An Analytic Approach to Significance Assessment in Local Sequence Alignment with Gaps

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

October 25, 1999

Abstract

A detailed study of the Smith-Waterman alignment algorithm is performed in order to find an analytical approach to the problem of assessing the statistical significance of local alignments with gaps. The significance is shown to be given in terms of an eigenvalue equation which captures the dynamics of the much simpler global alignment algorithm. This eigenvalue equation is then explicitly solved for a simple scoring system and the resulting significance estimations are verified by a comparison to extensive numerical simulations.

Keywords: sequence alignment, statistical significance, Gumbel distribution

1 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 [32, 10]. There are two classes of alignment algorithms. The simpler gapless alignment as it was implemented, e.g., in the original BLAST [1] is very fast and theoretically very well understood. However, in order to detect weakly homologous sequences, gaps have to be allowed in an alignment [25] which leads to the more sophisticated Smith-Waterman algorithm [28].

Both types of alignment algorithms have the drawback that they will find an optimal alignment and an optimal score for *any* pair of sequences — even randomly chosen and thus completely unrelated ones. Thus, it is necessary to assess the significance of a resulting alignment. A popular approach to this problem is to compare the score of the optimal alignment to the scores generated by the optimal alignments of *randomly chosen* sequences. This is quantified by the p- or E-value. This comparison steadily becomes more important since with the increasing size of the databases the probability for obtaining a relatively large score just by chance increases dramatically.

In order to be able to quote a p-value the *distribution* of optimal alignment scores for alignments of random sequences has to be known. In the case of gapless alignment it has been worked out rigorously [17, 18, 19] that this distribution is a Gumbel or extreme value distribution [14]. It is characterized by two parameters which depend on the scoring system used and on the amino acid frequencies with which the random sequences are generated. For gapless alignment also this dependence of the two Gumbel parameters on the scoring system is completely known.

For the case of gapped alignment, there is no theory which describes the distribution of alignment scores of random sequences. However, there is a lot of numerical evidence that the distribution is still of the Gumbel form [29, 8, 20, 30, 31, 2]. Nevertheless, it has turned out to be a very challenging problem to find the two Gumbel parameters for a given scoring system. The straightforward method is to generate a large number of alignment scores by shuffling the two sequences which are to be compared and take a histogram of this distribution. But, because of the slow exponential tail of the Gumbel distribution, this method is extremely time consuming. In practice, the two Gumbel parameters have to be pre-computed for some few fixed scoring systems [2, 3].

The necessity of pre-computing the Gumbel parameters clearly becomes problematic, if adaptive schemes like, e.g., PSI-BLAST [3] are being used. These schemes change their scoring system recursively depending on the sequence data they are confronted with and thus have to be able to find the two Gumbel parameters after each update of the scoring system from scratch. Another aspect which crucially depends on the real-time availability of the Gumbel parameters is the assessment of the statistical significance taking into account the amino acid composition of the individual sequences compared. Pre-computed values of the two Gumbel parameters are only correct for sequences which follow the overall amino acid frequencies which have been used in the computation of the Gumbel parameters.

To remedy this problem, a more effective numerical method which estimates the two Gumbel parameters has recently been proposed [24]. There are also some analytical approximations [21, 27, 22] which are mainly valid for rather large gap costs where the influence of the gaps on the Gumbel parameters is not yet too strong. In this paper, we will present a new analytical scheme to calculate the Gumbel parameter λ which describes the *tail* of the Gumbel distribution. First, we will show, that although λ is intrinsically a quantity characterizing *local* alignment it can be calculated from solely studying the much simpler *global* alignment algorithm. Under some very moderate approximation we then reformulate the problem of finding λ in terms of an eigenvalue equation which can be thought of as being a generalization of the eigenvalue equation for the case of sequences generated from a Markov model [18]. We then show the feasibility of our novel approach by calculating a *closed formula* for λ which we show numerically to be very close to the real values over the *whole range* of gap costs. Since the Gumbel parameter λ characterizes the more difficult tail of the Gumbel distribution it should be easy to extract the other Gumbel parameter characterizing its mean numerically from, e.g., the island method [24], once λ is known.

2 Review of Sequence 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 [1].

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

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

which is used for DNA sequence comparisons [23]. For protein sequences, usually the more sophisticated 20×20 PAM [9] or BLOSSUM matrices [15] are used. The computational task is to find the i, j, and ℓ which give the *highest* total score

$$\Sigma \equiv \max_{\mathcal{A}} S[\mathcal{A}] \tag{2}$$

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 dynamic programming algorithm

$$S_{i,j} = \max\{S_{i-1,j-1} + s_{a_i, b_j}, 0\}, \tag{3}$$

with the initial condition $S_{0,k} = 0 = S_{k,0}$. Once the $S_{i,j}$ are calculated, the global optimal score is obtained as $\Sigma = \max_{1 \le i \le M, 1 \le j \le N} S_{i,j}$.

In order to characterize the statistical significance of the alignment, it is necessary to know the distribution of Σ for the gapless alignment 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

$$\Pr\{\Sigma < S\} = \exp(-\kappa e^{-\lambda S}). \tag{4}$$

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 mean. For gapless alignment, these 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

$$\langle \exp(\lambda s) \rangle \equiv \sum_{a,b} p_a p_b \exp(\lambda s_{a,b}) = 1. \tag{5}$$

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. (4) and its parameters, we will 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 (3) without loss of generality. Eq. (3) then becomes

$$S_{i,i} \equiv S(i) = \max\{S(i-1) + s(i), 0\}, \tag{6}$$

where the variables $s(i) \equiv s_{a,b}$ are independent random variables with the identical distribution $\Pr\{s_i > s\} = \sum_{\{a,b | s_{a,b} > s\}} p_a p_b$.

The dynamics of the evolution equation (6) can be in two distinct phases. The quantity which distinguishes these two phases is the expected local similarity score $\langle s \rangle \equiv \sum_{a,b} p_a p_b s_{a,b}$. If it is

positive, the score $S(i)$ will increase on average. Thus, the zero option in Eq. (6) can effectively be omitted in this case and the dynamics becomes 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 a Gaussian instead of an extreme value distribution.

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.

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

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 $S = 0$. 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 using the method of steepest descent and optimizing over the length L of the islands, we asymptotically obtain a Poisson distribution

$$\Pr\{\sigma_k > \sigma\} = e^{-\lambda\sigma} \tag{7}$$

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 (5).

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

$$\Sigma = \max_k\{\sigma_k\}, \tag{8}$$

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 [14, 13]. In the case of a large number κ of independent island peak scores each of which obeys the Poisson distribution Eq. (7), the connection is especially simple and we get

$$\Pr\{\Sigma < S\} = \Pr\{\max\{\sigma_1, \ldots, \sigma_\kappa\} < S\} = \Pr\{\sigma_1 < S\}^\kappa = (1 - e^{-\lambda S})^\kappa \approx [\exp(-e^{-\lambda S})]^\kappa = \exp(-\kappa e^{-\lambda S}), \tag{9}$$

i.e., the parameter λ of the island peak score distribution Eq. (7) is the same as the parameter λ in the Gumbel distribution Eq. (4) of the maximal alignment scores.

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 [25]. Here, we will specifically consider Smith-Waterman local alignment [28]. 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 $S[\mathcal{A}] = \sum_{(a,b)\in\mathcal{A}} s_{a,b} - \delta N_{\text{g}}$ 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". Although in practical applications usually the more complicated affine gap cost is used, we will in the following concentrate on this somewhat easier case.

The task of local alignment is again to find the alignment $\mathcal{A}$ with the highest score as in Eq. (2), in this enlarged class of possible alignments. This can be very efficiently done by a dynamic programming method which becomes obvious in the alignment path representation [23]. 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(a) where we chose for simplicity $M = N$. 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 $s(r,t) \equiv s_{a_i,b_j}$ 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(a). This figure also shows the way we use the coordinates r and t to address points on the alignment lattice. The coordinate transformation from the base numbers i and j to the "space" and "time" variables r and t will later turn out to be crucial for understanding the alignment algorithm as a "dynamic process" which in turn allows us to estimate the statistical significance.

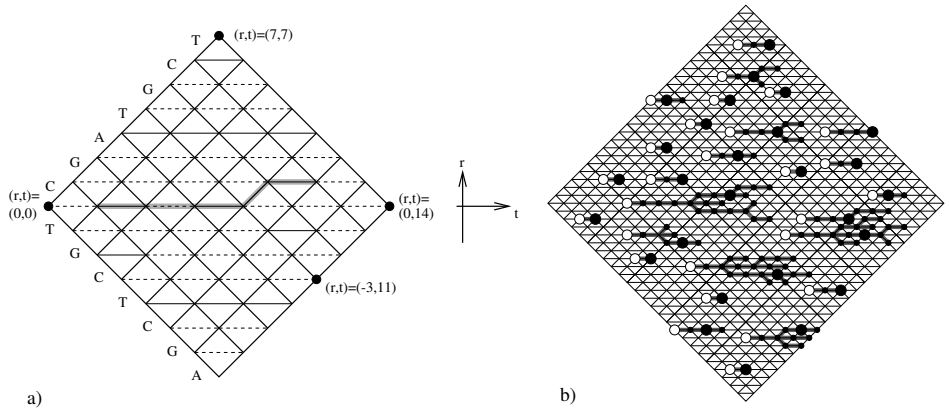

Figure 2: Local alignment of two sequences. (a) shows the alignment of $CGATGCT$ and $TGCTCGA$ represented as a directed path on the alignment lattice: the diagonal bonds correspond to gaps in the alignment. The horizontal bonds represent aligned pairs. Alignments of identical letters (matches) are shown 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. (b) is a 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 (11) are highlighted. Together they are the restricted optimal path associated with each point with a positive score. Each of these paths goes back on 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.

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 dynamic programming algorithm [23]

$$h(r,t+1) = \max\{h(r,t-1) + s(r,t),\ h(r+1,t) - \delta, h(r-1,t) - \delta\}. \tag{10}$$

If we are interested in *local* alignments, the same trick as in the gapless case (3) is used. Cutting off unfavorable scores by adding the choice of zero in the maximum of Eq. (10) leads to the Smith-Waterman algorithm [28]

$$S(r,t+1) = \max\{S(r,t-1) + s(r,t), S(r+1,t) - \delta, S(r-1,t) - \delta, 0\}. \tag{11}$$

The score of the best local alignment is then given by $\Sigma = \max_{r,t} S(r,t)$.

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 $u(\{s_{a,b}\}, \delta_c) + \langle s \rangle = 0$. The loci of the phase transition δ_c for alignment of random sequences is only known approximately [5] for the simple scoring system Eq. (1).

If the parameters are chosen such that $u + \langle s \rangle < 0$ 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 [29, 8, 20, 30, 31, 2]. The values of the parameters κ and λ are only known approximately for a few cases close to the gapless limit [21, 27, 22]. In practice, they have to be determined empirically by time consuming simulations. Below we will present an explicit calculation of the parameter λ for a simple scoring system.

3 Significance Estimation using Global Alignment

As a first and very crucial 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 (10). Later, we will derive an explicit formula by studying global alignment alone.

Let us define the generating function

$$Z_L(\lambda) \equiv \langle \exp[\lambda h(0, L)] \rangle \tag{12}$$

where the brackets $\langle \cdot \rangle$ denote the ensemble average 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. 3(a). It can be obtained from the recursion relation (10) with the initial condition $h(2k, 0) = h(2k + 1, 1) = 0$. It will turn out that the parameter λ of the Gumbel distribution is obtained from the condition

$$\lim_{L \to \infty} Z_L(\lambda) = 1. \tag{13}$$

Note, that this condition reduces simply to Eq. (5) in the case of gapless alignment, since for infinite gap cost δ, we have $\langle \exp[\lambda h(0, L)] \rangle = \langle \exp[\lambda \sum_{k=1}^{L/2} s(0, 2k - 1)] \rangle = \langle \exp[\lambda s] \rangle^{L/2}$.

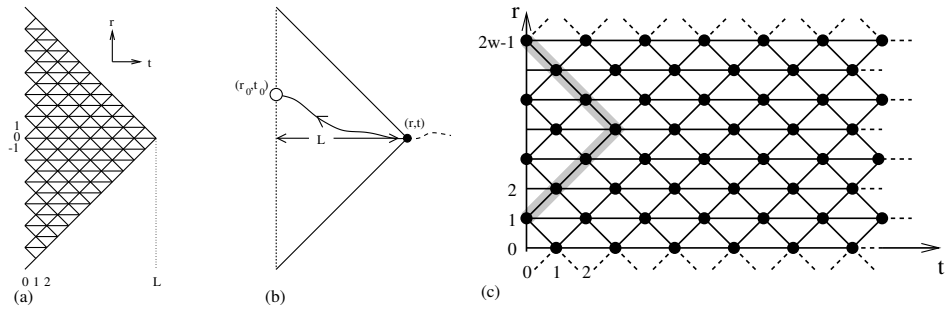

Figure 3: Global alignment lattice used for significance estimation. (a) shows the right half of the lattice from Fig. 2(a). It can represent all possible paths of length L which end at the point $(r, t) = (0, L)$ and start at $(r, 0)$ for an arbitrary r. (b) shows such a path schematically. It represents 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. (c) shows with the gray lines, how the triangular lattice similar to the one shown in (a) can be embedded into a rectangular alignment lattice of width $2W$ with periodic boundary conditions in the spatial (vertical) direction as long as $L < W$.

The key observation which leads to the result (13) 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* [24]. By the construction of the local alignment algorithm (11), 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 (11). 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(a). The path must start at some point (r_0, t_0) where a positive score is created out of the zero sea 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. 2(b). Each of these islands has a maximum score which we denote by σ_k as we did in the gapless case. By this definition, every lattice point with a positive score belongs to exactly one island. Thus, the maximal score Σ on the total lattice is given by Eq. (8). Since large islands are well separated by a sea of points with score zero, they can be treated as statistically independent clusters on the alignment lattice[1]. Thus, their maximal scores σ_k are again independent identically distributed random variables which yield a Gumbel distribution of Σ via Eq. (9). Our task is thus to calculate the distribution of the island peak scores σ_k in the presence of gaps.

This distribution of maximal island scores can be derived analogously to the gapless case (App. A.) Within an island we can neglect the lower cutoff 0 in the local alignment algorithm (11). Thus, an island with gaps corresponds to a *global gapped alignment* of some length L as the one shown schematically in Fig. 3(b). Analogous to the gapless case the length L of this island has to be optimized for each given peak score value σ. Then, the island peak distribution again has an asymptotically Poissonian form (7) with the decay constant λ given by Eq. (13). The condition $\lim_{L\to\infty}\langle\exp[\lambda h(0,L)]\rangle = 1$ can be understood as being the choice for λ where the exponential enhancement $\exp[\lambda h(0,L)]$ of events with $h(0,L) > 0$ exactly balances the exponential rareness of these events in the calculation of the expectation value $\langle\exp[\lambda h(0,L)]\rangle$.

4 Global Alignment as a Dynamic Process

From now on we will only study *global* alignment as described by Eq. (10) and finally use Eq. (13) to infer the value of the parameter λ characterizing local alignment. In order to simplify the discussion we will make three assumptions on the scoring parameters $s_{a,b}$ and δ. First, the differences between the possible values $s_{a,b}$ of the scoring matrix have to be multiples of some score unit Δ. This holds for nearly any practically used scoring system. For the match–mismatch scoring system as given by Eq. (1) it is trivial with $\Delta = 1 + \mu$. Protein scoring systems on the other hand usually consider for performance reasons only integer scores which automatically yields $\Delta = 1$ as the score unit. Additionally, the maximal entry $s_0 \equiv \max_{a,b}\{s_{a,b}\}$ of the scoring matrix $s_{a,b}$ is of special importance. We will only consider the discrete set of gap costs δ satisfying

$$2\delta = n_{\max}\Delta - s_0 \quad \text{with } n_{\max} \in \mathbf{N}. \tag{14}$$

In the case of protein sequence alignment, Eq. (14) is naturally satisfied, since δ and s_0 are usually chosen as integer numbers and $\Delta = 1$. For other scoring systems, the non integer case can in principle also be treated but becomes more complicated. Moreover, even if λ is only known for the discrete values of δ in Eq. (14), an interpolation should still give reasonable results for arbitrary gap costs [5].

Finally, 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 introduce uncorrelated random variables $\eta(r,t) \in \{0, 1, \ldots, n_{\max}\}$ such that $s(r,t) \equiv s_0 - \eta(r,t)\Delta$, i.e.,

$$\Pr\{\forall_{r,t}\, \eta(r,t) = \eta_{r,t}\} = \prod_{r,t} \Pr\{\eta(r,t) = \eta_{r,t}\} \tag{15}$$

[1] The independence of the peak scores of large islands has also been verified numerically [24].

with $\Pr\{\eta(r,t)=\eta\} = \sum_{a,b} p_a p_b \delta_{s_{a,b},s_0-\eta\Delta}$. The approximation (15) is known to change characteristic quantities of sequence alignment only slightly. We will confirm numerically at the end of this abstract, that this also holds for the values of λ which we are mainly interested in here.

Additionally, we will use instead of the triangular alignment lattice as shown in Fig. 3(a) a rectangular lattice of a fixed width of $2W$ lattice points as shown in Fig. 3(c). Across the lattice we apply periodic boundary conditions, i.e., we identify $h(0,t)$ and $h(2W,t)$ for all t. At the left we start with the initial conditions $h(r,t=1)=0$ for even r and $h(r,t=0)=\delta$ for odd r. The properties of global alignment for very long sequences naturally do not depend on the details of this choice of initial conditions, but these values will turn out to be especially convenient. This choice of the lattice is possible, since we can easily convince ourselves that the score $h(r,t)$ for all points with $t \le W$ will be identical with the corresponding score on the triangular lattice shown in Fig. 3(a) which can be seen as a sublattice of the rectangular lattice. At the end of our calculations we will take the limit of infinite W.

Now we will reformulate the sequence alignment algorithm Eq. (10) on the lattice shown in Fig. 3(c) step by step and transform condition (13) into an eigenvalue equation similar to the equation for λ given in [18] for the case of protein sequences generated by a Markov process instead of independently chosen amino acids. First, we will introduce the *score differences* between neighboring lattice points as new variables. We will parameterize these score differences by the bond variables $n(r,t)$. With the choice of coordinates as illustrated in Fig. 4(a), we define them to be

$$n(r,t) \equiv \begin{cases} \frac{1}{\Delta}[h(r+1,t) - h(r,t+1) + \delta + s_0] & r+t \text{ even} \\ \frac{1}{\Delta}[h(r+1,t+1) - h(r,t) + \delta] & r+t \text{ odd} \end{cases} \tag{16}$$

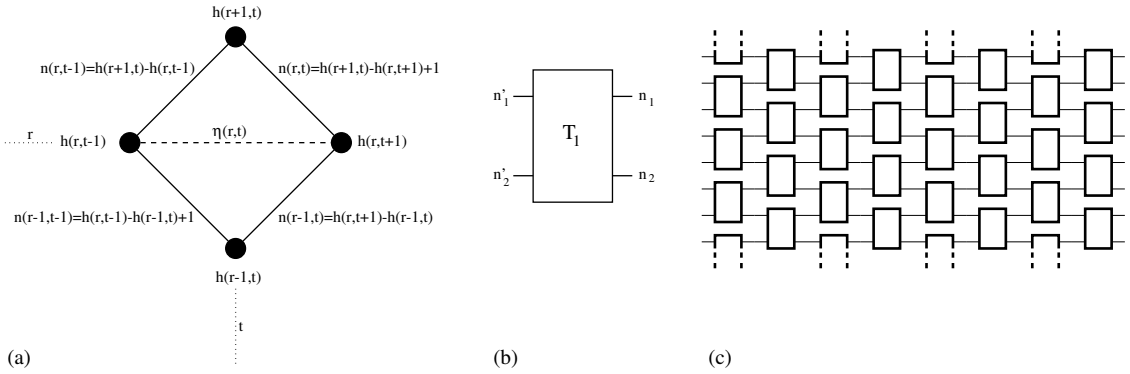

Figure 4: Building blocks of the alignment lattice. By our numbering scheme of the lattice r and t are either both even or both odd. (a) shows the scores at the lattice points and the bond variables $n(r,t)$. (b) shows this building block as an "device", which takes two incoming bond variables n_1' and n_2' and transforms them into the new bond variables n_1 and n_2. (c) shows schematically how these building blocks are "connected" in the alignment lattice of Fig. 3(c). 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.

The important feature of these quantities is, that the evolution equation (10) can be completely rewritten in terms of these bond variables without reference to the absolute scores $h(r,t)$. Since the lattice is composed of the building blocks shown in Fig. 4(a), the recursion equation transforms pairs of neighboring $n(r,t)$ at some $t-1$ into the same pairs at time t. The detailed calculation is shown in App. B. For even $r+t$ which corresponds to Fig. 4(a) the result is

$$n(r-1,t) = n(r-1,t-1) - j(r,t) \quad \text{and} \quad n(r,t) = n(r,t-1) + j(r,t) \tag{17}$$

where $r+t$ is even and

$$j(r,t) \equiv \min\{\eta(r,t), n_{\max} - n(r,t-1), n(r-1,t)\}. \tag{18}$$

By induction it is obvious from these equations, that the variables $n(r,t)$ will always remain integer numbers between 0 and $n_{\max}$, if they are integers at $t=0$ as it is the case for our choice of initial

conditions[2]. The transformation of a pair of variables $(n(r-1,t-1), n(r,t-1))$ into a pair of variables $(n(r-1,t), n(r,t))$ happens according to the lattice structure shown in Fig. 3(c) at odd t for all odd r and at even t for all even r. We can thus think of the dynamics in terms of the variables $n(r,t)$ as a network of elements of the form shown in Fig. 4(b) acting on pairs of variables $n(r,t)$ interconnected in the way schematically depicted in Fig. 4(c). This process is also known [6] as an asymmetric exclusion process which is a well studied model for surface growth and highway traffic [16, 26].

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 score differences $n(r,t)$ turn out to lead to meaningful periodic boundary conditions $h(0,t) = h(2W,t)$ for the scores only under the additional constraint

$$\sum_{r=0}^{2W-1} n(r,t) = W n_{\max}. \tag{19}$$

Since the quantity on the left hand side is obviously conserved under the dynamics Eqs. (17) this will not be an issue, if the initial conditions are chosen in accordance to Eq. (19).

So far, we transformed the dynamics of the sequence alignment algorithm as given by Eq. (10). We still have to express our main quantity of interest, $\langle \exp[\lambda h(0,N)] \rangle$ in terms of the dynamics of the variables $n(r,t)$. This can be done again by explicitly expressing the change of $h(0,N)$ under the dynamics given by Eqs.(17) and (18). The details of this calculation are worked out in App. B. It yields

$$\langle \exp[\lambda h(0,N)] \rangle_0 = \exp[\lambda s_0 N/2] \langle \exp[-\lambda \Delta J] \rangle_0, \tag{20}$$

where $\langle \cdot \rangle_0$ denotes the average over the independent random variables $\eta(r,t)$ and

$$J \equiv \frac{1}{2W} \sum_{l=1}^{N/2} \sum_{k=0}^{W-1} [j(2k+1, 2l-1) + j(2k, 2l)]. \tag{21}$$

In order to be able to calculate λ from conditions (13) we thus have to determine the generating function $\langle \exp[\omega J] \rangle_0$ of J.

As it is worked out in App. C this generating function can be expressed by the Nth power of some generalized transfer matrix. Its behavior for large N is thus given by the largest eigenvalue $\rho_W(\omega)$ of this generalized transfer matrix. The transfer matrix is built from the $(n_{\max}+1)^2$ dimensional matrix $\mathsf{T}_1(\omega/W)$ which describes the contribution to J from each of the elements as the one shown in Fig. 4(a). Its matrix elements are given by

$$(\mathsf{T}_1)_{(n_1,n_2),(n_1',n_2')}\left(\frac{\omega}{W}\right) \equiv \sum_{\eta} \Pr\{\eta(r,t)=\eta\} \delta_{n_1, n_1' - j(n_1',n_2',\eta)} \delta_{n_2, n_2' + j(n_1',n_2',\eta)} \exp[\frac{\omega}{2W} j(n_1', n_2', \eta)] \tag{22}$$

with $j(n_1', n_2', \eta) \equiv \min\{\eta, n_1', n - n_2'\}$. For $\omega = 0$ this reduces to the usual transfer matrix which just contains the probabilities, that the configuration (n_1', n_2') of two neighboring lattice sites is transformed by Eqs. (17) and (18) into the configuration (n_1, n_2).

Since at each time step W of the elements shown in Fig. 4(a) act in parallel, W of these matrices have to be multiplied together as $\mathsf{T}_W(\omega) \equiv \bigotimes_{k=1}^{W} \mathsf{T}_1(\omega/W)$. Additional to this matrix we need the matrix C which shifts all the variables $n(r,t)$ from r to $r+1$, i.e., $\mathsf{C}|n_0 n_1 \ldots n_{2W-1}\rangle \equiv |n_1 \ldots n_{2W-1} n_0\rangle$. With these definitions, the generating function of J at a given width of the lattice W turns out to be given by

$$\langle \exp[\omega J] \rangle_0 = [\rho_W(\omega)]^N \tag{23}$$

in terms of the largest eigenvalue $\rho_W(\omega)$ of the matrix $\mathsf{T}_W(\omega)\mathsf{C}$ restricted onto the space of vectors $|\psi\rangle$ fulfilling condition (19) and being translationally invariant, i.e., $\mathsf{C}^2|\psi\rangle = |\psi\rangle$. If we know the

[2]Even if the initial values of the $n(r, t=0)$ are not integer they will under the dynamics Eqs. (17) and (18) eventually try to take on values less than zero or larger than $n_{\max}$. The minimum in Eq. (18) then resets them to the integer values zero or $n_{\max}$. Thus, after some startup phase, the $n(r,t)$ will be integer even if their initial values are chosen to be non-integer.

large width limit $\rho(\omega) \equiv \lim_{W\to\infty} \rho_W(\omega)$ of this largest eigenvalue we can use Eqs. (13), (20), and (23) in order to calculate λ from the condition

$$\rho(-\lambda\Delta)\exp[\lambda s_0/2] = 1. \tag{24}$$

Additionally, the typical slope α of an island which is important for taking into account the effects of different sequence lengths is according to App. A given by

$$\alpha^* = s_0 - 2\Delta\rho'(-\lambda\Delta)\exp[\lambda s_0/2] \tag{25}$$

5 Results for a Simple Scoring System

Since finding the largest eigenvalue of the matrix $\mathsf{CT}(\omega)$ for a general scoring system is a rather difficult task, we will now restrict ourselves to an especially simple case. We will use the scoring matrix for DNA sequence comparisons as defined in Eq. (1), which implies $s_0 = 1$ and $\Delta = 1 + \mu$. Moreover, we will use the minimal[3] gap cost value $\delta = \mu/2$. This choice of scoring parameters implies $n_{\max} = 1$. One important special case of this scoring system is the longest common subsequence (LCS) problem [7] for $\mu = \delta = 0$ which has been of interest to mathematicians for a long time.

If we denote by p the probability for a match, e.g., $p = 1/c$ for an alphabet of size c with equal probabilities for all letters, the basic matrix $\mathsf{T}_1(\omega/W)$ is given by

$$\mathsf{T}_1(\omega/W) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & (1-p)e^{\frac{\omega}{2W}} & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{26}$$

in the basis $|00\rangle, |01\rangle, |10\rangle, |11\rangle$. The largest eigenvalue of the corresponding matrix $\mathsf{T}_W(\omega)\mathsf{C}$ can be calculated for small widths W using computer algebra. Generalizing the result to arbitrary W as explained in App. D yields

$$\rho(\omega) = \lim_{W\to\infty} \rho_W(\omega) = \frac{\sqrt{p} + e^{\frac{\omega}{2}}}{1 + \sqrt{p}e^{\frac{\omega}{2}}}. \tag{27}$$

Eq. (24) then leads to the condition

$$\frac{1 + \sqrt{p}\exp[\frac{\lambda}{2}(1+\mu)]}{1 + \sqrt{p}\exp[-\frac{\lambda}{2}(1+\mu)]}\exp[-\frac{\lambda}{2}\mu] = 1. \tag{28}$$

For large values of the mismatch (and gap) cost μ it converges towards the value $\lambda = -\log p$ which is to be expected from Eq. (5) in the gapless limit. Moreover, since $\lambda = 0$ is always trivially a solution of this equation, the implied value of λ tends towards zero, if the derivative of the left hand side vanishes. A vanishing λ corresponds to a breakdown of local alignment (all island sizes are equally probable) and indicates the phase transition between the logarithmic and the linear alignment phase. Equating the derivative of the left hand side of Eq. (28) to zero, correctly reproduces the exact result $\mu_c = 2\sqrt{p}/(1-\sqrt{p})$ for the mismatch cost at which the phase transition takes place under the assumption (15) of uncorrelated disorder [5]. Expanding Eq. (28) for small λ yields that λ vanishes proportional to $(\mu - \mu_c)^{1/2}$ as μ approaches its critical value μ_c as it has been predicted already on more general grounds in [11].

In order to test the approximation of uncorrelated local disorder (15) and the heuristic elements of the derivation of Eq. (28), we performed extensive numerical simulations to corrobate 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 [24] to find the values of λ as a function of μ numerically. For each value of δ several billion

[3] For any smaller gap cost $\delta < \mu/2$ two gaps become favorable compared to a mismatch. Thus, the alignment does not depend on the mismatch cost μ any more and is equivalent to the case where μ is adjusted to fulfill $\delta = 2\mu$.

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 to be one with probability p and $-\mu$ with probability $1-p$. The resulting values of λ are shown in Fig. 5. The solid line is the solution of Eq. (28) and the circles represent the values of λ for uncorrelated local scores (15). As shown in Fig. 5 the observed λ's follow the analytic solution very closely, thereby confirming Eq. (28). We also included the values of λ which result from correlated local scores generated from aligning randomly chosen sequences according to Eq. (1). 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 $\mu = 2$. The difference of $\sim 2\%$ in μ_c between the correlated and the uncorrelated case rapidly becomes much smaller for larger alphabet sizes c [4].

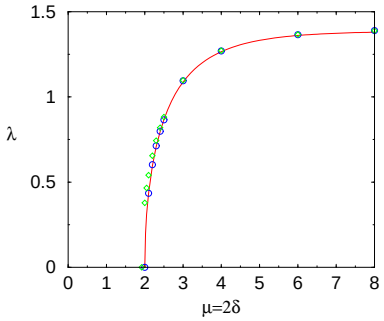

Figure 5: Dependence of the significance parameter λ on the scoring parameter μ. The circles represent the numerically obtained values of λ for uncorrelated local disorder (15) with match probability $p = 1/4$ for which Eq. (28) (the solid line) has been derived. They agree well with the analytical result. 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 μ_c.

6 Concluding Remarks

We have shown, how the important problem of assessing the statistical significance of local Smith-Waterman alignments can be solved by studying the simpler global alignment algorithm. Using the approximation of uncorrelated local similarity scores we have reformulated the problem of finding the Gumbel parameter λ in terms of an explicit eigenvalue equation. We have demonstrated the feasibility of this approach by deriving an explicit formula for λ for a simple scoring system and we have numerically verified that the expression for λ obtained under the approximation of uncorrelated local similarity scores describes the real distribution of optimal alignment scores of random sequences very well.

While we demonstrated the feasibility of our approach here for the simplest possible scoring system, it is necessary to solve the eigenvalue equation presented in this work for more general scoring systems. Even if this becomes more and more difficult analytically, it is still possible to compute the necessary eigenvalue numerically for systems of finite width. Since the largest eigenvalue numerically seems to converge rather fast to its infinite width limit this should still give reliable values for the parameter λ. Moreover the derivation should be extended to the case of affine gap costs.

Acknowledgments

The author gratefully acknowledges discussions with S. Altschul, T. Hwa, R. Olsen, and N. Rajewsky, and the hospitality of the Center for Physics and Biology at Rockefeller University, New York, and the National Center for Biotechnology Information, Bethesda, where this work has been completed. This work is supported in part by a Hochschulsonderprogramm III fellowship of the DAAD and by the National Science Foundation through Grant No. DMR-9971456.

Appendices

A Island High Score Distribution

In this appendix we derive heuristically the Poisson distribution of maximal island scores. We first treat the gapless case [12] 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 by definition given by

$$p(\sigma|L) = \langle \delta(\sigma - \sum_{i=1}^{L} s(i)) \rangle. \tag{29}$$

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

$$p(\sigma|L) = \frac{1}{2\pi} \int \exp(-ik\sigma) \langle \exp(iks) \rangle^L \mathrm{d}k. \tag{30}$$

For a given peak score σ this probability is maximal at a certain length $L_0(\sigma)$. It is reasonable to assume that the distribution of island lengths at a given peak score σ is sufficiently peaked around the length $L_0(\sigma)$ at which the probability is maximized. Thus, the probability of finding an island with peak score σ with *any* length is dominated by the probability of finding an island of length $L_0(\sigma)$. If we further assume that this length is proportional to the peak score, i.e., $L_0(\sigma) = \alpha^{-1}\sigma$, we get

$$p(\sigma) \sim \int \exp(-ik\sigma) \langle \exp(iks) \rangle^{\alpha^{-1}\sigma} \mathrm{d}k. \tag{31}$$

For large σ, Eq. (31) can be evaluated via the method of steepest descent as

$$p(\sigma) \sim \exp(-\lambda\sigma) \tag{32}$$

with

$$\lambda = ik^*(\alpha) - \log[\langle \exp(ik^*(\alpha)s) \rangle]/\alpha. \tag{33}$$

The saddle point $k^*(\alpha)$ is given by the saddle point equation

$$\frac{\langle s \exp(ik^*(\alpha)s) \rangle}{\langle \exp(ik^*(\alpha)s) \rangle \alpha} = 1. \tag{34}$$

The slope α of a typical island is so far unknown. It is fixed by the requirement that it maximizes the probability to find an island of the given peak score σ. Thus, we minimize Eq. (33) with respect to α and get together with Eq. (34)

$$\langle \exp(ik^*(\alpha^*)s) \rangle = 1. \tag{35}$$

Inserting this into Eq. (33) yields condition (5). Additionally we get from Eq. (34) the typical slope α^* of an island as

$$\alpha^* = \langle s \exp(\lambda s) \rangle. \tag{36}$$

It is related to the relative entropy H of the scoring system by

$$H \equiv \lambda \langle s \exp(\lambda s) \rangle = \lambda\alpha^* \tag{37}$$

which is the crucial quantity for the correction of the sequence length dependence of the statistical significance [2].

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 on a triangular lattice as the one shown in Fig. 3(a) taking into account all possible insertions of gaps. We can still use the Fourier transformation to get

$$p(\sigma|L) = \langle \delta(\sigma - h(0, L)) \rangle = \frac{1}{2\pi} \int \exp(-ik\sigma) \langle \exp(ikh(0, L)) \rangle \mathrm{d}k. \tag{38}$$

In App. C 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 $\widetilde{\rho}(\lambda)$ by

$$\langle \exp[\lambda h(0,L)]\rangle \equiv \widetilde{\rho}^L(\lambda) \tag{39}$$

and again assume a linear slope α of the islands where the triangular shape of Fig. 3(a) with only $L/2$ matches or mismatches at the central line $r=0$ makes the definition $\sigma \equiv \alpha L_0(\sigma)/2$ the most natural definition for α. Eq. (38) then becomes

$$p(\sigma) \sim \int \exp\{[-ik + 2\alpha^{-1}\log\widetilde{\rho}(ik)]\sigma\}\mathrm{d}k. \tag{40}$$

Applying as in the gapless case the method of steepest descent and maximizing with respect to the slope of the island α yields Eq. (13). Moreover it gives the typical slope of an island as

$$\alpha^* = 2\frac{\widetilde{\rho}'(\lambda)}{\widetilde{\rho}(\lambda)} = \frac{2}{L}\langle h(0,L)\exp[\lambda h(0,L)])\rangle. \tag{41}$$

B Expression of the score dynamics in terms of particle occupation numbers

In this appendix we will derive the evolution equations (17) and (18) from the recursion equation (10) of the original global alignment algorithm. To this end we apply Eq. (10) to the definition of $n(r,t)$, where we assume by convention that $r+t$ is even as in Fig. 4(a). We get

$$\begin{aligned}
n(r-1,t) &= \frac{1}{\Delta}[h(r,t+1) - h(r-1,t) + \delta] \\
&= \frac{1}{\Delta}[\max\{h(r,t-1) + s_0 - \eta(r,t)\Delta, h(r-1,t) - \delta, h(r+1,t) - \delta\} - h(r-1,t) + \delta] \\
&= \frac{1}{\Delta}[h(r,t-1) - h(r-1,t) + \delta + s_0 + \\
&\qquad + \max\{-\eta(r,t)\Delta, h(r-1,t) - h(r,t-1) - s_0 - \delta, h(r+1,t) - h(r,t-1) - s_0 - \delta\}] \\
&= n(r-1,t-1) + \frac{1}{\Delta}\max\{-\eta(r,t)\Delta, -n(r-1,t-1)\Delta, n(r,t-1)\Delta - s_0 - 2\delta\} \\
&= n(r-1,t-1) - \min\{\eta(r,t), n(r-1,t-1), n_{\max} - n(r,t-1)\}
\end{aligned}$$

and analogously

$$\begin{aligned}
n(r,t) &= \frac{1}{\Delta}[h(r+1,t) - h(r,t+1) + \delta + s_0] \\
&= \frac{1}{\Delta}[h(r+1,t) - \max\{h(r,t-1) + s_0 - \eta(r,t)\Delta, h(r-1,t) - \delta, h(r+1,t) - \delta\} + \delta + s_0] \\
&= \frac{1}{\Delta}[h(r+1,t) - h(r,t-1) + \delta - \\
&\qquad - \max\{-\eta(r,t)\Delta, h(r-1,t) - h(r,t-1) - s_0 - \delta, h(r+1,t) - h(r,t-1) - s_0 - \delta\}] \\
&= n(r,t-1) - \frac{1}{\Delta}\max\{-\eta(r,t)\Delta, -n(r-1,t-1)\Delta, n(r,t-1)\Delta - s_0 - 2\delta\} \\
&= n(r,t-1) + \min\{\eta(r,t), n(r-1,t-1), n_{\max} - n(r,t-1)\}.
\end{aligned}$$

With the definition Eq. (18) this yields Eq. (17).

In a similar way we can express the quantity $\langle \exp[\lambda h(0,N)]\rangle_0$ in terms of the variables $n(r,t)$. To achieve this, we first define for any "time" t the average score

$$\overline{h}(t) \equiv \begin{cases} \frac{1}{2W}\sum_{k=0}^{W-1}[h(2k,t-1) + h(2k+1,t)] & t \text{ even} \\ \frac{1}{2W}\sum_{k=0}^{W-1}[h(2k,t) + h(2k+1,t-1)] & t \text{ odd} \end{cases} \tag{42}$$

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

$$\langle \exp[\lambda h(0,N)]\rangle_0 = \langle \exp[\lambda \overline{h}(N)]\rangle_0. \tag{43}$$

Thus, we can restrict ourselves to calculating the large N behavior of the latter quantity. The change in the average score $\overline{h}(t)$ is given by

$$\overline{h}(t+1) - \overline{h}(t) = \begin{cases} \frac{1}{2W}\sum_{k=0}^{W-1}[h(2k,t+1) - h(2k,t-1)] & t \text{ even} \\ \frac{1}{2W}\sum_{k=0}^{W-1}[h(2k+1,t+1) - h(2k+1,t-1)] & t \text{ odd} \end{cases}. \tag{44}$$

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

$$\begin{aligned} h(r,t+1) - h(r,t-1) &= \max\{h(r,t-1) + s_0 - \eta(r,t)\Delta, h(r+1,t) - \delta, h(r-1,t) - \delta\} \\ &\quad -h(r,t-1) \\ &= s_0 + \max\{-\eta(r,t)\Delta, n(r,t-1)\Delta - 2\delta - s_0 - n(r-1,t-1)\Delta\} \\ &= s_0 - \Delta \min\{\eta(r,t), n_{\max} - n(r,t-1), n(r-1,t-1)\} \\ &= s_0 - j(r,t)\Delta. \end{aligned}$$

Inserting this into Eq. (44) yields

$$\overline{h}(t+1) - \overline{h}(t) = \frac{s_0}{2} - \frac{\Delta}{2W}\begin{cases} \sum_{k=0}^{W-1} j(2k,t) & t \text{ even} \\ \sum_{k=0}^{W-1} j(2k+1,t) & t \text{ odd} \end{cases}. \tag{45}$$

Combining Eqs. (43) and (45) finally yields

$$\begin{aligned} \langle \exp[\lambda h(0,N)]\rangle_0 &= \langle \exp[\lambda \overline{h}(N)]\rangle_0 = \langle \exp[\lambda \sum_{t=0}^{N-1}(\overline{h}(t+1) - \overline{h}(t))]\rangle_0 \qquad (46) \\ &= \exp[\lambda s_0 N/2]\langle \exp[-\frac{\lambda\Delta}{2W}\sum_{l=1}^{N/2}\sum_{k=0}^{W-1}(j(2k+1,2l-1) + j(2k,2l))]\rangle_0 \\ &= \exp[\lambda s_0 N/2]\langle \exp[-\lambda\Delta J]\rangle_0, \end{aligned}$$

with J as defined in Eq. (21).

C An eigenvalue equation

In this appendix we will reformulate the calculation of the generating function $\langle \exp[\omega J]\rangle_0$ as an eigenvalue problem. We start from the definition

$$\langle \exp[\omega J]\rangle_0 = \langle \prod_{l=1}^{N/2}\prod_{k=0}^{W-1} e^{\frac{\omega}{2W}j(2k+1,2l-1)} e^{\frac{\omega}{2W}j(2k,2l)}\rangle_0. \tag{47}$$

Since, the values of the variables $n(r,t)$ must be integers between 0 and $n_{\max}$ at any time, we do not change the expectation value, if we introduce ones of the form

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

at each fixed time t. This yields

$$\begin{aligned} &\langle \exp[\omega J]\rangle_0 = \qquad (48) \\ &\sum_{\{n_{r,0}\}}\cdots\sum_{\{n_{r,N}\}}\langle \prod_{r=0}^{2W-1}\delta_{n(r,0),n_{r,0}} \prod_{l=1}^{N/2}\left(\prod_{r=0}^{2W-1}\delta_{n(r,2l-1),n_{r,2l-1}}\right)\left(\prod_{k=0}^{W-1} e^{\frac{\omega}{2W}j(2k+1,2l-1)}\right)\times \end{aligned}$$

$$\times \left(\prod_{r=0}^{2W-1} \delta_{n(r,2l),n_{r,2l}} \right) \left(\prod_{k=0}^{W-1} e^{\frac{\omega}{2w} j(2k,2l)} \right) \rangle_0$$

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

$$\begin{aligned}
&\langle \prod_{r=0}^{2W-1} \delta_{n(r,0),n_{r,0}} \prod_{l=1}^{N/2} \left(\prod_{r=0}^{2W-1} \delta_{n(r,2l-1),n_{r,2z-1}} \right) \left(\prod_{k=0}^{W-1} e^{\frac{\omega}{2W} j(2k+1,2l-1)} \right) \times \\
&\times \left(\prod_{r=0}^{2W-1} \delta_{n(r,2l),n_{r,2l}} \right) \left(\prod_{l=0}^{W-1} e^{\frac{\omega}{2W} j(2k,2l)} \right) \rangle_0 = \\
&\quad \prod_{r=0}^{2W-1} \delta_{n(r,0),n_{r,0}} \times \\
&\quad \times \prod_{l=1}^{N/2} \prod_{k=0}^{W-1} \langle \delta_{n(2k,2l-2),n_{2k,2l-2}} \delta_{n(2k+1,2l-2),n_{2k+1,2l-2}} e^{\frac{\omega}{2W} j(2k+1,2l-1)} \times \\
&\qquad \times \delta_{n(2k,2l-1),n+2k,2l-1} \delta_{n(2k,2l-1),n_{2k+1,2l-1}} \rangle_0 \times \\
&\quad \times \prod_{k=0}^{W-1} \langle \delta_{n(2k-1,2l-1),n_{2k-1,2l-1}} \delta_{n(2k,2l-1),n_{2k,2l-1}} e^{\frac{\omega}{2W} j(2k,2l)} \delta_{n(2k-1,2l),n_{2k-1,2l}} \delta_{n(2k,2l),n_{2k+1,2z}} \rangle_0 \times 1.
\end{aligned}$$

Inserting this into Eq. (48) we can interpret the summation over the possible configurations of the $n(r,t)$ at each time step as the summation of 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 $(n_{\max}+1)^{2W}$ dimensional vector space indexed by all possible 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. It stands for the summation over all possible final configurations at $t=N$ and we call this vector $\langle\psi_1|$. All the other factors represent matrices. They can be written as tensor products of the $(n_{\max}+1)^2$ dimensional matrix $\mathsf{T}_1(\omega/W)$ which describes an elementary process as given by Eqs. (17) and (18). We can read off its elements to be

$$(\mathsf{T}_1(\omega/W))_{(n_1,n_2),(n_1',n_2')}(\omega) \equiv \langle \delta_{n(r-1,t-1),n_1'} \delta_{n(r,t-1),n_2'} \exp[\frac{\omega}{2W} j(r,t)] \delta_{n(r-1,t),n_1} \delta_{n(r,t),n_2} \rangle_0. \quad (49)$$

Note that the disorder average for each of these matrix elements is only an average over the single random variable $\eta(r,t)$. The matrix elements can thus be rewritten in the form given in Eq. (22).

Since the lattice of width $2W$ is at each time step decomposed into W of the building blocks described by $\mathsf{T}_1(\omega/W)$, the total system is described by the matrix

$$\mathsf{T}_W^{\text{even}}(\omega) \equiv \mathsf{T}_W(\omega) \equiv \bigotimes_{k=1}^{W} \mathsf{T}_1(\omega/W). \quad (50)$$

If $\mathsf{T}_W(\omega)$ describes the time evolution at even time steps, we can according to Fig. 4(c) generate the time evolution on odd time steps by shifting all variables to the right, applying the dynamics of even time steps and then shifting all variables back to the left. With the translation matrix C defined in the main text this can be written as $\mathsf{T}_W^{\text{odd}}(\omega) = \mathsf{C}\mathsf{T}_W(\omega)\mathsf{C}^{-1}$. The structure of the lattice as depicted by Fig. 4(c), finally leads to

$$\langle \exp[\omega J] \rangle_0 = \langle \psi_1 | (\mathsf{T}_W^{\text{even}}(\omega) \mathsf{T}_W^{\text{odd}}(\omega))^{N/2} | \psi_0 \rangle = \langle \psi_1 | (\mathsf{T}_W(\omega) \mathsf{C} \mathsf{T}_W(\omega) \mathsf{C}^{-1})^{N/2} | \psi_0 \rangle \quad (51)$$

In the limit of large N this obviously becomes Eq. (23) of the main text where $\rho_W^2(\omega)$ is the eigenvalue of $\mathsf{T}_W(\omega)\mathsf{C}\mathsf{T}_W(\omega)\mathsf{C}^{-1}$ with the largest real part. Since this matrix has no negative entries and is (restricted to the sector defined by condition (19)) for non-pathological choices of the

scoring matrix irreducible, the largest eigenvalue of this matrix is guaranteed to be non degenerate, real, and its eigenvector can be chosen without negative entries by the Perron Frobenius theorem.

Due to the lattice symmetry, the eigenvector $|\psi\rangle$ for the largest eigenvalue of $\mathsf{T}_W(\omega)\mathsf{C}\mathsf{T}_W(\omega)\mathsf{C}^{-1}$ should be translationally invariant, i.e., $\mathsf{C}^2|\psi\rangle = |\psi\rangle$ should hold. We can thus restrict the search for the largest eigenvalue to the subspace

$$\mathcal{C} \equiv \left\{ |\psi\rangle \middle| C^2|\psi\rangle = |\psi\rangle \right\}. \tag{52}$$

of all translationally invariant eigenvectors. On this subspace by definition $\mathsf{C} = \mathsf{C}^{-1}$ so that instead of looking for the largest eigenvalue $\rho_W^2(\omega)$ of $\mathsf{T}_W(\omega)\mathsf{C}\mathsf{T}_W(\omega)\mathsf{C}^{-1}$ we can also look for the largest eigenvalue $\rho_W(\omega)$ of $\mathsf{T}_W(\omega)\mathsf{C}$ on this subspace. Additionally $\mathsf{T}_W(\omega)\mathsf{C}$ has to be restricted onto the subspace of valid configurations as defined by condition (19). These reductions in the size and complexity of the matrices make an explicit calculation of $\rho_W(\omega)$ possible.

D Calculating the largest eigenvalue

For small W the largest eigenvalue $\rho_W(\omega)$ of $\mathsf{T}_W(\omega)\mathsf{C}$ restricted to the subspace of valid translationally invariant configurations as defined by condition (19) and Eq. (52) can be calculated using computer algebra. Although the matrix $\mathsf{T}_W(\omega)$ depends on $e^{\omega/2W}$ the largest eigenvalue of $\mathsf{T}_W(\omega)\mathsf{C}$ only contain terms of the form $e^{\omega/2}$. This is a consequence of the translational invariance of the lattice. In order to reveal the underlying structure of the largest eigenvalues $\rho_W(\omega)$ for different W it turns out to be very useful to expand them in powers of $e^{\omega/2}$. We get

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

where the $O((e^{\omega/2})^k)$ terms denote terms of the given order with prefactors which are different for different W. We can see, that the coefficients up to order $(e^{\omega/2})^{W-1}$ remain unchanged upon increasing W and that they moreover constitute the beginning of a simple geometric series. We can thus extrapolate this behavior to arbitrary W by assuming that this pattern holds for arbitrary orders. Resumming the geometric series then yields Eq. (27).

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403–410.

[2] Altschul, S.F., and Gish, W. 1996. Local Alignment Statistics. *Methods in Enzymology* **266**, 460–480.

[3] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

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

[5] Bundschuh, R., and Hwa, T. 1999. An Analytic Study of the Phase Transition Line in Local Sequence Alignment with Gaps. *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, S. Istrail *et al.* eds., 70–76, ACM press, (New York, NY); Bundschuh, R., and Hwa, T., to appear in *Disc. Appl. Math.*.

[6] Bundschuh, R. 1999. The Asymmetric Exclusion Process and the Extremal Statistics of Random Sequences. in preparation.

[7] Chvátal, V., and Sankoff, D. 1975. Longest common subsequences of two random sequences. *J. Applied Probab.* **12**, 306–315.

[8] Collins, J.F., Coulson, A.F.W., and Lyall, A. 1988. The significance of protein sequence similarities. *CABIOS* **4**, 67–71.

[9] Dayhoff, M.O., Schwartz, R.M., and Orcutt B.C. 1978. A Model of Evolutionary Change in Proteins. In *Atlas of Protein Sequence and Structure*, Dayhoff, M.O., and Eck, R.V., eds., **5** supp. 3, 345–358.

[10] Doolittle, R.F. 1996. *Methods in Enzymology* **266**, San Diego, Calif.: Academic Press.

[11] Drasdo, D., Hwa, T., and Lässig, M. 1998. A statistical theory of sequence alignment with gaps. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, J. Glasgow *et al.*, eds., 52–58, AAAI Press, (Menlo Park, CA).

[12] Fisher, D. 1999. Private communication.

[13] Galambos, J. 1978. *The Asymptotic Theory of Extreme Order Statistics*, John Wiley & Sons, (New York, NY).

[14] Gumbel, E.J. 1958. *Statistics of Extremes*, Columbia University Press, (New York, NY).

[15] Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919.

[16] Kandel, D., Domany, E., and Nienhuis, B. 1990. A six-vertex model as a diffusion problem – derivation of correlation functions. *J. Phys. A* **23**, L755–L762.

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

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

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

[20] Mott, R. 1992. Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59–75.

[21] Mott, R., and Tribe, R. 1999. Approximate statistics of gapped alignments. *J. Comp. Biol.* **6**, 91–112.

[22] Mott, R. 1999. Accurate estimate of p-values for gapped local sequence alignment. Private communication.

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

[24] Olsen, R., Bundschuh, R., and Hwa, T. 1999. Rapid Assessment of Extremal Statistics for Gapped Local Alignment. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, T. Lengauer *et al.*, eds., 211–222, AAAI Press, (Menlo Park, CA).

[25] Pearson, W.R. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**, 635–650.

[26] Rajewsky, N., Santen, L., Schadschneider, A., and Schreckenberg, M. 1998. The asymmetric exclusion process: Comparison of update procedures. *J. Stat. Phys.* **92** 151–194.

[27] Siegmund, D., and Yakir, B. 1999. Approximate p-values for Sequence Alignments. preprint.

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

[29] Smith, T.F., Waterman, M.S., and Burks, C. 1985. The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* **13**, 645–656.

[30] Waterman, M.S., and Vingron, M. 1994. Sequence Comparison Significance and Poisson Approximation. *Stat. Sci.* **9**, 367–381.

[31] Waterman, M.S., and Vingron, M. 1994. Rapid and accurate estimates of statistical significance for sequence database searches. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4625–4628.

[32] Waterman, M.S. 1994. *Introduction to Computational Biology.* London, UK: Chapman & Hall.