Asymmetric exclusion process and extremal statistics of random sequences

R. Bundschuh
*Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319, U.S.A.**
(Dated: 10 April 2001)

A mapping is established between sequence alignment, one of the most commonly used tools of computational biology, at a certain choice of scoring parameters and the asymmetric exclusion process, one of the few exactly solvable models of nonequilibrium physics. The statistical significance of sequence alignments is characterized through studying the total hopping current of the discrete time and space version of the asymmetric exclusion process.

PACS numbers: 05.45.-a, 87.10.+e, 02.50.-r

I. INTRODUCTION

Sequence alignment is one of the most commonly used computational tools of molecular biology. Its applications range from the identification of the function of newly sequenced genes to the construction of phylogenic trees [1, 2]. Beyond its practical importance, it is one of the simplest model systems for pattern matching. In computational biology, sequences are routinely compared via a transfer matrix algorithm to find the "optimal" alignment. Recently, it has been noted that this transfer matrix algorithm is the same as the one used to calculate the partition function or optimal energy of a directed polymer in a random medium [3]. This problem is known to belong to the universality class of surface growth as described by the Kardar-Parisi-Zhang (KPZ) equation [4]. From the assignment of sequence alignment to the KPZ universality class various *scaling laws* characterizing sequence alignment have been deduced. They have been used in order to answer questions of practical importance to sequence alignment, e.g., the optimal choice of alignment parameters [5–7]. But there are also *non universal* features which are of great importance for practical applications. They cannot be extracted from the knowledge of the universality class alone, but have to be evaluated by a microscopic study taking into account all the details of the given sequence alignment algorithm. In this paper, we will perform such a study for a certain choice of parameters for which sequence alignment maps onto the asymmetric exclusion process [8, 9], which is the best studied nonequilibrium system of the KPZ universality class, equivalent also to the six vertex model [10, 11]. The only approximation taken in this mapping is neglecting some subtle correlations in the hopping probabilities of the asymmetric exclusion process. We confirm numerically that neglecting these correlations introduces only minor deviations in the final results.

We will apply this mapping to address the central question in the biological application of sequence alignment, namely the assessment of alignment significance: The problem is that an "optimal" alignment, i.e., the best possible alignment of two given sequences according to some scoring function, does not necessarily reflect sequence homology. A sequence alignment algorithm will produce an "optimal" alignment for any pair of sequences, including randomly chosen ones. The important question is whether the alignment produced reflects an underlying similarity of the two sequences compared. A common way to address this question is to evaluate the probability of getting a certain alignment score by chance. This requires the knowledge of the distribution of alignment scores for random sequences. This distribution turns out to obey a universal (Gumbel) form with two non-universal parameters. In this paper, we will derive the Gumbel distribution and characterize some of its properties by relating them to the corresponding asymmetric exclusion process. In particular, we show how the tail of this distribution can be obtained from the generating function for the total number of hopped particles. The latter is also the generating function of the average surface height in the equivalent surface growth formulation of the asymmetric exclusion process. This important quantity has been calculated for the case of continuous time and continuous space using the replica trick [12] a long time ago. More recently, it has been obtained for the case of continuous time and discrete space in the scaling regime [13]. Here, we will calculate this quantity in discrete time and discrete space as necessary for the mapping to sequence alignment, in the asymptotic large size limit which is *beyond* the scaling limit. Our calculation does not make use of the replica trick and leads to a very simple closed form expression. It explicitly contains the anomalous $t^{1/3}$ scaling of the surface height fluctuations of KPZ surface growth in one dimension. We use this generating function to give an explicit expression for the significance of sequence alignments.

---

*current address: Department of Physics, The Ohio State University, 174 West 18th Avenue, Columbus, OH 43210-1106, U.S.A.

The paper is organized as follows: First, we will give a self-contained introduction to sequence alignment in Sec. II. This familiarizes the reader with the sequence alignment algorithm and gives us a chance to develop the notations to be used later. In Sec. III, we will reduce the problem of assessing the statistical significance of the widely used *local* alignment to a quantity defined in terms of the simpler *global* alignment. Readers more interested in the properties of the discrete asymmetric exclusion process can skip these two sections and go directly to Sec. IV, which describes the simplest version of the global alignment problem. Here, the mapping to the asymmetric exclusion process in discrete time and space with sublattice-parallel updating is described. Sec. V is devoted to the calculation of the generating function of interest for the asymmetric exclusion process. In Sec. VI, we discuss the result obtained, apply it to the assessment of alignment significance, and verify the analytical predictions numerically. In Sec. VII, we consider more general scoring systems and map them onto a generalized asymmetric exclusion processes. The final section gives a short summary of the paper and points towards several future directions. A number of technical details are given in the appendices.

II. REVIEW OF SEQUENCE ALIGNMENT

A. Gapless Alignment

Sequence alignment algorithms come in different levels of sophistication. The simplest alignment algorithm is *gapless* alignment. It is not only extremely fast but also very well understood theoretically. Thus, it has been very widely used, e.g., in its implementation of the program BLAST [14].

Gapless alignment looks for similarities between two sequences $\vec{a} = \{a_1 a_2 \dots a_M\}$, and $\vec{b} = \{b_1 b_2 \dots b_N\}$ of length M and $N \sim M$ respectively. The letters a_i and b_j are taken from an alphabet of size c. This may be the four letter alphabet $\{A, C, G, T\}$ of DNA sequences or the twenty letter alphabet of protein sequences with the letters distributed according to the natural frequencies of the twenty amino acids. A local gapless alignment $\mathcal{A}$ of these two sequences consists of a substring $a_{i-\ell+1} \dots a_{i-1} a_i$ of length ℓ of sequence $\vec{a}$ and a substring $b_{j-\ell+1} \dots b_{j-1} b_j$ of sequence $\vec{b}$ of the same length. Each such alignment is assigned a score

$$S[\mathcal{A}] = S(i, j, \ell) = \sum_{k=0}^{\ell-1} s_{a_{i-k}, b_{j-k}}, \tag{1}$$

where $s_{a,b}$ is some given "scoring matrix" measuring the mutual degree of similarity of the different letters of the alphabet. A simple example of such a scoring matrix is the match–mismatch matrix

$$s_{a,b} = \begin{cases} 1 & a = b \\ -\mu & a \neq b \end{cases} \tag{2}$$

which is used for DNA sequence comparisons [15]. For protein sequences, the more complicated 20×20 PAM [16] or BLOSUM matrices [17] are used to account for the variable degrees of similarity (e.g., hydrophobicity, size) among the 20 amino acids. The computational task is to find the i, j, and ℓ which give the *highest* total score

$$\Sigma \equiv \max_{\mathcal{A}} S[\mathcal{A}] \tag{3}$$

for a given scoring matrix $s_{a,b}$.

The optimization task called for in gapless alignment can be easily accomplished by introducing an auxiliary quantity, $S_{i,j}$, which is the optimal score of the above consecutive subsequences ending at (i, j) (optimized over ℓ.) It can be conveniently calculated in $O(N^2)$ instead of the expected $O(N^3)$ steps using the transfer matrix algorithm

$$S_{i,j} = \max\{S_{i-1,j-1} + s_{a_i, b_j}, 0\}, \tag{4}$$

with the initial condition $S_{0,k} = 0 = S_{k,0}$. This recursion equation reflects that for a given (i, j) the optimal ℓ is either zero or larger than zero. If the optimal ℓ is zero the corresponding score is zero as well. If the optimal ℓ is at least one, the pair (a_i, b_j) certainly belongs to the optimal alignment together with whatever has been chosen to be optimal up to the point $(i-1, j-1)$. Eq. (4) is basically a random walk with increments $s_{a,b}$ which is cut off if it falls below zero. The global optimal score is obtained as

$$\Sigma = \max_{1 \leq i \leq M, 1 \leq j \leq N} S_{i,j}. \tag{5}$$

In order to characterize the statistical significance of the alignment, it is necessary to know the distribution of Σ for gapless alignments of two *random* sequences, whose elements a_k's are generated independently from the same frequencies p_a as the query sequences, and scored with the same matrix $s_{a,b}$. This distribution of Σ has been worked out rigorously [18, 19]. For suitable scoring parameters, it is a Gumbel or extreme value distribution given by

$$\Pr\{\Sigma < S\} = \exp(-\kappa e^{-\lambda S}). \tag{6}$$

This distribution is characterized by the two parameters λ and κ with λ giving the tail of the distribution and $\lambda^{-1}\log\kappa$ describing the mode. For gapless alignment, these non universal parameters can be explicitly calculated [18, 19] from the scoring matrix $s_{a,b}$ and the letter frequencies p_a. For example, λ is the unique positive solution of the equation

$$\langle \exp(\lambda s)\rangle \equiv \sum_{a,b} p_a p_b \exp(\lambda s_{a,b}) = 1. \tag{7}$$

The other parameter κ is given by $\kappa = KMN$, where K is a more complicated function of the scoring matrix and the letter frequencies. Instead of reviewing the full derivation of the distribution Eq. (6) and its parameters, we will below give some heuristic arguments which yield the known result. These can later be generalized to the more relevant case of alignment with gaps.

For random sequences, one can take $j = i$ in (4) without loss of generality. Eq. (4) then becomes a discrete Langevin equation, with

$$S_{i,i} \equiv S(i) = \max\{S(i-1) + s(i), 0\}, \tag{8}$$

where the "noise" $s(i) \equiv s_{a,b}$ is uncorrelated and given by the distribution

$$\Pr\{s_i > s\} = \sum_{\{a,b|s_{a,b}>s\}} p_a p_b. \tag{9}$$

The dynamics of the evolution equation (8) can be in two distinct phases. The quantity which distinguishes these two phases is the expected local similarity score

$$\langle s\rangle \equiv \sum_{a,b} p_a p_b s_{a,b}. \tag{10}$$

If it is positive, the score $S(i)$ will increase on average. After a while, it becomes positive enough that the maximum in Eq. (8) will never be given by the zero option. This option could thus be omitted which corresponds to *global* gapless alignment. The dynamics is then a random walk $S(i) = S(i-1) + s(i)$ with an average upward drift $\langle s\rangle$. The maximal score will be close to the end of the sequences and will be given by $\Sigma \approx N \cdot \langle s\rangle$. Since it is linear in the length of the sequences, this is called the *linear phase* of local alignment. It is obviously not suited to identify matches of *subsequences*, and the distribution of the maximal score Σ is not an extreme value distribution. (It is just a sum of many independent local scores $s(i)$ and therefore obeys a Gaussian distribution according to the central limit theorem.)

The situation is dramatically different if $\langle s\rangle$ is negative. In this case the dynamics is qualitatively as follows: The score $S(i)$ starts at zero. If the next local score $s(i+1)$ is negative — which is the more typical case in this regime — then S remains zero. But if the next local score is positive, then S will increase by that amount. Once it is positive, $S(i)$ performs a random walk with independent increments $s(i)$. Since $\langle s\rangle$ is negative, there is a *negative drift* which forces $S(i)$ to eventually return to zero. After it is reset to zero, the whole process starts over again. The qualitative "temporal" behavior of the score $S(i)$ is depicted in Fig. 1.

From the figure, it is clear that the score landscape can be divided into a series of "*islands*" of positive scores, separated by "oceans" where $S = 0$. Each such island originates from a single jump out of the zero-score state and terminates when the zero-score state is reached again. Since each of these islands depends on a different subset of independent random numbers $s(i)$, the islands are *statistically independent* of each other. If we let the maximal score of the k^{th} island be σ_k, then these σ_k are independent random variables. Calculating the probability for the maximum score σ_k of an island of length L in a saddle point approximation and optimizing over the length L of the islands, we asymptotically obtain an exponential distribution

$$\Pr(\sigma_k > \sigma) \approx C_* e^{-\lambda\sigma} \tag{11}$$

for the maximal island scores σ_k (see App. A.) The parameter λ which gives the typical scale of the maximal island score is given by the drift-diffusion balance of the underlying Brownian process. If the local scores $s(i)$ were Gaussian variables with average $v < 0$ and variance D, this drift-diffusion balance would yield

$$\lambda = 2\frac{|v|}{D}. \tag{12}$$

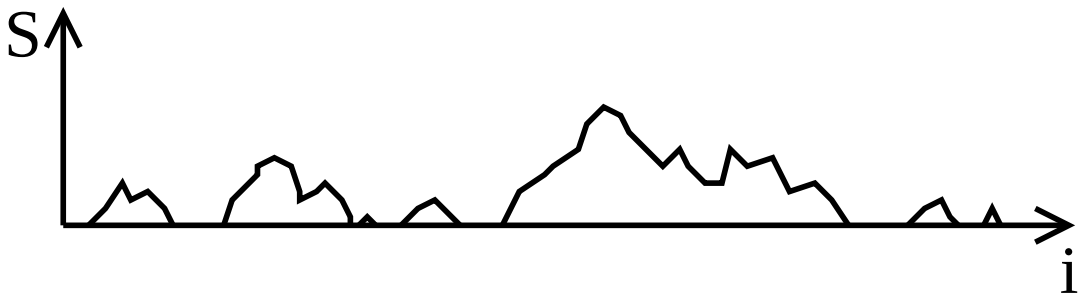

FIG. 1: Sketch of the total score as a function of sequence position in gapless local alignment.

For an arbitrary discrete or continuous distribution of the local scores $s(i)$, it turns out to be given by the more general condition (7), which reduces to Eq. (12) in the limit $\langle s\rangle \to 0^-$ where the central limit theorem takes hold.

Since the global optimal score Σ can be expressed by the maximal island scores as

$$\Sigma = \max_k\{\sigma_k\}, \tag{13}$$

the distribution of Σ can be calculated from the distribution of the σ_k. The connection is covered by the theory of extremal statistics as developed by Gumbel [20, 21]. In the case of a large number $K_* \sim N$ of independent island peak scores each of which asymptotically obeys the exponential distribution Eq. (11), the connection is especially simple and we get

$$\begin{aligned} \Pr\{\Sigma < S\} &= \Pr\{\max\{\sigma_1,\ldots,\sigma_{K_*}\} < S\} = \Pr\{\sigma_1 < S\}^{K_*} \\ &= (1 - C_* e^{-\lambda S})^{K_*} \approx [\exp(-C_* e^{-\lambda S})]^{K_*} = \exp(-\kappa e^{-\lambda S}) \end{aligned} \tag{14}$$

with $\kappa \equiv C_* K_*$, i.e., the parameter λ of the island peak score distribution Eq. (11) is the same as the parameter λ in the Gumbel distribution Eq. (6) of the maximal alignment scores.

B. Alignment with Gaps

In order to detect weak similarities between sequences separated by a large evolutionary distance, "gaps" have to be allowed within an alignment to compensate for insertions or deletions occurred during the course of evolution [22]. Here, we will specifically consider Smith-Waterman local alignment [23]. In this case, a possible alignment $\mathcal{A}$ still consists of two substrings of the two original sequences $\vec{a}$ and $\vec{b}$. But now, these subsequences may have different lengths, since gaps may be inserted in the alignment. For example the two subsequences `GATGC` and `GCTC` may be aligned as `GATGC` and `GCT-C` using one gap. Each such alignment $\mathcal{A}$ is assigned a score according to

$$S[\mathcal{A}] = \sum_{(a,b)\in\mathcal{A}} s_{a,b} - \delta N_{\mathrm{g}} \tag{15}$$

where the sum is taken over all pairs of aligned letters, N_{g} is the total number of gaps in the alignment, and δ is an additional scoring parameter, the "gap cost". In practice more complicated gap scores may be used, but we will concentrate on this version.

The task of local alignment is again to find the alignment $\mathcal{A}$ with the highest score as in Eq. (3), in this enlarged class of possible alignments. This can be very efficiently done by a transfer matrix method which becomes obvious in the alignment path representation [15]. In this representation, the two sequences to be compared are written on the edges of a square lattice as the one shown in Fig. 2 where we chose for simplicity $N = M$. Each directed path on this lattice represents one possible alignment. The score of this alignment is the sum over the local scores of the traversed bonds. Diagonal bonds correspond to gaps and carry the score $-\delta$. Horizontal bonds are assigned the similarity scores

$$s(r,t) \equiv s_{a_i,b_j} \tag{16}$$

where a_i and b_j are the letters of the two sequences belonging to the position $(r,t) = (i-j, i+j-1)$ as shown in Fig. 2.

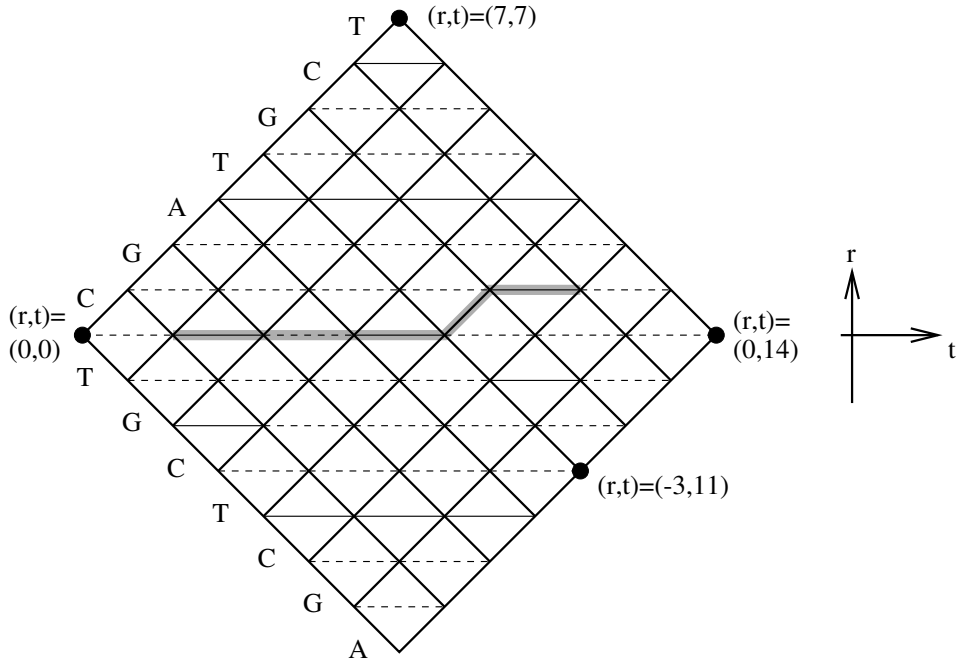

FIG. 2: Local alignment of two sequences $CGATGCT$ and $TGCTCGA$ represented as a directed path on the alignment lattice: the diagonal bonds correspond to gaps in the alignment. The horizontal bonds represent aligned pairs. Alignments of identical letters (matches) are shown as solid lines; alignments of different letters (mismatches) are shown dashed. The highlighted alignment path $r(t)$ corresponds to one possible alignment of two subsequences, $GATGC$ to $GCT-C$. This path contains one gap. It is also shown how the coordinates r and t are used to identify the nodes of the lattice.

If we were interested in finding the highest scoring *global* alignment of the two sequences $\vec{a}$ and $\vec{b}$, this corresponds to finding the best scoring path connecting the beginning $(0,0)$ with the end $(0,2N)$ of the lattice. To find this path effectively, we define the auxiliary quantity $h(r,t)$ to be the score of the best path ending in the lattice point (r,t). This quantity can be calculated by the Needleman-Wunsch transfer matrix algorithm [15]

$$h(r,t+1) = \max\{h(r,t-1) + s(r,t),\ h(r+1,t) - \delta, h(r-1,t) - \delta\}. \tag{17}$$

This is easily recognized [3] as the algorithm used to calculate the zero temperature configuration and energy of a directed polymer in a random potential given by the local scores $s(r,t)$. The scores $h(r,t)$ represent the (negative) energy of the optimally chosen polymer configuration ending in the point (r,t). Alternatively, the $h(r,t)$ can also be interpreted as the spatial height profile of a growing surface through the well known relation between the directed polymer and the KPZ equation.

If we are interested in *local* alignments, we can use the same trick as in the gapless case (4). Cutting off unfavorable scores by adding the choice of zero in the maximum of Eq. (17) leads to the Smith-Waterman algorithm [23]

$$S(r,t+1) = \max \left\{ \begin{array}{l} S(r,t-1) + s(r,t) \\ S(r+1,t) - \delta \\ S(r-1,t) - \delta \\ 0 \end{array} \right\}. \tag{18}$$

The score of the best local alignment is then given by

$$\Sigma = \max_{r,t} S(r,t). \tag{19}$$

In the presence of gaps, we can still distinguish a linear and a logarithmic phase. If the global alignment score tends to grow, the zero option of the local alignment algorithm does not play any role. We effectively revert to global alignment and get a maximum score which is linear in the length of the sequences. Contrary to gapless alignment, it is not enough to have a negative expectation value of the local scores $\langle s \rangle$ in order to prevent this. This is due to the fact that the alignment algorithm uses gaps to connect random stretches of good matches to optimize the score.

The average score grows by a gap dependent amount $u(\{s_{a,b}\},\delta)$ faster compared to the expectation value $\langle s\rangle$. The log-linear transition occurs now at $u(\{s_{a,b}\},\delta_c)+\langle s\rangle=0$. For the simple scoring system Eq. (2) this corresponds to a line $\delta_c(\mu)$ in the two dimensional space of the parameters μ and δ shown in Fig. 3. Even for this simple scoring system, the loci of the phase transition are only known approximately [24]; for more complicated scoring systems, only numerical results are available.

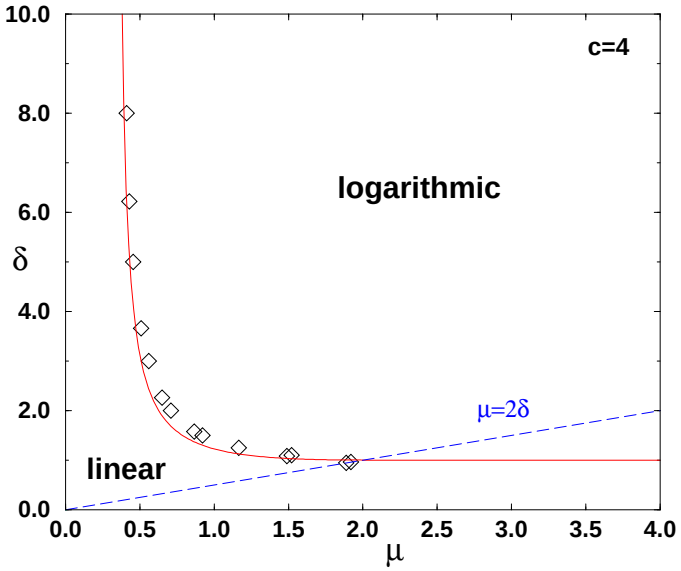

FIG. 3: Loci of the log-linear phase transition for alignment with the scoring system Eq. (2) for an alphabet of $c=4$ letters in terms of the mismatch cost μ and the gap cost δ. Useful alignments can only be obtained in the logarithmic phase above the phase transition line. The diamonds are numerically estimated points on the phase transition line; the solid line is the approximate locus calculated in [24]. Below the dashed line the alignments do not depend on the mismatch cost μ any more and the phase transition line is known to be strictly horizontal.

If the parameters are chosen such that $u+\langle s\rangle<0$, i.e., such that the expected global alignment score drifts downwards on average, then the average maximum score $\langle\Sigma\rangle$ is proportional to the logarithm of the sequence length as in the logarithmic phase of gapless alignment. The reduced value of $\langle\Sigma\rangle$ in the logarithmic phase makes it the regime of choice for the purpose of homology detection. Again, the distribution of Σ must be known for local alignments of random sequences in order to characterize the statistical significance of local alignment. There is no rigorous theory of this distribution in the presence of gaps. However, there is a lot of empirical evidence that the distribution is again of the Gumbel form [25–31]. The values of the parameters κ and λ are only known approximately for a few cases close to the gapless limit [32–34]. In practice, they are determined empirically by time consuming simulations. Below we will present an explicit calculation of the parameter λ for a simple scoring system.

III. SIGNIFICANCE ESTIMATION USING GLOBAL ALIGNMENT

As a first step, we want to show that the parameter λ, which describes the tail of the Gumbel distribution, can be derived solely from studying the much simpler *global* alignment governed by the recursion Eq. (17). Later, we will see that global alignment is in certain cases approximately equivalent to the asymmetric exclusion process. We will derive an explicit formula for λ by studying the corresponding asymmetric exclusion process.

A. An Expression for λ in Terms of Global Alignment

Let us define the generating function

$$Z(\lambda;L)\equiv\langle\exp[\lambda h(r=0,L)]\rangle \tag{20}$$

where the brackets $\langle\cdot\rangle$ denote the ensemble average over all possible realizations of the disorder, i.e., over all choices of random sequences $\vec{a}$ and $\vec{b}$ and $h(0,L)$ is the *global* alignment score at the end of a lattice of length L as shown in

Fig. 4(a). It can be obtained from the recursion relation (17) with the initial condition $h(2k, t = 0) = h(2k+1, t=1) = 0$. We claim that the parameter λ of the Gumbel distribution is obtained from

$$\lim_{L\to\infty} Z(\lambda; L) = 1. \tag{21}$$

Note, that this reduces simply to Eq. (7) in the case of gapless alignment, since for infinite gap cost δ, we have

$$\langle \exp[\lambda h(0, L)]\rangle = \langle \exp[\lambda \sum_{k=1}^{L/2} s(0, 2k-1)]\rangle = \langle \exp[\lambda s]\rangle^{L/2}. \tag{22}$$

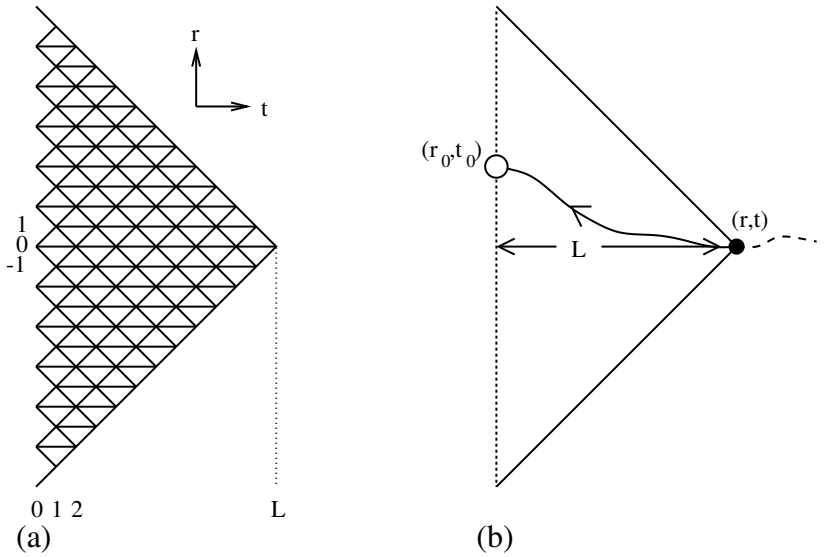

FIG. 4: Global alignment lattice used for significance estimation. (a) shows the right half of the lattice from Fig. 2. It can represent all possible paths of length L which end at the point $(r, t) = (0, L)$ and start at $(r, 0)$ for an arbitrary r. (b) shows such a path schematically. It represents the "rim" of an island with its high score denoted by the filled dot at the tip of the triangle. The open dot at (r_0, t_0) represents the corresponding island initiation event.

While we are not able to rigorously prove the condition on λ put forward in Eqs. (20) and (21) we will in the following give some heuristic arguments for its validity. One possible derivation uses two assumptions and otherwise applies some rigorous mathematical results. The second derivation is more intuitive and gives some feeling where the score distribution of local alignment comes from. In addition to these heuristic arguments we will verify in Sec. VI C that the equation for the Gumbel parameter λ that we will derive from Eqs. (20) and (21) indeed yields the correct statistics of local sequence alignment.

B. Derivation under the Assumption of a Gumbel Distribution

In this first derivation we will start from the assumption that the distribution of the local alignment score $\Sigma(L)$ for comparisons of two sequences of equal length L is of the Gumbel form Eq. (6) with $\kappa = KL^2$. This has been established by many numerical studies [25–31]. Under this assumption, a simple calculation shows that

$$\lim_{L\to\infty} \frac{\langle \Sigma(L)\rangle}{\log L} = \frac{2}{\lambda}. \tag{23}$$

Thus, we only have to calculate the asymptotic expectation value on the left hand side of Eq. (23) in order to determine the value of the Gumbel parameter λ.

The existence of this asymptotic expectation value has been rigorously established by Arratia and Waterman [35]. Its numerical value has been studied by Zhang [36] and we will reformulate Zhang's result in our notation. To this end, we will consider the global alignment score $\widehat{h}(r, t)$ calculated through the recursion Eq. (17) on the diamond-shaped

lattice shown in Fig. 2, i.e., with the initial conditions $\widehat{h}(r=t,t)=\widehat{h}(r=-t,t)=-t\delta$. With this quantity we can define $\widehat{\lambda}_L$ as the unique solution of $\langle\exp[\widehat{\lambda}_L\widehat{h}(r=0,2L)]\rangle=1$. Then, our Eq. (23) together with Theorem 1 and Eqs. (2.15) and (2.16) of Ref. [36] imply that given $\varepsilon>0$ and large enough L and n the inequality

$$\frac{\Sigma(Ln)}{\log n}+\varepsilon\geq\frac{2}{\widehat{\lambda}_L}\geq 2(\frac{1}{\lambda}-\varepsilon)\left(1-\frac{\varepsilon}{r(0)}\right) \tag{24}$$

holds where $r(0)$ is a positive constant independent of ε, n, and L. Thus, in the limit $n\to\infty$ we get almost surely

$$\frac{2}{\lambda}+\varepsilon\geq\frac{2}{\widehat{\lambda}_L}\geq 2(\frac{1}{\lambda}-\varepsilon)\left(1-\frac{\varepsilon}{r(0)}\right). \tag{25}$$

This implies that $\lim_{L\to\infty}\widehat{\lambda}_L=\lambda$ or in other words λ is given by the condition

$$\lim_{L\to\infty}\widehat{Z}(\lambda;L)=1 \tag{26}$$

on the generating function

$$\widehat{Z}(\lambda;L)\equiv\langle\exp[\lambda\widehat{h}(r=0,L)]\rangle. \tag{27}$$

To connect this to the conditions (20) and (21) we have to assume that $h(r=0,L)\approx\widehat{h}(r=0,L)$ in the limit of large L. The difference between these two scores are the boundary conditions. While the optimal path corresponding to $h(r=0,L)$ is allowed to start at any r_0 as indicated in Fig. 4 the optimal path corresponding to $\widehat{h}(r=0,L)$ has to start at $r_0=0$. However, the optimal path for $h(r=0,L)$ is expected to start at a distance $|r_0|$ that is sublinear in L. Thus, it is at least plausible to use $h(r=0,L)$ and $\widehat{h}(r=0,L)$ interchangeably at least as far as the growth behavior of a quantity like $\langle\exp[\lambda\widehat{h}(r=0,L)]\rangle$ for large L is concerned. This transforms Eqs. (26) and (27) into conditions (20) and (21).

C. Intuitive derivation

The key observation which allows us to understand the result Eqs. (20) and (21) intuitively is the fact that similar to the case of gapless alignment discussed in the last section, the points on the alignment lattice can be grouped together as *islands* [31]. By the construction of the local alignment algorithm (18), many points on the alignment lattice have a score of zero in the logarithmic alignment regime. As for gapless alignment, a positive score will be generated out of this "sea" of zeroes, if a good match occurs by chance. This positive score can then imply further positive scores via the recursion relation (18). For every point (r,t) on the lattice which has a positive score, we can define a restricted optimal path $\widehat{r}^*_{r,t}(\tau)$, which is the highest scoring path out of all paths $\widehat{r}(\tau)$ with an end fixed at $\widehat{r}(t)=r$; see the example in Fig. 2. This highest scoring path is uniquely defined for each point (r,t) if a convention of how to handle degeneracies in the maximization procedure Eq. (18) is chosen. While the specific choice of a convention should not matter[1] we can, e.g., declare that the first option that maximizes the right hand side of Eq. (18) locally defines the highest scoring path. This uniquely defined path must start at some point (r_0,t_0) where a positive score is created from a zero score by a good match. An island is then defined to be the collection of points (r,t) with positive score, i.e., $S(r,t)>0$, and whose restricted optimal path $\widehat{r}^*_{r,t}(\tau)$ originates at the same point (r_0,t_0). A sketch of these islands is shown in Fig. 5. By this definition, every lattice point with a positive score belongs to exactly one island. Each of these islands has a maximum score which we denote by σ_k as we did in the gapless case. Thus, the maximal score Σ on the total lattice is given by Eq. (13).

Although the positively scoring sites of the lattice are uniquely assigned to islands by this definition, islands do not necessarily have to be surrounded by zero scores. It is possible for two neighboring lattice points to belong to two different islands, i.e., for two islands to "touch" each other (see Fig. 5.) However, the higher the peak score of an island the less probable the configuration of the $s(r,t)$ that leads to such a high scoring island. Thus, if we restrict our attention only to islands the peak score σ_k of which is larger than some threshold σ_0, these islands

---

[1] The value of the Gumbel paramter λ should depend continuously on the scoring parameters $s_{a,b}$. Since a degeneracy in the maximization procedure Eq. (18) usually can be resolved through slightly varying the scoring parameters $s_{a,b}$ the choice of a procedure to handle these degeneracies cannot influence the final value of the Gumbel parameter λ.

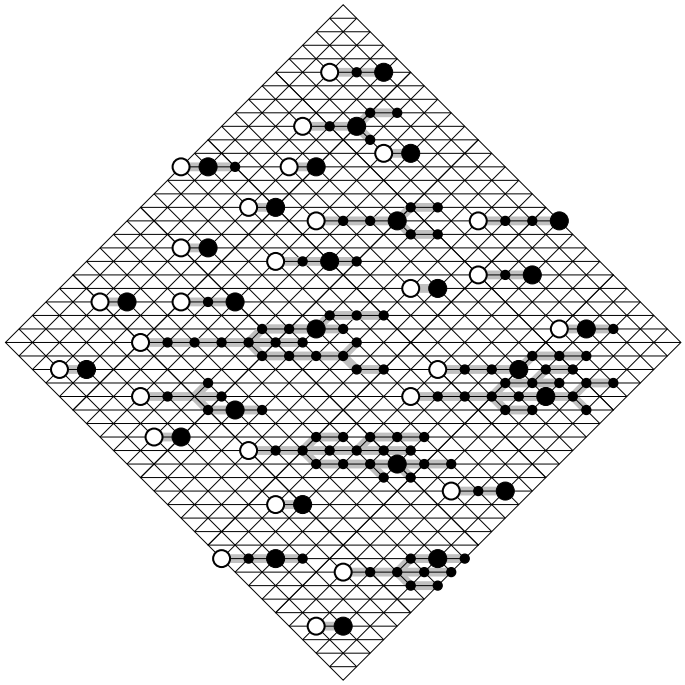

FIG. 5: Sketch of some islands on the local alignment lattice. The lattice sites with a positive score are marked with dots. The bonds which have been chosen in the maximization process (18) are highlighted. Together they are the restricted optimal path associated with each point with a positive score. Each of these paths goes back to an island initiation event which is marked by an open dot. The large filled dots mark the positions of the highest scoring point on each island. As exemplified by the two islands close to the right tip of the lattice islands do not have to be separated by lattice points with zero scores.

will occur at areas of the scoring lattice that are the further apart from each other (with lower scoring islands interspersed) the larger the threshold σ_0. Also the probability of not being separated by zero scores becomes small with increasing separation. Thus, the island peak scores σ_k of those islands exceeding a threshold σ_0 are expected to become statistically independent random variables, i.e., changes in the configuration of the $s(r,t)$ that affect the peak score of one of these high scoring islands do not affect the peak score of another of these high scoring islands. While this is an assumption, it can be numerically verified. The independence can be quantified by the correlation coefficient

$$R = \frac{\langle \sigma\sigma' \rangle - \langle \sigma \rangle^2}{\langle \sigma^2 \rangle - \langle \sigma \rangle^2} \tag{28}$$

where σ and σ' are the peak scores of two neighboring islands on the alignment lattice exceeding a threshold score σ_0. In Ref. [31] this quantity has been studied by averaging over 300 pairs of random sequences with an alphabet size of 20 and a gap cost $\delta = 2.9$ using the PAM-250 [16] scoring matrix for $s_{a,b}$. At $\sigma_0 = 7.5$ the correlation coefficient was estimated to be $R \approx -0.001$ indicating the statistical independence of these large islands. It is not to be expected that this independence should break down for the simpler local scoring matix Eq. (2).

Thus, we will in the following assume that the islands peak scores σ_k of sufficiently high scoring islands are statistically independent random variables. The islands with smaller scores do not contribute to the maximum in Eq. (13) and the fact that their island peak scores are not really uncorrelated only rescales the effective number of islands. Thus, we again observe a Gumbel distribution of Σ via Eq. (14) for very long sequences. The crossover sequence length at which a Gumbel distribution is a good description of the distribution of Σ depends on the scoring system. According to the above considerations, it is only valid if sufficiently many of the large independent islands occur on the scoring lattice. If the typical size of a single island is comparable to the length of the sequence we will not expect any Gumbel like distribution. This can easily happen as the log-linear phase transition is approached since the typical island sizes diverges at the transition. For a scoring system very close to the transition, the Gumbel distribution may be observed only for very long sequences. However, all practically useful scoring systems are far enough away from the phase transition to ensure a sufficient number of large islands on a scoring lattice for two sequences of realistic lengths, i.e., a few hundred letters each.

Our task is thus to calculate the distribution of the island peak scores σ_k for very large islands in the presence of gaps. This distribution of maximal island scores can be derived analogously to the gapless case (App. A.) While a single gapless island is described by a random walk of some optimized length L, an island with gaps corresponds to a *global gapped alignment* of some optimized length L as the one shown schematically in Fig. 4(b). Using this replacement, the maximal island distribution again has an asymptotically exponential form (11) with the decay constant λ given by Eq. (21). An approximate interpretation for the result (21) is the following: Due to the choice of scoring parameters in the logarithmic phase of local alignment, the average score $\langle h(0,L)\rangle$ of global alignment with the same choice of parameters decreases linearly with the length L of the alignment. Thus, typical configurations of the disorder have a strongly negative score $h(0,L)$ and hardly contribute to $Z(\lambda;L) = \langle \exp[\lambda h(0,L)]\rangle$. Only on very rare occasions, $h(0,L)$ is positive for large L and contributes significantly to $Z(\lambda;L)$. The fact that there is a choice of λ with $Z(\lambda;L) = 1$ for large L implies that these configurations with positive $h(0,L)$ are *exponentially rare*. It is thus necessary to weight these configurations with the exponential factor $\exp[\lambda h(0,L)]$, and choose λ to match the decay constant of the probability of finding such rare events.

D. Interpretation of Z

As already noted in the analogy between the directed polymer and sequence alignment, the score h corresponds to the (negative of the) free energy. Thus the quantity $Z(\lambda;L) = \langle \exp[\lambda h(0,L)]\rangle$ can be interpreted as the disorder-averaged (zero temperature) partition function[2] of λ "replicas" of a directed polymer of length L. Note that the replica number given by λ need not be integer. In the surface growth interpretation, $Z(\lambda;L)$ is the generating function for the space averaged surface height. While many of the universal features of global and local sequence alignment (e.g., its scaling behavior in the logarithmic phase and upon approaching the phase transition line) can be understood merely from the knowledge that sequence alignment belongs to the KPZ universality class [3, 5–7] or from the limit $Z(\lambda \to 0;L)$, a solution of Eq. (21) for the non universal quantity λ requires the knowledge of the large L behavior of the entire function $Z(\lambda;L)$ and hence a more detailed microscopic calculation for the given model. This is what we will undertake in the following sections.

IV. GLOBAL ALIGNMENT AS AN ASYMMETRIC EXCLUSION PROCESS

A. A Simple Model of Sequence Alignment

From now on we will focus on *global* alignment as described by Eq. (17), and use Eq. (21) to infer the value of the parameter λ characterizing local alignment. We restrict ourselves here to a very simple scoring system. In applications of sequence alignment this scoring system is not very useful since it allows more gaps than naturally related sequences would show and since it is much too restrictive as far as taking different degrees of similarity between different letters of the alphabet is concerned. However, as we will point out in Sec. VII, the mapping presented in this section between alignment with the simple scoring system and the asymmetric exclusion process can be generalized to a mapping between alignment with more complicated scoring systems and generalizations of the asymmetric exclusion process. As far as this mapping is concerned, restricting ourselves to the simple scoring system is solely a matter of convenience since it avoids lengthy expressions which would make the spirit of the mapping less accessible.

Although this mapping can be generalized to more realistic scoring systems, we will see in Sec. V that calculating the parameter λ involves solving explicitely for the largest eigenvalue of a generalized transfer matrix of the asymmetric exclusion process. This second step is only readily possible for this simple scoring system. Thus, our explicit expression for λ is only valid for this simple scoring system which is not practically used. However, being able to solve for λ even for unrealistic scoring parameters is still very valuable as a test bed for numerical estimation methods for the value λ.

Specifically, we will study the scoring system in which the local similarity scores $s_{a,b}$ can take on only two possible values,

$$s_{a,b} = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases} . \tag{29}$$

[2] However, $Z(\lambda;L)$ should *not* be interpreted as the partition function at temperature λ^{-1}.

Moreover we will choose the gap cost to be $\delta = 0$. With this choice of the scoring parameters, the score h has the additional interpretation of being the length of the *longest common subsequence* of the two sequences $\vec{a}$ and $\vec{b}$. This longest common subsequence problem has a long history as a toy model for sequence comparisons [38–40].

Additionally, we will neglect correlations between the local scores $s(r,t)$, which arise from the fact that all $M \times N$ local scores are generated by the $M+N$ randomly drawn letters. Instead of taking these correlations into account, we will assume that $s(r,t) = \eta(r,t)$ with independent random variables $\eta(r,t)$ given by

$$\eta(r,t) = \begin{cases} 1 & \text{with probab. } p \\ 0 & \text{with probab. } 1-p \end{cases} \tag{30}$$

with

$$\Pr\{\forall_{r,t}\, \eta(r,t) = \eta_{r,t}\} = \prod_{r,t} \Pr\{\eta(r,t) = \eta_{r,t}\}. \tag{31}$$

To model sequences randomly drawn with equal probability from an alphabet of size c, we take $p = 1/c$. The approximation (31) is known to change characteristic quantities of sequence alignment only slightly [5]. We will confirm numerically at the end of this paper, that this also holds for the values of λ which we are mainly interested in here. For our choices of parameters, the global alignment algorithm (17) reads

$$h(r,t+1) = \max\{h(r,t-1) + \eta(r,t), h(r+1,t), h(r-1,t)\}. \tag{32}$$

B. Choice of the alignment lattice geometry

In order to handle finite size effects better, we will use a rectangular geometry (Fig. 6) for the alignment lattice, instead of the triangular geometry shown in Fig. 4(a). We will further apply periodic boundary condition to the top and bottom edges of the lattice, i.e., $h(0,t) = h(2W,t)$ for a rectangular lattice of width $2W$, and will start on the left edge with the initial conditions $h(2k+1, t=0) = h(2k, t=1) = 0$. Note that despite the different lattice geometries, the score $h(r,t)$ for all points with $t \leq W$ on the rectangular lattice will be *identical* to the score at the same (r,t) coordinate on the triangular lattice[3]; see Fig. 6.

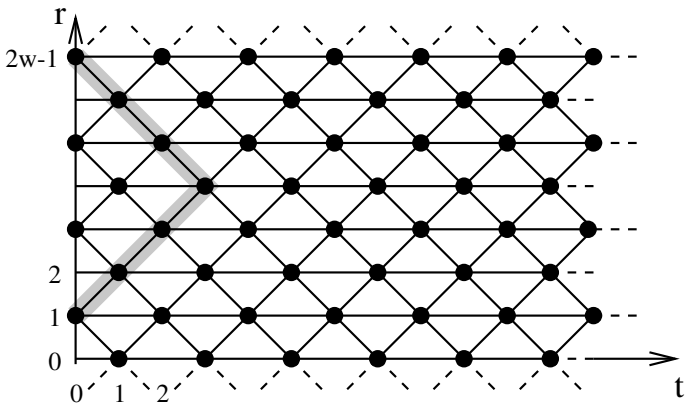

FIG. 6: Rectangular alignment lattice of width $2W$ with periodic boundary conditions in the spatial (vertical) direction. We use this lattice instead of the triangular lattice shown in Fig. 4(a) in order to simplify the handling of finite-size effects. As indicated by the thick gray lines, the score at a point with $t \leq W$ as the one at the tip of the triangle is identical with the corresponding score calculated on a triangular lattice as the one shown in Fig. 4(a).

[3] Since directed polymers in a random medium are known to have a wandering exponent $\zeta = 2/3$ this actually still holds for $t < W^{3/2}$.

C. The dynamics of sequence alignment as an asymmetric exclusion process

In this section we will perform a change of variables on the sequence alignment algorithm (32) for the rectangular lattice shown in Fig. 6. We will find that the resulting problem is equivalent to an asymmetric exclusion process on a one-dimensional lattice of width $2W$. As a guidance towards the choice of suitable variables, we take the knowledge from the (continuous) KPZ equation that the gradient of the surface height is an especially simple quantity. At a fixed time, the gradients at different positions become uncorrelated and Gaussian distributed [4, 41]. Thus, we will look at their discrete analogs in the alignment problem. They are the score differences between neighboring lattice points and thus located on the diagonal bonds of the lattice. We will parameterize these score differences by the bond variables $n(r,t)$. They will later turn out to be the occupation numbers of the sites of an asymmetric exclusion process. With the choice of coordinates as illustrated in Fig. 7(a), we define them to be[4]

$$n(r,t) \equiv \begin{cases} h(r+1,t) - h(r,t+1) + 1 & \text{for } r+t \text{ even} \\ h(r+1,t+1) - h(r,t) & \text{for } r+t \text{ odd} \end{cases} \tag{33}$$

As explained in detail in App. B, rewriting the time evolution equation (32) in terms of the variables $n(r,t)$ leads to a time evolution equation of $n(r,t)$ alone, without any reference to the absolute scores $h(r,t)$. Moreover, this time evolution equation implies that the score differences take only the values $n(r,t) \in \{0,1\}$. By the structure of the alignment lattice as a composition of elements as the one shown in Fig. 7(a), the resulting time evolution for the $n(r,t)$ transforms a pair $(n(r-1,t-1), n(r,t-1)) \in \{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$ into the new pair $(n(r-1,t), n(r,t)) \in \{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$ independently from all the other $n(r',t-1)$. This transformation only depends on the single random variable $\eta(r,t)$ and can be expressed by the transfer matrix

$$\mathsf{T}_1(0) \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1-p & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{34}$$

in the basis $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$. We can thus interpret the action of the lattice element shown in Fig. 7(a) as a "device" like the one shown in Fig. 7(b) which takes a pair of variables (n_1', n_2') as its inputs, applies the transfer matrix $\mathsf{T}_1(0)$, and generates a new pair of variables (n_1, n_2) as its outputs.

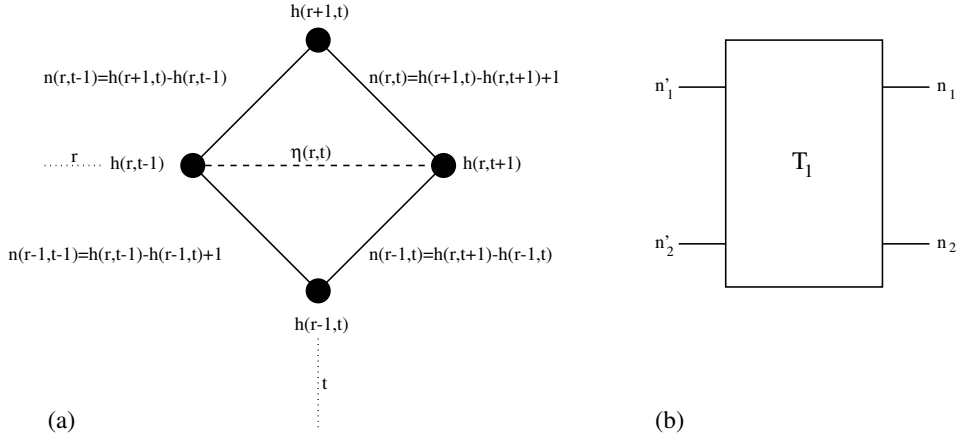

FIG. 7: One building block of the alignment lattice. By our numbering scheme of the lattice r and t are either both even or both odd. (a) shows the scores at the lattice points and the bond variables $n(r,t)$. (b) shows this building block as a "device", which takes two incoming bond variables n_1' and n_2' and transforms them with the help of the transfer matrix T_1 into the new bond variables n_1 and n_2.

[4] Note, that the $n(r,t)$ are not literally score differences but suitably chosen parameterizations of these score differences. This complication is necessary in order to enable the interpretation as the particle occupation numbers in the asymmetric exclusion process.

We recognize the action of the transfer matrix $\mathsf{T}_1(0)$ as the elementary time step of an asymmetric exclusion process, if we interpret the $n(r,t)$ as particle occupation numbers on a one-dimensional lattice of $2W$ sites with periodic boundary conditions as the one shown in Fig. 8. Each of these sites can either be empty or occupied by a single particle. In each time step for each pair of neighboring sites, a particle hops to the right with some probability $1-p$, if the site to its right is empty according to the non vanishing entry $|10\rangle \to |01\rangle$ of the transfer matrix $\mathsf{T}_1(0)$. If there is no particle or if the site on the right is already occupied the configuration remains unchanged.

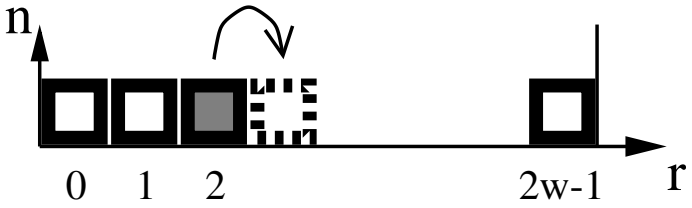

FIG. 8: Interpretation of the transfer matrix $\mathsf{T}_1(0)$ as given in Eq. (34) as an asymmetric exclusion process. A configuration of the local score differences is represented by particles on a one-dimensional lattice of width $2W$. At an odd time step for each even site $r-1$ a particle hop is attempted with probability $1-p$. In our example, the particle at site 0 cannot hop, since site 1 is already occupied. The particle on site 2 can hop to site 3 as indicated by the dashed square.

In terms of the elementary devices shown in Fig. 7(b) the lattice structure of Fig. 6 can be depicted schematically as shown in Fig. 9. Thus, the process of hopping a particle to the right is attempted for each even numbered site at odd time steps and for each odd numbered site at even time steps. This hopping dynamics is exactly the asymmetric exclusion process with sublattice-parallel updating with periodic boundary conditions[5] [10, 42].

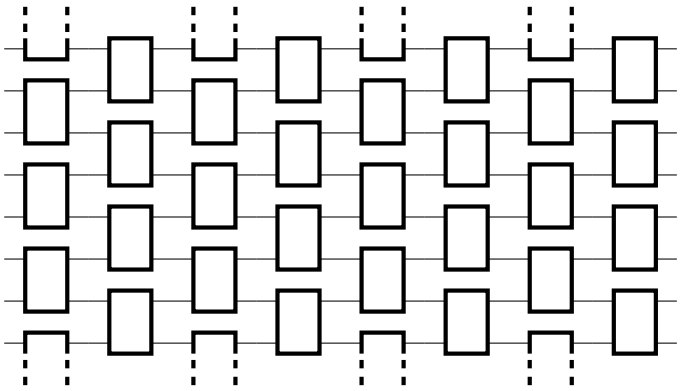

FIG. 9: Schematic representation of the alignment lattice of Fig. 6 as an "electric circuit". The boxes represent elements of the type shown in Fig. 7(b). They take two particle occupation numbers as their "inputs" and generate two new particle occupation numbers as their "outputs". Their interconnection into a layered structure as shown here with a shifted pairing scheme in every time step leads to the non-trivial behavior of sequence alignment.

In reducing the dynamics from a dynamics of scores into a dynamics of the occupation numbers $n(r,t)$, one has to pay attention to the boundary conditions. Periodic boundary conditions for the $n(r,t)$ do not automatically lead to meaningful periodic boundary conditions for the scores $h(r,t)$. We thus have to impose the additional constraint that the total sum of the local score differences across the whole lattice vanishes. In terms of our bond variables $n(r,t)$ this translates into the condition

$$\frac{1}{2W}\sum_{r=0}^{2W-1} n(r,t) = \frac{1}{2}, \tag{35}$$

[5] If we had chosen the "hard wall" boundary conditions $h(-1,t) = h(2W,t) = \infty$ instead of the periodic boundary conditions $h(2W,t) = h(0,t)$ for the score, we would have arrived at the asymmetric exclusion process with sublattice-parallel updating and *open* boundary conditions at a feeding and extinction rate of $\alpha = \beta = 1-p$ at the two ends of the lattice with $2W-1$ sites respectively.

i.e., the system of hopping particles is at half filling. Since the number of particles is conserved under the dynamics described by the transfer matrix $\mathsf{T}_1(0)$, the condition (35) is guaranteed to hold if we choose the initial conditions $\sum_{r=0}^{2W-1} n(r,t=0)/2W = 1/2$. Particle densities different from one half would correspond to a tilted "score profile" $h(r,t)$ at each fixed time t.

V. THE GENERATING FUNCTION

A. Expressing the generating function in terms of the hopping process

We now want to apply the mapping between sequence alignment and the asymmetric exclusion process to the practical problem of assessing alignment significance. As noted in Sec. III, this amounts to calculating the generating function

$$Z_0(\lambda;N) \equiv \langle \exp[\lambda h(0,N)] \rangle_0, \tag{36}$$

where $\langle \ldots \rangle_0$ denoted the average over the ensemble of uncorrelated disorder defined by Eqs. (30) and (31). Thus, we first need to express the total score $h(0,N)$ in terms of the occupation numbers $n(r,t)$. As explained in more detail in App. B, $h(0,t)$ is on average incremented by $1/2W$ every time the transfer matrix $\mathsf{T}_1(0)$ is applied except for the transition $|01\rangle \to |10\rangle$. Thus, $Z_0(\lambda;N)$ can be expressed as

$$Z_0(\lambda;N) = \exp[\lambda N/2] \langle \exp[-\lambda J] \rangle_0 \tag{37}$$

in terms of the total number of particle hops per lattice site

$$J \equiv \frac{1}{2W} \sum_{l=1}^{N/2} \sum_{k=0}^{W-1} (j(2k+1,2l-1) + j(2k,2l)), \tag{38}$$

where $j(r,t) \in \{0,1\}$ is the number of particle hops at lattice site (r,t). We thus need to determine the generating function

$$Q(\lambda;W,N) \equiv \langle \exp[-\lambda J] \rangle_0 \tag{39}$$

for the asymmetric exclusion process. Note, that this is different from the generating function of the local current $j(r,t)$: since J/N is the *time and space averaged* current, Q contains information on spatial and temporal *correlations* in the number of hopping particles which the generating function for the local current does not.

B. The Generating function as an eigenvalue problem

Now we will reformulate the calculation of the generating function $Q(\lambda;W,N)$ for the asymmetric exclusion process as an eigenvalue problem. As already mentioned, $\exp[-\lambda J]$ is a product of factors $\exp[-\lambda/2W]$ for every particle that hops. Since the dynamics of the hopping process is described by the transfer matrix $\mathsf{T}_1(0)$ defined in Eq. (34), we can calculate $Q(\lambda;W,N)$ by associating a weight $\exp[-\lambda/2W]$ to the element of the transfer matrix $\mathsf{T}_1(0)$ which corresponds to a hop. This can be derived more formally from a dynamics path integral representation of $Z_0(\lambda;N)$ as detailed in App. C. We get the modified local transfer matrix

$$\mathsf{T}_1\left(\frac{\lambda}{W}\right) \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & (1-p)e^{-\frac{\lambda}{2W}} & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{40}$$

in the basis $|00\rangle, |01\rangle, |10\rangle, |11\rangle$ of a pair of neighboring lattice sites.

Next, we need to take into account the special lattice structure of Fig. 9. We note that at every even time step the lattice is decomposed into W of the building blocks described by T_1. Thus, a single time step of the total system at even time is described by the matrix

$$\mathsf{T}_{even} \equiv \mathsf{T}_W(\lambda) \equiv \bigotimes_{k=1}^{W} \mathsf{T}_1\left(\frac{\lambda}{W}\right). \tag{41}$$

At odd times the dynamics is the same, but the pairing of neighboring sites is shifted. To generate the time evolution at odd time steps, we can thus shift all particles to the right, apply the dynamics of even time steps and then shift all particles back to the left. Let C be the translation operator such that

$$\mathsf{C}|n_0 n_1 \dots n_{2W-1}\rangle \equiv |n_1 \dots n_{2W-1} n_0\rangle \tag{42}$$

which shifts all particles by one site to the left taking into account the periodic boundary conditions. With this definition we can write $\mathsf{T}_{odd} = \mathsf{C}\mathsf{T}_W(\lambda)\mathsf{C}^{-1}$.

The sublattice-parallel updating procedure (i.e., the structure of the lattice as depicted by Fig. 9) finally leads to

$$Q(\lambda; W, N) = \langle\psi_1|(\mathsf{T}_{even}\mathsf{T}_{odd})^{N/2}|\psi_0\rangle = \langle\psi_1|(\mathsf{T}_W(\lambda)\mathsf{C}\mathsf{T}_W(\lambda)\mathsf{C}^{-1})^{N/2}|\psi_0\rangle \tag{43}$$

where $|\psi_0\rangle$ is a 4^W dimensional state vector representing the initial conditions, and $\langle\psi_1|$ is the 4^W dimensional vector whose entries are all 1, used here to denote a summation over all possible final configurations. In the limit of large $N \gg W$, this obviously becomes

$$Q(\lambda; W, N) = \rho_W^N(\lambda) \tag{44}$$

where $\rho_W^2(\lambda)$ is the eigenvalue of $\mathsf{T}_W(\lambda)\mathsf{C}\mathsf{T}_W(\lambda)\mathsf{C}^{-1}$ with the largest real part. Since this matrix has no negative entries and is irreducible for non-pathological choices of the scoring matrix (while restricted to the physical sector of half filling), the largest eigenvalue of this matrix is guaranteed by the Perron Frobenius theorem to be non degenerate and real, and its eigenvector can be chosen without negative entries. When $\lambda = 0$, we have $\rho(0) = 1$ and its eigenvector is the stationary distribution of the asymmetric exclusion process, which is a simple tensor product of independent occupation numbers. This is no longer the case for $\lambda \neq 0$.

C. Calculating the largest eigenvalue

For a finite W, it is in principle possible to solve for the largest eigenvalue of the 4^W dimensional matrix $\mathsf{T}_W(\lambda)\mathsf{C}\mathsf{T}_W(\lambda)\mathsf{C}^{-1}$ by directly diagonalizing the matrix. It is convenient to reduce the size of this matrix by exploiting some symmetries. Since the lattice is translationally invariant with respect to shifts in r by 2, we expect the same symmetry of the largest eigenvalue of $\mathsf{T}_W(\lambda)\mathsf{C}\mathsf{T}_W(\lambda)\mathsf{C}^{-1}$. Thus, for the purpose of computing the largest eigenvalue we can restrict ourselves to the subspace $\mathcal{C}$ of translationally invariant vectors

$$\mathcal{C} \equiv \left\{ |\psi\rangle \middle| C^2|\psi\rangle = |\psi\rangle \right\}. \tag{45}$$

This corresponds to a discrete Fourier transform of the matrix $\mathsf{T}_W(\lambda)\mathsf{C}\mathsf{T}_W(\lambda)\mathsf{C}^{-1}$ and choosing the $k = 0$ component. On $\mathcal{C}$, we have $\mathsf{C}^{-1} = \mathsf{C}$ by definition. Thus, it is enough to look for the largest eigenvalue $\rho_W(\lambda)$ of the matrix $\mathsf{T}_W(\lambda)\mathsf{C}$ restricted to $\mathcal{C}$. A further restriction which helps reducing the size of the matrix is the mirror symmetry of the lattice which has to be respected by the eigenvector as well. Additionally, $\mathsf{T}_W(\lambda)\mathsf{C}$ has to be restricted onto the physical subspace of half filling.

After applying these simplifications, the largest eigenvalue can be calculated for small widths W using computer algebra. Although the matrix $\mathsf{T}_W(\lambda)\mathsf{C}$ explicitly contains the quantity $\exp[-\lambda/2W]$, it turns out that the characteristic polynomial depends only on $\exp[-\lambda/2]$. This is a consequence of the translational invariance of the lattice[6]. In order to reveal the underlying structure of the largest eigenvalues for different W, it is very useful to *expand* the resulting largest eigenvalues $\rho_W(\lambda)$ in powers of this quantity $e^{-\lambda/2}$. We get

$$\begin{aligned}
W = 1: \quad & \rho_1(\lambda) &&= \sqrt{p} + O(e^{-\frac{\lambda}{2}}) \\
W = 2: \quad & \rho_2(\lambda) &&= \sqrt{p} - (p-1)e^{-\frac{\lambda}{2}} + O((e^{-\frac{\lambda}{2}})^2) \\
W = 3: \quad & \rho_3(\lambda) &&= \sqrt{p} - (p-1)e^{-\frac{\lambda}{2}} + (p-1)\sqrt{p}(e^{-\frac{\lambda}{2}})^2 + O((e^{-\frac{\lambda}{2}})^3) \\
W = 4: \quad & \rho_4(\lambda) &&= \sqrt{p} - (p-1)e^{-\frac{\lambda}{2}} + (p-1)\sqrt{p}(e^{-\frac{\lambda}{2}})^2 - (p-1)\sqrt{p}^2(e^{-\frac{\lambda}{2}})^3 + O((e^{-\frac{\lambda}{2}})^4),
\end{aligned}$$

---

[6] Instead of looking at the average score $\overline{h}(N) = \frac{1}{2W}\sum_r h(r, N)$ as we do in the derivation of Eq. (43) in App. C, we could also have chosen a specific position, say $r = 0$ and $r = 1$, and monitored the behavior of the score $\widetilde{h}(N) \equiv \frac{1}{2}[h(1, N) + h(0, N-1)]$. Since the differences between scores at the same time are bounded, these two quantities must have the same generating function for large N. The transfer matrix which calculates the generating function for $\widetilde{h}(N)$ is $\widetilde{\mathsf{T}}(\lambda) \equiv \mathsf{T}_1(\lambda) \otimes \bigotimes_{k=2}^{W} \mathsf{T}_1(0)$ instead of $\mathsf{T}_W(\lambda)$. It has the technical disadvantage that it breaks the translational invariance, but it explicitly depends only on $\exp[-\lambda/2]$ instead of $\exp[-\lambda/2W]$.

where the $O((e^{-\lambda/2})^k)$ terms denote terms of the given order with prefactors which are different for different W. We can see that the coefficients up to order $(e^{-\lambda/2})^{W-1}$ remain unchanged upon increasing W and they constitute the beginning of a simple geometric series. Assuming that this pattern holds for arbitrary orders, we can resum the series for any *fixed* $\lambda > 0$ and get

$$\rho(\lambda) \equiv \lim_{W\to\infty} \rho_W(\lambda) = \frac{\sqrt{p} + e^{-\frac{\lambda}{2}}}{1 + \sqrt{p}e^{-\frac{\lambda}{2}}}. \tag{46}$$

Combined with Eqs. (37), (39), and (44) this yields the generating function

$$Z_0(\lambda; N) = \exp[\lambda N/2]\rho^N(\lambda) = (\exp[\lambda/2]\rho(\lambda))^N = \left(\frac{1 + \sqrt{p}\exp[\frac{\lambda}{2}]}{1 + \sqrt{p}\exp[-\frac{\lambda}{2}]}\exp[-\frac{\lambda}{2}]\right)^N \tag{47}$$

in the limit of large N.

Eq. (47) can be easily generalized to the match-mismatch scoring system given in Eq. (2) with a gap cost $\delta = \mu/2$ for an arbitrary value of μ. If we denote the score in this scoring system by $h'(r,t)$ it is connected to the score $h(r,t)$ of the scoring system with $\mu = \delta = 0$ by the simple global rescaling and shifting

$$h'(r,t) = (1+\mu)h(r,t) - \frac{\mu}{2}t. \tag{48}$$

Thus the corresponding generating function is given by

$$Z(\lambda, \mu; N) \equiv \langle e^{\lambda h'(0,N)}\rangle = e^{-\mu N}\langle e^{\lambda(1+\mu)h(0,N)}\rangle. \tag{49}$$

If we again neglect correlations and use uncorrelated random variables

$$\eta(r,t) = \begin{cases} 1 & \text{with probab. } p \\ -\mu & \text{with probab. } 1-p \end{cases} \tag{50}$$

the same rescaling and shifting leads to

$$Z_0(\lambda, \mu; N) \equiv \langle e^{\lambda h'(0,N)}\rangle_0 = \left(\frac{1 + \sqrt{p}\exp[\frac{\lambda}{2}(1+\mu)]}{1 + \sqrt{p}\exp[-\frac{\lambda}{2}(1+\mu)]}\exp[-\frac{\lambda}{2}\mu]\right)^N. \tag{51}$$

D. Connections to related work

The distribution of the height of a surface governed by KPZ dynamics has been of quite some recent interest. On the one hand, a generating function very closely related to Eq. (51) has been calculated [13] in the context of an asymmetric exclusion process. While Derrida *et al.* are able to calculate the full dependence on the finite width W, they restrict themselves to the simpler case of *continuous time* which is not an option for our problem since we are given the discrete lattice.

On the other hand, an explicit distribution of the height distribution in specific growth models has been derived [37] and shown to be connected to the eigenvalue distributions of random matrix ensembles. Prähofer and Spohn use a mapping between the surface height of a growth model and the length of the *longest increasing subsequence* of a *random permutation*. The longest increasing subsequence problem can be interpreted as the alignment problem of a permutation of the numbers $1, 2, 3, \ldots, N$ to the sequence of the ordered numbers $1, 2, 3, \ldots, N$. Thus, there are only N matches on a lattice of size $N \times N$ and no symbol of one sequence matches to more than one symbol of the other sequence. Interpreting the N matches as nucleation events, a growing surface can be constructed the height of which is precisely the length of the longest increasing subsequence. Applied to disorder $\eta(r,t)$ which fulfills the constraints of the longest increasing subsequence problem, i.e., exactly one match for every symbol in each of the sequences, the mapping presented in this paper essentially reduces to the mapping used by Prähofer and Spohn. In this case, the vanishing density of matches in the limit $N \to \infty$ allows Prähofer and Spohn to use a continuum limit which again simplifies the calculations. However, the alignment problem deals with a finite alphabet and the order of possible matches is proportional to N^2. Moreover, each letter in one sequence can (and will) match an extensive number of letters in the other sequence. In this case, the detailed mapping presented in this paper has to be used.

As far as results are concerned, the studies by Derrida *et al.* and by Prähofer and Spohn both come to the conclusion that the generating function or the distribution of surface heights respectively takes a *universal* form in

the limit $W \to \infty$ which we are interested in. However, this form is much more complicated than our simple result Eq. (51). This is due to a different order of taking limits. Derrida *et al.* take the limit $W \to \infty$ of the generating function while keeping $\lambda W^{1/2}$ constant in order to obtain their universal distribution, i.e., they simultaneously take the limits $W \to \infty$ *and* $\lambda \to 0$ in some controlled way. Prähofer and Spohn directly look at the distribution of the surface height which is defined by the properties of the generating function at $\lambda = 0$. However, the expansion of $\rho_W(\lambda)$ in terms of $e^{-\lambda/2}$ that we used is *not valid any more* in the limit $\lambda \to 0$. Since our main interest is in solving Eq. (21) for λ which results in a *finite* result of λ our expression Eq. (51) is appropriate. It is an expression for the generating function *beyond* the regime in which it was found to be universal by Derrida *et al.*. Similarly, the universal infinite W surface height distribution found by Prähofer and Spohn, corresponds to the same scaling limit as Derrida *et al.*'s result after exchanging the regularization through a finite width W by a regularization through a finite time t. It also contains all the terms which vanish in the limit $W \to \infty$ at fixed λ but come into play if λ vanishes simultaneously. There is no reason to assume the result Eq. (51) to be universal. This is supported by the explicit dependence of Eq. (51) on the parameter p. Eq. (51) has to be calculated taking the discreteness of the lattice into full account as shown in this publication.

VI. IMPLICATIONS ON DIRECTED POLYMERS AND SEQUENCE ALIGNMENT

Now, we will study the consequences of our main result Eq. (51). First, we will discuss the general properties of the generating function and its implications on the physics of directed polymers in a random medium. Then, we will come back to our original question of the assessment of sequence alignment significance. We find, that Eq. (51) is an explicit expression for the significance assessment parameter λ. It reproduces known limiting cases and we will demonstrate that our result agrees well with numerical simulations.

A. Properties of the generating function

The most notable property of the generating function of the connected moments of the average score (or average height)

$$\log\langle \exp[\lambda h'(0,N)]\rangle_0 = \log Z_0(\lambda,\mu;N) \tag{52}$$

is that it is an *odd* function of λ. The first two terms of its expansion are

$$\frac{\log Z_0(\lambda,\mu;N)}{N} = v(\mu)\lambda + \frac{1}{6}b(\mu)\lambda^3 + O(\lambda^5). \tag{53}$$

where

$$v(\mu) = \left.\frac{\mathrm{d}}{\mathrm{d}\lambda}\right|_{\lambda=0} [Z_0(\lambda,\mu;N)]^{\frac{1}{N}} = -\frac{\mu}{2} + (1+\mu)\frac{\sqrt{p}}{1+\sqrt{p}} \tag{54}$$

and $b(\mu) = \left(\frac{1+\mu}{1+\sqrt{p}}\right)^3 \frac{(1-\sqrt{p})\sqrt{p}}{4} > 0$. As already mentioned, we can regard the generating function $Z_0(\lambda,\mu;N)$ as the ensemble averaged partition function of λ replicas of a directed polymer in a random medium. In this sense, Eq. (53) is the free energy per length of this λ replica system. It has the same form (with a vanishing quadratic term) as the result of an earlier explicit replica calculation in continuous time and continuous space [12]. However, our analysis is directly on the discrete model and is not plagued by the difficulty of taking the continuum limit in [12].

The vanishing of the second order term in λ will not even be affected by the universal contributions to our result for small λ which have been found in [13] using the explicit dependence on the width W, since its second order coefficient vanishes as $W^{-1/2}$ in the limit of large width. The consequence of this vanishing second order term in λ is that the second connected moment of the average height, i.e., the height fluctuations, scales sublinear in N. Instead the *third* moment of the height fluctuations scales linearly with N. This is a signature of the presence of the anomalous $N^{1/3}$ fluctuations of the average surface height characteristic for the KPZ universality class.

B. Statistical significance and the log-linear transition

According to Eqs. (21) and (51) the parameter λ which characterizes the statistical significance of local alignments with the match-mismatch scoring scheme Eq. (2) and gap cost $\delta = \mu/2$ is given by the unique positive solution of the

equation

$$\frac{1+\sqrt{p}\exp[\frac{\lambda}{2}(1+\mu)]}{1+\sqrt{p}\exp[-\frac{\lambda}{2}(1+\mu)]}\exp[-\frac{\lambda}{2}\mu]=1. \tag{55}$$

In the limit of large μ, the solution of Eq. (55) converges to $\lambda=-\log p$. This is the value which we expect since this limit corresponds to the case of gapless alignment (recall that $\delta=\mu/2$ here), and $\lambda=-\log p$ is the solution of the large μ limit of Eq. (7). If the gap cost is decreased, λ is reduced, too. At some critical value of μ there will not be any positive solution of Eq. (55) any more, i.e., islands of all sizes are equally probable. This indicates a phase transition between the logarithmic and the linear alignment phase. The approach of this phase transition is especially interesting.

Close to the phase transition, we can use the expansion (53) and rewrite Eq. (55) as

$$v(\mu)\lambda+\frac{1}{6}b(\mu)\lambda^3+O(\lambda^5)=0. \tag{56}$$

From this expansion the origin of the phase transition is very clear: If $v(\mu)>0$, the right hand side of Eq. (56) is a monotonously increasing function of λ. Thus, $\lambda=0$ is the only solution of Eq. (56). This corresponds to a flat distribution of island sizes, i.e., the linear alignment phase. If $v(\mu<0)$, the shape of the right hand side of Eq. (56) changes and there are three roots, one of which is the positive solution

$$\lambda\approx\left(-6\frac{v(\mu)}{b(\mu)}\right)^{1/2}. \tag{57}$$

This indicates that we are in the logarithmic alignment phase. Thus, the phase transition occurs at the critical mismatch cost μ_c which is defined by the condition

$$v(\mu_c)=0. \tag{58}$$

Using the explicit form (54) of $v(\mu)$, we get the critical mismatch cost

$$\mu_c=\frac{2\sqrt{p}}{1-\sqrt{p}}. \tag{59}$$

This reproduces the already known result [24] for the phase transition point of this model. As the mismatch cost μ approaches this critical value from above, λ vanishes as

$$\lambda\approx\left(\frac{6(1-\sqrt{p})^3}{\sqrt{p}(1+\sqrt{p})}\right)^{1/2}(\mu-\mu_c)^{1/2}. \tag{60}$$

In the case of finite width W, the above expression is valid down to $\lambda\sim W^{-1}$. This confirms the characteristic universal power law $|\mu-\mu_c|^{1/2}$ proposed previously [7] by scaling arguments.

C. Numerical Verification

In order to test the approximation of uncorrelated local disorder (31) and the heuristic elements of the derivation of Eq. (55), we performed extensive numerical simulations to corroborate our result. We used the DNA alphabet of size $c=4$ with identical frequencies for all four letters, i.e., $p=1/4$. For different choices of the mismatch cost μ with corresponding gap cost $\delta=\mu/2$, we used the island method [31] to find the values of λ as a function of μ numerically. For each value of δ several billion islands have been generated using sequences of $N=25,000$ in order to achieve relative errors of approximately 1%. We used completely uncorrelated local scores chosen as

$$s(r,t)=\begin{cases}1 & \text{with probab. } p\\ -\mu & \text{with probab. } 1-p\end{cases} \tag{61}$$

with $p=1/4$. The resulting values of λ are shown in Fig. 10. The solid line is the solution of Eq. (55) and the circles represent the values of λ for uncorrelated local scores (31). As shown in Fig. 10 the observed λ's follow the analytic solution very closely, thereby confirming Eq. (55). We also included the values of λ which result from correlated local scores generated from aligning randomly chosen sequences according to Eq. (2). As one can see, they deviate only slightly from the analytical result for uncorrelated disorder. This deviation is strongest close to the log-linear phase transition, which for uncorrelated disorder happens at $\mu_c=2$. The difference of $\sim 2\%$ in μ_c between the correlated and the uncorrelated case rapidly becomes much smaller for larger alphabet sizes c [40].

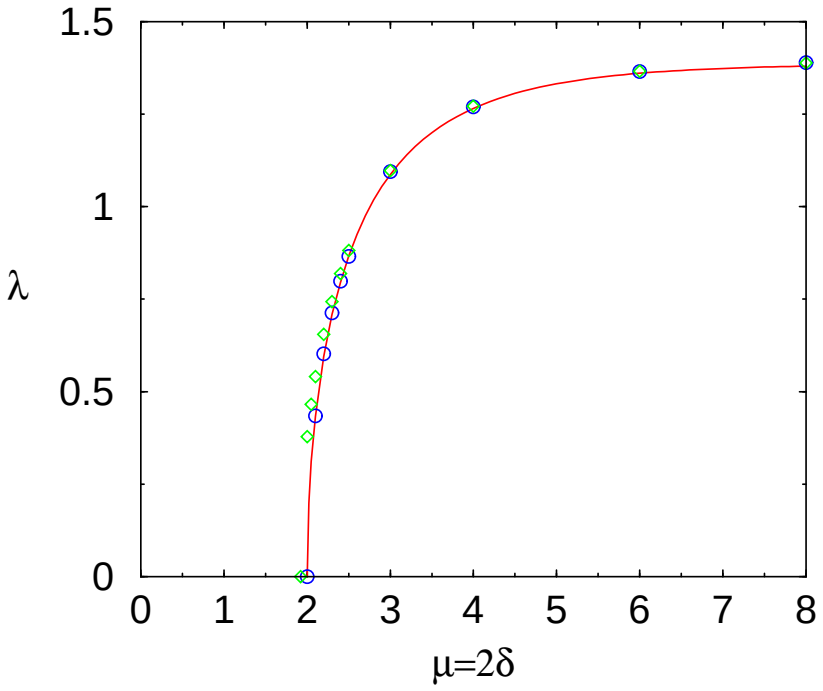

FIG. 10: Dependence of the significance parameter λ on the scoring parameter μ. The circles represent the numerically obtained values of λ for uncorrelated local disorder (31) with match probability $p = 1/4$ for which Eq. (55) (the solid line) has been derived. They agree well with the analytical result. The diamonds correspond to local disorder generated by comparing two randomly chosen sequences over an alphabet of size $c = 4$. The values of λ obtained from the two ensembles differ from each other only very close to the phase transition point $\mu_c = 2$.

VII. MORE GENERAL SCORING SYSTEMS

While the approximation of the ensemble of random sequences by the ensemble of independent local scores appears to have negligible effects, our treatment is so far limited to the special scoring system Eq. (30). While the computation of the generating function $\langle \exp[-\lambda J] \rangle_0$ seems feasible only for this special scoring system, the mapping to an asymmetric exclusion process and the reformulation as an eigenvalue problem is still possible for more general scoring systems.

We consider here scoring systems satisfying the following two conditions: First, the differences between the possible values $s_{a,b}$ of the scoring matrix are multiples of some score unit Δ. Second, the gap costs δ is such that $2\delta + s_0$ is also an integer multiple of Δ, with

$$s_0 \equiv \max_{a,b}\{s_{a,b}\} \tag{62}$$

being the maximal entry of the scoring matrix $s_{a,b}$. These two conditions are easily satisfied (with $\Delta = 1$) by the most frequently used protein scoring systems [16, 17] which use integer scores and gap costs for performance reasons. For the match–mismatch scoring system (2), the first condition is satisfied with $\Delta = 1 + \mu$, while the second condition applies only to a discrete set of δ's. However, it is possible in principle to interpolate to arbitrary gap costs [24].

Mapping to an asymmetric exclusion process is possible for scoring systems satisfying the above two conditions. It will be convenient to express the gap cost δ in the following way,

$$2\delta = n_{max}\Delta - s_0 \quad \text{with } n_{max} \in \mathbf{N}. \tag{63}$$

As before, we shall ignore correlations between the local scores $s(r,t)$ and introduce uncorrelated random variables $\eta(r,t) \in \{0, 1, \ldots\}$ such that

$$s(r,t) \equiv s_0 - \eta(r,t)\Delta, \tag{64}$$

i.e.,

$$\Pr\{\forall_{r,t}\, \eta(r,t) = \eta_{r,t}\} = \prod_{r,t} \Pr\{\eta(r,t) = \eta_{r,t}\} \tag{65}$$

with

$$\Pr\{\eta(r,t)=\eta\}=\sum_{a,b}p_a p_b \delta_{s_{a,b},s_0-\eta\Delta}. \tag{66}$$

Note, that these random variables $\eta(r,t)$ only take on a finite number of different positive integer values, since the scoring matrix $s_{a,b}$ itself has only a finite number of entries.

A derivation analogous to the one given above for the longest common subsequence problem again maps the dynamics of the alignment algorithm onto the dynamics of particles on a one-dimensional lattice. The state of the system is still given by the number of particles $n(r,t)$ at each lattice site, but now these occupation numbers are defined as

$$n(r,t)\equiv\begin{cases}\frac{1}{\Delta}[h(r+1,t)-h(r,t+1)+\delta+s_0] & r+t\ even\\ \frac{1}{\Delta}[h(r+1,t+1)-h(r,t)+\delta] & r+t\ odd\end{cases} \tag{67}$$

and can take any integer value between 0 and n_{max}. The dynamics is given by the relations

$$n(r-1,t) = n(r-1,t-1)-j(r,t) \text{ and} \tag{68}$$

$$n(r,t) = n(r,t-1)+j(r,t) \tag{69}$$

for even $r+t$, where

$$j(r,t)\equiv\min\{\eta(r,t),n_{max}-n(r,t-1),n(r-1,t-1)\} \tag{70}$$

and the total number of particles is fixed to be

$$\frac{1}{2W}\sum_{r=0}^{2W-1}n(r,t)=\frac{n_{max}}{2}. \tag{71}$$

Eqs. (68)-(70) can be equally expressed as the following cellular automata: For each time step and for each pair of neighboring sites of the one-dimensional lattice the particles live on,

1. choose an integer number $\eta\geq 0$ of particles to hop from site $r-1$ to site r according to the distribution (66)
2. if there are fewer particles than η on site $r-1$, then reduce η to the number of particles on site $r-1$
3. if there are fewer free spaces than η on site r, then reduce η to the number of free spaces on site r
4. move η particles from site $r-1$ to site r

This updating rule is to be applied sublattice-parallel as for the simpler scoring system. The process is illustrated in Fig. 11.

The more complicated hopping process is reflected in a different matrix $\mathsf{T}_1(\lambda/W)$ without changing anything else in the calculations. Thus, the significance assessment constant λ is still given by the generating function of the space and time averaged current as

$$\exp[\lambda s_0/2]\langle\exp[-\lambda\Delta J]\rangle_0^{\frac{1}{N}}=1 \tag{72}$$

but the calculation of this generating function for an arbitrary distribution (66) becomes much more difficult for the generalized asymmetric exclusion process than for the case $n_{max}=1$ of the original asymmetric exclusion process.

However, already the knowledge of the dependence of the average current on the scoring parameters would be very helpful to biologists, since this determines the position of the log-linear phase transition. As discussed in the case of the simpler scoring system, the phase transition occurs, if the first moment of the score distribution vanishes, i.e., for

$$0=\left.\frac{\mathrm{d}}{\mathrm{d}\lambda}\right|_{\lambda=0}\exp[\lambda s_0/2]\langle\exp[-\lambda\Delta J]\rangle_0^{\frac{1}{N}}=s_0/2-\frac{\langle J\rangle_0}{N}\Delta=s_0/2-\langle j\rangle_0\Delta \tag{73}$$

The average current is much easier to calculate, since in contrast to the generating function, it is independent of temporal correlations. Thus, it can be calculated from the knowledge of the stationary state alone. For the original asymmetric exclusion process, the occupation numbers of the stationary state become independent random variables. For the generalized asymmetric exclusion process presented here, this is not the case any more. If the number of particles which hop in one move is at most one (as for the scoring system (2) with arbitrary gap costs) approximating the stationary state as a product state still yields reasonable values of $\langle j\rangle_0$ and hence the phase transition point (δ_c,μ_c) [24] (see Fig. 3.) Nevertheless, exact results or at least systematic improvements taking into account the spatial correlations of the occupation numbers would be desirable. For the more general case allowing for an arbitrary number of particles to hop at a given time, no analytical result is known.

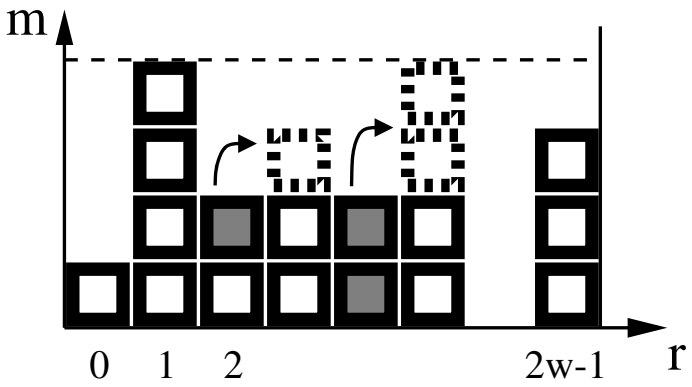

FIG. 11: Interpretation of Eqs. (68)-(70) as a generalized asymmetric exclusion process. A configuration of the local score differences is represented by particles on a one-dimensional lattice of width $2W$. Each lattice site can accommodate up to n_{max} particles (here $n_{max} = 4$.) At an odd time step for each even site $r-1$, a number of particles is chosen to attempt hopping to the right. If there are enough particles at site $r-1$ and enough space on site r, the chosen number hops. In the example shown, the filled particles are the ones to hop and the dashed boxes show their positions after the time step. No particle which could hop is on site 6. The particle on site 0 cannot hop since its destination site is already fully occupied. For site 2, one particle has been chosen. On site 4, at least two particles tried to hop. If the number chosen was larger, it would have been cut down to two since there are only two particles on site 4 and since there are only two spaces left at site 5.

VIII. CONCLUDING REMARKS

In this paper, we have shown how a question of great practical importance to molecular biologists, like the significance assessment of local sequence alignment results, can be answered by studying the asymmetric exclusion process, an exactly solvable model of the KPZ universality class. Conversely, in trying to answer this question for biologists, we derived an important physical quantity like the generating function Z for the corresponding physical system in discrete time and discrete space. This complements the existing solutions in continuous time and space [12] and in continuous time and discrete space [13]. Our result is the first successful analytical approach to assessing the statistical significance of sequence alignment with gaps.

Future work of practical importance includes solving the generalizations of the asymmetric exclusion process described in Sec. VII and studying the effect of the widely used "affine gap cost", where a contiguous gap of length ℓ is assigned some gap cost $\delta + (\ell - 1)\epsilon$ instead of simply $\delta\ell$. A general expression which gives λ as a function of an arbitrary scoring system should finally give rise to a deeper understanding of the role of the gap cost and lead to better choices of scoring systems for alignments of biological sequences.

Acknowledgments

The author gratefully acknowledges discussions with S. Altschul, A. Dembo, T. Hwa, R. Olsen, and N. Rajewsky, and the hospitality of the Center for Studies in Physics and Biology at Rockefeller University where this work was completed. This work is supported in part by a Hochschulsonderprogramm III fellowship of the DAAD and by the NSF through Grant No. DMR-9971456.

APPENDIX A: ISLAND HIGH SCORE DISTRIBUTION

In this appendix we derive heuristically the exponential distribution of maximal island scores. We first treat the gapless case [43] and then generalize the derivation to alignment with gaps. In the gapless case, the distribution of large islands of length L measured from their beginning to their peak point at height σ is given by

$$p(\sigma) = \langle \delta(\sigma - \sum_{i=1}^{L} s(i)) \rangle. \tag{A1}$$

Using the Fourier representation of the delta function and the statistical independence of the $s(i)$ this yields

$$p(\sigma) = \frac{1}{2\pi} \int \exp(-ik\sigma) \langle \exp(iks) \rangle^L \mathrm{d}k. \tag{A2}$$

If we assume that the peak score of the island is proportional to its length, i.e., that an island has on average a linear slope α, we get

$$p(\sigma) = \frac{1}{2\pi} \int \exp(-ik\alpha L) \langle \exp(iks) \rangle^L \mathrm{d}k, \tag{A3}$$

which can be evaluated in a saddle point approximation as

$$p(\sigma) \sim \exp(-\lambda\sigma) \tag{A4}$$

with

$$\lambda = ik^* - \log[\langle \exp(ik^* s) \rangle]/\alpha. \tag{A5}$$

The saddle point k^* is given by the saddle point equation

$$\frac{\langle s \exp(ik^* s) \rangle}{\langle \exp(ik^* s) \rangle \alpha} = 1. \tag{A6}$$

This k^* is itself a function of the so far unknown slope α. To find the correct value of α, we minimize Eq. (A5) with respect to α and get together with Eq. (A6)

$$\langle \exp(ik^* s) \rangle = 1. \tag{A7}$$

Inserting this into Eq. (A5) yields condition (7). Additionally we get from Eq. (A6) the typical slope α of an island as

$$\alpha = \langle s \exp(\lambda s) \rangle. \tag{A8}$$

For alignment with gaps, the high score of an island of length L from its beginning to its peak point is not just the sum of local scores any more. Instead, it is given by the final score $h(0, L)$ of a global alignment of two sequences of length L taking into account all possible insertions of gaps. We can still use the Fourier transformation to get

$$p(\sigma) = \langle \delta(\sigma - h(0, L)) \rangle = \frac{1}{2\pi} \int \exp(-ik\sigma) \langle \exp(ikh(0, L)) \rangle \mathrm{d}k. \tag{A9}$$

In Sec. V B we will see, that $\langle \exp(\lambda h(0, L)) \rangle$ is for large L the L'th power of the eigenvalue of some matrix. We thus define $\rho(\lambda)$ by

$$\langle \exp[\lambda h(0, L)] \rangle \equiv \rho^L(\lambda) \tag{A10}$$

and again assume a linear slope α of the islands which we conveniently define by $\sigma = \alpha L/2$ in order to take into account that the lattice of length L actually only contains $L/2$ matches or mismatches in a row. We then get

$$p(\sigma) = \frac{1}{2\pi} \int \exp[(-ik\alpha/2 + \log \rho(ik))L] \mathrm{d}k. \tag{A11}$$

Applying the above saddle point approximation and maximization with respect to the slope of the island α yields Eq. (21). Moreover it gives the typical slope of an island as

$$\alpha = 2\frac{\rho'(\lambda)}{\rho(\lambda)} = \frac{2}{L} \langle h(0, L) \exp[\lambda h(0, L)] \rangle. \tag{A12}$$

APPENDIX B: EXPRESSION OF THE SCORE DYNAMICS IN TERMS OF PARTICLE OCCUPATION NUMBERS

In this appendix we describe the mapping from the evolution equation (32) of the sequence alignment scores onto the asymmetric exclusion process with the $n(r,t)$ as the particle occupation numbers in detail. To this end we apply Eq. (32) to the definition Eq. (33) of $n(r,t)$, where we assume by convention that $r+t$ is even as in Fig. 7(a). We get

$$\begin{aligned} n(r-1,t) &= h(r,t+1) - h(r-1,t) \\ &= \max\{h(r,t-1)+\eta(r,t), h(r-1,t), h(r+1,t)\} - h(r-1,t)\} \\ &= h(r,t-1) - h(r-1,t) + 1 + \max\{\eta(r,t)-1, h(r-1,t)-h(r,t-1)-1, h(r+1,t)-h(r,t-1)-1\} \\ &= n(r-1,t-1) + \max\{\eta(r,t)-1, -n(r-1,t-1), n(r,t-1)-1\} \\ &= n(r-1,t-1) - \min\{1-\eta(r,t), n(r-1,t-1), 1-n(r,t-1)\} \end{aligned}$$

and analogously

$$\begin{aligned} n(r,t) &= h(r+1,t) - h(r,t+1) + 1 \\ &= h(r+1,t) - \max\{h(r,t-1)+\eta(r,t), h(r-1,t), h(r+1,t)\} + 1 \\ &= h(r+1,t) - h(r,t-1) - \max\{\eta(r,t)-1, h(r-1,t)-h(r,t-1)-1, h(r+1,t)-h(r,t-1)-1\} \\ &= n(r,t-1) - \max\{\eta(r,t)-1, -n(r-1,t-1), n(r,t-1)-1\} \\ &= n(r,t-1) + \min\{1-\eta(r,t), n(r-1,t-1), 1-n(r,t-1)\}. \end{aligned}$$

This can be summarized in the form

$$n(r-1,t) = n(r-1,t-1) - j(r,t) \quad \text{and} \tag{B1}$$

$$n(r,t) = n(r,t-1) + j(r,t), \tag{B2}$$

where

$$j(r,t) \equiv \min\{1-\eta(r,t), 1-n(r,t-1), n(r-1,t-1)\}. \tag{B3}$$

As we can see, there is no reference to the actual alignment scores $h(r,t)$ in these equations. As a first consequence of these equations we note that they imply that the variables $n(r,t)$ can only take on the values zero and one. This is obvious by induction, if it is fulfilled at $t=0$ as it is the case for our choice of initial conditions[7]. Thus, it is reasonable to interpret the $n(r,t)$ as particle occupation numbers.

Moreover, we note that a pair of neighboring occupation numbers $(n(r-1,t), n(r,t))$ at time t depends only on the corresponding pair $(n(r-1,t-1), n(r,t-1))$ at time $t-1$ and the random variable $\eta(r,t)$. Thus, the elements as the one shown in Fig. 7 perform these transformations of a pair of neighboring occupation numbers into a new pair of neighboring occupation numbers completely independently from each other.

Looking at Eqs. (B1)-(B3) more closely, we see that $j(r,t)=0$ whenever $(n(r-1,t-1), n(r,t-1)) \in \{|00\rangle, |01\rangle, |11\rangle\}$. Thus, $(n(r-1,t), n(r,t)) = (n(r-1,t-1), n(r,t-1))$ in these cases. Only if site $r-1$ is occupied and site r is empty, the number $j(r,t)$ of transfered particles can be one with probability $\Pr\{\eta(r,t)=0\} = 1-p$. This leads to the interpretation of the dynamics given by Eqs. (B1)-(B3) as an asymmetric exclusion process described by the transfer matrix $\mathsf{T}_1(0)$ defined in Eq. (34) of the main text.

So far we transformed the dynamics of the sequence alignment algorithm as given by Eq. (32) into an asymmetric exclusion process. We still have to express $Z_0(\lambda;N)$ in terms of this asymmetric exclusion process. To achieve this, we first define for any "time" t the average score (or space-averaged surface height)

$$\overline{h}(t) \equiv \begin{cases} \frac{1}{2W}\sum_{k=0}^{W-1}[h(2k,t-1)+h(2k+1,t)] & t\ even \\ \frac{1}{2W}\sum_{k=0}^{W-1}[h(2k,t)+h(2k+1,t-1)] & t\ odd \end{cases} \tag{B4}$$

Because of the translational invariance of the system in the spatial (r) direction we get

$$Z_0(\lambda;N) = \langle \exp[\lambda h(0,N)]\rangle_0 = \langle \exp[\lambda \overline{h}(N)]\rangle_0. \tag{B5}$$

---

[7] Even if the initial values of the $n(r,t=0)$ are not zero or one they will under the dynamics Eqs. (B1)-(B3) eventually try to take on values less than zero or larger than one. The minimum in Eq. (B3) then resets them to zero or one. Thus, after some startup phase, the $n(r,t)$ will be integer even if their initial values are chosen to be non-integer.

Thus, we can restrict ourselves to calculating the large N behavior of the latter quantity.

The change in the average score $\overline{h}(t)$ can be expressed in terms of the occupation numbers $n(r,t)$ via Eqs. (32) and (33). It is given by

$$\overline{h}(t+1)-\overline{h}(t)=\begin{cases}\frac{1}{2W}\sum_{k=0}^{W-1}[h(2k,t+1)-h(2k,t-1)] & t\ even\\ \frac{1}{2W}\sum_{k=0}^{W-1}[h(2k+1,t+1)-h(2k+1,t-1)] & t\ odd\end{cases}. \tag{B6}$$

The local score differences in this equation can for even $r+t$ be expressed as

$$\begin{aligned}h(r,t+1)-h(r,t-1) &= \max\{h(r,t-1)+\eta(r,t),h(r+1,t),h(r-1,t)\}-h(r,t-1)\\ &= 1+\max\{\eta(r,t)-1,n(r,t-1)-1-n(r-1,t-1)\}\\ &= 1-\min\{1-\eta(r,t),1-n(r,t-1),n(r-1,t-1)\}\\ &= 1-j(r,t).\end{aligned}$$

Inserting this into Eq. (B6) yields

$$\overline{h}(t+1)-\overline{h}(t)=\frac{1}{2}-\frac{1}{2W}\begin{cases}\sum_{k=0}^{W-1}j(2k,t) & t\ even\\ \sum_{k=0}^{W-1}j(2k+1,t) & t\ odd\end{cases}. \tag{B7}$$

Combining Eqs. (B5) and (B7) finally yields

$$\begin{aligned}Z_0(\lambda;N) &= \langle\exp[\lambda\overline{h}(N)]\rangle_0=\langle\exp[\lambda\sum_{t=0}^{N-1}(\overline{h}(t+1)-\overline{h}(t))]\rangle_0 \qquad \text{(B8)}\\ &= \exp[\lambda N]\langle\exp[-\frac{\lambda}{2W}\sum_{l=1}^{N/2}\sum_{k=0}^{W-1}(j(2k+1,2l-1)+j(2k,2l))]\rangle_0\\ &= \exp[\lambda N]\langle\exp[-\lambda J]\rangle_0,\end{aligned}$$

where

$$J\equiv\frac{1}{2W}\sum_{l=1}^{N/2}\sum_{k=0}^{W-1}(j(2k+1,2l-1)+j(2k,2l)) \tag{B9}$$

is the total number of particles hopped divided by the number of sites. This is Eq. (39) of the main text.

APPENDIX C: DYNAMIC PATH INTEGRAL REPRESENTATION

In this appendix we want to show that the generating function $Q(\lambda;W,N)$ can be expressed as a product of some 4^W dimensional matrices as stated in Eq. (43) in the main text. This rewriting is crucial in transforming the calculation of the generating function into an eigenvalue problem. We start from the definition

$$Q(\lambda;W,N)=\langle\exp[-\lambda J]\rangle_0=\langle\prod_{l=1}^{N/2}\prod_{k=0}^{W-1}e^{-\frac{\lambda}{2W}j(2k+1,2l-1)}e^{-\frac{\lambda}{2W}j(2k,2l)}\rangle_0. \tag{C1}$$

Since, the number of particles in each bin must be either 0 or 1 at any time, we do not change the expectation value, if we introduce ones of the form

$$1=\sum_{\{n_{r,t}\}\in\{0,1\}^{2W}}\prod_{r=0}^{2W-1}\delta_{n(r,t),n_{r,t}} \tag{C2}$$

at each fixed time t. This corresponds to a path integral formulation of the quantity $Q(\lambda;W,N)$ and yields

$$\begin{aligned}&\langle\exp[-\lambda J]\rangle_0= \qquad \text{(C3)}\\ &\sum_{\{n_{r,0}\}}\cdots\sum_{\{n_{r,N}\}}\langle\prod_{r=0}^{2W-1}\delta_{n(r,0),n_{r,0}}\prod_{l=1}^{N/2}\left(\prod_{r=0}^{2W-1}\delta_{n(r,2l-1),n_{r,2l-1}}\right)\left(\prod_{k=0}^{W-1}e^{-\frac{\lambda}{2W}j(2k+1,2l-1)}\right)\times\\ &\qquad\times\left(\prod_{r=0}^{2W-1}\delta_{n(r,2l),n_{r,2l}}\right)\left(\prod_{k=0}^{W-1}e^{-\frac{\lambda}{2W}j(2k,2l)}\right)\rangle_0\end{aligned}$$

Once a configuration of the particles at each time step is fixed, the expectation value can be factorized into the parts which contain only a single random variable $\eta(r,t)$

$$\langle \prod_{r=0}^{2W-1} \delta_{n(r,0),n_{r,0}} \prod_{l=1}^{N/2} \left(\prod_{r=0}^{2W-1} \delta_{n(r,2l-1),n_{r,2z-1}} \right) \left(\prod_{k=0}^{W-1} e^{-\frac{\lambda}{2W} j(2k+1,2l-1)} \right) \left(\prod_{r=0}^{2W-1} \delta_{n(r,2l),n_{r,2l}} \right) \left(\prod_{l=0}^{W-1} e^{-\frac{\lambda}{2W} j(2k,2l)} \right) \rangle_0 =$$

$$\prod_{r=0}^{2W-1} \delta_{n(r,0),n_{r,0}} \times$$

$$\times \prod_{l=1}^{N/2} \prod_{k=0}^{W-1} \langle \delta_{n(2k,2l-2),n_{2k,2l-2}} \delta_{n(2k+1,2l-2),n_{2k+1,2l-2}} e^{-\frac{\lambda}{2W} j(2k+1,2l-1)} \delta_{n(2k,2l-1),n+2k,2l-1} \delta_{n(2k,2l-1),n_{2k+1,2l-1}} \rangle_0 \times$$

$$\times \prod_{k=0}^{W-1} \langle \delta_{n(2k-1,2l-1),n_{2k-1,2l-1}} \delta_{n(2k,2l-1),n_{2k,2l-1}} e^{-\frac{\lambda}{2W} j(2k,2l)} \delta_{n(2k-1,2l),n_{2k-1,2l}} \delta_{n(2k,2l),n_{2k+1,2z}} \rangle_0 \times 1.$$

Inserting this into Eq. (C3) we can interpret the summation over the possible configurations of the particles at each time step as the summation over inner indices in a matrix multiplication. In this language the first term $\prod_{r=0}^{2W-1} \delta_{n(r,0),n_{r,0}}$ is a vector on the 4^W dimensional vector space indexed by all possible particle configurations. This vector has exactly one non vanishing entry at the configuration which is chosen as the initial configuration at $t = 0$. This non vanishing entry is one and we call this vector $|\psi_0\rangle$. The factor of one which we added for the sake of clarity also plays the role of a vector the entries of which are all one. We call this vector $\langle\psi_1|$. All the other factors represent matrices. There is one matrix for every time step and each of these matrices is a tensor product of W identical matrices describing an elementary hopping process. Their matrix elements are

$$\left(\mathsf{T}_1 \left(\frac{\lambda}{W} \right) \right)_{(n_1,n_2),(n_1',n_2')} \equiv \langle \delta_{n(r-1,t-1),n_1'} \delta_{n(r,t-1),n_2'} \exp[-\frac{\lambda}{2W} j(r,t)] \delta_{n(r-1,t),n_1} \delta_{n(r,t),n_2} \rangle_0. \tag{C4}$$

The disorder average here is over one single random variable $\eta(r,t)$. Performing this disorder average yields the matrix $\mathsf{T}_1(\lambda/W)$ as defined in Eq. (40). The matrices for even time steps and the matrices for odd time steps are shifted against each other by one lattice unit which finally leads to the expression of Eq. (43).

---

[1] M.S. Waterman, *Introduction to Computational Biology*, Chapman & Hall, (London, UK, 1994).
[2] D.F. Feng and R.F. Doolittle, *Methods in Enzymology* **266**, 368–382 (1996).
[3] T. Hwa and M. Lässig, *Phys. Rev. Lett.* **76**, 2591–2594 (1996).
[4] M. Kardar, G. Parisi, and Y.-C. Zhang, *Phys. Rev. Lett.* **56**, 889–892 (1986).
[5] D. Drasdo, T. Hwa, and M. Lässig, *J. Comp. Biol.* **7**, 115–141 (2000).
[6] T. Hwa and M. Lässig, *Proceedings of the Second Annual International Conference on Computational Molecular Biology*, S. Istrail *et al.* eds., 109–116, ACM press, (New York, NY, 1998).
[7] D. Drasdo, T. Hwa, and M. Lässig, *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, J. Glasgow *et al.*, eds., 52–58, AAAI Press, (Menlo Park, CA, 1998).
[8] J. Krug, *Phys. Rev. Lett.* **67**, 1882-1885 (1991).
[9] B. Derrida, *Phys. Rep.* **301**, 65–83 (1998) and references therein.
[10] D. Kandel, E. Domany, and B. Nienhuis, *J. Phys. A* **23**, L755–L762 (1990).
[11] L.H. Gwa and H. Spohn, *Phys. Rev. Lett.* **68**, 725–728 (1992); *Phys. Rev. A* **46**, 844–854 (1992).
[12] M. Kardar, *Phys. Rev. Lett.* **55**, 2235–2238 (1985); *Nucl. Phys. B* **290**, 582–602 (1987).
[13] B. Derrida and J.L. Lebowitz, *Phys. Rev. Lett.* **80**, 209–213 (1998); B. Derrida and C. Appert, *J. Stat. Phys.* **94**, 1–30 (1999).
[14] S.F. Altschul *et al.*, *J. Mol. Biol.* **215**, 403–410 (1990).
[15] S.B. Needleman and C.D. Wunsch, *J. Mol. Biol.* **48**, 443–453 (1970).
[16] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt, in *Atlas of Protein Sequence and Structure*, M.O. Dayhoff and R.V. Eck, eds., **5** supp. 3, 345–358 (1978).
[17] S. Henikoff and J.G. Henikoff, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919 (1992).
[18] S. Karlin and A. Dembo, *Adv. Appl. Prob.* **24**, 113–140 (1992).
[19] S. Karlin and S.F. Altschul, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 5873–5877 (1993).
[20] E.J. Gumbel, *Statistics of Extremes*, Columbia University Press, (New York, NY, 1958).
[21] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, John Wiley & Sons, (New York, NY, 1978).
[22] W.R. Pearson, *Genomics* **11**, 635–650 (1991).

[23] T.F. Smith and M.S. Waterman, *Adv. Appl. Math.* **2**, 482–489 (1981).
[24] R. Bundschuh and T. Hwa, *Disc. Appl. Math.* **104**, 113–142 (2000).
[25] T.F. Smith, M.S. Waterman, and C. Burks, *Nucleic Acids Research* **13**, 645–656 (1985).
[26] J.F. Collins, A.F.W. Coulson, and A. Lyall, *CABIOS* **4**, 67–71 (1988).
[27] R. Mott, *Bull. Math. Biol.* **54**, 59–75 (1992).
[28] M.S. Waterman and M. Vingron, *Stat. Sci.* **9**, 367–381 (1994).
[29] M.S. Waterman and M. Vingron, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4625–4628 (1994).
[30] S.F. Altschul and W. Gish, *Methods in Enzymology* **266**, 460–480 (1996).
[31] R. Olsen, R. Bundschuh, and T. Hwa, *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, T. Lengauer *et al.*, eds., 211–222, AAAI Press, (Menlo Park, CA, 1999).
[32] R. Mott and R. Tribe, *J. Comp. Biol.* **6**, 91–112 (1999).
[33] R. Mott, *J. Mol. Biol.* **300**, 649–659, (2000).
[34] D. Siegmund and B. Yakir, *Ann. Stat.* **28** 657–680 (2000).
[35] R. Arratia and M.S. Waterman, *Ann. Appl. Probab.* **4**, 200–225 (1994).
[36] Y. Zhang, *Ann. App. Probab.* **5**, 1236–1240 (1995).
[37] M. Prähofer and H. Spohn, *Physica A* **279**, 342–352 (2000); *Phys. Rev. Lett.* **84**, 4882–4885 (2000).
[38] V. Chvátal and D. Sankoff, *J. Applied Probab.* **12**, 306–315 (1975).
[39] V. Dančík, *Expected Length of longest common subsequences*, PhD thesis, University of Warwick (1994), and references therein.
[40] J. Boutet de Monvel, *Europ. Phys. J. B* **7**, 293–308 (1999).
[41] J. Krug and H. Spohn, in *Solids far from Equilibrium: Growth, Morphology, and Defects*, C. Godreche, ed., 479, Cambridge University Press (Cambridge, UK, 1991).
[42] N. Rajewsky, L. Santen, A. Schadschneider, and M. Schreckenberg, *J. Stat. Phys.* **92** 151–194 (1998).
[43] D. Fisher, private communication.