c. We argue in Appendix A that

$$a_c(\infty) \equiv \lim_{W \to \infty} a_c(W) = a_c, \qquad (8)$$

i.e., that calculating $a_c(W)$ for fixed W and taking the limit $W \to \infty$ actually yields the Chvátal-Sankoff constant a_c. The advantage of this procedure over other numerical approaches is that we know the functional form of the finite size dependence of $a_c(W)$ in advance and can extract the Chvátal-Sankoff constant a_c by fitting numerically estimated $a_c(W)$ to the functional form

$$a_c(W) \approx a_c - \frac{b_c}{W} . \qquad (9)$$

At this point we want to remark for the reader familiar with the surface growth analogy of the LCS problem that in the surface growth language there are *two* finite size correction formulae known [21]. If the surface grows for a very long time t but is restricted to a width $W \ll t^{2/3}$ equation (7) applies. In the opposite regime, *i.e.*, for $W \gg t^{2/3}$, there is a different scaling behavior which now depends on t instead of W. It appears that the diamond shaped lattice of the original LCS problem is in this transient phase since we have $t = 2W$. However, the known finite size correction formula in this regime applies only if the surface grows from a *flat* substrate while the diamond lattice forms a *wedged* substrate leading to the more difficult finite size effects observed numerically by Boutet de Monvel [12,13]. It turns out, though, that the flat substrate can also be translated into the LCS language: It requires cutting the diamond shaped lattice of size t into two triangles as indicated by the long dashed line in Figure 1a, applying the recursion equation (5) with the diamond boundary conditions equation (4) within the left triangle, and then calculating

$$\widetilde{a}_c(t) \equiv \frac{2}{t} \left\langle \max_{r \in \{-t, -t+2, \dots t-2, t\}} \ell(r, t) \right\rangle . \qquad (10)$$

It can be shown [19] that $\lim_{t\to\infty}\widetilde{a}_c(t)=a_c$ and the finite size correction should for this geometry be [21]

$$\widetilde{a}_c(t)=a_c-\frac{\widetilde{b}_c}{t^{2/3}}+\text{higher order terms.}\qquad(11)$$

This explicit finite size dependence formula could also be used in order to get high precision estimates of the Chvátal-Sankoff constant. However, as we will see below the rectangular geometry is advantageous in terms of the numerical evaluation of the recursion equation (5). Thus, we will use the rectangular geometry in our current study.

For a better theoretical understanding of the longest common subsequence problem on the rectangular lattice and in order to formulate a very efficient algorithm for its simulation it is helpful to concentrate on the differences between neighboring $\ell(r,t)$. In order to preserve the symmetry of the lattice shown in Figure 1b we define them as

$$\Delta(r,t)\equiv\begin{cases}\ell(r,t+1)-\ell(r+1,t) & \text{for } r+t \text{ even}\\ \ell(r+1,t+1)-\ell(r,t) & \text{for } r+t \text{ odd.}\end{cases}$$
$$(12)$$

From the recursion equation (5) it can be easily checked by induction that $\Delta(r,t)\in\{0,1\}$. They can thus be considered *Boolean variables*. In terms of these quantities, the recursion equation (5) can be rewritten as

$$\Delta(r,t)=\max\begin{cases}\eta(r,t)-\Delta(r,t-1)\\ \Delta(r-1,t-1)-\Delta(r,t-1)\\ 0\end{cases}\quad(13)$$

$$\Delta(r-1,t)=\max\begin{cases}\eta(r,t)-\Delta(r-1,t-1)\\ \Delta(r,t-1)-\Delta(r-1,t-1)\\ 0\end{cases}\quad(14)$$

for even $r+t$. Note, that this formulation does not make any reference to the absolute lengths $\ell(r,t)$ of the longest common subsequences. The dynamics of the LCS problem is thus completely captured by the dynamics of the length differences as given by equations (13) and (14). Every configuration of the $\ell(r,t)$ can be equally well represented by a configuration of the $\Delta(r,t)$. Reversely, a configuration of the $\Delta(r,t)$ represents a configuration of the $\ell(r,t)$, if the consistency condition

$$\sum_{r=0}^{2W-1}(-1)^r\Delta(r,t)=0\qquad(15)$$

for the validity of the periodic boundary conditions in $\ell(r,t)$ is fulfilled. The alternating sum on the left hand side of equation (15) is conserved under the dynamics given by equations (13) and (14) and is thus always fulfilled, if it is fulfilled for the initial choice of the $\Delta(r,t=0)$.

Slaved to the dynamics of the Δ is the evolution of the *average length*

$$\overline{\ell}(t)\equiv\begin{cases}\dfrac{1}{2W}\displaystyle\sum_{k=0}^{W-1}[\ell(2k,t-1)+\ell(2k+1,t)] & \text{for } t \text{ even}\\[18pt]\dfrac{1}{2W}\displaystyle\sum_{k=0}^{W-1}[\ell(2k,t)+\ell(2k+1,t-1)] & \text{for } t \text{ odd}\end{cases}$$

of the longest common subsequence which changes by

$$\overline{\ell}(t+1)-\overline{\ell}(t)=\frac{1}{2W}\sum_{k=0}^{W-1}\begin{cases}\delta(2k,t) & \text{for } t \text{ even}\\ \delta(2k+1,t) & \text{for } t \text{ odd}\end{cases}\quad(16)$$

with

$$\delta(r,t)\equiv\ell(r,t+1)-\ell(r,t-1)=\max\begin{cases}\eta(r,t)\\ \Delta(r,t-1)\\ \Delta(r-1,t-1)\end{cases}$$
$$(17)$$

in every time step. Together with equation (6) it thus yields

$$a_c(W)=\langle\delta(r,t)\rangle\cdot\qquad(18)$$

The rewriting of $a_c(W)$ in terms of $\langle\delta\rangle$ opens the possibility of a very fast implementation of the longest common subsequence problem. We note that equations (13) and (14) transform a pair of length differences into another pair of length differences independent of the other pairs at the same time t. According to the lattice structure shown in Figure 1b this happens for different ways of pairing at different times. If we interpret the $\Delta(r,t)$ and $\eta(r,t)$ as Boolean variables, we can give equations (13) and (14) the form

$$\Delta(r,t)=[\neg\,\Delta(r,t-1)]\wedge[\eta(r,t)\vee\Delta(r-1,t-1)]\quad(19)$$

$$\Delta(r-1,t)=[\neg\,\Delta(r-1,t-1)]\wedge[\eta(r,t)\vee\Delta(r,t-1)].\quad(20)$$

Thus, we store 32 of the $\Delta(r,t)$ for even r and the corresponding 32 of the $\Delta(r,t)$ for odd r simultaneously as the single bits of one integer variable each. 32 of the match-mismatch variables $\eta(r,t)$ can be stored in the same way. Then, equations (19) and (20) can be performed in parallel on these 64 lattice sites by bitwise logical operations. After these operations the odd sites are shifted by one bit to the right so that they pair up with their new even site partners and the same procedure is repeated. After shifting the odd sites one bit back to the left the algorithm can proceed from the beginning

In order to measure $a_c(W)$ we note that the increase $\delta(r,t)$ in equation (16) is given by the Boolean equation

$$\delta(r,t)=\eta(r,t)\vee\Delta(r,t-1)\vee\Delta(r-1,t-1)\qquad(21)$$

and can thus also be easily calculated for 32 pairs of sites with two bitwise or operations. A bit-count then gives the value of $a_c(W)$ after averaging over many time steps and configurations of the disorder $\eta(r,t)$.

4 First passage percolation

We verify the algorithm and the finite size behavior given by equation (9) on the first passage percolation version of the LCS problem. For this system we are not only able to prove explicitly the scaling form equation (9) but we can calculate the full finite size correction and specifically the prefactor $\widehat{b}_c$ of its leading term. The first passage percolation version of the LCS problem also obeys the dynamics given by equations (13) and (14). The only difference is, that instead of choosing random sequences $x_1 \ldots x_N$ and $y_1 \ldots y_N$ and calculating the local matches $\eta(r,t)$ from these sequences, we use the variables $\widehat{\eta}(r,t)$ which are taken to be *independent* identically distributed random variables with

$$\widehat{\eta}(r,t) = \begin{cases} 1 \text{ with probab. } 1/c \\ 0 \text{ with probab. } 1 - 1/c. \end{cases} \quad (22)$$

The analog $\widehat{a}_c$ of the Chvátal-Sankoff constant for this problem is known to be $\widehat{a}_c = 2/(\sqrt{c} + 1)$ which had been conjectured by Arratia for the a_c of the LCS problem. This result follows, *e.g.*, as a consequence of an exact mapping [22] between the LCS problem and the discrete time asymmetric exclusion process [23,24], a well studied exactly solvable model of the KPZ surface growth universality class. (The mapping basically interprets the length differences $\Delta(r,t)$ as the site occupation numbers of the asymmetric exclusion process). Below, we will derive the result $\widehat{a}_c = 2/(\sqrt{c}+1)$ and its finite size correction without explicit reference to this mapping.

  Equations (19) and (20) can be interpreted in terms of a pair $(\Delta(r-1,t-1), \Delta(r,t-1)) \in \{(0,0),(0,1),(1,0),(1,1)\}$ mapped onto another pair $(\Delta(r-1,t), \Delta(r,t)) \in \{(0,0),\ (0,1),\ (1,0),\ (1,1)\}$ with probabilities depending on the possible values of $\widehat{\eta}(r,t)$. This mapping can be written as the matrix

$$T_0 \equiv \begin{pmatrix} 1-\frac{1}{c} & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{c} & 0 & 0 & 0 \end{pmatrix}, \quad (23)$$

in the basis $|00\rangle, |01\rangle, |10\rangle, |11\rangle$. A distribution of possible states of the system can be interpreted as a normalized vector $|\psi\rangle$ in the 2^{2W} dimensional space of all possible configurations of the $2W$ variables $\Delta(r,t)$ at a fixed time. In a time step with even t, each neighboring pair of length differences is transformed according to the matrix T_0. Thus, the whole state distribution $|\psi\rangle$ is transformed into the vector $T|\psi\rangle$ with

$$T \equiv \bigotimes_{i=1}^{W} T_0. \quad (24)$$

At odd time steps the pairing of the length differences is different. In order to describe the transformation at odd

time steps, we introduce the translation operator C, which shifts all length differences $\Delta(r,t)$ by one spatial position to the right, *i.e.*,

$$C|\Delta_{2W-1} \ldots \Delta_1 \Delta_0\rangle \equiv |\Delta_0 \Delta_{2W-1} \ldots \Delta_1\rangle \cdot \quad (25)$$

The time evolution at odd times is then given by a shift of all length differences to the left, application of the even time evolution operator T and a shift of all length differences back to the right, *i.e.*, by CTC^{-1}. After two time steps, the state vector $|\psi\rangle$ is thus transformed into the state vector $CTC^{-1}T|\psi\rangle$. The stationary state is the eigenvector $|\psi_0\rangle$ of $CTC^{-1}T$ with the eigenvalue one which is consistent with the condition (15). It is easy to check, that this eigenvector is given by

$$|\psi_0\rangle \equiv \frac{1}{\mathcal{N}_W} \sum_{\left\{\Delta_j : \sum\limits_{i=0}^{2W-1} (-1)^i \Delta_i = 0\right\}} \left(\frac{1}{\sqrt{c}}\right)^{\sum\limits_{i=0}^{2W-1} \Delta_i} |\Delta_{2W-1} \ldots \Delta_0\rangle \cdot$$
$$(26)$$

with the combinatorial normalization factor

$$\mathcal{N}_W = \sum_{\mu=0}^{W} \left(\frac{1}{c}\right)^{-\mu} \binom{W}{\mu}^2 = \left(1 - \frac{1}{c}\right)^W P_W\left(\frac{c+1}{c-1}\right), \quad (27)$$

where P_W is the Legendre polynomial of degree W. Since $CTC^{-1}T$ restricted to the subspace of state vectors consistent with condition (15) is irreducible and neither the eigenvector $|\psi_0\rangle$ nor $CTC^{-1}T$ have any negative entries, it is the unique stationary state of the system according to the Perron Frobenius theorem [25].

  The analog $\widehat{a}_c(W)$ of the Chvátal-Sankoff constant for this modified system can be calculated from the knowledge of this stationary state *via* equations (18) and (21). Since $\widehat{\eta}(r,t)$ is uncorrelated from $\Delta(r-1,t-1)$ and $\Delta(r,t-1)$ we get

$$\widehat{a}_c(W) = \langle \widehat{\eta}(r,t) \vee \Delta(r,t-1) \vee \Delta(r-1,t-1)\rangle \quad (28)$$

$$= 1 - \left(1 - \frac{1}{c}\right) \langle [\neg \Delta(r,t-1)] \wedge [\neg \Delta(r-1,t-1)]\rangle$$

$$= 1 - \left(1 - \frac{1}{c}\right) \Pr\{\Delta(r,t-1) = \Delta(r-1,t-1) = 0\} \cdot$$

The probability that two neighboring length differences vanish simultaneously can be directly calculated from the explicitly known stationary state equation (26). Since fixing a pair of the length differences to zero still leaves all the possibilities to the $W-1$ remaining pairs, it is

$$\Pr\{\Delta(r,t-1) = \Delta(r-1,t-1) = 0\} = \frac{\mathcal{N}_{W-1}}{\mathcal{N}_W} \cdot \quad (29)$$

Inserting this into equation (28) yields the *full finite size dependence* of the first passage percolation version of the LCS problem in terms of Legendre polynomials. Expanding these Legendre polynomials for large degrees W finally

yields the first order terms

$$\widehat{a}_c(W) = \frac{2}{\sqrt{c}+1} - \frac{\sqrt{c}-1}{2(\sqrt{c}+1)}\frac{1}{W} + O\left(\frac{1}{W^2}\right) \quad (30)$$

of this exact finite size correction.

This can be directly compared to the numerical results obtained by the multi-spin coding algorithm described above. To this end we choose 1 000 random initial conditions of the $\Delta(r, t = 0)$ (that fulfil Eq. (15)) and the same number of random configurations of the $\widehat{\eta}(r,t)$. Then, we proceed as described at the end of Section 3, *i.e.*, we apply the recursion equations (19) and (20) a large number of times, evaluate equation (21) in every time step, and generate estimates of $\widehat{a}_c(W)$ according to equation (18) by averaging over many time steps and the different configurations of the disorder.

An important question is how close these estimates of $\widehat{a}_c(W)$ come to its true value. In answering this question, we have to consider *statistical* and *systematic* deviations. The statistical deviations can be measured by looking at the sample-to-sample fluctuations between the different configurations of the disorder $\widehat{\eta}(r,t)$ and made as small as we wish by increasing the number of disorder configurations and the number of recursion steps we average over. The systematic deviations have to be studied more carefully. They arise because the Markov dynamics of the differences $\Delta(r,t)$ reaches its stationary state only in the limit of infinitely many recursion steps if we start from a random initial state. For a finite number t_1 of recursions steps there is a systematic correction that decays exponentially in t_1 with some decay "time" t_0.

In order to verify if this exponential approach to the stationary state plays a role in our numerics, we iterate equations (19) and (20) 10^7 times for each configuration of the disorder $\widehat{\eta}(r,t)$ and calculate 10 different estimates of $\widehat{a}_c(W)$ by averaging in equation (18) over $t \in]t_1, t_1 + 10^6]$ for $t_1 \in \{0, 10^6, 2 \times 10^6, \dots, 9 \times 10^6\}$. If there is a systematic deviation this should manifest itself in a systematic dependence of these estimates on the number t_1 of recursion steps performed before the averaging starts.

Since the approach to the stationary state has to become the slower the wider the lattice we have to study this effect for the largest width $W = 4096$ we plan on using in our numerics. Figure 2a shows the ten estimates of $\widehat{a}_2(W = 4096)$ which are obtained as described above. Clearly, there is no noticeable systematic dependence on the number t_1 of recursion steps. All fluctuations of the individual points are within the statistical uncertainties of each other. While the statistical uncertainties of the individual points are already on the order of $\pm 10^{-6}$ we can further improve them by averaging over all the points. To be absolutely certain that the approach to the stationary state does not influence our result, we take this average only over the last nine points, *i.e.*, we average over all $t \in]10^6, 10^7]$. This average is indicated as the dashed line in Figure 2a.

After we have convinced ourselves, that we can estimate the $\widehat{a}_c(W)$ to within a precision of $\pm 10^{-6}$ for all W, we can study the dependence of $\widehat{a}_c(W)$ on the width W

of the lattice. Figure 3a shows these numerically obtained values of $\widehat{a}_2(W)$ for an alphabet of size $c = 2$ as a function of $1/W$. The points clearly fall on a straight line over the whole range of lattice widths from $W = 128$ to $W = 4096$. A simple least square fit to a straight line yields

$$\widehat{a}_2(W) \approx (0.828\,427\,0 \pm 0.000\,000\,9)$$
$$- (0.084\,8 \pm 0.000\,2)\frac{1}{W}. \quad (31)$$

This has to be compared to the exact results

$$\frac{2}{\sqrt{2}+1} \approx 0.828\,427\,12 \quad \text{and} \quad \frac{\sqrt{2}-1}{2(\sqrt{2}+1)} \approx 0.085\,79.$$

We conclude that the value for infinite system size is extremely well reproduced by taking into account the $1/W$ finite size correction. Even the prefactor of the finite size correction can be extracted with a reasonably high precision, although the error of the fit for the finite size coefficient as given in equation (31) underestimates the true error. This is to be expected since the finite size coefficient itself is subject to further finite size corrections which yield an additional *systematic* error in its estimation.

5 The Chvátal-Sankoff constant

In order to get a good numerical estimate of the real Chvátal-Sankoff constant, we perform the same simulations for the subtly correlated disorder $\eta(r,t)$ generated from pairs of randomly chosen sequences. In the presence of these disorder correlations, the dynamics of the $\Delta(r,t)$ at a given width W can still be interpreted as a Markov process. However, the applicable state space comprises at a given t the $\Delta(r,t)$ for $r \in \{0, \dots, 2W - 1\}$ *and* the last W letters chosen in each of the two sequences. It is much larger than in the first passage percolation version. Thus, although the dynamics still converges exponentially to the stationary state, the initial phase can be much longer than in the first passage percolation case and more care has to be taken to minimize the systematic error of numerical estimates introduced by this effect.

Again, we collect estimates from 1 000 pairs of random sequences and apply the recursion equations (19) and (20) 10^7 times for each sequence pair. Then, we generate the same 10 estimates of $a_2(W = 4096)$ as in Section 4 for the largest width $W = 4096$ we plan to use in our simulations. Figure 2b shows the dependence of these estimates on the number t_1 of recursion steps performed before starting the averaging. We note two important differences to the first passage percolation case: (i) The estimate with $t_1 = 0$ is significantly different from the estimates with $t_1 \geq 10^6$ and (ii) the individual statistical fluctuations are larger than in the first passage percolation case (note the difference in scale between Figures 2a and b). Both findings are to be expected due to the slower approach to the stationary state. Since the first estimate is different from the other nine we conclude that the characteristic number of steps t_0 necessary to approach the stationary state is smaller than
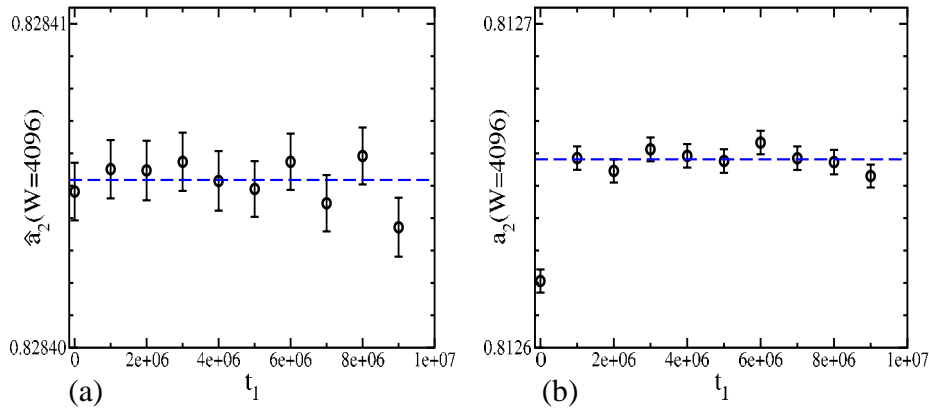
Fig. 2. Dependence of estimates of the Chvátal-Sankoff constant on the number t_1 of recursion steps performed before the averaging starts. (a) shows estimates of $\widehat{a}_2(W)$ for completely uncorrelated disorder, *i.e.*, for the first-passage percolation problem. They do coincide within their statistical fluctuations for all t_1. The dashed line shows the average over all points but the first. (b) shows estimates of $a_2(W)$ for the full LCS problem including the disorder correlations. They have a larger statistical fluctuation than in the first passage percolation case and the first point deviates significantly from the others. Both effects stem from the much slower approach to the stationary state in the presence of the correlations. However, there is no apparent dependence on t_1 for $t_1 \geq 10^6$ in this case either. The dashed line shows the average over the last eight data points.
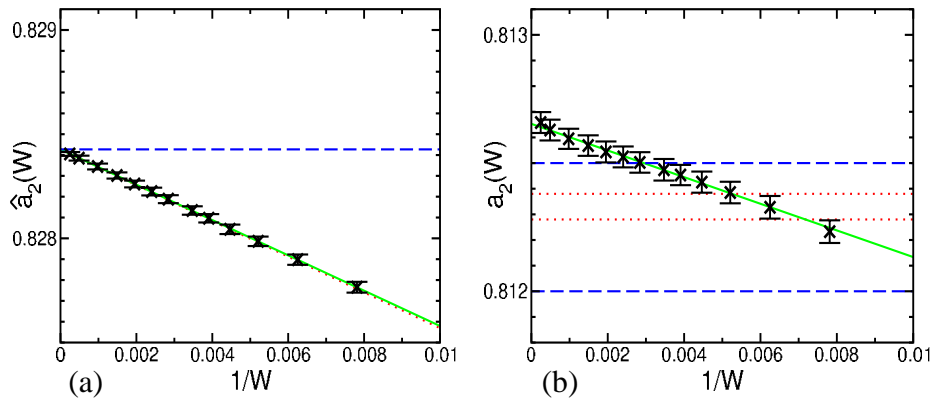

Fig. 3. Numerical estimates of the Chvátal-Sankoff constant for a two letter alphabet as a function of $1/W$. The solid lines are in both cases best fits to a linear function. (a) shows values $\widehat{a}_2(W)$ for completely uncorrelated disorder, *i.e.*, for the first-passage percolation problem. The dashed line indicates the exact value for infinite width, while the dotted line includes the predicted finite size correction. It is nearly indistinguishable from the best fit to a linear function (solid line). (b) shows the values $a_2(W)$ for the true longest common subsequence problem including the disorder correlations. Here, the dashed and dotted lines show the upper and lower bounds for the infinite system given by Dančík [11] and Boutet de Monvel [12,13] respectively.

but comparable to 10^6. This slow decorrelation property also renders the average over many recursion steps less effective yielding the larger statistical fluctuations which we observe. These statistical fluctuations are of the order of $\pm 3.5 \times 10^{-6}$. Since the systematic deviation of the estimate for $t_1 = 10^6$ is already smaller than that and since the characteristics number of steps t_0 is smaller than 10^6, we can be sure that the systematic deviations of all points with $t_1 \geq 2 \times 10^6$ are less than 10^{-6} due to the exponential decay of these deviations. Then, averaging in equation (18) over all $t \in]2 \times 10^6, 10^7]$ yields estimates of $a_c(W)$ with a statistical and systematic error of $\pm 10^{-6}$.

The results of this procedure for all finite system sizes W are shown in Figure 3b. In addition to the larger statis-

tical error bars of the individual points slight corrections to the $1/W$ behavior are visible for $W < 160$. Nevertheless, we get a good fit to a straight line in the regime $W \geq 160$ with the result

$$a_2(W) \approx (0.812\,653 \pm 0.000\,004)$$
$$- (0.052\,0 \pm 0.001\,2)\frac{1}{W} \cdot \quad (32)$$

Varying the fitting range towards larger W changes the infinite W value of the measured Chvátal-Sankoff constant only within the quoted error bars. This result is significantly larger than the results of Dančík [11] and of Boutet de Monvel [12,13] obtained for 4 lattices of size

Table 1. Numerical values for the Chvátal-Sankoff constant a_c and its finite size correction b_c for different alphabet sizes c. The errors in parentheses are statistical errors on the last digit. For comparison the counterpart $\widehat{a}_c = 2/(\sqrt{c}+1)$ of the Chvátal-Sankoff constant for first-passage percolation is also given.

c	a_c	b_c	$\widehat{a}_c$
2	0.812 653(4)	0.052(1)	0.828 427
4	0.654 361(2)	0.122(1)	0.666 667
8	0.515 143(4)	0.197(2)	0.522 408
16	0.396 316(2)	0.268(1)	0.4

up to $10^6 \times 10^6$ and 10 000 lattices of size $10^4 \times 10^4$ respectively. As shown in Figure 3b the value of the finite width Chvátal-Sankoff constant already significantly exceeds their estimates of the infinite system size value at moderate $W \approx 256$.

In the same way we get high precision values for the Chvátal-Sankoff constant for larger alphabet sizes c. However, the finite size corrections for larger alphabet size are larger and fits to a linear law only settle to within the statistical error estimate for $W \geq 352$. The fitted values are summarized in Table 1. While our method produces estimates with a much higher precision than the traditional method due to the fact that we know the form of the finite size dependence, all our estimates for $c \geq 4$ are consistent with Boutet de Monvel's within his variations between different assumed finite size corrections [12]. We observe that the differences between the Chvátal-Sankoff constant a_c and its counterpart $\widehat{a}_c$ for the uncorrelated disorder are significant but become smaller with increasing alphabet size. This is to be expected intuitively since the correlations between the $\eta(r,t)$ become weaker with increasing alphabet size (if we know the letter in one sequence and the information that we have a mismatch this fixes the letter in the other sequence only up to $c-1$ possible choices). This dependence on the alphabet size is also consistent with the fact that it apparently becomes more difficult to take the correlations into account numerically and to get good numerical estimates of the a_c as the alphabet size decreases. This materializes in the discrepancies between our high precision results for a_2 and earlier estimates of the Chvátal-Sankoff constant which are not present for the larger alphabet sizes.

6 Conclusions

We conclude that in order to get numerical values of the Chvátal-Sankoff constant with a high precision a proper handling of the finite size effects is absolutely necessary. We showed how concepts from statistical physics can help to tackle this problem. We moreover presented a highly effective multi-spin coding scheme which further increases the precision of the simulations. The numerical results

given will be valuable in guiding and evaluating theoretical approaches to understand the small but noticeable effect of the subtle disorder correlations which are responsible for the difference between the Chvátal-Sankoff constant and its counterpart in the first-passage percolation problem.

Appendix A: Change of lattice

In this appendix we will argue that the limiting quantity

$$a_c(\infty) = \lim_{W \to \infty} a_c(W) = \lim_{W \to \infty} \lim_{t \to \infty} \frac{1}{t} \langle \ell(W-1, 2t) \rangle \tag{33}$$

defined on the rectangular lattice shown in Figure 1b equals the Chvátal-Sankoff constant

$$a_c \equiv \lim_{N \to \infty} \frac{\langle L(N) \rangle}{N} \tag{34}$$

defined on the diamond shaped lattice shown in Figure 1a. To this end, we first consider a path from the left edge of the rectangular lattice to the point $(W-1, 2t)$ as indicated in Figure 4a which is the optimal path in each of the diamonds. Clearly, $\ell(W-1, 2t)$ can only be larger than the number of matches along this specific path which implies

$$\langle \ell(W-1, 2t) \rangle \geq \frac{t}{W} \langle L(W) \rangle \tag{35}$$

or after dividing both sides by t and taking the limit of $t \to \infty$

$$a_c(W) \geq \frac{\langle L(W) \rangle}{W}. \tag{36}$$

Taking the limit $W \to \infty$ of this inequality yields

$$a_c(\infty) \geq a_c. \tag{37}$$

On the other hand, let us consider the best path from the left edge of the rectangular lattice to the point $(W-1, 2t)$, *i.e.*, the path with $\ell(W-1, 2t)$ matches. As shown in Figure 4b, we can construct a path which connects the two end points in a diamond shaped lattice of size $(t+W) \times (t+W)$ from this path. Since the length of the longest common subsequence of the two associated sequences of length $t+W$ can only be longer than the common subsequence corresponding to this path we get

$$\langle L(t+W) \rangle \geq \langle \ell(W-1, 2t) \rangle - W \langle n_{\mathrm{wrap}}(t) \rangle, \tag{38}$$
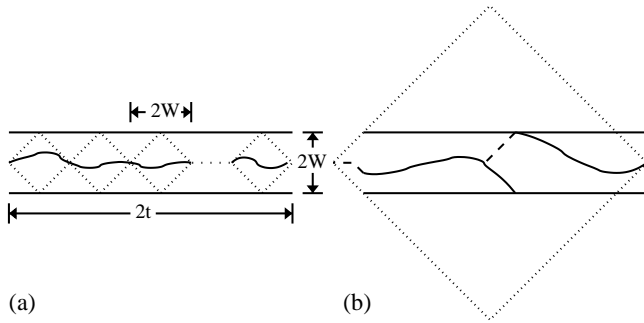
Fig. 4. Relation between optimal paths on the rectangular and the diamond shaped lattice. (a) shows how a good path for the rectangular lattice can be constructed by stringing together optimal paths on diamond shaped lattices of size $W \times W$. In (b) a good path for a diamond shaped lattice of size $(t+W) \times (t+W)$ is constructed from the optimal path on the rectangular lattice. The left end of the path is connected to the tip of the diamond which can only increase the number of matches. For each time the optimal path wraps around the periodic boundary conditions the path has to be modified as shown by the diagonal dashed line. For each such event, at most W matches along the solid path can be missed.

where $n_{\mathrm{wrap}}(t)$ is the number of times that the best path from the left edge to $(W-1, 2t)$ "wraps around" the periodic boundary conditions. Dividing by t yields

$$\frac{t+W}{t} \frac{\langle L(t+W) \rangle}{t+W} \geq \frac{\langle \ell(W-1, 2t) \rangle}{t} - W \frac{\langle n_{\mathrm{wrap}}(t) \rangle}{t} \tag{39}$$

which in the limit $t \to \infty$ leads to

$$a_c \geq a_c(W) - W \lim_{t \to \infty} \frac{\langle n_{\mathrm{wrap}}(t) \rangle}{t} . \tag{40}$$

From the analogy with the directed polymer in a random medium it is well known that the displacement of a typical optimal path of length t scales like $t^{2/3}$ [15, 20, 26, 27]. Thus, we expect one wrapping around every $W^{3/2}$ steps along the lattice, *i.e.*,

$$\langle n_{\mathrm{wrap}}(t) \rangle \sim W^{-3/2} t. \tag{41}$$

Therefore, if we take the limit $W \to \infty$ in equation (40) the second term on the right hand side vanishes and we are left with

$$a_c \geq a_c(\infty). \tag{42}$$

Together with equation (37) this establishes the equality between the infinite width limit of the $a_c(W)$ and the Chvátal-Sankoff constant a_c.

References

1. In the context of sequence comparison, see, *e.g.*, S. Karlin, S. Altschul, Proc. Natl. Acad. Sci. USA **87**, 2264 (1990).
2. M.S. Waterman, *Introduction to Computational Biology* (Chapman & Hall, London, UK, 1994).
3. V. Chvátal, D. Sankoff, J. Appl. Prob. **12**, 306 (1975).
4. D.S. Hirschberg, Inf. Proc. Lett. **7**, 40 (1978).
5. J. Deken, Disc. Math. **26**, 17 (1979).
6. J. Steele, SIAM J. Appl. Math. **42**, 731 (1982).
7. V. Chvátal, D. Sankoff, in *Time Warps, String Edits, and Macromolecules: The theory and practice of sequence comparison*, edited by D. Sankoff, J.B. Kruskal (Addison-Wesley, Reading, Mass., 1983), p. 353.
8. J.G. Deken, in *Time Warps, String Edits, and Macromolecules: The theory and practice of sequence comparison*, edited by D. Sankoff, J.B. Kruskal (Addison-Wesley, Reading, Mass., 1983), p. 359.
9. A. Apostolico, C. Guerra, Algorithmica **2**, 315 (1987).
10. V. Dančík, M. Paterson, in *STACS94, Lecture Notes in Computer Science*, **775** (Springer, Berlin, 1994), p. 669.
11. V. Dančík, Ph.D. thesis, University of Warwick, 1994.
12. J. Boutet de Monvel, Eur. Phys. J. B **7**, 293 (1999).
13. J. Boutet de Monvel, Phys. Rev. E **62**, 204 (2000).
14. J. Steele, *Probability Theory and Combinatorial Optimization* (SIAM, Philadelphia, Penn., 1997).
15. T. Hwa, M. Lässig, Phys. Rev. Lett. **76**, 2591 (1996).
16. R. Bundschuh, T. Hwa, Disc. Appl. Math. **104** 113 (2000).
17. M. Kardar, G. Parisi, Y. Zhang, Phys. Rev. Lett. **56**, 889 (1986).
18. S.B. Needleman, C.D. Wunsch, J. Mol. Biol. **48**, 443 (1970).
19. K.S. Alexander, Ann. Appl. Probab. **4**, 1074 (1994).
20. for a review, see J. Krug, H. Spohn, in *Solids far from Equilibrium: Growth, Morphology, and Defects*, edited by C. Godreche (Cambridge University Press, Cambridge, UK, 1991), p. 479.
21. J. Krug, P. Meakin, J. Phys. A **32**, L987 (1990).
22. R. Bundschuh, submitted to Phys. Rev. E, `cond-mat/9911386`.
23. J. Krug, Phys. Rev. Lett. **67**, 1882 (1991).
24. B. Derrida, Phys. Rep. **301**, 65 (1998); and references therein.
25. see, *e.g.*, R.B. Bapat, T.E.S. Raghavan, *Nonnegative Matrices and Applications* (Cambridge University Press, Cambridge, UK, 1997).
26. D. Drasdo, T. Hwa, M. Lässig, Mat. Res. Soc. Symp. Proc. **263**, 75 (1997).
27. D. Drasdo, T. Hwa, M. Lässig, in *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, edited by J. Glasgow *et al.* (AAAI Press, Menlo Park, Calif., 1998), p. 52.