An Analytic Study of the Phase Transition Line in Local Sequence Alignment with Gaps

R. Bundschuh and T. Hwa

Department of Physics
University of California at San Diego
La Jolla, CA 92093-0319 U.S.A.
E-mail: rbund@ucsd.edu, hwa@ucsd.edu

August 23, 1999

Abstract

A detailed analytic study of the log-linear phase transition of the Smith-Waterman local alignment algorithm is presented. A rectangular alignment lattice is introduced to facilitate the statistical analysis for alignment with gaps. With a few simplifying assumptions, we obtain an analytic expression for the loci of the phase transition line. Our result reproduces the exact and conjectured values for the very large and very small gap costs; the latter corresponds to the related problem of the longest common subsequence. For intermediate values of gap costs, our result is not exact, although a comparison to numerical results yielded a difference of no more than several percent.

Keywords: sequence alignment, phase transition, first-passage percolation, longest common subsequence

1 Introduction

Sequence alignment is one of the most widely used computational tools in molecular biology. It has made a strong impact on the functional identification of newly sequenced genes and on the reconstruction of phylogenic trees (Doolittle, 1996). There are two types of sequence alignment algorithms used. Gapless alignment such as BLAST (Altschul *et al.*, 1990) or FASTA (Lipman and Pearson, 1985; Pearson and Lipman, 1988) is well understood (Arratia *et al.*, 1988; Karlin and Altschul, 1990, 1993; Dembo and Karlin, 1991) and is extensively used as a "first cut" in large scale database searches. However due to the occurrence of insertions and deletions in the evolution of biological sequences, *weak* sequence homologies can only be detected by alignment algorithms which allow for gaps, e.g., the Needleman-Wunsch (Needleman and Wunsch, 1970) or the Smith-Waterman (Smith and Waterman, 1981; Waterman, 1989, 1994a) algorithm. Unlike gapless alignment however, statistical properties of alignment with gaps are more complicated and little understood.

Recently, properties of alignment with gaps have been examined using concepts developed in statistical physics (Hwa and Lässig, 1996, 1998; Drasdo *et al.*, 1997, 1998). Various scaling laws, i.e. power laws, have been identified for global alignment of random sequences. These laws have been exploited to deduce properties of local alignment of random sequences in the vicinity of the transition line between the two modes of local alignment, the so-called linear and the logarithmic phases. They have also been applied to understand the global and local alignment of *mutually correlated* sequences quantitatively. The knowledge gained has then been used to guide the optimal detection of unknown sequence homologies (Drasdo *et al.*, 1998; Olsen *et al.*, 1998).

This series of statistical studies relies on the correspondence between the sequence alignment problem and certain well-studied problems of statistical physics[1]. They focus on the so-called "universal" properties (e.g. scaling laws) which are laws of large numbers and hence do not depend on many details of the problem. On the other hand, certain important characteristics, e.g., the precise location of the phase transition line, do depend on the specifics of the algorithm, and are in general more difficult to characterize. The location of the phase transition line is nevertheless crucial to biological applications, since optimal detection of weak homologies is claimed to occur close to this phase transition line (Drasdo *et al.*, 1998; Olsen *et al.*, 1998).

In this publication, we present an approximation scheme to calculate the position of the log-linear phase transition line for gapped local alignment of random sequences. We focus on the simplest version of the Smith-Waterman algorithm with linear gap cost. Our calculation reproduces respectively the exact and conjectured result for very large and small gap costs; the latter yields directly (Waterman *et al.*, 1987) the important Chvátal-Sankoff constant of the longest-common-subsequence problem (Chvátal and Sankoff, 1975). Our results are not exact for intermediate values of gap costs; however, a comparison with numerical estimates for random nucleotide sequences indicates that our approximate result is off by only several percent in the worst case. Therefore, our approximation scheme is a reasonable starting point for more refined calculations.

This paper is outlined as follows: In Section 2, we review the alignment algorithm with gaps and discuss its properties relevant to our study. New boundary conditions are introduced to simplify the calculations without affecting the position of the phase transition line. Section 3 serves mainly pedagogical purposes. It discusses alignment which allows only gaps alternating between the two sequences. This problem also has a log-linear phase transition and can be solved exactly. By describing its solution, we develop the necessary notation and techniques which are applied to the full alignment problem in Section 4. There, we first describe in detail the assumptions introduced to handle the full alignment problem. Next we show that one of the assumptions is actually exact for special parameter values which correspond to the longest common subsequence problem. Then, we use the scheme developed in Section 3 to calculate the phase transition line of the full alignment problem. Finally, we compare our results with numerical estimates and discuss their implications for the longest common subsequence problem. Some detailed calculations are relegated to the Appendices: In Appendix A, we describe the full calculation of the exact phase transition line for the case considered in Section 3. In Appendix B, we examine the effect of the approximation scheme used in Section 4, and indicate ways of improving the calculation.

2 Review of Alignment with Gaps

We study the simplest version of the Smith-Waterman local alignment algorithm with a linear gap cost. Specifically, we consider the ensemble of pairwise local alignments of sequences, each consisting of N elements[2] drawn randomly from an alphabet of c letters (e.g. $c = 4$ for nucleotides.) A score $S[\mathcal{A}]$ is associated with each alignment $\mathcal{A}$ of two sequences by adding up the local score of $+1$ for each match, $-\mu$ for each mismatch, and $-\delta \cdot \ell$ for each contiguous gap of length ℓ. For an alignment with $N_{\rm m}$ matches, $N_{\rm mm}$ mismatches and $N_{\rm g}$ gaps[3] then, the score is

$$S[\mathcal{A}] = N_{\rm m} - \mu \cdot N_{\rm mm} - \delta \cdot N_{\rm g}. \tag{1}$$

Of interest here are the statistical properties of the optimal alignment, i.e., the alignment with the highest score, $\Sigma_N(\mu, \delta) \equiv \max_{\mathcal{A}} S[\mathcal{A}]$.

2.1 The log-linear phase transition

It is well known from the study by Waterman *et al.* (1987) and Arratia and Waterman (1994) that gapped local alignment of random sequences exhibits a log-linear phase transition along a critical

[1] A review of some of the related physics problems can be found in Krug and Spohn (1991).

[2] Our results are applicable (Bundschuh and Hwa, to be published) to the alignment of sequences of different lengths N and M, as long as $|N - M| \ll (N + M)^{2/3}$.

[3] $N_{\rm g}$ denotes here the number of single gaps, i.e. a contiguous gap of length ℓ is counted as ℓ gaps.

line of parameters $\mu_c(\delta)$ in the parameter space (μ, δ). For $\mu < \mu_c(\delta)$, $\langle \Sigma_N \rangle$ depends linearly on N while for $\mu > \mu_c(\delta)$, $\langle \Sigma_N \rangle$ depends logarithmically on N. The angular brackets denote here the average over the whole ensemble of randomly drawn sequences.

In the gapless limit ($\delta \to \infty$), the statistics of Σ is known exactly: The phase transition occurs at $\mu_c(N) = 1/(c-1)$ and Σ is distributed according to a Gumbel distribution in the logarithmic phase (Karlin and Dembo, 1992). However, once gaps are allowed, even the position of the critical line $\mu_c(\delta)$ is known only empirically not to mention the distribution of Σ.

In the following sections, we present a scheme to calculate the critical line $\mu_c(\delta)$ for the range $\delta_0 < \delta < \infty$, where δ_0 is defined by the intersection of $\mu_c(\delta)$ with the special line[4] $\mu = 2\delta$. Our result recovers the exact result $\mu_c(\infty) = 1/(c-1)$ for gapless alignment and the conjectured result $\mu_c(\delta_0) = 2/(\sqrt{c}-1)$ (Arratia, unpublished; see also Steele (1986)) along the line $\mu = 2\delta$.

Our approach is based on the observation of Arratia and Waterman (1994) that the position of the critical line corresponds to the parameter values $(\mu_c(\delta), \delta)$ for which the expected score of *global* alignment vanishes, i.e., $\langle \Sigma_N^G(\mu_c(\delta), \delta) \rangle = 0$. Since this expected score is known to have an asymptotic linear dependence on N for all parameters (Arratia and Waterman, 1994), our task amounts to finding the zero of the proportionality factor

$$a(\mu, \delta) \equiv \lim_{N \to \infty} \frac{1}{N} \langle \Sigma_N^G(\mu, \delta) \rangle \tag{2}$$

for global alignment. We shall thus focus on global alignment from here on.

Before proceeding, we recall that the $(\mu = 0, \delta = 0)$ limit of global alignment corresponds to the well-known problem of the longest common subsequence (LCS) introduced by Chvátal and Sankoff (1975). Here Σ_N^G is is the length of the LCS which is proportional to N as in (2). The proportionality constant a_0 is known as the Chvátal-Sankoff constant. There have been many extensive numerical and analytical studies of the LCS problem (Deken, 1979; Dančík and Paterson, 1994; Paterson and Dančík, 1994; Waterman, 1994b). The numerical value of the Chvátal-Sankoff constant is given approximately by the expression

$$a_0 = \frac{2}{\sqrt{c}+1} \tag{3}$$

conjectured first by R. Arratia (unpublished; see also Steele (1986)). Through a simple relation between a_0 and $\mu_c(\delta_0)$ pointed out first by Waterman *et al.* (1987), our work gives an indirect calculation of the Chvátal-Sankoff constant.

2.2 Alignment paths and the dynamic programming algorithm

It will be convenient to adopt the directed path representation for global alignment. An example of such an alignment is shown in Fig. 1 for a specific pair of sequences. In this figure we rotated the alignment lattice by 45 degrees with respect to its commonly used representation in which the path is directed down the diagonal of a square lattice. This is meant to stress the fact that alignment can be seen as a *dynamical process* which progresses from left to right in the way we have drawn the lattice. The description of alignment in terms of this dynamical process will turn out to be crucial for our calculations. Consistently, we identify lattice nodes by coordinates r and z (see Fig. 1) instead of the more commonly used letter index. For this way of drawing the alignment lattice, all the diagonal bonds correspond to gaps. So the score of an alignment path gets a contribution $-\delta$ for *each* diagonal bond along the (high-lighted) alignment path. The horizontal bonds of the lattice correspond to pairings of one letter from the first sequence with one letter of the second sequence. Such a bond contributes a score $s(r,z) \in \{1, -\mu\}$ depending on whether the specific letters paired up at a given bond (r,z) match or not. The task of global alignment is then to find the highest scoring path connecting the left and right extremities at $(r,z) = (0,0)$ and $(r,z) = (0,2N)$ for a given set of $\{s(r,z)\}$. Let $S(r,z)$ be the maximal score of any alignment path ending at the lattice

[4] For $\mu > 2\delta$, it is always advantageous to use two gaps to circumvent a mismatch. The alignment therefore becomes independent of μ and is completely characterized by the alignment at $\mu = 2\delta$.

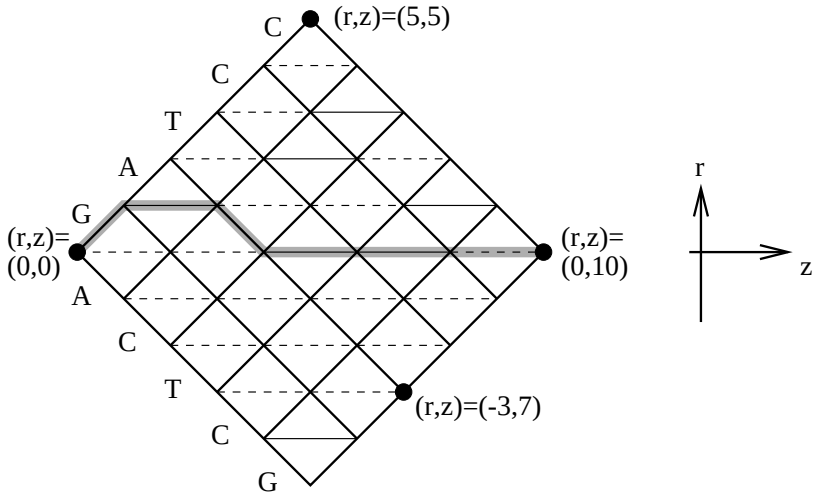

Figure 1: Global alignment of two sequences $GATCC$ and $ACTCG$ represented as a directed path on the alignment lattice: the diagonal bonds correspond to gaps in the alignment. The horizontal bonds are matches (solid lines) and mismatches (dashed lines). The highlighted alignment path $r(z)$ therefore corresponds to one possible alignment, $GA-TCC$ to $-ACTCG$. This path contains two gaps, three matches and one mismatch. It is also shown how the coordinates r and z are used to identify the nodes of the lattice.

point (r, z) and starting from $(0, 0)$. This quantity can be calculated recursively by the dynamic programming algorithm

$$S(r, z+1) = \max \left\{ \begin{array}{l} S(r+1, z) - \delta \\ S(r-1, z) - \delta \\ S(r, z-1) + s(r, z) \end{array} \right\}, \tag{4}$$

supplemented by the global conditions that $S(0, 0) = 0$ and $S(r, z) = -\infty$ beyond the boundaries of the diamond-shaped lattice. The score of the optimal alignment path is then $\Sigma_N^G = S(r = 0, z = 2N)$.

2.3 A simplified scoring scheme

While the scoring scheme (1) is convenient for the numerical implementation in terms of the simple form of the recursion relation (4), it leads to cumbersome notations in the analytic treatment to be described below. In what follows, we shall use instead a transformed set of parameters $\widetilde{\mu}(\mu, \delta)$ and $\widetilde{\delta}(\mu, \delta)$, chosen such that

$$S[\mathcal{A}] = \widetilde{S}_0 \left[N_{\rm m} - \widetilde{\mu} \left(N_{\rm m} + N_{\rm mm} + \frac{1}{2} N_{\rm g} \right) - \widetilde{\delta} \cdot N_{\rm g} \right]. \tag{5}$$

The scoring scheme (5) is motivated by the observation that for global alignment, $N_{\rm m} + N_{\rm mm} + \frac{1}{2} N_{\rm g} = N$ is *independent* of the alignment $\mathcal{A}$. Hence, the second term in $[\ldots]$ of Eq. (5) is merely an additive constant which does not affect the alignment, and global alignment depends on one *single* parameter, $\widetilde{\delta}$. [The prefactor $\widetilde{S}_0 = 1 + \mu$ fixes the overall scale of S and is therefore completely arbitrary.] Equating Eqs. (1) and (5), we have

$$\widetilde{\mu} = \frac{\mu}{1+\mu} \quad \text{and} \quad \widetilde{\delta} = \frac{\delta - \mu/2}{1+\mu}. \tag{6}$$

With the modified scoring scheme (5), the recursion relation (4) should be replaced by

$$h(r,z+1) = \max \left\{ \begin{array}{l} h(r+1,z) - \widetilde{\delta} \\ h(r-1,z) - \widetilde{\delta} \\ h(r,z-1) + \eta(r,z) \end{array} \right\}, \tag{7}$$

with $\eta(r,z) \in \{0,1\}$, and the boundary conditions $h(0,0) = 0$ and $h(r,z) = -\infty$ beyond the boundaries of the diamond lattice. In terms of h, the optimal score is $\langle \Sigma_N^G \rangle = \widetilde{S}_0 \cdot [-\widetilde{\mu} N + \langle h(0,2N) \rangle]$. Let $v_0(\widetilde{\delta}) = \lim_{N\to\infty} \frac{1}{N} \langle h(r=0, z=2N) \rangle$, then the proportionality factor $a(\mu,\delta)$ defined in (2) becomes

$$a(\mu,\delta) = \widetilde{S}_0 \cdot [-\widetilde{\mu} + v_0(\widetilde{\delta})], \tag{8}$$

and the condition $a(\mu,\delta) = 0$ for the occurrence of the log-linear phase transition in local alignment is simply

$$\widetilde{\mu}_c = v_0(\widetilde{\delta}). \tag{9}$$

The position of the critical line $\mu_c(\delta)$ in the original parameter space is then obtained from (9) by inverting the relations in Eq. (6).

Note that the LCS problem ($\mu = 0, \delta = 0$) corresponds to $\widetilde{\mu} = 0$, $\widetilde{\delta} = 0$, and $\widetilde{S}_0 = 1$. Thus the Chvátal-Sankoff constant $a_0 = a(0,0)$ is given by

$$a_0 = v_0(0). \tag{10}$$

The knowledge of a_0 then also gives the terminal point of the critical line $\mu_c(\delta)$

$$\mu_c(\delta_0) = 2\delta_0 = \frac{a_0}{1-a_0}, \tag{11}$$

as pointed out by Waterman *et al.* (1987).

2.4 Alignment geometry

Instead of computing $\langle h(0,z) \rangle$, which gives the average score of the optimal path connecting $(0,0)$ and $(0,z)$, we shall compute the average score $\langle \max_r \{h(r,z)\} \rangle$ of the optimal path connecting $(0,0)$ and *any* point at the distance z (see Fig. 2). This amounts to changing the "boundary condition"

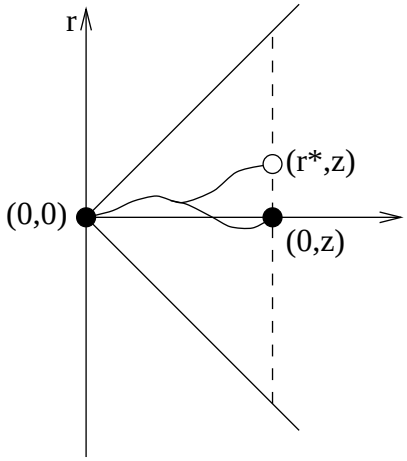

Figure 2: Optimal paths connecting $(0,0)$ with the *fixed* end point $(0,z)$ or the *free* end point (r^*,z) respectively. r^* is chosen such that the score of the path is optimal among all paths from $(0,0)$ ending at (r,z) for any r. The solid circles represent the fixed ends and the open circle represents the free end.

on the terminal of the path $r(z)$ from fixed end to free end and does not change the asymptotics of the average score in the limit of infinitely long sequences, i.e.,

$$v(\widetilde{\delta}) \equiv \lim_{z\to\infty} \frac{1}{z}\langle \max_r\{h(r,2z)\}\rangle = v_0(\widetilde{\delta}) \tag{12}$$

as shown by Alexander (1994).

The quantity $v(\widetilde{\delta})$ can be conveniently calculated in a *rectangular* geometry consisting of L rows of lattice points; see Fig. 3. If we apply the recursion relation (7) on the rectangular lattice, with

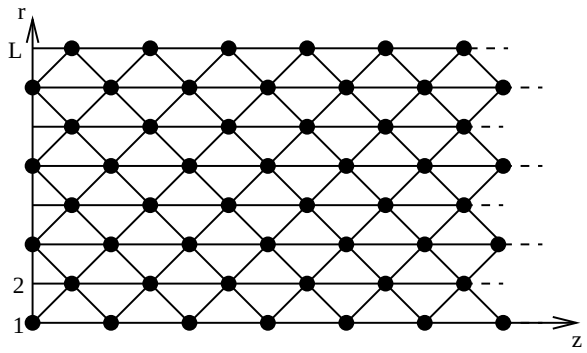

Figure 3: Rectangular alignment lattice of L rows. The simpler shape of this lattice (compared to the one shown in Fig. 1) will facilitate the analytical calculations to be presented below.

the "initial conditions" $h(r,0) = 0$ and $h(r,-1) = -\infty$, the score $h(r,z)$ obtained simply represents the score of the optimal path connecting the point (r,z) to any point a distance z away to the left (see Fig. 4.) In the limit $L \gg z$, the optimal path in Fig. 4 becomes the same as the optimal path

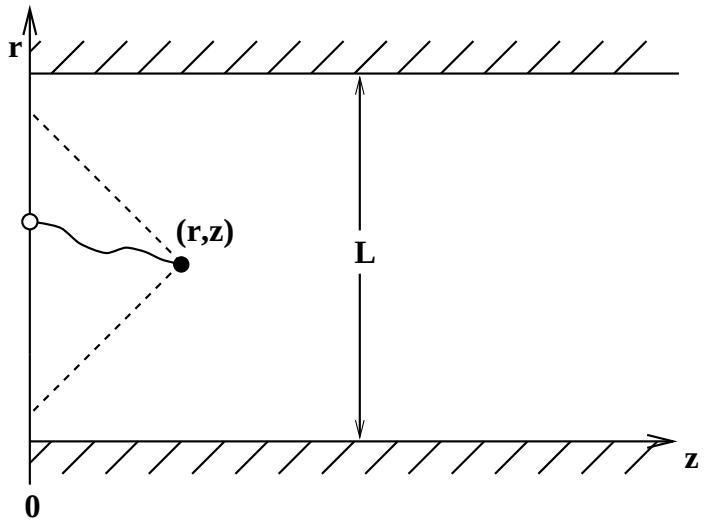

Figure 4: The optimal path leading to (r,z) from any point $(r',0)$ in the rectangular geometry is the mirror image of the optimal path with a free end point (see Fig. 2).

with a free end depicted in Fig. 2, provided we *reverse* the order of both sequences being aligned. Since reversing the direction of random sequences does not change ensemble averaged quantities, we have

$$v(\widetilde{\delta}) = \lim_{z\to\infty}\lim_{L\to\infty} \frac{1}{z}\langle h(r,2z)\rangle_L, \tag{13}$$

where $\langle \ldots \rangle_L$ indicates ensemble average over the rectangular geometry of width L. The rectangular geometry (with $L \to \infty$) offers a *statistical translational symmetry* in r which simplifies the calculations. For example, $\langle h(r,z) \rangle_\infty$ is independent of r. The quantity $v(\widetilde{\delta})$ can now be interpreted also as the "drift rate" of the *spatial average* of the score profile $h(r,z)$. We shall present a scheme to compute the drift rate $v(\widetilde{\delta})$ for the $L = \infty$ system in Section 4. Before doing so, we shall first calculate exactly the drift rate for the much simpler $L = 2$ system in Section 3. This will set up the notation and the main procedures used for the $L = \infty$ system.

3 Alignment on a Single Strip

For pedagogical purposes, we study in this section alignment on a single strip (corresponding to $L = 2$) as shown in Fig. 5. Let the score on the lower and upper nodes be $h(0,t)$ and $h(1,t)$ respectively. Note that due to the symmetry of this problem, it is more convenient to combine the

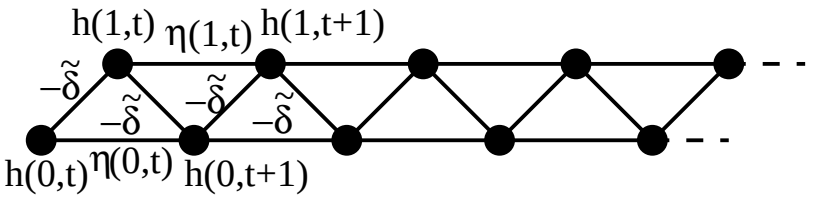

Figure 5: Alignment in a strip geometry. For each time t, there are two alignment scores $h(0,t)$ and $h(1,t)$ which are associated with the nodes of the lattice. They depend on the values on the bonds of the lattice. The values $\eta(r,t)$ of the horizontal bonds are random variables given according to (15); the diagonal bonds always contribute a score of $-\widetilde{\delta}$ each.

scores at two adjacent values of z in (7) into one "time step" t. (This leads to a correspondence $z = 2t$ between the general formulation of the dynamic programming algorithm as given in Section 2 and the notion of "time" used here.)

Applying the algorithm (7) to this strip geometry as described in Section 2.4, with the boundary conditions $h(r,t) = -\infty$ outside the strip, we obtain the following recursion relation for the single strip,

$$\begin{aligned} h(0,t+1) &= \max\{h(0,t) + \eta(0,t), h(1,t) - \widetilde{\delta}\} \\ h(1,t+1) &= \max\{h(1,t) + \eta(1,t), h(0,t+1) - \widetilde{\delta}\} \\ &= \max\{h(1,t) + \eta(1,t), h(0,t) + \eta(0,t) - \widetilde{\delta}\}, \end{aligned} \tag{14}$$

for the regime $\widetilde{\delta} \geq 0$ where a mismatch cannot be circumvented by two gaps. For the single strip system, the values of the $\eta(r,t)$ become uncorrelated random variables, i.e.,

$$\eta(r,t) = \begin{cases} 1 & \text{with probability } \frac{1}{c} \\ 0 & \text{with probability } 1 - \frac{1}{c} \end{cases} . \tag{15}$$

It will be convenient to introduce the quantities

$$\widetilde{h}(t) \equiv h(0,t) - h(1,t) \qquad \text{and} \qquad \bar{h}(t) \equiv \frac{1}{2}[h(0,t) + h(1,t)]. \tag{16}$$

The evolution equations (14) expressed in these quantities become

$$\widetilde{h}(t+1) = \widetilde{h}(t) + [\eta(0,t) - \eta(1,t)] + [m(0,t) - m(1,t)] \quad \text{and} \tag{17}$$

$$\bar{h}(t+1) = \bar{h}(t) + \frac{1}{2}[\eta(0,t) + \eta(1,t)] + \frac{1}{2}[m(0,t) + m(1,t)] \tag{18}$$

with

$$\begin{aligned} m(0,t) &\equiv \max\{0, -\widetilde{h}(t) - \eta(0,t) - \widetilde{\delta}\} \\ \text{and} \qquad m(1,t) &\equiv \max\{0, \widetilde{h}(t) + \eta(0,t) - \eta(1,t) - \widetilde{\delta}\}. \end{aligned} \tag{19}$$

A remarkable property of these equations is that the equation for $\bar{h}$ is *slaved* to $\widetilde{h}$. This allows us to study the evolution of $\widetilde{h}$ first. We will show that the distribution of $\widetilde{h}(t)$ becomes t-independent for large t. Since $\widetilde{h}(t)$ does not depend on $\eta(r,t)$ at the same t, we can calculate the drift

$$v^{(2)}(\widetilde{\delta}) \equiv \langle \bar{h}(t+1) - \bar{h}(t) \rangle_{L=2} \tag{20}$$

from Eqs. (18) and (19) if we know the distribution of $\widetilde{h}(t)$. This will then give us the phase transition line $\widetilde{\mu}_{\rm c}^{(2)} = v^{(2)}(\widetilde{\delta})$ for the $L=2$ system.

An important observation about the evolution equation (17) is that $\widetilde{\eta}(t) \equiv \eta(0,t) - \eta(1,t)$ is *symmetrically* distributed with

$$\widetilde{\eta}(t) = \begin{cases} -1 & \text{with prob. } \frac{1}{c}\left(1 - \frac{1}{c}\right) \\ 0 & \text{with prob. } 1 - \frac{2}{c} + \frac{2}{c^2} \\ 1 & \text{with prob. } \frac{1}{c}\left(1 - \frac{1}{c}\right) \end{cases} . \tag{21}$$

For $-\widetilde{\delta} \le \widetilde{h}(t) \le \widetilde{\delta} - 1$, there is no contribution from either $m(0,t)$ or $m(1,t)$ in Eq. (17). Consequently, $\widetilde{h}$ performs a *symmetric random walk* in this interval, with "step size" 1 and a "step probability" of $\frac{1}{c}(1 - \frac{1}{c})$ in either direction.

It is easy to verify that $\widetilde{h}(t+1) = \widetilde{\delta}$ if $\widetilde{h}(t) \ge \widetilde{\delta} + 1$, while for $\widetilde{h}(t) \le -\widetilde{\delta} - 1$, $\widetilde{h}(t+1) = -\widetilde{\delta} - \eta(1,t) \in \{-\widetilde{\delta} - 1, -\widetilde{\delta}\}$. More careful inspection shows $-\widetilde{\delta} - 1 \le \widetilde{h}(t+1) \le \widetilde{\delta}$ for any $\widetilde{h}(t)$. So we can conclude that after some finite "equilibration period" during which $\widetilde{h}(t)$ performs a random walk starting from an arbitrary initial value, it will reach or exceed the boundaries $\widetilde{\delta}$ or $-\widetilde{\delta} - 1$, and consequently be confined thereafter to the values $\{\widetilde{\delta}, \widetilde{\delta} - 1, \widetilde{\delta} - 2, \ldots\}$ and $\{-\widetilde{\delta} - 1, -\widetilde{\delta}, -\widetilde{\delta} + 1 \ldots\}$, between $-\widetilde{\delta} - 1$ and $\widetilde{\delta}$.

We now focus on the case where $2\widetilde{\delta}$ is an integer number. The solution for the more cumbersome non-integer case is provided in Appendix A. The deviations from the integer results turn out to be small for small $\widetilde{\delta}$ and completely negligible for large $\widetilde{\delta}$, as will be shown in Fig. 7. [Note that the value $\widetilde{\delta} = 0$ on the special line belongs to the integer case.] For $2\widetilde{\delta} \in \mathbf{N}$, $\widetilde{h}(t)$ can only take on $2\widetilde{\delta} + 2$ discrete values after equilibration, i.e.,

$$\widetilde{h}(t \gg 1) \in \{-\widetilde{\delta} - 1, -\widetilde{\delta}, -\widetilde{\delta} + 1, \ldots, \widetilde{\delta} - 1, \widetilde{\delta}\}, \tag{22}$$

as illustrated in Fig. 6. Since this is a finite set, we expect there to exist a stationary distribution

$$P(\widetilde{h}) \equiv \Pr\{\widetilde{h}(t) = \widetilde{h}\}. \tag{23}$$

As a reminder of the random walk nature of the evolution of $\widetilde{h}$ we will refer to $\widetilde{h}(t) = \widetilde{h}$ by the notion that the score difference is at "position" $\widetilde{h}$. The two values $\widetilde{\delta}$ and $-\widetilde{\delta} - 1$ are called the "boundaries" and all other values in (22) are "interior positions".

As we already noted, from each interior position, $\widetilde{h}$ can jump to the neighboring position $\widetilde{h} \pm 1$ with a probability

$$w_{\pm} = \frac{1}{c}\left(1 - \frac{1}{c}\right), \tag{24}$$

or remain at position $\widetilde{h}$ with probability $1 - w_- - w_+$. At the right boundary $\widetilde{h}(t) = \widetilde{\delta}$, Eq. (17) gives

$$\widetilde{h}(t+1) = \widetilde{\delta} + [\eta(0,t) - \eta(1,t)] - \max\{0, \eta(0,t) - \eta(1,t)\}, \tag{25}$$

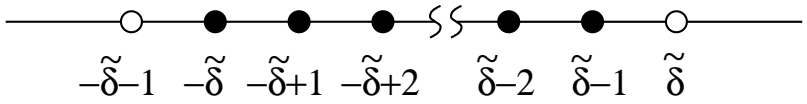

Figure 6: Possible values of the score difference $\widetilde{h}(t)$ for $2\widetilde{\delta} \in \mathbf{N}$ and $t \gg 1$. The solid circles are the "interior positions" while the open circles indicate the "boundaries". The evolution equation (17) combined with the distribution of $\widetilde{\eta}$ in (21) does not allow $\widetilde{h}$ to take on any other values as discussed in the text.

revealing that there is a probability w_- that $\widetilde{h}$ jumps back to the interior position $\widetilde{\delta}-1$ and a probability $1-w_-$ that it remains at the boundary at $\widetilde{\delta}$. These combine to yield

$$\begin{aligned} & P(\widetilde{\delta}) &=& (1-w_-)P(\widetilde{\delta}) + w_- P(\widetilde{\delta}-1) \\ \text{and} \quad & P(\widetilde{h}) &=& w_+ P(\widetilde{h}-1) + (1-w_+-w_-)P(\widetilde{h}) + w_- P(\widetilde{h}+1) \end{aligned}$$

for the interior positions $\widetilde{h} \in \{-\widetilde{\delta}+1, \ldots, \widetilde{\delta}-1\}$. Since $w_+ = w_-$, the above two equations immediately enforce the solution

$$P(-\widetilde{\delta}) = P(-\widetilde{\delta}+1) = \ldots = P(\widetilde{\delta}).$$

To find the behavior at the left boundary $-\widetilde{\delta}-1$, we insert $\widetilde{h}(t) = -\widetilde{\delta}-1$ into (17) and obtain $\widetilde{h}(t+1) = -\widetilde{\delta} - \eta(1,t)$. $\widetilde{h}$ therefore jumps with a probability $1-\frac{1}{c}$ to the interior position $-\widetilde{\delta}$ and stays with probability $\frac{1}{c}$ at the left boundary $-\widetilde{\delta}-1$. This gives us the two additional equations

$$\begin{aligned} & P(-\widetilde{\delta}-1) &=& w_- P(-\widetilde{\delta}) + \frac{1}{c} P(-\widetilde{\delta}-1) \\ \text{and} \quad & P(-\widetilde{\delta}) &=& \left(1-\frac{1}{c}\right) P(-\widetilde{\delta}-1) + (1-w_+-w_-)P(-\widetilde{\delta}) + w_- P(-\widetilde{\delta}+1). \end{aligned}$$

These equations can be solved *simultaneously* by $P(-\widetilde{\delta}-1) = \frac{1}{c}\, P(-\widetilde{\delta})$. The normalization condition $\sum_{\widetilde{h}=-\widetilde{\delta}-1}^{\widetilde{\delta}} P(\widetilde{h}) = 1$ finally yields

$$P(\widetilde{h}) = \begin{cases} \dfrac{1}{c(2\widetilde{\delta}+1)+1} & \text{for } \widetilde{h} = -\widetilde{\delta}-1 \\ \dfrac{c}{c(2\widetilde{\delta}+1)+1} & \text{for } \widetilde{h} \in \{-\widetilde{\delta}, \ldots, \widetilde{\delta}\} \end{cases} . \tag{26}$$

With the knowledge of $P(\widetilde{h})$, it is now straightforward to compute the drift of $\bar{h}$ from (20). First of all, we have in the gapless limit ($\widetilde{\delta} \to \infty$) that $m(0,t) = m(1,t) = 0$. Hence, $v^{(2)}(\widetilde{\delta} \to \infty) = \frac{1}{2}\langle \eta(0,t) + \eta(1,t) \rangle = \frac{1}{c}$. For finite (half-integer) values of $\widetilde{\delta}$'s, we have $v^{(2)}(\widetilde{\delta}) = \frac{1}{c} + u(\widetilde{\delta})$ where $u(\widetilde{\delta}) \equiv \frac{1}{2}\langle m(0,t) + m(1,t) \rangle$. Now observe that $\widetilde{h}(t)$ only depends on $\eta(r,t')$ for $t' < t$; hence there are no correlations between $\widetilde{h}(t)$ and $\eta(r,t'=t)$. Since $\bar{h}$ receives an extra "push" if and *only* if $\widetilde{h}$ is at one of the boundaries, i.e., at position $-\widetilde{\delta}-1$ or $\widetilde{\delta}$, we only need to add up contributions from these two cases in the calculation of u. Weighting these by the respective probabilities, we obtain

$$\begin{aligned} u(\widetilde{\delta}) &= \frac{1}{2} P(-\widetilde{\delta}-1) \langle \max\{0, 1-\eta(0,t)\} \rangle + \frac{1}{2} P(\widetilde{\delta}) \langle \max\{0, \eta(0,t)-\eta(1,t)\} \rangle \qquad (27) \\ &= \frac{1}{2} \frac{1}{c(2\widetilde{\delta}+1)+1} \left(1-\frac{1}{c}\right) + \frac{1}{2} \frac{c}{c(2\widetilde{\delta}+1)+1} \left[\frac{1}{c}\left(1-\frac{1}{c}\right)\right] \\ &= \left(1-\frac{1}{c}\right) \frac{1}{c(2\widetilde{\delta}+1)+1}. \end{aligned}$$

This result immediately yields the phase transition points

$$\widetilde{\mu}_{\rm c}^{(2)}(\widetilde{\delta}) = v^{(2)}(\widetilde{\delta}) = \frac{1}{c} + \left(1 - \frac{1}{c}\right) \frac{1}{c(2\widetilde{\delta}+1)+1}, \quad 2\widetilde{\delta} \in \mathbf{N} \tag{28}$$

between the linear and the logarithmic phase of local alignment for this $L = 2$ system. Translating (28) via (6) back to the original parameters μ and δ, we obtain

$$\mu_{\rm c}^{(2)}(\delta) = 2\frac{\delta+1}{2\delta(c-1)+(c-2)}. \tag{29}$$

The expression (29) is plotted as the dashed line in Fig. 7 for $c = 4$. The points corresponding

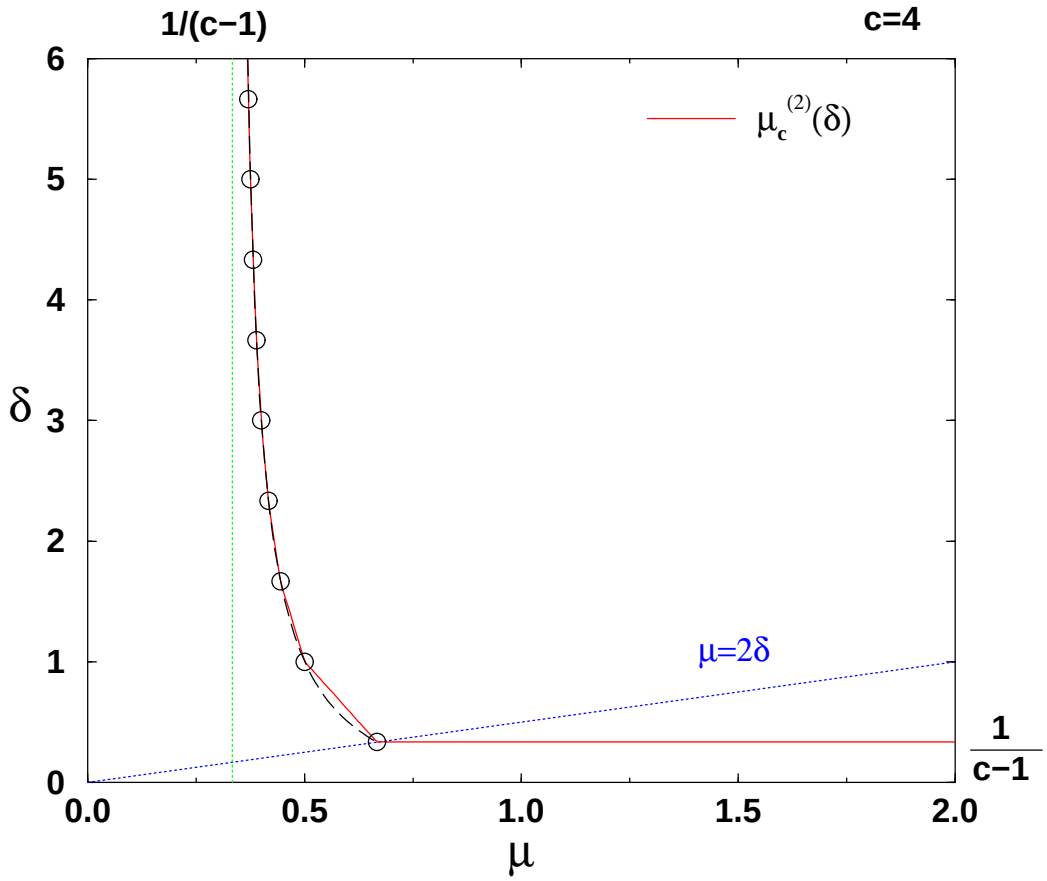

Figure 7: The phase transition line between the phases of linear and logarithmic alignment for alignment on a single strip. The solid curve is the exact phase transition line; the dashed line is the expression (29), whose values corresponding to $2\widetilde{\delta} \in \mathbf{N}$ are marked by the circles. The dotted lines show the boundary of the region $2\delta \geq \mu$ and $\mu \geq 1/(c-1)$. The critical line $\mu_{\rm c}^{(2)}(\delta)$ terminates on the line $2\delta = \mu$ at $\delta_0^{(2)} = 1/(c-1)$.

to integer values of $2\widetilde{\delta}$ are indicated by the circles. The full result including non-integer $2\widetilde{\delta}$'s (from Appendix A) is shown as the solid line. Note that the solid and dashed lines become essentially indistinguishable for $\delta > 2$. The shape of the phase transition line obtained (either the solid or dashed) for this single-strip system already resembles qualitatively the transition line of the full Smith-Waterman problem (Waterman *et al.*, 1987; Arratia and Waterman, 1994). In particular, it approaches the gapless limit of $\mu_{\rm c} = 1/(c-1)$ proportional to $1/\delta$ for large δ, and terminates on the line $\mu = 2\delta$ at a c-dependent value, $\delta_0(c)$.

4 Alignment on the Full Lattice

In this section, we generalize the calculation described in Sec. 3 for the single strip to the full lattice with $L \to \infty$. While all of the results obtained up to this point are exact, the complexity of the calculation for the full lattice necessitates two simplifying approximations:

First, we will assume that the variables $\eta(r,t)$ in the dynamic programming algorithm (7) are *statistically independent* from each other at different sites of the lattice, i.e.,

$$\Pr\{\eta(r,t)=\eta_1, \eta(r',t')=\eta_2\} = \Pr\{\eta(r,t)=\eta_1\} \cdot \Pr\{\eta(r',t')=\eta_2\} \tag{30}$$

if $r \neq r'$ or $t \neq t'$, while the distribution of each of the $\eta(r,t)$ is still given by (15). This uncorrelated distribution of $\eta(r,t)$ describes the related first-passage percolation problem (Kesten, 1986). It is not exactly valid for the sequence alignment problem at hand, since all $\eta(r,t)$ are generated by only $2N$ (rather than N^2) random letters. Nevertheless, extensive numerical investigations indicate that for many ensemble averaged quantities, including the subject of interest here, the drift rate v, the results obtained from these two different distributions of $\eta(r,t)$ are virtually indistinguishable. Thus the independent-η approximation, which greatly simplifies the calculation, is a reasonable starting point for obtaining the critical line $\mu_c(\delta)$.

The other key approximation we make is known as the "self-consistent" approximation in statistical physics. This is an Ansatz in which we assume that the score difference between neighboring rows, defined as $\widetilde{h}(r,t) = (-1)^r[h(r,t) - h(r+1,t)]$, is *statistically independent* for different r's in the large-t limit (see details below.) This Ansatz allows us to solve the stationary distribution $P(\widetilde{h})$ explicitly, the knowledge of which can be used to compute the drift rate v as demonstrated for the $L=2$ system in Section 3. On the special line $\widetilde{\delta}=0$, we will show below that the self-consistent Ansatz becomes *exact* for the case of independent η's. Our result $v(\widetilde{\delta}=0) = 2/(1+\sqrt{c})$ coincides with the conjectured expression by R. Arratia (unpublished; see also Steele (1986)) for the Chvátal-Sankoff constant of the LCS problem (see Eqs. (3) and (10)). For $\widetilde{\delta} > 0$, our Ansatz does *not* give the exact result, since correlations in score difference do exist as shown in Appendix B. However, these correlations are *short-ranged*, a result well known from studies of the related physics problems ; see e.g., Krug and Spohn (1991). Consequently, the error made due to this Ansatz is expected to be small. In fact, comparison with numerical estimates yields a discrepancy in $\mu_c(\delta)$ of only a few percent in the worst case (at intermediate values of δ's.) Our Ansatz thus provides a reasonable approximation of $\mu_c(\delta)$ and a good starting point for more refined calculations. The correlation effects can be treated exactly in principle for $\widetilde{\delta} = \frac{1}{2}, 1, \frac{3}{2}, \ldots$ as indicated in Appendix B. The details become more and more cumbersome and will be discussed elsewhere.

4.1 Properties of score differences

In the infinite lattice shown in Fig. 8, we focus on the local score differences

$$\widetilde{h}(r,t) \equiv (-1)^r[h(r,t) - h(r+1,t)]. \tag{31}$$

The way these variables are defined reflects the symmetry of the infinite lattice and ensures that each of the $\widetilde{h}(r,t)$ is comparable to $\widetilde{h}(t)$ studied in Sec. 3.

The even ($r=2n$) and odd ($r=2n+1$) rows have the following (different) recursion relations in the regime $\widetilde{\delta} \geq 0$

$$\begin{aligned} h(2n,t+1) = & \\ &= \max\{h(2n,t)+\eta(2n,t), h(2n+1,t)-\widetilde{\delta}, h(2n-1,t)-\widetilde{\delta}\} \end{aligned} \tag{32}$$

and

$$\begin{aligned} h(2n+1,t+1) = & \\ &= \max\{h(2n+1,t)+\eta(2n+1,t), h(2n,t+1)-\widetilde{\delta}, h(2n+2,t+1)-\widetilde{\delta}\} \\ &= \max\{h(2n+1,t)+\eta(2n+1,t), h(2n,t)+\eta(2n,t)-\widetilde{\delta}, h(2n-1,t)-2\widetilde{\delta}, \\ &\qquad h(2n+2,t)+\eta(2n+2,t)-\widetilde{\delta}, h(2n+3,t)-2\widetilde{\delta}\}. \end{aligned} \tag{33}$$

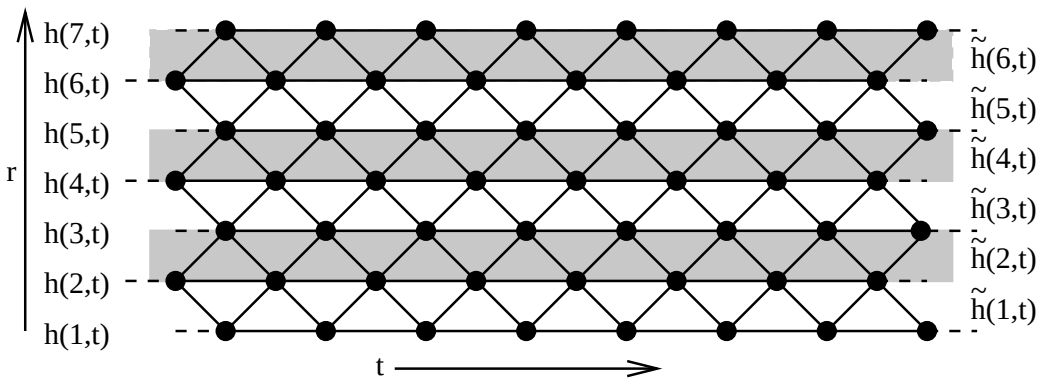

Figure 8: The infinite alignment lattice consists of a stack of many strips. Each strip is characterized by a variable $\widetilde{h}(r,t) = (-1)^r[h(r,t) - h(r+1,t)]$ similar to that studied in Section 3.

These lead to the following evolution equation for $\widetilde{h}(2n,t)$,

$$\begin{aligned} \widetilde{h}(2n,t+1) = & \qquad (34) \\ = \; & \widetilde{h}(2n,t) + \eta(2n,t) - \eta(2n{+}1,t) + m(2n,t) - m(2n{+}1,t), \end{aligned}$$

where we defined

$$\begin{aligned} m(2n,t) \;\equiv\; & \max\{0, -\widetilde{h}(2n,t) - \eta(2n,t) - \widetilde{\delta}, \qquad (35) \\ & -\widetilde{h}(2n{-}1,t) - \eta(2n,t) - \widetilde{\delta}\} \\ m(2n{+}1,t) \;\equiv\; & \max\{0, \widetilde{h}(2n,t) + \eta(2n,t) - \eta(2n{+}1,t) - \widetilde{\delta}, \qquad (36) \\ & \widetilde{h}(2n,t) - \widetilde{h}(2n{-}1,t) - \eta(2n{+}1,t) - 2\widetilde{\delta}, \\ & \widetilde{h}(2n{+}1,t) + \eta(2n{+}2,t) - \eta(2n{+}1,t) - \widetilde{\delta}, \\ & \widetilde{h}(2n{+}1,t) - \widetilde{h}(2n{+}2,t) - \eta(2n{+}1,t) - 2\widetilde{\delta}\}. \end{aligned}$$

The evolution equation for $\widetilde{h}(2n+1,t)$ is similar and can be obtained from the symmetry of the lattice.

For simplicity, we will consider here only those values of the scoring parameters such that $2\widetilde{\delta}$ is an integer. Under this condition, the allowed values of each of the $\widetilde{h}(r,t)$ are the same as for the $L = 2$ problem in the stationary state, i.e.,

$$\widetilde{h}(r,t) \in \{-\widetilde{\delta} - 1, -\widetilde{\delta}, \ldots, \widetilde{\delta} - 1, \widetilde{\delta}\}, \qquad (37)$$

with the interior positions at $-\widetilde{\delta}, \ldots, \widetilde{\delta} - 1$, and boundaries at $-\widetilde{\delta} - 1$ and $\widetilde{\delta}$ as shown in Fig. 6. To see this, we first note that $\widetilde{h}(r,t)$ cannot exceed the boundary at $\widetilde{\delta}$ since the basic action of the alignment algorithm (7) guarantees that

$$\widetilde{h}(2n,t) = h(2n,t) - h(2n+1,t) \leq \widetilde{\delta} \qquad (38)$$

for any $r = 2n$ [and analogously for odd r.] The remaining part of (37) can be shown by induction: Our choice of initial conditions (i.e., $h(2n, t=0) = 0$ and $h(2n+1, t=-1) = -\infty$) implies $\widetilde{h}(r,0) = \widetilde{\delta}$. If we now assume $\widetilde{h}(r,t) \in \{-\widetilde{\delta}-1, -\widetilde{\delta}, \ldots, \widetilde{\delta}-1, \widetilde{\delta}\}$, it is easy to see that $\eta(r,t) + m(r,t)$ can only take on the values 0 and 1 for any r. This immediately implies that $\widetilde{h}(r,t+1) \in \widetilde{\delta} + \mathbf{Z}$, and that for even $r = 2n$,

$$\begin{aligned} \widetilde{h}(2n,t{+}1) = & \\ = \; & \widetilde{h}(2n,t) + \eta(2n,t) + m(2n,t) - \eta(2n{+}1,t) - m(2n{+}1,t) \\ \geq \; & \widetilde{h}(2n,t) + \eta(2n,t) - \widetilde{h}(2n,t) - \eta(2n,t) - \widetilde{\delta} - [\eta(2n{+}1,t) + m(2n{+}1,t)] \\ \geq \; & -\widetilde{\delta} - 1. \end{aligned}$$

[For odd r this still holds due to the symmetry of the lattice.] Thus, $-\widetilde{\delta}-1$ and $\widetilde{\delta}$ are the boundary positions and Eq. (37) is verified.

From Eqs. (35)-(37), we also get

$$\begin{aligned}\widetilde{h}(2n,t+1)-\widetilde{h}(2n,t) &= \\ &= \eta(2n,t)-\eta(2n+1,t)+m(2n,t)-m(2n+1,t)\in\{-1,0,1\},\end{aligned} \tag{39}$$

indicating that in the full alignment problem, each $\widetilde{h}(r,t)$ can still change at most by a single step of size 1 in one time step.

4.2 Jump probabilities

The difference between alignment on a single strip studied in Sec. 3 and the infinite alignment lattice studied here lies in the jump rate of $\widetilde{h}$. It is affected by whether or not one of the neighboring $\widetilde{h}$'s is at the boundary $\widetilde{\delta}$ or $-\widetilde{\delta}-1$. To see this, we look at the interior points $-\widetilde{\delta}\le\widetilde{h}(2n,t)\le\widetilde{\delta}-1$. In this case, $\widetilde{h}(2n,t)$ does not enter into the expressions for $m(2n,t)$ or $m(2n+1,t)$. The two quantities which modify the single-strip jump probabilities are

$$\begin{aligned}q_-(\widetilde{h}) &\equiv \Pr\{\max\{0,-\widetilde{h}(2n-1,t)+1-\widetilde{\delta}\}=1\,|\,\widetilde{h}(2n,t)=\widetilde{h}\}\\ q_+(\widetilde{h}) &\equiv \Pr\{\max\{0,\widetilde{h}(2n+1,t)+\eta(2n+2,t)+1-\widetilde{\delta},\\ &\qquad \widetilde{h}(2n+1,t)-\widetilde{h}(2n+2,t)+1-2\widetilde{\delta}\}=1\,|\,\widetilde{h}(2n,t)=\widetilde{h}\}.\end{aligned}$$

They can be expressed by the conditional expectation values

$$q_-(\widetilde{h}) = \Pr\{\widetilde{h}(2n-1,t)=-\widetilde{\delta}-1\,|\,\widetilde{h}(2n,t)=\widetilde{h}\} \tag{40}$$

$$\begin{aligned}\text{and}\quad q_+(\widetilde{h}) &= \Pr\{\widetilde{h}(2n+1,t)=\widetilde{\delta}\,|\,\widetilde{h}(2n,t)=\widetilde{h}\}\left(\frac{1}{c}+\left(1-\frac{1}{c}\right)\times\right.\\ &\qquad \left.\times\Pr\{\widetilde{h}(2n+2,t)=-\widetilde{\delta}-1\,|\,\widetilde{h}(2n+1,t)=\widetilde{\delta}\wedge\widetilde{h}(2n,t)=\widetilde{h}\}\right).\end{aligned} \tag{41}$$

Due to *translational symmetry* of the lattice, these conditional expectation values do not depend on n. Taking those contributions from the neighboring differences into account, the jump probabilities w_- and w_+ are no longer equal as in Eq. (24). Instead, we find the jump size $\eta(2n,t)-\eta(2n+1,t)+m(2n,t)-m(2n+1,t)$ to be -1 with probability

$$w_-=\frac{1}{c}\left(1-\frac{1}{c}\right)\left[1-q_-(\widetilde{h})\right]+\left(1-\frac{1}{c}\right)^2\left[1-q_-(\widetilde{h})\right]q_+(\widetilde{h}), \tag{42}$$

and $+1$ with probability

$$w_+\equiv\frac{1}{c}\left(1-\frac{1}{c}\right)\left[1-q_+(\widetilde{h})\right]+\left(1-\frac{1}{c}\right)^2\left[1-q_+(\widetilde{h})\right]q_-(\widetilde{h}). \tag{43}$$

The probability of not jumping is of course $1-w_+-w_-$.

4.3 Independence of $\widetilde{h}$ on the special line $\mu=2\delta$

The key difficulty in computing the above jump rates is the joint distribution of the $\widetilde{h}(r,t)$ that appear in Eqs. (40) and (41). Here, we show that the joint distribution actually *factorizes*, i.e. neighboring $\widetilde{h}$'s are uncorrelated along the special line $\mu=2\delta$ or $\widetilde{\delta}=0$; this result will greatly simplify the subsequent calculations. As noted already, this special line corresponds to the LCS problem. However the reader should be cautioned that our result of independent $\widetilde{h}$'s is subject to the independent-η approximation.

In order to show the statistical independence of $\widetilde{h}(r,t)$ in the limit of large t for $\widetilde{\delta}=0$, we will need to prove that the distribution of independent $\widetilde{h}$'s is stationary under the dynamics given by Eqs. (34)-(36). Specifically, we will show that if $\widetilde{h}(r,t)$ are statistically independent, i.e. given by the distribution $\mathcal{P}(\widetilde{h}(1,t),\widetilde{h}(2,t),\ldots)=\prod_r P(\widetilde{h}(r,t))$ at some t, then the distribution of $\widetilde{h}(r,t+1)$ according to the evolution Eqs. (34)-(36) is given by the *same* distribution $\prod_r P(\widetilde{h}(r,t+1))$ for some *suitable choice* of the single-strip distribution $P(\widetilde{h})$, provided $\eta(r,t)$ at different (r,t) are also statistically independent. It should be remarked that we have not yet attempted an analytical study of the *stability* of this stationary solution. However, extensive numerical investigations indicate quite convincingly that this stationary state is the global attractor of the dynamics.

To proceed with the proof, we observe first that for $\widetilde{\delta}=0$, $\widetilde{h}(r,t)$ can take on only two possible values, -1 or 0. Since $\eta(r,t)\in\{0,1\}$, the evolution equations (34)-(36) can be rewritten as

$$\begin{aligned}\widetilde{h}(2n,t+1) &= -f\Big(f\Big(-\widetilde{h}(2n,t),-\widetilde{h}(2n-1,t),\eta(2n,t)\Big), \\ &\qquad f\Big(-\widetilde{h}(2n+1,t),-\widetilde{h}(2n+2,t),\eta(2n+2,t)\Big),\eta(2n+1,t)\Big)\end{aligned} \tag{44}$$

where the function $f(a,b,\eta)\equiv 1-a[1-b(1-\eta)]$ has the Boolean representation

$$f=\overline{a}\vee(b\wedge\overline{\eta}). \tag{45}$$

Eq. (44) can be verified by direct inspection of all 128 possible combinations of its variables. By symmetry, the evolution equation for odd r reads

$$\begin{aligned}\widetilde{h}(2n+1,t+1) &= -f\Big(f\Big(-\widetilde{h}(2n+1,t),-\widetilde{h}(2n+2,t),\eta(2n+2,t)\Big), \\ &\qquad f\Big(-\widetilde{h}(2n,t),-\widetilde{h}(2n-1,t),\eta(2n,t)\Big),\eta(2n+1,t)\Big),\end{aligned} \tag{46}$$

which is almost the same as (44) except that the first two arguments of the outer f function have switched order.

Our task is to show the statistical independence of the left hand sides of Eqs. (44) and (46) assuming an independent distribution for the $\widetilde{h}$'s that enter on the right hand side. Since $\widetilde{h}$ takes on only two values, $P(\widetilde{h})$ is simply parameterized by one parameter $p\equiv\Pr\{\widetilde{h}=-1\}$, with $\Pr\{\widetilde{h}=0\}=1-p$. Recalling that $\Pr\{\eta=1\}=1/c$ and $\Pr\{\eta=0\}=1-1/c$, and assuming that $\widetilde{h}(r,t)$ and $\eta(r,t)$ are independent random variables according to our strategy, we easily find that

$$\begin{aligned}&\Pr\{f(-\widetilde{h}(2n,t),-\widetilde{h}(2n-1,t),\eta(2n,t))=1\}\equiv \\ &\qquad\equiv\quad \widehat{p}=1-p\left[\frac{1}{c}+\left(1-\frac{1}{c}\right)(1-p)\right].\end{aligned} \tag{47}$$

Since the same holds for $f(-\widetilde{h}(2n+1,t),-\widetilde{h}(2n+2,t),\eta(2n+2,t))$, the quantities

$$\widehat{h}(2n,t) \equiv f(-\widetilde{h}(2n,t),-\widetilde{h}(2n-1,t),\eta(2n,t)) \tag{48}$$

$$\text{and}\quad \widehat{h}(2n+1,t) \equiv f(-\widetilde{h}(2n+1,t),-\widetilde{h}(2n+2,t),\eta(2n+2,t)) \tag{49}$$

are independent random Bernoulli variables which take the value 1 with probability $\widehat{p}$ and 0 with probability $1-\widehat{p}$. Using this result, Eqs. (44) and (46) can be more succinctly written as

$$\widetilde{h}(2n,t+1) = -f(\widehat{h}(2n,t),\widehat{h}(2n+1,t),\eta(2n+1,t)) \tag{50}$$

$$\text{and}\quad \widetilde{h}(2n+1,t+1) = -f(\widehat{h}(2n+1,t),\widehat{h}(2n,t),\eta(2n+1,t)), \tag{51}$$

with the distribution of $\widetilde{h}(2n,t+1)$ given by

$$\Pr\{\widetilde{h}(2n,t+1)=-1\}=1-\widehat{p}\left[\frac{1}{c}+\left(1-\frac{1}{c}\right)(1-\widehat{p})\right]. \tag{52}$$

Using Eq. (47) for $\widehat{p}$ in (52), we find that the *stationarity condition* $\Pr\{\widetilde{h}(2n,t+1)=-1\}=p$ can only be fulfilled if we choose

$$p=\frac{\sqrt{c}}{1+\sqrt{c}}. \tag{53}$$

Note that with Eq. (53), we also have $\widehat{p}=p$.

It remains to be shown that all $\widetilde{h}(r,t+1)$ as computed from Eqs. (44) and (46) are uncorrelated from each other. This is by no means automatic since $\widetilde{h}(r,t+1)$ for the four neighboring r's are functions of common $\widetilde{h}(r,t)$'s and $\eta(r,t)$'s. We now make use of a remarkable and crucial property of the function f that if a, b, and η are independent Bernoulli random variables with $\Pr\{a=1\}=\Pr\{b=1\}=p$ and $\Pr\{\eta=1\}=1/c$, then the two expressions $f(a,b,\eta)$ and $f(b,a,\eta)$ are also *independent* Bernoulli random variables, with $\Pr\{f(a,b,\eta)=1\}=\Pr\{f(b,a,\eta)=1\}=p$ themselves, i.e.,

$$\Pr\{f(a,b,\eta)=x, f(b,a,\eta)=y\}=\Pr\{f(a,b,\eta)=x\}\cdot\Pr\{f(b,a,\eta)=y\}. \tag{54}$$

This can be easily checked by inserting all eight possible input combinations of the function f and calculating the joint probabilities for the four possible outcomes of $f(a,b,\eta)$ and $f(b,a,\eta)$.

We are now ready to show the statistical independence of $\widetilde{h}(r,t+1)$. First, we show the independence of $\widehat{h}(r,t)$: From the definitions of the $\widehat{h}$'s [Eqs. (48) and (49)], and the assumption of independent $\widetilde{h}(r,t)$ and η, we note first of all that $\widehat{h}(2n,t)$ and $\widehat{h}(r,t)$ are statistically independent for any $r\neq 2n-1$ for the trivial reason that they depend on different and statistically independent variables. On the other hand, $\widehat{h}(2n,t)$ and $\widehat{h}(2n-1,t)=f(-\widetilde{h}(2n-1,t),-\widetilde{h}(2n,t),\eta(2n,t))$ can in principle be correlated, since they depend on the same $\widetilde{h}$ and η's. However, comparison of the above expression for $\widehat{h}(2n-1,t)$ and Eq. (48) for $\widehat{h}(2n,t)$ shows that the two expressions are just of the form $f(a,b,\eta)$ and $f(b,a,\eta)$. Thus, the two are also statistically independent in light of the special property (54), with $\Pr\{\widehat{h}=1\}=\Pr\{\widetilde{h}=-1\}=p$.

We next use $\widehat{h}(r,t)$ as statistically independent input variables for $\widetilde{h}(r,t+1)$, as specified by Eqs. (50) and (51). The right-hand sides of Eqs. (50) and (51) are again of the form $f(a,b,\eta)$ and $f(b,a,\eta)$. This follows from the distribution of $\widehat{h}(r,t)$ derived above, and the fact that $\eta(2n+1,t)$ are uncorrelated with the $\widehat{h}(r,t)$'s since only $\eta(r,t)$ with even r appeared in the calculation of the latter; see Eqs. (48) and (49). Therefore, the property (54) can be used again, with $a=\widehat{h}(2n,t)$ and $b=\widehat{h}(2n+1,t)$, leading to the final result

$$\Pr\{\widetilde{h}(2n,t+1)=\widetilde{h}', \widetilde{h}(2n+1,t+1)=\widetilde{h}''\}=P(\widetilde{h}')\cdot P(\widetilde{h}''),$$

which is easily generalized to $\widetilde{h}(r,t+1)$ for all values of r. It then follows that

$$\mathcal{P}(\widetilde{h}(1,t),\widetilde{h}(2,t),\ldots)=\prod_r P(\widetilde{h}(r,t))$$

is indeed a stationary distribution as was claimed.

4.4 The independent-$\widetilde{h}$ Ansatz for generic parameters

We now take the result of Sec. 4.3 and *assume* that the score differences $\widetilde{h}(r,t)$ are all statistically independent from each other at different r's even for $\widetilde{\delta}>0$. This is not quite true, since the correlation between $\widetilde{h}$'s, as measured by the correlation function

$$C(|r-r'|)\equiv\frac{\langle\widetilde{h}(r,t)\widetilde{h}(r',t)\rangle-\langle\widetilde{h}(r,t)\rangle^2}{\langle\widetilde{h}(r,t)^2\rangle} \tag{55}$$

does *not* vanish for $r\neq r'$ (see Fig. 9(a)). However the correlation function shown in Fig. 9(a) is *short-ranged*, i.e., it decays rapidly for $|r-r'|>1$. Although the range of correlations is expected

to increase for increasing $\widetilde{\delta}$ (there is after all no correlation for $\widetilde{\delta} = 0$), we find it to saturate for $\widetilde{\delta} > 10$. Fig. 9(a) is in fact a "worst case" situation taken at a rather large value of $\widetilde{\delta}$. But even there, we see that the correlation function essentially vanishes beyond a few lattice spacings. Also, the difference between the numerically calculated stationary single-site distribution of $\widetilde{h}$ (the circles in Fig. 9(b)) and the distribution we will derive from our simplifying Ansatz (the crosses) is not dramatic. Therefore, these correlations are expected to produce only small changes in the precise loci

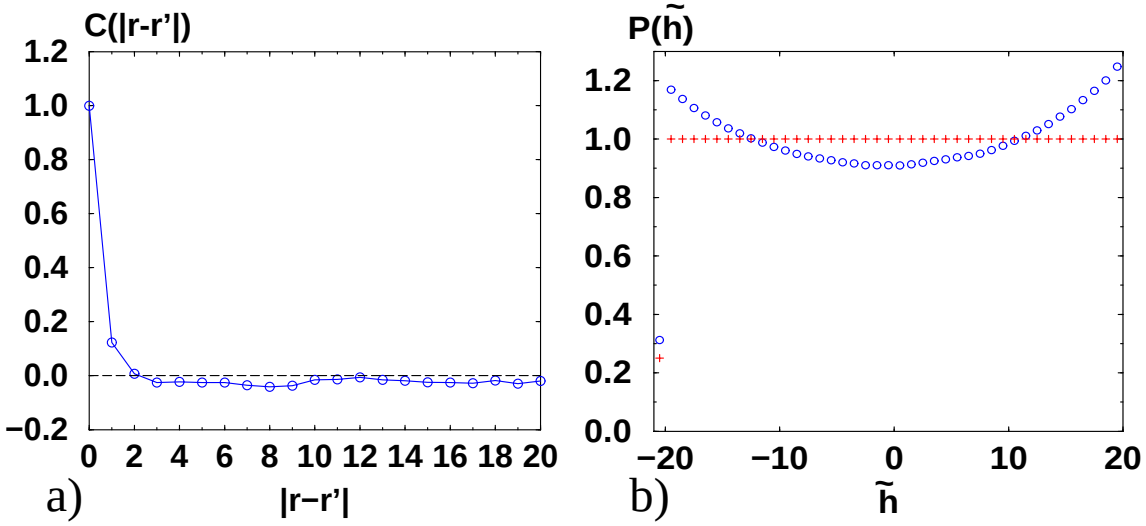

Figure 9: Effects of correlations between neighboring $\widetilde{h}$'s for large $\widetilde{\delta}$. The numerical results shown here correspond to $c = 4$ and $\widetilde{\delta} = 19.5$, where the correlations appear to have saturated to the largest extent. (a) shows the fast decay of the correlations in $\widetilde{h}$'s over the separation distance $|r - r'|$. In (b), the (rescaled) stationary distribution of $\widetilde{h}$ is shown. The circles are numerical data; the crosses represent our analytical result obtained by neglecting the correlations.

of the phase transition line, but not in the gross form of the parameter dependences. In Appendix B, we address in some detail the source of these correlations and why they do not affect the qualitative properties of the stationary distribution.

Our strategy will be the following: Under the assumption that $\widetilde{h}(r,t)$ are independent for all r's at the same t, the large-t behavior of the infinite system is completely characterized by the single-strip stationary distribution $P(\widetilde{h}) = \Pr\{\widetilde{h}(r,t) = \widetilde{h}\}$, similar to (23). To find this distribution, we need to know the jump probabilities for $\widetilde{h}(2n,t)$, given by the allowed values of $\eta(2n,t) + m(2n,t) - \eta(2n+1,t) - m(2n+1,t)$ according to Eq. (34). The m's are in turn affected (through Eqs. (35) and (36)) by the values of the neighboring $\widetilde{h}$'s, if the latter are at one of the boundaries, $-\widetilde{\delta} - 1$ or $\widetilde{\delta}$. Thus the probabilities

$$p_+ = \Pr\{\widetilde{h} = \widetilde{\delta}\} \quad \text{and} \quad p_- = \Pr\{\widetilde{h} = -\widetilde{\delta} - 1\} \tag{56}$$

enter into the jump probabilities for $\widetilde{h}(2n,t)$. Here, we will derive the distribution $P(\widetilde{h})$ using p_+ and p_- as *free* parameters. At the end, the conditions $P(-\widetilde{\delta} - 1) = p_-$ and $P(\widetilde{\delta}) = p_+$ will fix the values of these parameters which in turn completely specifies $P(\widetilde{h})$. We emphasize that there is no additional approximation involved in this way of finding the distribution $P(\widetilde{h})$. The same solution can be obtained directly from the independent-$\widetilde{h}$ assumption, by detailed considerations of the various jump probabilities. The strategy adopted here merely shortcuts the algebraic complexity.

We start by establishing a relation between p_+ and p_-, so that we only need to deal with one of these free parameters. We find this relation by considering the drift of $h(r,t)$. Eqs. (32), (33),

(35) and (36) imply that for even as well as odd r's, the change in $h(r,t)$ is given by $h(r,t+1) = h(r,t) + \eta(r,t) + m(r,t)$. Thus we have

$$\langle h(r,t)\rangle = (\langle \eta(r,t)\rangle + \langle m(r,t)\rangle) \cdot t = \left(\frac{1}{c} + \langle m(r,t)\rangle\right) \cdot t. \tag{57}$$

Under the independent-$\widetilde{h}$ assumption, the expectation value $\langle m(r,t)\rangle$ can be expressed by $p_\pm$ as

$$\langle m(r=2n,t)\rangle = \left(1-\frac{1}{c}\right)(2-p_-)p_- \tag{58}$$

for even r, and

$$\begin{aligned} \langle m(r=2n+1,t)\rangle = & \qquad (59) \\ = & \left(1-\frac{1}{c}\right) p_+ \left[\frac{1}{c} + \left(1-\frac{1}{c}\right) p_-\right] \left\{2 - p_+ \left[\frac{1}{c} + \left(1-\frac{1}{c}\right) p_-\right]\right\} \end{aligned}$$

for odd r. We know that for all t, the score difference $|h(r,t) - h(r+1,t)|$ is less than or equal to $\widetilde{\delta}+1$. From (57), we then conclude $|\langle m(2n,t)\rangle - \langle m(2n+1,t)\rangle| \leq (\widetilde{\delta}+1)/t$. Since the expectation values of the $m(r,t)$ do not depend on t in the stationary state, the limit $t \to \infty$ gives us

$$\langle m(2n,t)\rangle = \langle m(2n+1,t)\rangle. \tag{60}$$

This is nothing else than the statement that in the stationary state, the drift rate must be the same for every r. Equating (58) and (59) gives as the only reasonable solution[5]

$$p_+ \left[\frac{1}{c} + \left(1-\frac{1}{c}\right) p_-\right] = p_-. \tag{61}$$

In the following, we can therefore regard p_- as the only free parameter.

Now we calculate the stationary distribution $P(\widetilde{h})$ for a given value of p_-. If $\widetilde{h}(2n,t)$ is at an interior position, the probabilities to jump to one of its neighbors or to remain at the same position are given by the quantities $w_\pm$ in Eqs. (42) and (43). By the independent-$\widetilde{h}$ assumption, the conditional expectation values in Eqs. (40) and (41) become ordinary expectation values independent of $\widetilde{h}$, and can therefore be replaced by p_+ and p_- respectively. With the help of Eq. (61), we then find the result $q_-(\widetilde{h}) = q_+(\widetilde{h}) = p_-$. This simplifies the jump probabilities (42) and (43) to

$$w_- = w_+ = \left(1-\frac{1}{c}\right)(1-p_-)\left[\frac{1}{c} + \left(1-\frac{1}{c}\right) p_-\right]. \tag{62}$$

At the right boundary $\widetilde{h}(2n,t) = \widetilde{\delta}$, Eq. (34) along with the independent-$\widetilde{h}$ assumption yield the result that $\widetilde{h}(2n,t+1)$ remains at the boundary with probability $1-w_-$ and jumps to the interior position $\widetilde{\delta}-1$ with probability w_-. Combining this with the random walk of $\widetilde{h}$ at the interior positions, we have

$$\begin{aligned} P(\widetilde{\delta}) &= (1-w_-)P(\widetilde{\delta}) + w_+ P(\widetilde{\delta}-1) \\ \text{and} \qquad P(\widetilde{h}) &= w_+ P(\widetilde{h}-1) + (1-w_+-w_-)P(\widetilde{h}) + w_- P(\widetilde{h}+1) \end{aligned}$$

for the interior positions $\widetilde{h} \in \{-\widetilde{\delta}+1, -\widetilde{\delta}+2, \ldots, \widetilde{\delta}-1\}$. The second of these equations can be recast as

$$w_-[P(\widetilde{h}+1) - P(\widetilde{h})] = w_+[P(\widetilde{h}) - P(\widetilde{h}-1)]. \tag{63}$$

Since $w_+ = w_-$, it immediately follows that $P(-\widetilde{\delta}) = P(-\widetilde{\delta}+1) = \ldots = P(\widetilde{\delta})$ as in the case of alignment on a single strip.

[5]The other possible solution would lead to probabilities larger than one.

Examining Eq. (34) at the left boundary $\widetilde{h}(2n,t) = -\widetilde{\delta} - 1$, we find that $\widetilde{h}(2n,t+1)$ remains at the boundary $-\widetilde{\delta}-1$ with probability $q \equiv \left[\frac{1}{c} + \left(1-\frac{1}{c}\right) p_-\right]$ and jumps to the interior position $-\widetilde{\delta}$ with probability $1-q$. Together with $P(-\widetilde{\delta}+1) = P(-\widetilde{\delta})$ we obtain the remaining equations

$$\begin{aligned} P(-\widetilde{\delta}-1) &= qP(-\widetilde{\delta}-1) + w_- P(-\widetilde{\delta}) \\ \text{and} \qquad P(-\widetilde{\delta}) &= (1-w_+)P(-\widetilde{\delta}) + (1-q)P(-\widetilde{\delta}-1). \end{aligned}$$

They are simultaneously solved by

$$P(-\widetilde{\delta}) = \frac{1-q}{w_-} P(-\widetilde{\delta}-1). \tag{64}$$

The normalization condition of $P(\widetilde{h})$ gives us

$$1 = (2\widetilde{\delta}+1)P(\widetilde{\delta}) + P(-\widetilde{\delta}-1) = \left[(2\widetilde{\delta}+1)\frac{1-q}{w_-} + 1\right] P(-\widetilde{\delta}-1). \tag{65}$$

Together with the consistency condition (56) we obtain

$$p_- = \frac{\sqrt{c^2\widetilde{\delta}^2 + 2c\widetilde{\delta} + c} - c\widetilde{\delta} - 1}{c-1} \tag{66}$$

and hence the entire distribution

$$P(\widetilde{h}) = \begin{cases} p_- = \dfrac{\sqrt{c^2\widetilde{\delta}^2 + 2c\widetilde{\delta} + c} - c\widetilde{\delta} - 1}{c-1} & \text{for } \widetilde{h} = -\widetilde{\delta} - 1 \\ \\ p_+ = \dfrac{c\widetilde{\delta} + c - \sqrt{c^2\widetilde{\delta}^2 + 2c\widetilde{\delta} + c}}{(c-1)(2\widetilde{\delta}+1)} & \text{for } \widetilde{h} \in \{-\widetilde{\delta}, \ldots, \widetilde{\delta}\} \end{cases} . \tag{67}$$

[Note that $p_+ = P(\widetilde{\delta})$ is automatically fulfilled due to (61) and (64).] For $\widetilde{\delta} = 0$ this of course simplifies to (53) with $p = p_-$.

Using this result in Eqs. (57) and (58), we find the drift rate

$$v(\widetilde{\delta}) = \frac{1}{c} + \langle m(2n,t)\rangle = 2\frac{\widetilde{\delta}+1}{c-1}\left(\sqrt{c^2\widetilde{\delta}^2 + 2c\widetilde{\delta} + c} - c\widetilde{\delta} - 1\right) \tag{68}$$

which immediately gives the phase transition line $\widetilde{\mu}_c = v(\widetilde{\delta})$. Translating the result into the original parameters δ and μ using Eq. (6), we finally obtain

$$\mu_c(\delta) = \frac{2(\delta+1)}{c}\left[c\,\delta - (\delta+1) - \sqrt{(c-1)[\delta^2(c-1) - 2\delta - 1]}\right] \tag{69}$$

which is plotted in Fig. 10 for $c = 4$. As the comparison with numerical data shows, the phase transition line obtained from the independent-$\widetilde{h}$ and independent-η assumptions approximates the numerical results rather well.

4.5 The Chvátal-Sankoff constant

As mentioned previously, the case of $\mu = 2\delta$ or $\widetilde{\delta} = 0$ corresponds to the longest common subsequence problem (Chvátal and Sankoff, 1975); the Chvátal-Sankoff constant is given by $a_0 = v(\widetilde{\delta} = 0)$ (see Eq. (10)). Our result

$$a_0 = v(0) = \frac{2}{\sqrt{c}+1} \tag{70}$$

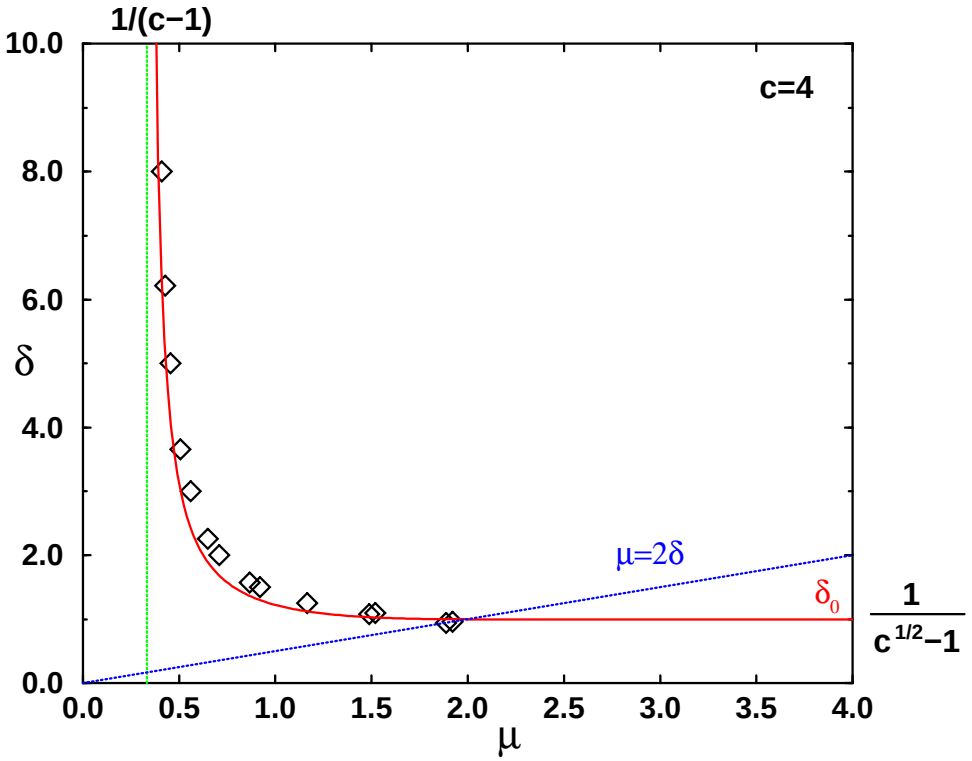

Figure 10: The log-linear phase transition line $\mu_c(\delta)$ for local alignment with a four-letter alphabet. The solid line is our result (69). The phase transition line intersects the special line $\mu = 2\delta$ *exactly* at $\delta_0 = \frac{1}{\sqrt{c}-1}$. Below the line $\mu = 2\delta$, the alignment is independent of μ. Above this line, our result is only an approximation but agrees well with the numerical results (shown as the diamonds). In the limit $\delta \to \infty$, our result converges toward the gapless limit of $\mu_c(\infty) = 1/(c-1)$.

has been conjectured previously by Arratia (unpublished; see also Steele (1986)) based on numerical observations. A corollary first pointed out by Waterman *et al.* (1987) is that the phase transition line would end exactly at the point

$$\delta_0 = \frac{1}{\sqrt{c}-1} \tag{71}$$

along the line $\mu = 2\delta$; see Fig. 10.

As shown in Sec. 4.3, our result on $v(\widetilde{\delta} = 0)$, depends only on the assumption of statistical independence of the η's. It is thus an *exact* result for the first-passage percolation version of the LCS problem. During the preparation of this manuscript, we became aware of a related study (Boutet de Monvel, 1999) which also obtained the Arratia conjecture (3) using an approximate "cavity" method. Based on extensive numerical simulations, Boutet de Monvel (1999) finds additionally that the actual value of the Chvátal-Sankoff constant and its counterpart for the problem with independent η's are somewhat different, by approximately 2% for an alphabet of size $c = 2$ and smaller differences for larger size alphabets. This difference has been noticed earlier by Dančík (1994). These authors reached their conclusions by using contradictory methods of accounting for the influences of the *finite* length of the sequences. The appropriate correction form has been given heuristically and verified numerically in recent statistical studies of sequence alignment (Drasdo *et al.*, 1998) and will be proved elsewhere (Bundschuh, to be published). It differs significantly from the empirical

correction forms used by both Dančík and Boutet de Monvel. Nevertheless, their general conclusion that the Chvátal-Sankoff constant of the LCS problem is slightly different from its counterpart for statistically independent η's turns out to be correct, as we have verified using the correct form of the sequence length dependence while exploiting our Eq. (44) in order to implement an extremely fast computer code for generating longest common subsequences (Bundschuh, to be published).

5 Concluding Remarks

We have studied local sequence alignment with gaps in a rectangular geometry as the first extension of the theory of gapless alignment. This geometry permits detailed calculations and is therefore well-suited for detailed studies of the influence of gaps. The assumption that neighboring score differences in the alignment lattice are statistically independent allowed us to calculate the loci of the log-linear phase transition line for the Smith-Waterman local alignment algorithm. For $\mu = 2\delta$, our results are actually exact for independent local scores η's, and reproduce the conjectured Chvátal-Sankoff constant for the longest common subsequence problem. Overall, the phase transition line obtained is in reasonable agreement with Monte Carlo data on alignment of random sequences. As we will illustrate in separate publications, the results described here can be used to obtain the important scaling laws given heuristically in the recent series of statistical studies. Our calculations can also be improved systematically to include the effect of the neglected correlations, thereby providing an even more accurate description of the phase transition line for a broad range of scoring parameters.

Acknowledgments

Discussions with S.F. Altschul, R. Arratia, and M. Lässig have contributed to this work. R.B. is supported by a Hochschulsonderprogramm III fellowship of the DAAD; T.H. acknowledges a Young Investigator Award by the Arnold and Mabel Beckman Foundation and a Sloan Research Fellowship.

Appendix A: Alignment on a Single Strip — Non-integer Case

If $2\widetilde{\delta}$ is not an integer, the difference between $\widetilde{\delta}$ and $-\widetilde{\delta}-1$, respectively the largest and smallest values $\widetilde{h}$ can take on, is not a multiple of the step size 1 of the random walk. In this case, $\widetilde{h}$ takes on values from the two sets $\widetilde{\delta}-j$ and $-\widetilde{\delta}-1+j$, with integer j. If the value of $\widetilde{h}$ is sufficiently far away from the boundaries at $-\widetilde{\delta}-1$ and $\widetilde{\delta}$, it will remain in one of the two sets according to (21). Only close to the boundaries are transitions between the two sets possible. Since the evolution equation (17) does not allow for values of $\widetilde{h}$ larger than $\widetilde{\delta}$ or smaller than $-\widetilde{\delta}-1$, the possible values of $\widetilde{h}$ are actually $\{\widetilde{\delta},\widetilde{\delta}-1,\ldots,\widetilde{\delta}-n^*\}$ or $\{-\widetilde{\delta}-1,-\widetilde{\delta},\ldots,-\widetilde{\delta}+n^*-1\}$, where n^* denotes the largest integer smaller or equal to $2\widetilde{\delta}+1$.

Our task is to solve the stationary distribution $P(\widetilde{h})=\Pr\{\widetilde{h}(t)=\widetilde{h}\}$ for all the $2n^*+2$ values of $\widetilde{h}$. Since $m(0,t)$ and $m(1,t)$ in (17) vanish at the interior positions $-\widetilde{\delta}\le\widetilde{h}(t)\le\widetilde{\delta}-1$, the probability for a jump from $\widetilde{h}$ to $\widetilde{h}\pm1$ is $w_\pm$ for $\widetilde{h}\in\{\widetilde{\delta}-(n^*-1),\ldots,\widetilde{\delta}-1\}$ or $\widetilde{h}\in\{-\widetilde{\delta},\ldots,-\widetilde{\delta}+n^*-2\}$, with $w_\pm$ given by Eq. (24). The corresponding probability to remain at $\widetilde{h}$ is $1-w_+-w_-$. This implies

$$P(\widetilde{h}) \;=\; w_+P(\widetilde{h}-1)+(1-w_+-w_-)P(\widetilde{h})+w_-P(\widetilde{h}+1)$$

for each $\widetilde{h}\in\{\widetilde{\delta}-(n^*-2),\ldots,\widetilde{\delta}-2\}$ and each $\widetilde{h}\in\{-\widetilde{\delta}+1,\ldots,-\widetilde{\delta}+n^*-3\}$. The solution is of the form

$$P(\widetilde{\delta}-j)=aj+b \qquad \text{and} \qquad P(-\widetilde{\delta}-1+j)=cj+d$$

for every $j\in\{1,\ldots,n^*-1\}$, with some yet undetermined constants a, b, c, and d. These constants can be fixed by the four remaining values of $P(\widetilde{h})$'s at the boundaries. Two of these values at $\widetilde{h}(t)=\widetilde{\delta}$ and $\widetilde{h}(t)=-\widetilde{\delta}-1$ are already known from the main text. Of the two remaining values, we have at position $\widetilde{h}(t)=\widetilde{\delta}-n^*$,

$$\widetilde{h}(t+1)=\widetilde{\delta}-n^*+[\eta(0,t)-\eta(1,t)]+\max\{0,-2\widetilde{\delta}+n^*-\eta(0,t)\}$$

from Eq. (17). Thus, $\widetilde{h}$ will stay at $\widetilde{\delta}-n^*$ with probability $\frac{1}{c^2}$, jump to $\widetilde{\delta}-(n^*-1)$ with probability $\frac{1}{c}(1-\frac{1}{c})$, to $-\widetilde{\delta}-1$ with probability $\frac{1}{c}(1-\frac{1}{c})$, and to $-\widetilde{\delta}$ with probability $(1-\frac{1}{c})^2$. In the same spirit, we get

$$\begin{aligned}\widetilde{h}(t+1) \;=\;& -\widetilde{\delta}+n^*-1+[\eta(0,t)-\eta(1,t)]\\ &-\max\{0,-2\widetilde{\delta}-n^*+1+\eta(0,t)-\eta(1,t)\}\end{aligned}$$

upon inserting $\widetilde{h}(t)=-\widetilde{\delta}+n^*-1$ into (17). This yields $\widetilde{h}=-\widetilde{\delta}+n^*-1$ with probability $1-\frac{2}{c}+\frac{2}{c^2}$, with jumps to $\widetilde{\delta}$ with probability $\frac{1}{c}(1-\frac{1}{c})$, and to $-\widetilde{\delta}+n^*-2$ with probability $\frac{1}{c}(1-\frac{1}{c})$.

Collecting the above together, we get the equations

$$\begin{aligned}
P(\widetilde{\delta}) &= (1-w_-)P(\widetilde{\delta})+w_+P(\widetilde{\delta}-1)+\frac{1}{c}\left(1-\frac{1}{c}\right)P(-\widetilde{\delta}+n^*-1),\\
P(\widetilde{\delta}-n^*) &= \frac{1}{c^2}P(\widetilde{\delta}-n^*)+w_-P(\widetilde{\delta}-n^*+1),\\
P(-\widetilde{\delta}-1) &= \frac{1}{c}P(-\widetilde{\delta}-1)+w_-P(-\widetilde{\delta})+\frac{1}{c}\left(1-\frac{1}{c}\right)P(\widetilde{\delta}-n^*),\\
P(-\widetilde{\delta}) &= (1-w_+-w_-)P(-\widetilde{\delta})+\left(1-\frac{1}{c}\right)P(-\widetilde{\delta}-1)+\\
&\qquad +\left(1-\frac{1}{c}\right)^2P(\widetilde{\delta}-n^*)+w_-P(-\widetilde{\delta}+1),\\
P(-\widetilde{\delta}+n^*-1) &= (1-w_+-w_-)P(-\widetilde{\delta}+n^*-1)+w_+P(-\widetilde{\delta}+n^*-2).
\end{aligned}$$

The (already normalized) solution of these equations is the stationary distribution

$$P(\widetilde{h}) = \begin{cases} \dfrac{n^* + \widetilde{h} - \widetilde{\delta} + \frac{1}{c}}{\left(n^* + \frac{1}{c}\right)\left(1 + n^* + \frac{1}{c}\right)} & \text{for } \widetilde{h} \in \{\widetilde{\delta} - n^*, \ldots, \widetilde{\delta}\} \\ \dfrac{1}{c(1+n^*)+1} & \text{for } \widetilde{h} = -\widetilde{\delta} - 1 \\ \dfrac{n^* - \widetilde{\delta} - \widetilde{h}}{\left(n^* + \frac{1}{c}\right)\left(1 + n^* + \frac{1}{c}\right)} & \text{for } \widetilde{h} \in \{-\widetilde{\delta}, \ldots, -\widetilde{\delta} + n^* - 1\} \end{cases} . \tag{72}$$

We can now calculate the drift rate $v^{(2)}(\widetilde{\delta}) = \frac{1}{c} + \frac{1}{2}\langle m(0,t) + m(1,t)\rangle$ using Eqs. (19) and (72). Contributions only arise from the four boundary values $\widetilde{h}(t) = \widetilde{\delta}$, $\widetilde{h}(t) = \widetilde{\delta} - n^*$, $\widetilde{h}(t) = -\widetilde{\delta} - 1$, and $\widetilde{h}(t) = -\widetilde{\delta} + n^* - 1$. We get

$$\begin{aligned} &\langle \max\{0, -\widetilde{h}(t) - \eta(0,t) - \widetilde{\delta}\}\rangle \\ &\quad = P(-\widetilde{\delta}-1)\langle \max\{0, 1 - \eta(0,t)\}\rangle \\ &\qquad + P(\widetilde{\delta}-n^*)\langle \max\{0, -2\widetilde{\delta} + n^* - \eta(0,t)\}\rangle \\ &\quad = P(-\widetilde{\delta}-1)\left(1-\frac{1}{c}\right) + P(\widetilde{\delta}-n^*)\left(1-\frac{1}{c}\right)\left(n^* - 2\widetilde{\delta}\right), \\ &\langle \max\{0, \widetilde{h}(t) + \eta(0,t) - \eta(1,t) - \widetilde{\delta}\}\rangle \\ &\quad = P(\widetilde{\delta}) \max\{0, \eta(0,t) - \eta(1,t)\} \\ &\qquad + P(-\widetilde{\delta}+n^*-1) \max\{0, -2\widetilde{\delta} + n^* - 1 + \eta(0,t) - \eta(1,t)\} \\ &\quad = P(\widetilde{\delta})\frac{1}{c}\left(1-\frac{1}{c}\right) + P(-\widetilde{\delta}+n^*-1)\frac{1}{c}\left(1-\frac{1}{c}\right)\left(n^* - 2\widetilde{\delta}\right). \end{aligned}$$

Inserting the values of P's from the distribution (72) and averaging over the above two contributions, we obtain the drift rate

$$v^{(2)}(\widetilde{\delta}) = \frac{\frac{1}{c}\left(1-\frac{1}{c}\right)}{\left(n^* + \frac{1}{c}\right)\left(1 + n^* + \frac{1}{c}\right)}\left(2n^* - 2\widetilde{\delta} + \frac{1}{c}\right), \tag{73}$$

and from it, the phase transition line $\mu_c^{(2)} = v^{(2)}(\widetilde{\delta})$ for this single-strip system. The result is plotted as the solid curve in Fig. 7.

As can be seen from the figure, the full result (73) and the analytic continuation of the result (29) obtained for integer values of $2\widetilde{\delta}$ approach each other very quickly. To see this behavior analytically, we factor out the integer result (29) from (73), i.e.,

$$v^{(2)}(\widetilde{\delta}) = \left(1 - \frac{1}{c}\right)\frac{1}{c(2\widetilde{\delta}+1)+1} f\left(2\widetilde{\delta}+1, n^*\right)$$

to obtain a correction factor

$$f(x, n^*) = \frac{x + 1/c}{n^* + 1/c}\left(1 - \frac{x - n^*}{1 + n^* + \frac{1}{c}}\right).$$

For fixed n^*, this factor has a maximum at $x = n^* + 1/2$ and is therefore bounded as

$$\begin{aligned} 1 \le \mathcal{F}(x, n^*) &\le \frac{n^* + 1/2 + 1/c}{n^* + 1/c}\left(1 - \frac{1}{2 + 2n^* + \frac{2}{c}}\right) \\ &\approx 1 + \frac{1}{4n^{*2}} + O\left(\frac{1}{n^{*3}}\right). \end{aligned}$$

Thus, it rapidly becomes negligible for $n^* > 1$ as manifested in Fig. (7) for $c = 4$. Of course for integer $2\widetilde{\delta}$, i.e., $n^* = 2\widetilde{\delta} + 1$, we have $f(2\widetilde{\delta}+1, n^*) = 1$.

Appendix B: Effect of Correlations in $\widetilde{h}$

In this appendix, we discuss the effects due to correlations between the neighboring $\widetilde{h}$'s neglected in the treatment of Section 4. Let us consider the quantities $q_-(\widetilde{h})$ and $q_+(\widetilde{h})$ defined in (40) and (41). These quantities describe the probabilities that one of the *neighboring* $\widetilde{h}$'s is at the boundary while $\widetilde{h}$ is at some given position. Correlations between neighboring $\widetilde{h}$'s lead to a non-trivial dependence of $q_\pm(\widetilde{h})$ on $\widetilde{h}$. Hence, the jump probabilities w_- and w_+ as defined in (42) and (43) for jumps to the left and right do not need to be equal to each other any more. According to (63), the distribution $P(\widetilde{h})$ then acquires a curvature in the bulk. In the following, we will show quantitatively how the distribution $P(\widetilde{h})$ is influenced by the correlations. An estimation of these correlations then leads to a qualitative understanding of the modified $P(\widetilde{h})$ and its effect on the position of the phase transition line.

Since our independent-$\widetilde{h}$ assumption is exact in the limit $\widetilde{\delta} = 0$ as shown in Sec. 4.3, we will consider here the opposite limit of large $\widetilde{\delta}$ where this assumption is the worst. As mentioned, the modifiers $q_\pm(\widetilde{h})$ are proportional to the probabilities that a neighboring $\widetilde{h}$ is at one of the boundaries. Since there are $2\widetilde{\delta}+2$ different values possible for the neighboring $\widetilde{h}$, these modifications are of the order $1/\widetilde{\delta}$ by normalization. Therefore on first sight one might take them to be negligible for large $\widetilde{\delta}$. But on the other hand, a given $\widetilde{h}$ will experience on the order of $\widetilde{\delta}$ such modifications, from $\widetilde{h}=0$ until it reaches its boundary values. Thus these modifications lead to a *finite* contribution in the limit of large $\widetilde{\delta}$. We can, however, neglect terms of order $1/\widetilde{\delta}^2$. These include, for example, the product $q_-(\widetilde{h})\cdot q_+(\widetilde{h})$, and the second term in the expression for q_+ itself. Therefore, in the limit of large δ, we can study a somewhat simplified system with $\widetilde{\eta}$ replaced by

$$\widehat{\eta}(t)=\begin{cases} -1 & \text{with prob. } \left(1-\frac{1}{c}\right)\left[\frac{1}{c}\left(1-\widehat{q}_-(\widetilde{h})\right)+\left(1-\frac{1}{c}\right)\widehat{q}_+(\widetilde{h})\right] \\ 0 & \text{with prob. } 1-\frac{2}{c}+\frac{2}{c^2}-\left(1-\frac{1}{c}\right)\left(1-\frac{2}{c}\right)[\widehat{q}_-(\widetilde{h})+\widehat{q}_+(\widetilde{h})] \\ 1 & \text{with prob. } \left(1-\frac{1}{c}\right)\left[\frac{1}{c}\left(1-\widehat{q}_+(\widetilde{h})\right)+\left(1-\frac{1}{c}\right)\widehat{q}_-(\widetilde{h})\right] \end{cases},$$

with

$$\begin{aligned} & \widehat{q}_-(\widetilde{h}) \equiv \Pr\{\widetilde{h}(2n-1,t)=-\widetilde{\delta}-1\,|\,\widetilde{h}(2n,t)=\widetilde{h}\}, \\ \text{and} \quad & \widehat{q}_+(\widetilde{h}) \equiv \frac{1}{c}\Pr\{\widetilde{h}(2n+1,t)=\widetilde{\delta}\,|\,\widetilde{h}(2n,t)=\widetilde{h}\}. \end{aligned}$$

These quantities are completely characterized by the *joint* probability distribution of two neighboring strips,

$$P_2(\widetilde{h},\widetilde{h}') \equiv \Pr\{\widetilde{h}(2n,t)=\widetilde{h}\wedge\widetilde{h}(2n+1,t)=\widetilde{h}'\}. \tag{74}$$

While we will provide below a qualitative description of this joint distribution function, let us assume for the moment that it is given, and study which effects the correlations have on the single strip distribution function $P(\widetilde{h})$. The conditional expectation values $\widehat{q}_\pm$ are expressed by $P_2(\widetilde{h},\widetilde{h}')$ as

$$\begin{aligned} \widehat{q}_-(\widetilde{h}) &= \frac{P_2(\widetilde{h},-\widetilde{\delta}-1)}{\sum_{\widetilde{h}'=-\widetilde{\delta}-1}^{\widetilde{\delta}} P_2(\widetilde{h},\widetilde{h}')} \\ \text{and} \quad \widehat{q}_+(\widetilde{h}) &= \frac{1}{c}\frac{P_2(\widetilde{h},\widetilde{\delta})}{\sum_{\widetilde{h}'=-\widetilde{\delta}-1}^{\widetilde{\delta}} P_2(\widetilde{h},\widetilde{h}')}. \end{aligned}$$

They enter into the stationarity equation

$$P(\widetilde{h}) = \left[1-\frac{2}{c}+\frac{2}{c^2}-\left(1-\frac{1}{c}\right)\left(1-\frac{2}{c}\right)[\widehat{q}_-(\widetilde{h})+\widehat{q}_+(\widetilde{h})]\right]P(\widetilde{h})$$

$$+ \quad \left[\left(1-\frac{1}{c}\right)\left(\frac{1}{c}(1-\widehat{q}_-(\widetilde{h}+1))+\left(1-\frac{1}{c}\right)\widehat{q}_+(\widetilde{h}+1)\right)\right]P(\widetilde{h}+1) \tag{75}$$
$$+ \quad \left[\left(1-\frac{1}{c}\right)\left(\frac{1}{c}(1-\widehat{q}_+(\widetilde{h}-1))+\left(1-\frac{1}{c}\right)\widehat{q}_-(\widetilde{h}-1)\right)\right]P(\widetilde{h}-1)$$

for the single strip probability distribution $P(\widetilde{h})$ for all $\widetilde{h}$ except at the boundaries. For large $\widetilde{\delta}$, we can assume that the probability distribution does not vary too rapidly over $\widetilde{h}$ and therefore approximate it by a smooth function

$$P(\widetilde{h}) \approx \frac{1}{2\widetilde{\delta}}\, p(\widetilde{h}/\widetilde{\delta}).$$

The same is true for the joint probability distribution

$$P_2(\widetilde{h},\widetilde{h}') \approx \frac{1}{4\widetilde{\delta}^2} p_2(\widetilde{h}/\widetilde{\delta},\widetilde{h}'/\widetilde{\delta}).$$

Recalling from Section 3 that $P(\widetilde{h}=-\widetilde{\delta}-1)$ — and therefore also $P_2(-\widetilde{\delta}-1,\widetilde{h})$ — is reduced by a factor $\frac{1}{c}$ with respect to $P(\widetilde{h}=\widetilde{\delta})$, and taking into account the symmetry of the lattice, we obtain in the limit of large $\widetilde{\delta}$,

$$\widehat{q}_+(\widetilde{h}) = \frac{1}{2c\widetilde{\delta}}\, q(\widetilde{h}/\widetilde{\delta}) \quad \text{and} \quad \widehat{q}_-(\widetilde{h}) = \frac{1}{2c\widetilde{\delta}}\, q(-\widetilde{h}/\widetilde{\delta}),$$

with

$$q(x) \equiv \frac{p_2(x,1)}{\int\limits_{-1}^{1} p_2(x,y)\mathrm{d}y}.$$

Expanding everything to the leading non-vanishing order in $1/\widetilde{\delta}$, we obtain (after some lengthy but straightforward calculation) the following *differential equation* from the stationarity equation (75),

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}p(x) + \frac{\mathrm{d}}{\mathrm{d}x}\big\{[q(x)-q(-x)]p(x)\big\} = 0, \tag{76}$$

with the solution

$$p(x) = \frac{\exp\left[\int\limits_0^x q(-y)-q(y)\mathrm{d}y\right]}{\int\limits_{-1}^{1}\exp\left[\int\limits_0^x q(-y)-q(y)\mathrm{d}y\right]\mathrm{d}x}. \tag{77}$$

Eq. (77) expresses the single-strip distribution p in terms of the joint distribution function p_2 or q (which is unknown so far) in a way that will be more useful, for the purpose of inferring the qualitative effects of correlations, than just integrating out one of the variables from the joint distribution function p_2. Let us now understand qualitatively how correlations in the $\widetilde{h}$'s make $q(y)$ a non-trivial function of y [$q(y)$ is constant under the independent-$\widetilde{h}$ assumption.] Consider the different ways the last gaps in two neighboring strips can be arranged: The four possibilities are shown in Fig. 11. Case (a) stems from configurations in which the middle row of the alignment lattice contains many matches. Therefore the optimal paths from both its neighbors eventually go back onto it. The configurations (b) and (c) correspond to cases where either the top or the bottom row has most of the matches. Configuration (d) appears if the middle row has many mismatches so that its score is first pulled up by the top row, but still has many mismatches so that its score is pulled up again by the bottom row. This implies that there are two degenerate (i.e., nearly equally scored) paths associated with the fixed end point of the middle row as shown in Fig. 11. Obviously, the last configuration is less probable than the configurations shown in (a), (b) and (c): Two neighboring $\widetilde{h}$'s tend to be both $\widetilde{\delta}$, or one $\widetilde{\delta}$ and the other $-\widetilde{\delta}-1$, but more rarely both $-\widetilde{\delta}-1$, in the limit of

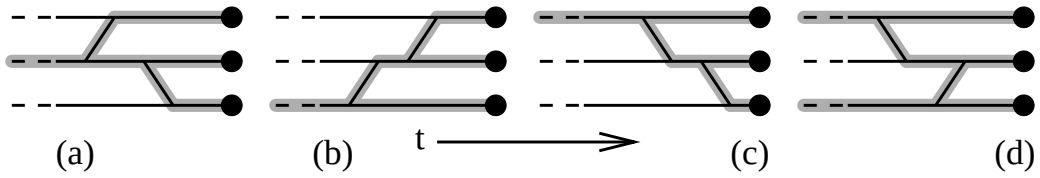

Figure 11: Different configurations of the last gaps inserted in two neighboring strips on the alignment lattice. The highlighted paths are the corresponding optimal paths associated with fixed end points (solid circles) on the right. As argued in the text, the configurations (a), (b), and (c) are more probable than the configuration (d), which produces an anti-correlation between the neighboring $\widetilde{h}$'s.

large $\widetilde{\delta}$. So if one of the $\widetilde{h}$'s is close to $-\widetilde{\delta}-1$, it is more likely that its neighbor is at $\widetilde{\delta}$ than close to 0. Therefore, $q(y)$ will be a decreasing function of y, leading to positive integrands in (77) for positive x. Thus, the distribution $p(x)$ takes on larger values at the boundaries $x=\pm 1$ than in the middle $x=0$ which qualitatively explains the observed form as shown in Fig. 9(b).

We have seen that the effect of correlations is to *enhance* the probability of finding $\widetilde{h}$'s near the boundary values $\widetilde{\delta}$ and $-\widetilde{\delta}-1$. Since the correction to the jump probabilities is of the order $1/\widetilde{\delta}$, while there are of the order of $2\widetilde{\delta}$ possible values for the score difference, this correction will be a *finite* numerical factor even in the worst case scenario of $\widetilde{\delta}\gg 1$. This can be seen in Fig. 9. Remembering that only the probability to be at the boundaries enters into the calculation of the total drift rate, we find that the drift rate will *increase* due to this correlation effect. According to the relation between drift rate and phase transition line given in Section 2.3, this increase will shift the phase transition line (69) obtained in Section 4 towards the log-side by some finite numerical factor.

Computing precisely this numerical factor requires the full solution of the joint distribution $P_2(\widetilde{h},\widetilde{h}')$ in (74). This task is difficult for large $\widetilde{\delta}\gg 1$, but not unsurmountable for smaller $\widetilde{\delta}$'s (e.g., $\widetilde{\delta}=\frac{1}{2},1,\frac{2}{3},\ldots$) for which $\widetilde{h}$ can take on only a few discrete values. In these cases, one can systematically find approximations to the finite correlation matrix $P_2(\widetilde{h},\widetilde{h}')$, thereby incorporating more and more correlations effects. The single-strip distribution $P(\widetilde{h})$ readily follows. The result can be used to compute the drift rate $v(\widetilde{\delta})$ as done in Sections 3 and 4.

References

Alexander, K.S., 1994. The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.* 4, 1074–1082.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Arratia, R., Morris, P., and Waterman, M.S. 1988. Stochastic scrabbles: a law of large numbers for sequence matching with scores. *J. Appl. Probab.* 25, 106–119.

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

Boutet de Monvel, J., 1999. Extensive Simulations for Longest Common Subsequences. *Europ. Phys. J. B* 7, 293–308.

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

Dančík, V., 1994. Expected Length of Longest Common Subsequences. PhD thesis, University of Warwick.

Dančík, V., and Paterson, M., 1994. Upper bounds for expected length of a longest common subsequence of two random sequences. In: *Proceedings of STACS94. Lecture Notes in Computer Science* 775, 669–678, Springer, Berlin.

Deken, J., 1979. Some Limit Results for Longest Common Subsequences. *Disc. Math.* 26, 17–31.

Dembo, A., and Karlin, S. 1991. Strong limit theorems of empirical functionals for large exceedances of partial sums of iid variables. *Ann. Probab.* 19, 1737–1755.

Doolittle, R.F. 1996. *Methods in Enzymology* 266. Academic Press, San Diego.

Drasdo, D., Hwa, T., and Lässig, M. 1997. DNA sequence alignment and critical phenomena. *Mat. Res. Soc. Symp. Proc.* 263, 75–80.

Drasdo, D., Hwa, T., and Lässig, M. 1998. Scaling laws and similarity detection in sequence alignment with gaps. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, Glasgow, J., *et al.* eds, 52–58, AAAI Press, Menlo Park.

Hwa, T., and Lässig, M. 1996. Similarity detection and localization. *Phys. Rev. Lett.* 76, 2591–2594.

Hwa, T., and Lässig, M. 1998. Optimal detection of sequence similarity by local alignment. In *Proceedings of the Second Annual International Conference on Computational Molecular Biology*, Israil S., *et al.*, eds, 109–116, ACM Press.

Karlin, S., and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87, 2264–2268.

Karlin, S., and Dembo, A. 1992. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Prob.* 24, 113–140.

Karlin, S., and Altschul, S.F. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA* 90, 5873–5877.

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

Krug, J., and Spohn, H. 1991. Kinetic Roughening of Growing Surfaces. In *Solids Far From Equilibrium: Growth, Morphology, and Defects*, Godreche, C. ed, 479, Cambridge University Press.

Lipman, D.J., and Pearson, W.R. 1985. Rapid and sensitive protein similarity searches. *Science* 227, 1435–1441.

Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

Olsen, R., Hwa, T., and Lässig, M. 1998. Optimizing Smith-Waterman Alignments. To appear in *Proceedings of the Forth Pacific Symposium on Biocomputing*

Paterson, M., and Dančík, V., 1994. Longest Common Subsequences. In: *Proceedings of the 19th international Symposium of Mathematical Foundations of Computer Science 1994, MFSC'94.*, 127–142, Springer, Berlin.

Pearson, W.R., and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.

Smith, T.F., and Waterman, M.S. 1981. Comparison of biosequences. *Adv. Appl. Math.* 2, 482–489.

Steele, J.M. 1986. An Efron-Stein inequality for nonsymmetric statistics. *Ann. Stat.* 14, 753–758.

Steele, J. 1997. *Probability Theory and Combinatorial Optimization*, SIAM, Philadelphia.

Waterman, M.S., Gordon, L., and Arratia R. 1987. Phase transitions in sequence matches and nucleic acid structure. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 1239–1243.

Waterman, M.S. 1989. In *Mathematical Methods for DNA Sequences*, Waterman, M.S., ed., CRC Press.

Waterman, M.S. 1994a. *Introduction to Computational Biology*, Chapman & Hall, London.

Waterman, M.S. 1994b. Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B* **344**, 383–390.