A Practical Approach to Significance Assessment in Alignment with Gaps

Nicholas Chia and Ralf Bundschuh[1]

Ohio State University,
Columbus, OH 43210, USA

Abstract. Current numerical methods for assessing the statistical significance of local alignments with gaps are time consuming. Analytical solutions thus far have been limited to specific cases. Here, we present a new line of attack to the problem of statistical significance assessment. We combine this new approach with known properties of the dynamics of the global alignment algorithm and high performance numerical techniques and present a novel method for assessing significance of gaps within practical time scales. The results and performance of these new methods test very well against tried methods with drastically less effort.

Keywords: pairwise sequence alignment, Markov models and/or hidden Markov models, statistics of motifs or strings, statistical significance, Gumbel distribution, extreme value distribution, Kardar-Parisi-Zhang universality class, asymmetric exclusion process.

1 Introduction

Sequence alignment is one of the most commonly used computational tools of molecular biology. Its applications range from identifying the function of newly sequenced genes to the construction of phylogenic trees [43, 18]. Its importance is epitomized by the popularity of the program BLAST [1, 3] which is currently used 300,000 times a day on the NCBI's web site alone.

All 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 reliably quote a p-value, the *distribution* of optimal alignment scores for alignments of random sequences must be known. In the case of alignment without "gaps", it has been worked out rigorously [24–26] that this distribution is a Gumbel or extreme value distribution [20]. This distribution is characterized by two parameters that depend on the scoring system used and on

the amino acid frequencies with which the random sequences are generated. For gapless alignment, the dependence of the two Gumbel parameters on the scoring system is completely known.

However, in order to detect weakly homologous sequences, gaps must be allowed in an alignment [35]. Unfortunately, for the case of gapped alignment, there currently exists no theory that describes the distribution of alignment scores for random sequences. However, there remains a lot of numerical evidence as well as a number of heuristic arguments that this distribution is still of the Gumbel form [39, 12, 30, 41, 42, 2]. Nevertheless, even assuming the correctness of the Gumbel form, finding the two Gumbel parameters for a given scoring system turns out to be a very challenging problem. The straightforward method generates a large number of alignment scores by shuffling the two sequences to be compared and taking a histogram of this distribution. But, because of the slow exponential tail of the Gumbel distribution, this method is extremely time consuming. Thus, in practice, the two Gumbel parameters have to be pre-computed for some few fixed scoring systems [2, 3].

Pre-computed Gumbel parameters have the disadvantage that they restrict the user to a few scoring systems (substitution matrices and gap costs) for which the Gumbel parameters have been pre-computed. The necessity of pre-computing the Gumbel parameters definitely becomes problematic if adaptive schemes, 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.

To remedy this problem, a more effective numerical method which estimates the two Gumbel parameters has been proposed [34, 4]. There are also some analytical approximations [31, 37, 32] 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 addition, an analytical scheme has been used to successfully calculate the Gumbel parameter λ, which describes the tail of the Gumbel distribution, for just one particular scoring system [8, 10]. In this paper, we will present a novel approach that calculates λ for a variety of scoring schemes while drastically reducing the time required to calculate λ and retaining a high degree of precision in the solution. This approach will expand upon and combine the different analytical works devised in [8, 10] and [16, 17], creating a new scheme for calculating λ using the numerical tools of [29]. Once λ is known, it then becomes a simple matter to extract the remaining Gumbel parameter, which characterizes the mean of the score distribution, numerically via e.g., the island method [34, 4] or direct simulation.

In section 2, we will present an abbreviated review of sequence alignment. We then point out that, although λ is intrinsically a quantity of *local* alignments, it may be calculated from solely studying the simpler *global* alignment algorithm. Under some very moderate approximation we then briefly reformulate the problem of finding λ in terms of an eigenvalue equation, as done with more detail in [8, 10]. We then show the feasibility of our novel approach by comparing the

results from this new method with established analytical [8] and numerical [4] methods for a variety of scoring systems.

2 Review of Sequence Alignment

In the vast majority of sequence alignment applications, gapped alignment is used as the fundamental alignment technique. Gapped alignment looks for similarities between two sequences $\boldsymbol{a} = a_1 a_2 \dots a_M$, and $\boldsymbol{b} = b_1 b_2 \dots b_N$ of length M and N 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 or the twenty letter amino-acid alphabet. Here, we consider Smith-Waterman local alignment [38]. In this case, a possible alignment $\mathcal{A}$ consists of two substrings of the two original sequences $\boldsymbol{a}$ and $\boldsymbol{b}$. 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 $S[\mathcal{A}] = \sum_{(a,b)\in\mathcal{A}} s_{a,b} - \delta N_{\mathrm{g}}$ where the sum is taken over all pairs of aligned letters, N_{g} is the total number of gaps in the alignment, δ is an additional scoring parameter, the "gap cost," and $s_{a,b}$ is some given "scoring matrix" measuring the mutual degree of similarity between the different letters of the alphabet. A simple example, the match-mismatch matrix

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

is used for DNA sequence comparisons [33]. For protein sequences, normally the 20 x 20 PAM [13] or BLOSUM matrices [21] are used. Practical applications usually use the more complicated affine gap cost. For the purpose of clarity, the following will only consider the case of linear gap cost. However, we want to stress that our approach is applicable to affine gap costs as well as discussed at the end of the manuscript. The computational task is to find the subsequences which give the *highest* total score for a given scoring matrix $s_{a,b}$

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

The task is to find the alignment $\mathcal{A}$ with the highest score as in Eq. (2). This can be very efficiently done by a dynamic programming method which becomes obvious in the alignment path representation [33]. In this representation, the two sequences to be compared are written on the edges of a square lattice as shown in Fig. 1 where we chose $L \equiv M = N$. 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 $s(r,t) \equiv s_{a,b}$ where a and b are the letters of the two sequences belonging to the position (r,t) as shown in Fig. 1.

If interested in finding the highest scoring *global* alignment of the two sequences $\boldsymbol{a}$ and $\boldsymbol{b}$, one finds the best scoring path connecting the beginning $(0,0)$

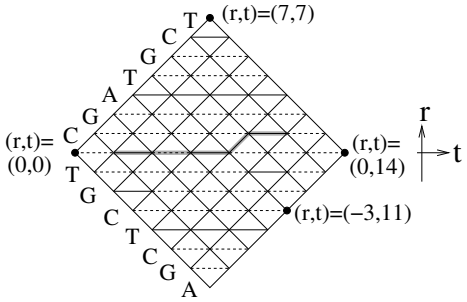

Fig. 1. Local alignment of two sequences. This figure shows the alignment of `CGATGCT` and `TGCTCGA` represented as a directed path on the alignment lattice. The highlighted alignment path $r(t)$ corresponds to one possible alignment of two subsequences, `GATGC` to `GCT-C`. This path contains one gap. It is also shown how the coordinates r and t are used to identify the nodes of the lattice.

to the end $(0, 2L)$. This path can be found efficiently by defining the auxiliary quantity $h(r,t)$ to be the score of the best path ending in the lattice point (r,t) with initial conditions $h(t,t) = -t\delta = h(-t,t)$. This quantity can be calculated recursively by the Needleman-Wunsch dynamic programming algorithm [33]

$$h(r,t+1) = \max\{h(r,t-1) + s(r,t),\ h(r \pm 1,t) - \delta\}. \tag{3}$$

For *local* alignments, the Smith-Waterman algorithm [38], supplemented by the initial conditions $S(t,t) = 0 = S(-t,t)$, describes the appropriate recursion

$$S(r,t+1) = \max\{S(r,t-1) + s(r,t), S(r \pm 1,t) - \delta, 0\}. \tag{4}$$

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

Characterizing the statistical significance of alignments requires the distribution of Σ for the alignment of two *random* sequences whose elements, a_k's and b_k's, are generated independently from the same frequencies p_a as the query sequences, and scored using the scoring matrix $s_{a,b}$. In the gapless limit where $\delta \to \infty$, this distribution of Σ has been worked out rigorously for the regime pertinent to significance assessment — i.e. in the *logarithmic phase* characterized by a negative $\langle s \rangle \equiv \Sigma_{a,b} p_a p_b s_{a,b}$ and $\Sigma \propto \log L$ [7, 25, 26]. For scoring parameters in the logarithmic phase, it is a Gumbel or extreme value distribution given by

$$\Pr\{\Sigma < S\} = \exp(-\kappa e^{-\lambda S}). \tag{5}$$

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 [25, 26] from the scoring matrix $s_{a,b}$ and the letter frequencies p_a. For example, λ is the unique positive solution of the equation

$$\langle \exp(\lambda s) \rangle \equiv \sum_{a,b} p_a p_b \exp(\lambda s_{a,b}) = 1. \tag{6}$$

In the presence of gaps, one can still distinguish a logarithmic phase [40]. If the parameters are chosen such that the expected *global* alignment score drifts downwards on average, then the average maximum score $\langle \Sigma \rangle$ for gapped *local* alignment remains 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 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 exists no rigorous theory for this distribution in the presence of gaps. However, a slew of empirical evidence strongly suggests that the distribution of local scores describes the Gumbel distribution [39, 12, 30, 41, 42, 2]. In practice, they have to be determined empirically by time consuming simulations [4]. In the absence of a more efficient means of calculating λ and κ, the use of adaptive schemes such as PSI-BLAST or more finely tuned significance assessment for various letter compositions remains elusive. Below we will present a new method to calculate the parameter λ, as well as an explicit calculation of this parameter for some simple scoring systems, that can resolve this dilemma. Since κ determines the mean and not tail of the distribution, κ can always be determined efficiently by simulation once λ is known. The method outlined here may also be applied directly to more complex scoring schemes, e.g., affine gap costs.

3 Review of Significance Estimation using Global Alignment as a Dynamic Process

As a first and very crucial step, we will use the fact that, accepting the empirical applicability of the Gumbel distribution to gapped local alignment, the parameter λ, describing the tail of the Gumbel distribution, can be derived solely from studying the much simpler *global* alignment (3). This has been shown in [8, 10]. For our purposes, we will recast the result from [8, 10] in the following form.

Let us define the generating function

$$Z_L(\gamma; \Omega) \equiv \langle \exp[\gamma h(0, L)] \rangle \tag{7}$$

where the brackets $\langle \cdot \rangle$ denote the ensemble average over all choices of random sequences $\boldsymbol{a}$, $\boldsymbol{b}$ and $h(0, L)$ is the *global* alignment score at the end of a lattice of length L as shown in Fig. 2(a), and Ω summarizes the parameters p_a, μ, and δ that contribute to the evaluation of $h(0, L)$. This score can be obtained from the recursion relation (3) with initial condition $h(r, 0) = 0$. Let us now define

$$\Phi(\gamma; \Omega) = \lim_{L \to \infty} \frac{1}{L} \log Z_L(\gamma; \Omega) \tag{8}$$

Then according to [8, 10] the parameter λ of the Gumbel distribution is obtained as the unique positive solution of the equation

$$\Phi(\lambda; \Omega) = 0. \tag{9}$$

Note that this condition reduces simply to Eq. (6) in the case of gapless alignment, since for infinite gap cost δ, we have $\langle \exp[\gamma h(0,L)]\rangle = \langle \exp[\gamma \sum_{k=1}^{L/2} s(0,2k-1)]\rangle = \langle \exp[\gamma s]\rangle^{\frac{L}{2}}$ and thus $\Phi(\gamma;\Omega) = \frac{1}{2}\log\langle \exp(\gamma S)\rangle$.

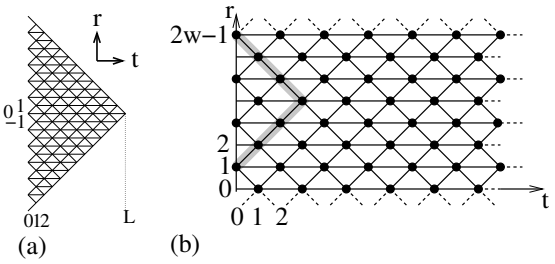

Fig. 2. Global alignment lattice used for significance estimation. (a) shows the right half of the lattice from Fig. 1. It can represent all possible paths of length L which end at the point $(r,t) = (0,L)$ and start at $(r,0)$ for an arbitrary r. (b) shows with the gray lines, how the triangular lattice 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 $L < W$.

In order to calculate $\Phi(\gamma;\Omega)$, instead of the triangular alignment lattice shown in Fig. 2(a), we utilize the rectangular lattice of $2W$ lattice points shown in Fig. 2(b). Across the lattice, we apply periodic boundary conditions $h(0,t) = h(2W,t)$ for all t. Defining the generating function of the finite width $Z_{L,W}(\gamma;\Omega)$ by Eq. (7) with $h(r,t)$ calculated on the lattice of width $2W$, we introduce

$$\Phi_W(\gamma;\Omega) = \lim_{L\to\infty} \frac{1}{L} Z_{L,W}(\gamma;\Omega). \tag{10}$$

The function $\Phi(\gamma;\Omega)$ on the original lattice is then given by

$$\Phi(\gamma;\Omega) = \lim_{W\to\infty} \Phi_W(\gamma;\Omega). \tag{11}$$

Thus, our approach will be to first calculate $\Phi_W(\gamma;\Omega)$ for some small W's and then take the limit for large W. Indeed, the major contribution of this work is the procedure for successfully extrapolating the infinite W limit $\Phi(\gamma;\Omega)$ from $\Phi_W(\gamma;\Omega)$ calculated for a few small W's. Further details will be given in section 4. Here, we stress that this methodology may be used in order to calculate $\Phi(\gamma;\Omega)$ from $\Phi_W(\gamma;\Omega)$ regardless of the specific scoring scheme and parameters including affine gap costs. Our method applies equally to *all* means available for calculating $\Phi_W(\gamma;\Omega)$. Ultimately once $\Phi(\gamma;\Omega)$ has been determined, we will use Eq. (9) to infer the value of the parameter λ characterizing local alignment.

In order to illustrate our method as clearly as possible, we will specialize the remaining discussion to the match–mismatch scoring system given by Eq. (1)

and even restrict the space of allowable scoring parameters further as discussed below. For this scoring scheme, we can utilize results from [8, 10] to calculate $\Phi_W(\gamma;\Omega)$ for small widths W. Thus, we will next review the appropriate results from [8, 10]. For the reader who is uninterested in the specifics of how $\Phi_W(\gamma;\Omega)$ is calculated here, we suggest skipping forward to the third paragraph of section 4.

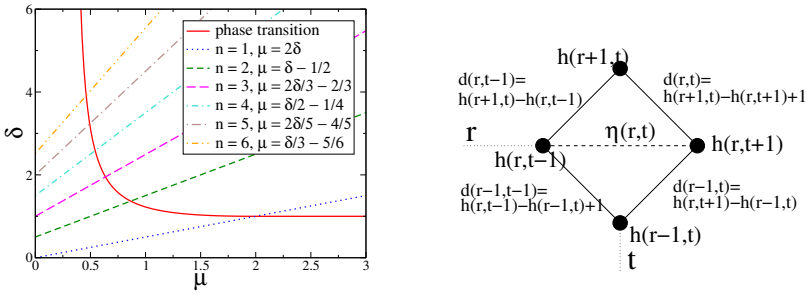

Fig. 3. (a)This figure illustrates the constraint given by Eq. (12). The straight lines plot those μ–δ values that obey the constraint. For any point on these lines, the solution for λ may be obtained using the method given in this presentation. The logarithmic phase is above the solid line denoting the phase transition. The phase transition line was obtained from [7] and has been supplied here for reference. (b)Building blocks of the alignment lattice. By construction r and t are either both even or both odd. This figure shows the relation between the scores at the lattice points and the bond variables $d(r,t)$.

In addition to specializing to the match–mismatch scoring system, we constrain μ and δ such that

$$2\delta = n(1+\mu) - 1 \quad \text{with } n \in \mathbf{N}. \tag{12}$$

This technical condition is necessary in order to utilize the results from [8, 10]. However, it is not a very severe condition since the (μ,δ)–pairs that fulfill this condition can be found all over the μ–δ plane as shown in Fig. 3(a).

The sole approximation neglects the correlations arising between the local scores $s(r,t)$ from the fact that all $M \times N$ local scores are generated by just $M+N$ randomly drawn letters. Instead of taking these correlations into account, we introduce uncorrelated random variables $\eta(r,t) \in \{1,-\mu\}$ replacing the $s(r,t)$ calculated from the letters in the sequences, i.e.,

$$\Pr\{\forall_{r,t}\,\eta(r,t) = \eta_{r,t}\} = \prod_{r,t} \Pr\{\eta(r,t) = \eta_{r,t}\} \tag{13}$$

with match probability $\Pr\{\eta(r,t) = 1\} = \sum_{a=b} p_a^2 \equiv p$ and mismatch probability $\Pr\{\eta(r,t) = -\mu\} = \sum_{a\neq b} p_a p_b = 1-p$. This approximation, also known

as the Bernoulli randomness approximation, is known to change characteristic quantities of sequence alignment only slightly [14, 15, 5, 6, 9, 19, 11]. This general property has been confirmed through numerical studies specifically for the quantity of interest here λ [8]. Numerical evidence for the similarity between the values for λ with and without this approximation [8] is reproduced in Fig. 4.

In [8, 10] it was argued that the calculation of $\Phi_W(\gamma;\Omega)$ can be cast as an eigenvalue problem. Realizing this requires that we introduce *score differences* $d(r,t)$ as defined in Fig. 3(b) and apply them to the finite width picture drawn in Fig. 2(b). Solely from the Needleman-Wunsch recursion relation given by Eq. (3), several important properties of these score differences can be derived [8, 10]: (i) the score differences can only have $n+1$ different values where n is the natural integer characterizing the choice of μ and δ according to Eq. (12); (ii) the score differences $d(r,t+1)$ can be calculated from the knowledge of the $d(r,t)$ and the random variables $\eta(r,t)$ without reference to the $h(r,t)$; (iii) the score increases $h(r,t+1)-h(r,t)$ can be calculated from the score differences $d(r,t)$ and the random variable $\eta(r,t)$. The first two statements together with the uncorrelated bonds $\eta(r,t)$ assumed in Eq. (13) imply that the dynamics of the score differences $d(r,t)$ can be viewed as a Markov process on the $(n+1)^{2W}$–dimensional state space of the equal time difference vector $(d(0,t), d(1,t), \ldots, d(2W,t))$. This Markov process may be described by a transfer matrix $\hat{\mathcal{T}}_W(0;\Omega)$. The entries of this transfer matrix encode the probabilities of the different configurations of the $\eta(r,t)$ in terms of the match probability p and the transitions between the state vectors that these configurations of the $\eta(r,t)$ imply. Finally, property (iii) allows us to modify the transfer matrix in such a way that it keeps track of the changes in the absolute score $h(r,t)$. The curious reader may refer to [10] which provides a detailed explanation of how this p-dependent modified transfer matrix $\hat{\mathcal{T}}_W(\gamma;\Omega)$ is obtained. This modified transfer matrix allows us to write

$$Z_{L,W}(\gamma;\Omega) = \boldsymbol{v}^T \hat{\mathcal{T}}_W(\gamma;\Omega)^L \boldsymbol{w} \cdot e^{\frac{\gamma L}{2}} \tag{14}$$

with some fixed $(n+1)^{2W}$–dimensional vectors $\boldsymbol{v}$ and $\boldsymbol{w}$. For large L the matrix product is dominated by the largest eigenvalue $\rho_W(\gamma;\Omega)$ which leads to

$$\Phi_W(\gamma;\Omega) = \log \rho_W(\gamma;\Omega) + \frac{\gamma}{2}. \tag{15}$$

For the very simplest scoring system consistent with condition (12), i.e., $n = 1$ where $\mu = 2\delta$, the analytical limit of $\lim_{W\to\infty} \rho_W(\gamma;\Omega)$ can be taken and Eq. (15) yields the closed analytical result [10]

$$\frac{1+\sqrt{p}\exp[\frac{\lambda}{2}(1+\mu)]}{1+\sqrt{p}\exp[-\frac{\lambda}{2}(1+\mu)]}\exp[-\frac{\lambda}{2}\mu] = 1. \tag{16}$$

For scoring systems of greater complexity, i.e., larger n, analytic solutions are not readily available for λ. Next, we will present an approach that combines the power of computational numerics and the known analytical properties of the dynamic process described above in order to calculate $\Phi(\gamma;\Omega)$.

4 Numerical Calculation for More Complex Scoring Systems

The main obstacle to obtaining the function $\Phi(\gamma;\Omega)$ (and consequently the Gumbel parameter λ) for more complex scoring systems is the extrapolation (11) of $\Phi(\gamma;\Omega)$ from its finite width counterparts $\Phi_W(\gamma;\Omega)$. In order to get a reliable estimate of the function $\Phi(\gamma;\Omega)$ we need two ingredients: First, we have to be able to calculate $\Phi_W(\gamma;\Omega)$ for as large W as reasonably possible. We will do this using the high performance numerical package ARPACK [29] as described in the next paragraph. Second, we have to extrapolate from as few finite width results as possible toward the infinite width limit $\Phi(\gamma;\Omega)$. The latter is done by using some results from statistical physics and is the main contribution of this manuscript.

The size of the state space, as well as the size of the characteristic matrix $\hat{\mathcal{T}}_W$ grows rapidly with the integer n and the width W. Even after exploiting various symmetries, the problem roughly behaves like $(n+1)^{2W}/nW$. Solving for all eigenvalues in order to discern the greatest quickly becomes exhaustively expensive for $n > 1$. However, two features of this eigenvalue problem succor this otherwise hopeless task for moderate values of n and W. First, the matrix $\hat{\mathcal{T}}_W$ is very sparse. The number of non-zero elements grows close to linearly, namely as $O(k \log k)$, where k represents the size of $\hat{\mathcal{T}}$. Second, this problem only requires the largest eigenvalue $\rho_W(\gamma;\Omega)$ and not all the eigenvalues. This makes it well suited for the implicitly restarted Arnoldi method (IRAM) [36, 28]. The numerical software package ARPACK [29], which implements IRAM, has been tested as the fastest and most dependable program for finding numerical eigenvalues and eigenvectors [27]. Indeed, in our context ARPACK allows for the quick and specific calculation of only the largest eigenvalue $\rho_W(\gamma;\Omega)$ of the sparse matrix $\hat{\mathcal{T}}_W$ for $n < 7$ for at least a few W.

Our accessible numerical solution for $\rho_W(\gamma;\Omega)$, gained via the use of the numerical software package ARPACK, directly gives $\Phi_W(\gamma;\Omega)$ by using Eq. (15). However, the solutions we can obtain for some few small widths W still skirt far from the limit of infinite W in Eq. (11). As such, $\Phi(\gamma;\Omega)$ cannot be straightforwardly approximated from the available $\Phi_W(\gamma;\Omega)$ with any real accuracy. In order to extrapolate from the $\Phi_W(\gamma;\Omega)$ for small finite widths to their infinite limit $\Phi(\gamma;\Omega)$, we make use of two results obtained in the statistical physics community. The first key result is that sequence alignment is a member of the so-called Kardar-Parisi-Zhang (KPZ) universality class [23, 22]. A universality class is a large class of problems that are known to share certain quantitative traits. The second result comes from work by Derrida *et al.*, who were able to calculate an exact solution for what amounts to our $\Phi_W(\gamma;\Omega)$ in a different system of the same universality class. Derrida *et al.* conjecture on general grounds that their exact result for the deviation function $\Phi_W(\gamma;\Omega) - \Phi(\gamma;\Omega)$ from the infinite system is given by a *universal scaling function*, i.e., that it's shape remains the same for *all* members of the KPZ universality class [16, 17]. Together, these two findings imply that our $\Phi_W(\gamma;\Omega) - \Phi(\gamma;\Omega)$ have the same functional form as the Derrida *et al.* deviation function. Expressed in our notation, this

means

$$\Phi_W(\gamma;\Omega) = \Phi(\gamma;\Omega) - \frac{a_\Omega G(\gamma W^{1/2} b_\Omega)}{W^{3/2}}. \tag{17}$$

where a_Ω and b_Ω are unknown scaling factors dependent on the particular parameters of the alignment Ω and the scaling function G has been explicitly solved [16, 17] (see appendix A). In order to use property (17) to extrapolate $\Phi(\gamma;\Omega)$ from $\Phi_W(\gamma;\Omega)$, a_Ω and b_Ω must be determined. To that end, we take the difference

$$\Phi_W(\gamma;\Omega) - \Phi_{W-1}(\gamma;\Omega) = \frac{a_\Omega G(\gamma W^{1/2} b_\Omega)}{W^{3/2}} - \frac{a_\Omega G(\gamma (W-1)^{1/2} b_\Omega)}{(W-1)^{3/2}} \tag{18}$$

allowing us to eliminate the unknown function $\Phi(\gamma;\Omega)$. We can numerically evaluate the left hand side of this equation as a function of γ. Knowing the exact form of G means that on the right hand side only the scales, controlled by a_Ω and b_Ω remain undetermined. The act of finding these two scaling factors then becomes a matter of fitting the left hand side to the right hand side of Eq. (18). Once a_Ω and b_Ω have been determined, all that remains is to solve for λ using Eqs. (9) and (17)

$$\Phi(\lambda;\Omega) = \Phi_W(\lambda;\Omega) - \frac{a_\Omega G(\lambda W^{1/2} b_\Omega)}{W^{3/2}} = 0. \tag{19}$$

The specifics of the computer algorithm used in determining λ follow. First, we use computer algebra to generate the structure of the transfer matrices $\hat{\mathcal{I}}_W$ for different n and W. This process consumes a great deal of time, however, once done for every of the discrete combinations of n and W, the form of the transfer matrices are recorded and can be reused for any choice of mismatch cost μ (which fixes the gap cost δ according to condition (12)) and match probability p. Once μ and p are supplied and the numerical transfer matrix is tabulated, the numerical tool ARPACK obtains the eigenvalues $\rho_W(\gamma;\Omega)$ and $\rho_{W-1}(\gamma;\Omega)$ for $\gamma = 0.8\lambda_{gapless}, 0.9\lambda_{gapless}$ and $\lambda_{gapless}$ where $\lambda_{gapless}$ is the Gumbel parameter λ for the same μ and p in the absence of gaps (as calculated by using Eq. (6)). These initial values, along with the tabulated function G, obtained from KPZ theory, allow for the first approximations of a_Ω and b_Ω to be calculated. The newly found scaling factors are then used along with Eq. (19) in order to choose a new $\gamma \approx \lambda$ by linear extrapolation toward the root. $\rho_W(\gamma;\Omega)$ and $\rho_{W-1}(\gamma;\Omega)$ for this new γ are then evaluated. The whole set of ρ-values then feeds into the reevaluation of a_Ω and b_Ω. This process iterates until γ converges to the solution for λ.

5 Results for More Complex Scoring Systems

Table 1 summarizes the performance of the computer program outlined in section 4. For each value of the integer n, a combination of μ and δ is chosen that leads to a gapped λ of approximately $\lambda \sim 0.8\lambda_{gapless}$. This is considered to be the most relevant region for similarity searches. Most importantly, table 1

n	μ	δ	W	time(seconds)	error(%)	n	μ	δ	W	time(seconds)	error(%)
1	2.2	1.1	2	0.4	0.1	3	0.9	2.35	2	<0.1	1.6
			3	0.4	0.5				3	0.1	0.3
			4	0.3	0.1				4	0.3	0.2
			5	0.2	<0.1				5	4.3	0.2
			6	0.2	<0.1				6	108.2	0.2
			7	0.2	<0.1	4	0.7	2.9	2	0.1	8.7
			8	0.3	<0.1				3	0.1	<0.1
			9	1.4	<0.1				4	1.0	0.2
			10	4.7	<0.1				5	39.0	0.2
			11	26.4	<0.1	5	0.7	3.75	2	0.1	<0.1
2	1.5	2.0	2	<0.1	0.2				3	0.2	<0.1
			3	0.1	0.4				4	5.3	<0.1
			4	0.3	0.4	6	0.55	4.15	2	0.1	3.2
			5	0.6	0.4				3	1.1	0.3
			6	2.2	0.4				4	39.5	<0.1
			7	26.5	0.4						

Table 1. This table shows the calculation time and precision with which our algorithm performs. This table was generated using a 2.4 GHz Intel®Xeon™processor. The error percentages are based on comparisons of results obtained in the range where $\lambda \sim 0.8\lambda_{gapless}$ using the island method [4]. The exception is for $n = 1$, where we have an analytical solution (16), we calculate the error based on the results obtained through the known equation. It should be noted that the percent error inherent in the island method for the simple scoring system described by Eq. (1) is 0.5%.

shows us that λ converges for $W \geq 4$. (Note that, except for $n = 1$ where the reference value for λ is determined from the exact equation (16), the statistical error on the numerically determined reference values is 0.5% in and of itself.) Our method lands almost all values within the error range of the numerically determined values. This result verifies just how reliably the finite size effects of W are taken into account by the scaling form presented by Derrida *et al.* Secondly, table 1 shows that the evaluation of the Gumbel parameter λ by our new method for all but the largest W (which are unnecessary), finishes in about a second or less. This compares very favorably with the fastest currently available alternatives for obtaining λ, i.e., the island method [4]. The major disadvantage of using the island method for DNA significance assessment lies in the amount of time needed in order to accurately evaluate each data point — the same machine that produced the times in table 1 requires approximately a fortnight in order to obtain an accuracy of 0.5%.

Fig. 4 gives an overview of the dependence of λ on the mismatch cost μ for different n. The lines show the values obtained by our method. In the $n = 1$ case, the solution of Eq. (16) is plotted as well. It is quasi-indistinguishable from the results of our new algorithm. For $n > 1$ the only way to obtain reference values for comparison is by the island method the results of which are shown as the points. Still, our method is within the statistical error of the numerical data of the island method over the whole parameter range.

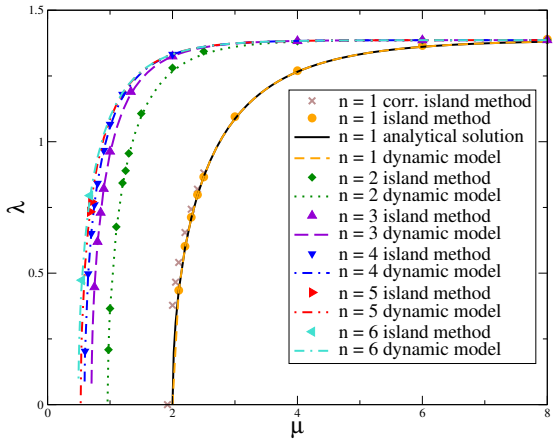

Fig. 4. Values of λ for the DNA alphabet ($p = 0.25$) as a function of the mismatch cost μ. The lines are the results of our new approach; the points are results from stochastic simulation with the island method. The λ-values for $n = 1$ verify well when plotted against solutions of Eq. (16) (also shown as a line barely distinguishable from the line representing the values calculated by our new method) and the island method in [4]. The λ-values for $n = 2, 3, 4, 5$, and 6 displayed here also match well with values obtained from the island method. For $n = 1$, we also include points obtained without use of the uncorrelated approximation Eq. (13). The correlated data obtained via the island method generally compares well with the uncorrelated points and only changes the value of λ slightly. The estimated error of the island method is approximately one quarter of the symbol sizes.

6 Conclusions

We have presented a new numerical method to reliably calculate λ with great accuracy and very little computational effort. The efficiency and dependability of this method in characterizing the difficult tail end of the Gumbel distribution removes the major impediment to gapped significance assessment. As previously stated, the remaining Gumbel parameter may be obtained from direct simulation Furthermore, this algorithm grants real time access to the Gumbel parameters and allows for the possibility of updating schemes such as PSI-BLAST to run without resorting to a small set of pre-computed values. The gains in the ability to calculate these parameters not only aids sequence comparison tools but also furthers our ability to discern the most appropriate scoring schemes. We believe that adaptation of these methods is possible for values of μ and δ that do not adhere to the technical condition we imposed for the purpose of our work. This includes the biologically practical and often used affine gap cost schemes. Indeed, future efforts will be directed at using these methods for the more complicated affine gap costs as well as for correlated sequence alignments.

7 Acknowledgments

RB gratefully acknowledges funding from the National Science Foundation through grants DBI-0317335 and DMR-0404615.

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. Altschul, S.F., Bundschuh, R., Olsen, R., and Hwa, T. 2001. The estimation of statistical parameters for local alignment score distributions. *Nucl. Acids Res.* **29**, 351–361.
5. Boutet de Monvel, J. 1999. Extensive Simulations for Longest Common Subsequences. *Europ. Phys. J. B* **7**, 293–308.
6. Boutet de Monvel, J. 2000. Mean-field Approximations to the Longest Common Subsequence Problem. *Phys. Rev. E* **62**, 204–209.
7. Bundschuh, R., and Hwa, T. 2000. An analytic study of the phase transition line in local sequence alignment with gaps. *Disc. Appl. Math.* **104**, 113–142.
8. Bundschuh, R., 2000. An analytic approach to significance assessment in local sequence alignment with gaps. *Proceedings of the fourth annual international conference on computational molecular biology (RECOMB2000)*, S. Istrail *et al.*, eds., ACM press, (New York, NY), 86–95.
9. Bundschuh, R. 2001. High Precision Simulations of the Longest Common Subsequence Problem. *Europ. Phys. J. B* **22**, 533–541.
10. Bundschuh, R., 2002. Asymmetric exclusion process and extremal statistics of random sequences. *Phys. Rev. E* **65** 031911.
11. Chia, N. and Bundschuh, R. 2004. Finite Width Model Sequence Comparison. *Phys. Rev. E* **70** 021906.
12. Collins, J.F., Coulson, A.F.W., and Lyall, A. 1988. The significance of protein sequence similarities. *CABIOS* **4**, 67–71.
13. 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.
14. Dančík, V., Paterson, M. 1994. Longest Common Subsequences. *Proceedings of 19th International Symposium Mathematical Foundations of Computer Science, Lecture Notes in Computer Science* **841**, 127–142.
15. Dančík, V. 1994. Expected Length of Longest Common Subsequences. *PhD thesis, University of Warwick*.
16. Derrida, B. and Lebowitz, J.L. 1998. Exact Large Deviation Function in the Asymmetric Exclusion Process, *Phys. Rev. Lett.* **80**, 209–213.
17. Derrida, B. and Appert, C. 1999. Universal Large-Deviation Function of the Karder-Parisi-Zhang Equation in One Dimension, *J. Stat. Phys.* **94**, 1–30.
18. Doolittle, R.F. 1996. *Methods in Enzymology* **266**, San Diego, Calif.: Academic Press.

19. Drasdo, D., Hwa, T., and Lassig, M. 2001. Scaling Laws and Similiarity Detection in Sequence Alignment with Gaps. *J. Comp. Biol.* **7**, 115–141.
20. Gumbel, E.J. 1958. *Statistics of Extremes*, Columbia University Press, (New York, NY).
21. Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919.
22. Hwa, T. and Lässig, M. 1996. Similiarity-Detection and Localization. *Phys. Rev. Lett.* **76**, 2591–2594.
23. Kardar, M., Parisi, G., and Zhang, Y.C. 1986. Dynamic Scaling of Growing Surfaces. *Phys. Rev. Lett.* **56**, 889–892.
24. 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.
25. Karlin, S., and Dembo, A. 1992. Limit distributions of the maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Prob.* **24**, 113–140.
26. 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.
27. Lehoucq, R.B. and Scott, J.A. 1996. An evaluation of software for computing eigenvalues of sparse nonsymmetric matrices. preprint MCS-P547-1195, Argonne National Laboratory, Argonne, IL.
28. Lehoucq, R.B. 1997. Truncated QR algorithms and the numerical solution of large scale eigenvalue problems. preprint MCS-P648-0297, Argonne National Laboratory, Argonne, IL.
29. Lehoucq, R.B., Sorensen, D.C., and Yang, C. 1997. *ARPACK Users' Guide: Solutions of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM,(Philadelphia, PA)
30. Mott, R. 1992. Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59–75.
31. Mott, R., and Tribe, R. 1999. Approximate statistics of gapped alignments. *J. Comp. Biol.* **6**, 91–112.
32. Mott, R. 1999. Accurate estimate of p-values for gapped local sequence alignment. Private communication.
33. 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.
34. 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).
35. 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.
36. Sorensen, D.C. 1992. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM J. Matrix Analysis and Applications* **13** 357–385.
37. Siegmund, D., and Yakir, B. 2000. Approximate p-values for Sequence Alignments. *Ann. Statist.* **28** 657–680
38. Smith, S.F., and Waterman, M.S., 1981. Comparison of biosequences.*Adv. Appl. Math.* **2**, 482–489.
39. Smith, T.F., Waterman, M.S., and Burks, C. 1985. The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* **13**, 645–656.

40. Waterman, M.S., Gordon, L., and Arratia, R. 1987. Phase transitions in sequence matches and nucleic acid structure, Proc. Natl. Acad. Sci. U.S.A. **84**, 1239–1243.
41. Waterman, M.S., and Vingron, M. 1994. Sequence Comparison Significance and Poisson Approximation. *Stat. Sci.* **9**, 367–381. v
42. 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.
43. Waterman, M.S. 1994. *Introduction to Computational Biology.* London, UK: Chapman & Hall.

A Deviation Function of the Particle Hopping

The deviation function G_D as solved by Derrida *et al.* is independent of the model parameters and has the following parametric form [16, 17]

$$\beta = \frac{2}{\sqrt{\pi}} \int_0^\infty \epsilon^{1/2} \frac{Ce^{-\epsilon} d\epsilon}{1 + Ce^{-\epsilon}} \tag{20}$$

$$G_D(\beta) = \frac{4}{3\sqrt{\pi}} \int_0^\infty \epsilon^{3/2} \frac{Ce^{-\epsilon} d\epsilon}{1 + Ce^{-\epsilon}}. \tag{21}$$

As C approaches -1 we require a new representation to go beyond $\beta_- = \lim_{C \to -1} \beta$. The analytical continuation of $G_D(\beta)$ is beyond β_- given by the parametric equations [16, 17]

$$\beta = -4\sqrt{\pi} \left[-\ln(-C) \right]^{1/2} - \sum_{q=1}^{\infty} (-C)^q q^{-3/2} \tag{22}$$

$$G_D(\beta) = \frac{8}{3}\sqrt{\pi} \left[-\ln(-C) \right]^{3/2} - \sum_{q=1}^{\infty} (-C)^q q^{-5/2}, \tag{23}$$

as C for these equation varies between 0 and -1, this gives the function $G_D(\beta)$ for all $\beta < \beta_-$.

In the limit as $\beta \to -\infty$ [16, 17],

$$G_D(\beta) \approx -\frac{\beta^3}{24\pi} \tag{24}$$

implying that for large γ,

$$W^{-3/2} G_D(\gamma W^{1/2}) \approx -\frac{\gamma^3}{24\pi}. \tag{25}$$

This term is independent of W, i.e., of finite size effects. In order to appropriately reflect this, we include this W-independent term in Φ. Therefore, the function G used in our methodology relates to the Derrida *et al.* solution for $G_D(\beta)$ via the equation $G(\beta) = G_D(\beta) + \beta^3/(24\pi)$.