Local Sequence Alignments Statistics: Deviations from Gumbel Statistics in the Rare-Event Tail

Stefan Wolfsheimer *[1], Bernd Burghardt [1] and Alexander K Hartmann [1]

[1] Institut für Theoretische Physik, Universität Göttingen, 37077 Göttingen, Friedrich-Hund-Platz 1, Germany

Email: Stefan Wolfsheimer *- wolfsh@theorie.physik.uni-goettingen.de; Bernd Burghardt - burghard@theorie.physik.uni-goettingen.de; Alexander K Hartmann - hartmann@physik.uni-goettingen.de;

*Corresponding author

Abstract

Background: The optimal score for ungapped local alignments of infinitly long random sequences is known to follow a Gumbel extreme value distribution. Less is known about the important case, where gaps are allowed. For this case, the distribution is only known empirically in the high-probability region, which is biologically less relevant.

Results: We provide a method to obtain numerically the biologically relevant rare-event tail of the distribution. The method, which has been outlined in an earlier work, is based on viewing a pair of sequences as the degrees of freedom of a physical system held at a finite temperature. Here, we first present the approach in detail and evaluate the convergence of the algorithm by considering a simple test case. In the earlier work, the method was just applied to one single example case. Therefore, we consider here a large set of parameters: We study the distributions for protein alignment with different substitution matrices (BLOSUM62 and PAM250) and affine gap costs with different parameter values. In the logarithmic phase (large gap costs) it was previously assumed that the Gumbel form still holds, hence the Gumbel distribution is usually used when evaluating p-values in databases. Here we show that for all cases, provided that the sequences are not too long ($L > 400$), a "modified" Gumbel distribution, i.e. a Gumbel distribution with an additional Gaussian factor is suitable to describe the data. We also provide a "scaling analysis" of the parameters used in the modified Gumbel distribution. Furthermore, via a comparison with BLAST parameters, we show that significance estimations change considerably when using the true distributions as presented here. Finally, we study also the distribution of the sum statistics of the k best alignments.

Conclusions: Our results show that the statistics of gapped and ungapped local alignments deviates significantly from Gumbel in the rare-event tail. We provide a Gaussian correction to the distribution and an analysis of its scaling behavior for several different scoring parameter sets, which are commonly used to search protein data bases. The case of sum statistics of k best alignments is included.

Background

Sequence alignment is a powerful tool in bioinformatics [1, 2] to detect evolutionarily related proteins by comparing their sequences of amino acids. Basically one wants to determine the "similarity" of the sequences. For example, given a protein in a database like PDB [3], such similarity analysis can be used to detect other proteins, which are evolutionary close to it. Related approaches are also used for the comparison of DNA sequences, i.e. shotgun DNA sequencing [4], but the application to DNA is not considered in this article.

Alignment algorithms find optimum alignments and maximum alignment scores S of two or more sequences for a given scoring system. Needleman and Wunsch suggested a method to compute global alignments [5], whereas the Smith-Waterman algorithm [6] aims at finding local similarities. Insertions and deletions of residues are taken into account by allowing for gaps in the alignment. Gaps yield a negative contribution to the alignment score and are usually modeled by a gap-length depending function $g(l)$. Widely used are affine gap costs because for two given sequences of length L and M, fast algorithms with running time $\mathcal{O}(LM)$ are available for this case [7]. Note that for database queries even this is too complex, hence fast heuristics like BLAST [8] are used here.

By itself, the alignment *score*, which measures the similarity of two given sequences, does not contain any information about the statistical significance of an alignment. One approach to quantify the statistical significance is to compute the probability that for a given optimum score S, this value or a higher one occurs by pure chance. Therefore, the probability density P of the optimal score S of random sequences is an essential quantity in the interpretation of the data. Analytic expressions for $P(S)$ are only known asymptotically in the case of gapless alignments of long sequences, where an *extreme value distribution* (also called *Gumbel distribution*) [9, 10] was found. For alignments with gaps, such analytical expressions are not available. Approximation for scenarios with gaps based on probabilistic alignment [11–13], large deviations [14] and a Possion model [15] had been developed. Altschul and Gish [16] investigated the score statistics of random sequences for a number of scoring systems and gap parameters by computer simulations: They obtained histograms of optimum scores for randomly sampled pairs of sequences by simple sampling. By curve fitting, they showed that in the region of high probability the extreme value distribution describes the data well, also for gapped alignments of finite sequences. Additionally, they found that the theoretical predictions for the relation between the scoring system on one side and the Gumbel parameters on the other side hold approximately for gapped alignments. In this context they obtained two improvements: Using a correction to account for finite sequence lengths and sum statistics of the k-best alignments, theoretical predictions for ungapped alignments could be applied more accurately to gapped alignments. Recently Olsen et al. introduced the "island method" [17,18], which accelerates sampling time. BLAST [8] uses precomputed data, generated with the island method, to estimate p-values. This means one wants to know the probability for the occurence of at least one hit with a score S greater than or equal to some given threshold value b

$$\mathbb{P}(S \geq b).$$

Often E-values are used instead. They describe the number of expected hits with a score greater than or equal to some threshold value. In any case, as already pointed out, the studies in Ref. [16] and [18] give reliable data in the region where $P(S)$ is large only. This is outside the region of biological interest because pairs of biologically related sequences have a higher similarity than pairs of purely randomly drawn sequences.

To overcome this drawback a rare-event sampling technique was proposed recently [19], which is based on methods from statistical physics. This general approach allows to obtain the distribution over a wide range down to 10^{-40}. So far this method has been applied to one relevant case only, namely protein alignment with the BLOSUM 62 score matrix [7] and affine gap costs with $\alpha = 12$ opening and $\beta = 1$ extension costs. It turned out that at least for one scoring matrix and one set of gap-cost parameters, the distribution deviates from the Gumbel form in the biologically relevant rare-event tail, where simple sampling methods fail. Empirically, a Gaussian correction to the original distribution was proposed for this case.

Results as in Ref. [19] are only useful if one obtains the distribution for a large range of parameter values which are commonly used in bioinformatics. It is the purpose of this work to study the distribution P for other relevant cases. Here we consider the `BLOSUM62` and the `PAM250` score matrices in connection with various parameters α, β of affine gap costs.

The paper is organized as follows. In the second section we define alignments formally and state a few main results on the statistics of local sequence alignment. Next, we explain some basic notions from statistical physics, which are necessary to understand the rare-event approach used here. In the fourth section we explain our approach in detail. We introduce some toy examples which are also used to evaluate the convergence properties of the algorithm. In the fifth section, we present our results for BLOSUM62 and PAM 250 matrices in conjunction with different affine gap costs. We show also our results for the sum statistics of the k largest alignments. In the last section, we summarize and discuss our results.

Statistics of local sequence alignment

In this section, we define sequence alignment, and state some analytical results for the distribution P of the optimum scores S over pairs of random sequences.

Let $\mathbf{x} = x_1 x_2 \ldots x_L$ and $\mathbf{y} = y_1 y_2 \ldots y_M$ be two sequences over a finite alphabet Σ with $r = |\Sigma|$ letters (e.g. nucleic acids or amino acids). An *alignment* $\mathcal{A}$ is a set $\mathcal{A} = \{(i_k, j_k)\}$ of K pairs of "non-crossing" indices ($k = 1, 2, \ldots, K$, $1 \leq i_k < i_{k+1} \leq L$ and $1 \leq j_k < j_{k+1} \leq M$) identifying pairs of letters from the two sequences. Letters, which are not paired are called *unpaired* or *gapped*. A *gap g of length l_g* is a substring of l_g gapped letters from one sequence. Note, that this representation [14] of an alignment is equivalent to an introduction of a gap symbol, as commonly used. Formally the gap cost function can be defined by considering the length of a gap beginning at the kth pairing in sequence (1) or sequence (2) respectively, in detail

$$
\begin{aligned}
l_g^{(1)}(k) &= i_{k+1} - i_k - 1 \\
l_g^{(2)}(k) &= j_{k+1} - j_k - 1.
\end{aligned}
$$

The *score* $\tilde{S}(\mathbf{x}, \mathbf{y}, \mathcal{A})$ of the local alignment of the two sequences is composed of a sum over all aligned pairs and a sum over all gaps of both sequences:

$$
\begin{aligned}
\tilde{S}(\mathbf{x}, \mathbf{y}, \mathcal{A}) &= \sum_{k=1}^{K} s(x_{i_k}, y_{j_k}) + \sum_{\text{gaps } g} g(l_g) \\
&= \sum_{k=1}^{K} s(x_{i_k}, y_{j_k}) + \sum_{k=1}^{K-1} \left\{ g(l_g^{(1)}(k)) + g(l_g^{(2)}(k)) \right\} \qquad (1)
\end{aligned}
$$

where $s(a, b)$ $a, b \in \mathcal{A}$ is the given *score matrix* (or *substitution matrix*) and $g(l)$ the *gap-cost function* with $g(0) = 0$. Note that the alignment is local, because the (possibly large) gaps at the beginning and the end of each sequence are not included in the scoring function. Otherwise the alignment would be global. Here, we consider the BLOSUM62 [20] and the PAM250 [21, 22] matrices and affine gap costs, i.e. $g(l) = \alpha + \beta(l - 1)$. The similarity of the sequences is the optimum alignment with the maximum score $S(\mathbf{x}, \mathbf{y}) = \max_{\mathcal{A}} \tilde{S}(\mathbf{x}, \mathbf{y}, \mathcal{A})$, which can be obtained in $\mathcal{O}(LM)$ time [7].

In the case of gapless optimum local alignments of two random sequences of L and M independent letters from Σ with frequencies $\{f_a\}(a \in \Sigma)$, the score statistics can be calculated analytically in the asymptotic regime of long sequences [9, 10]. In this case one obtains the Gumbel distribution (Karlin-Altschul statistics) [23]

$$
P(S) = \lambda \exp\left[-\lambda\left(S - S_0\right) - e^{-\lambda(S - S_0)} \right]. \qquad (2)
$$

The parameters λ and S_0 of Eq. (2) can be derived directly from the score matrix $s(a, b)$ and frequencies f_a [9, 10].

As pointed out by Altschul and Gish [16], in finite systems there occur edge effects: An alignment may extend to the end of either sequence and the score will be distorted towards lower values and high scores become less probable. Since this effect vanishes in the limit of infinite sequences, the tail of Eq. (2) can be understood as an upper bound for finite sequences.

Arratia and Waterman [24] predicted a phase transition between a linear phase and a logarithmic phase, i.e. a linear growth of the excepted score as a function of the sequence length, changing to a logarithmic growth with increasing gap costs. In the linear phase an optimum alignment may spread over a large range of the sequences and the statistical theory breaks down. However, only the logarithmic phase is of interest in biological questions because the alignment algorithm becomes more sensitive in this phase, especially near the threshold [25].

Often the sensitivity of an alignment algorithm can be increased by not only considering the best optimal alignment score, but also the k-best scores of non overlapping alignments. An $\mathcal{O}(LM)$ algorithm for this task, based on Sellers concept of local optimality, was developed [26, 27]. According to Karlin and Altschul [28] also the sum statistics of the k-best alignment scores for random sequences can be derived analytically for asymptotically long sequences. The probability density distribution f for the sum of the k-best normalized scores $T_k = \lambda \sum_i^k (S_i - S_0)$ (λ and S_0 are the corresponding Gumbel-parameters for the optimal alignment) is given by the integral

$$f(t) = \frac{e^{-t}}{k!\,(k-2)!} \int_0^\infty y^{k-2} \exp(-e^{(y-t)/k}) dy. \tag{3}$$

In the tail the distribution is well approximated by

$$f_{\text{tail}}(t) = \frac{e^{-t}}{k!(k-1)!} \left[t^{k-1} - (k-1)t^{k-2})\right]. \tag{4}$$

Below we will compare the results of our numerical studies to this distribution in the tail of the data for values $k = 2, \ldots, 5$.

Statistical mechanics and Monte Carlo simulations

As already pointed out, the main purpose of this paper is to calculate the tail of the distribution of optimum scores of gapped local alignments over pairs of sequences of finite lengths. Our approach is based on concepts of statistical mechanics. For many non-physical applications like economical models [29], combinatorial optimization problems [30], neural networks [31] and other biological problems [32], reformulations in terms of statistical mechanics systems have been proven to be useful. Also sequence alignment has attracted attention in the statistical physics community, because it can be seen as complex physical system, with interesting properties, allowing to derive some alignment properties analytically. One possible mapping [33, 34] is based on the equivalence to the model of *directed polymers in a random medium* , where the energy is computed by a shortest path algorithm in a similar way as the global alignment score. Additionally, there is another reason, why a statistical mechanics approach is useful for studying sequence alignment. At least approximately, the distribution of local alignment follows a Gumbel distribution, which exhibits an exponential behavior in the tail. This is the same functional form as the *Boltzmann distribution* (see below), which describes the behavior of a physical system held at a constant temperature, the so called *canonical ensemble*. This similarity allows us to adopt well established sampling methods for the case of local sequence alignment, which were originally developed for physical systems. Additionally, this permits us to easily derive an analytical expression, which allows, in principle, to obtain the *full* distribution of the optimum alignment score from the simulation data. To make our approach accessible to non-physicists, we next review the basic concepts from statistical mechanics we need here.[1]

Note that the energy E and the optimal alignment score S have different conventions: A *low* energy of a physical system corresponds to an ordered state with low entropy, i.e. a state which looks everywhere similar. On the other hand similar sequences are indicated by a *large* alignment score. Hence a natural interpretation of the alignment energy is $E = -S$.

[1] Note that one could formulate all algorithms we have used in terms of non-physical expressions and use just "a scaling parameter" instead of a "temperature". Nevertheless, since the approaches described here originate in physics and have a descriptive meaning, we use the statistical mechanics notations here as well.

A central quantity in statistical mechanics is the density of states (DOS) $g(E)$ (the number of micro states of energy E of a physical system). Other quantities, like the free energy of the system, can be computed directly from it. Note that the configurations of a system with the lowest energy $E_{\min}$ are called *ground states*. At finite temperature T the probability $\rho_T(E)$ in the canonical ensemble for the system to be in a configuration R of energy $E = E(R)$ follows the Boltzmann distribution, i.e.

$$\rho_T(E) = \frac{1}{Z_T} g(E) e^{-\beta E}, \tag{5}$$

with $\beta = \frac{1}{k_B T}$ and k_B the Boltzmann constant. For convenience, all temperatures are measured in terms of k_B, which means we can set $k_B = 1$.

The normalization constant in Eq. (5) $Z_T = \sum_E g(E) e^{-\beta E}$ is known as the *partition function.* [2] Hence, when lowering the temperatures $T \to 0$, only the ground states contribute to measurable quantities. In general, averages of physical quantities $A(E)$, which depend only on the energy, can be computed by the canonical average

$$\langle A \rangle = \sum_E \rho_T(E) A(E) = \frac{1}{Z_T} \sum_E A(E) g(E) e^{-\beta E}.$$

For small systems one may enumerate all configurations and derive the DOS. This is unpractical for big systems, because the number of configurations is usually increasing exponentially with system size. Monte Carlo methods are fruitful solutions to such problems. A naive sampling scheme, with uniform transition probabilities, just generates a number of random configurations of the underlying model. This leads to a high variance of the measurable quantities, because many points with a small Boltzmann factor $e^{-\beta E}$ are selected. The Metropolis algorithm [35] overcomes these difficulties, because the configurations are visited in the stationary state with a frequency proportional to their Boltzmann weight. The algorithm is based on a Markov chain where in each step (the current configuration being called R), first a *trial configuration* $R^\star$ is generated randomly (with uniform weight in the simplest case, or which might depend on R) and then *accepted* with acceptance probabilitiy $\tilde{p}(R \to R^\star)$, which depends suitably on the energy change $\Delta E = E(R^\star) - E(R)$ between the current and the trial configuration: In the standard case, to sample according to Eq. (5), the acceptance probabilities are given by

$$p(R \to R^\star) = \begin{cases} 1 & \text{if } \Delta E < 0 \\ e^{-\beta \Delta E} & \text{else} \end{cases}. \tag{6}$$

If the trial configuration is not accepted, the previous configuration is kept for the next time step. A generic version of the algorithm is shown in Algorithm 1, note that `RAND()` returns a random number in $[0, 1]$. Practically, since one wants to sample the configuration space according the equilibrium distribution, a Metropolis Monte Carlo simulation is divided into two parts: First an equilibrated state is approached by a long simulation starting from a given configuration (equilibration run). Then the simulation is continued and physical properties are measured and averaged over this run. The question, whether the equilibrium is reached after the equilibration run is often not easily answered, because correlations on very long timescales are often not visible in short equilibration runs. In the next section we will give a criterion, which we use here. In any case, one can speedup equilibration be using techniques like *parallel tempering*, which is explained next.

Parallel tempering and flat histogram methods

Parallel Tempering has become a standard tool in disordered systems with a rough free energy landscape [36]. These rough energy landscapes are characterized by high energy barriers and can be found for problems like protein folding [37–41], nucleation [42], spin-glasses [43, 44] and other models characterized by rare events [19, 45].

[2] We assume only discrete energy configurations here

```
initalize system with some configuration
compute energy E

FOREACH step
{
  generate trial configuration
  compute energy E_t of trial configuration
  IF (E_t < E) OR
     (EXP(-beta*(E_t-E) > RAND()) THEN
  {
    replcace current configuration with trial configuration
    set E=E_t
  }
  ELSE
  {
    reject trial conifguration (keep current one)
  }
}
```

Algorithm 1: The Metropolis algorithm

To elucidate how one benefits from this, we first consider the standard case of the Monte Carlo simulation at fixed temperature T. Usually one starts from a random (i.e. disordered) configuration and equilibrates the system. At the beginning of the simulation the system has a high energy and hence it can overcome many barriers in the landscape. If the temperature is low, typically the energy decreases or stays the same during the simulation with only few energy-increasing fluctuations. Hence, the system gets trapped into a local minimum. This means, within one run, one samples only the local vicinity of one local minimum. For this reason, usually the equilibration procedure is repeated many times, where the system ends up in a different local minimum each time. Note that if ground states are also known, i.e. the energetic minima of the system which usually exhibit a high degree of ordering, the reverse process is possible, i.e. starting from a deep minimum and sampling its local environment. One can use this fact to verify, whether a system has equilibrated on a larger scale, i.e. whether it is able to overcome the typical energy barriers. This is the case when the average behavior for two runs, one starting with a disordered configuration and one starting with an ordered configuration, is the same (within fluctuation). If the temperature is too small, this is usually not possible.

It turned out that the parallel tempering method [46,47] accelerates the equilibration process remarkably. In this approach N_T replica $R_1, ..., R_{N_T}$ of the system are simulated simultaneously with different temperatures $T_1, T_2, ...T_{N_T}$. Apart from conventional Monte Carlo steps in each replica one also attempts to induce flips between two neighboring temperatures. The acceptance probability for a replica flip step between R_i and R_{i+1} with inverse temperatures β_i and β_{i+1} and energies E_i and E_{i+1} is given by the product of the probability that the energy at β_i changes from E_i to E_{i+1} times the probability of change from E_{i+1} to E_i at β_{i+1}. This leads to the following Metropolis acceptance probability

$$p\left(R_i \leftrightarrow R_j\right) = \max\left(1, \exp(-\Delta\beta\Delta E\right), \quad (7)$$

where $\Delta E = E_i - E_{i+1}$ and $\Delta\beta = \beta_i - \beta_{i+1}$. The replica flip steps allow the replica to perform a random walk in temperature space, hence occasionally high temperatures are visited for all replica in a controlled manner. By using the Metropolis acceptance rate also for this replica flip, it is guaranteed that the stationary distribution is given by the Boltzmann distribution at all temperatures. At high temperatures, energy barriers can be overcome and hence equilibration is strongly accelerated with respect to the standard Metropolis algorithm. For completeness, we mention the methods of Berg and Neuhaus [48] developed in 1992 in the context of first-order phase transitions. They introduced so called "multicanonical sampling" to overcome high energy barriers in physical systems. The concept is very general and had influenced a number of modern "flat histogram" algorithms in statistical physics, like Wang-Landau sampling [49, 50] or our approach.

Obtaining the density of states

Here we describe how to compute the DOS for a physical system by the knowledge of finite temperature energy density $\rho_T(E)$, which are accessible from the results of parallel tempering at different temperatures. From Eq. (5) we know the DOS up to a normalization constant (the partition function Z_T)

$$g(E) = \frac{1}{Z_T}\rho_T(E)e^{\beta E}. \tag{8}$$

Different temperatures yield different energy windows of the distribution. If the temperatures are chosen such that the energy windows overlap, it is always possible to find a normalization constant for each temperature, such that the ordinate values of neighboring distributions match at the same abscissae. With this procedure one may construct the complete distribution window by window from the infinite temperature limit down to small temperatures, up to one single global (irrelevant) normalization constant. This allows to obtain the DOS even in the rare-event tail, where the probabilities get exponentially small. The method is limited by the lowest temperature, where equilibration is possible, i.e. the lowest temperature determines the smallest probability of $g(E)$ which is accessible.

Sampling of rare-events in local sequence alignments

As mentioned in the introduction, empirical studies of the probability $P(S)$ are essential for the interpretation of the score. Here we provide a method to access the tails of these probability distributions, which are out of reach for simple sampling methods. Our approach is based on mapping the sequence alignment-problem on a physical system with a pair of sequences $(\mathbf{x}, \mathbf{y})$ being a physical configuration and the negative optimum score being the energy of the system $E = -S$. In this picture $P(S)$ equals the density of states $g(-S)$, which is obtained from the numerical data using Eq. (8). In the tail the Gumbel distribution is dominated by the factor $P(S) = e^{-\lambda(S-S_0)}$, which is similar to the Boltzmann factor in Eq. (5). Hence, the rare event sampling method based on performing Monte Carlo simulations at different temperatures, as described in the previous section, is a reasonable method to obtain $P(S)$ over a wide range down to the tail.

Details of the method by a simple example

In this section we describe the details of the parallel tempering method used here. We considered a small toy model consisting of a 4-letter alphabet with pairs $(\mathbf{x}, \mathbf{y})$ of sequences of length $L = M$. Each letter $a \in \Sigma$ occurs independently with the probability $f_a = 1/4$. As explained above, each pair of sequences is a *configuration* of the corresponding physical system. The scoring matrix is just given by

$$s(a, b) = \begin{cases} +1 & \text{if } a = b \\ -3 & \text{else} \end{cases}. \tag{9}$$

We used affine gap costs with $\alpha = 4$ and $\beta = 2$ for computing the optimal local alignment scores, which is interpreted as the negative energy of a configuration. Obviously, this model has $Z = 4^{2L}$ configurations. The maximum score over the ensemble of all possible configurations is $S_{\max} = L$ (i.e. $E_{\min} = -L$ is the *ground state energy*). This corresponds to a pair of sequences with L equal letters $x_i = y_i$ ($i = 1 \dots L$). The degeneracy of the ground state, i.e. the number of configurations with the lowest energy, is 4^L. Hence, the probability to find a maximum score among all random sequences is $P(S_{\max}) = 4^L/4^{2L} = 4^{-L}$. Below, to benchmark

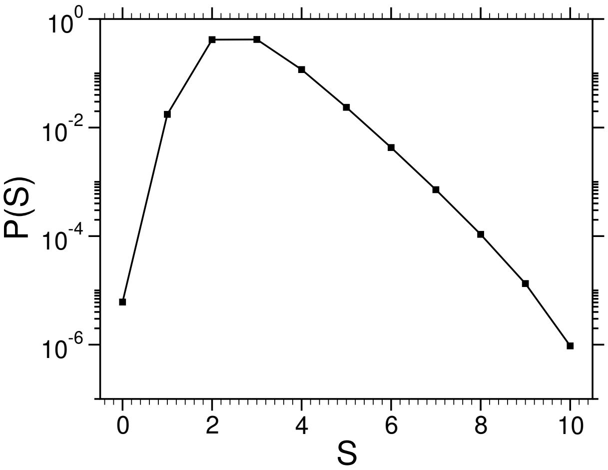

Figure 1: Score probability density for a 4-letter system with sequence-length 10 and scoring parameters Eq. (9) using affine gap costs ($\alpha = 4,\ \beta = 2$). The density had been obtained by exact enumeration of all $4^{2\times 10}$ configurations.

the Monte Carlo algorithm, we compare the convergence of the relative error $\epsilon(S) = \frac{|P_{\rm sample}(S_{\rm max})-4^{-L}|}{4^{-L}}$ for different sequence lengths, $P_{\rm sample}(S)$ being the corresponding probability obtained from the MC simulation.

For small systems one may enumerate all possible configurations and compare the complete distribution with the Monte Carlo data. Fig. 1 shows the exact result for $L = 10$, which is only accessible on a modern computer cluster, but still a very small system in contrast to real biological sequences, which are considered in section "Results".

Details of the Monte Carlo method to generate configurations are as follows: One starts with a somehow given configuration, e.g. two random sequences. For each conventional Monte Carlo step, which affects only one configuration at one temperature, first one of the sequences $\mathbf{x}$ and $\mathbf{y}$ is chosen randomly. Next, a random letter position i is selected and, finally, a letter a is chosen randomly from the 4-letter alphabet, each letter occuring with probability f_a. The ith letter in the selected sequence is then replaced by a, resulting in the trial configuration $(\mathbf{x}_t, \mathbf{y}_t)$. Then the new optimum local alignment score for $S(\mathbf{x}_t, \mathbf{y}_t)$ is computed. This trial is accepted with the Metropolis criterion Eq. (6).

Within the framework of a parallel tempering simulation, beside conventional Monte Carlo steps, i.e. the letter-exchange steps, also replica flips between the sets of temperatures $\{\infty, 1, 0.6, 0.4\}$ ($\{\infty, 1, 0.7, 0.6, 0.5\}$ for $L = 20$) are attempted according Eq. (7). One full iteration, denoted as on "time step" or one "Monte Carlo sweep", consists of L local Monte Carlo steps for each replica followed by $N_T - 1$ attempts for exchanges of replicas between neighboring temperatures.

Before starting data sampling, equilibrium has to be guaranteed for each temperature. For this purpose we simulated two kinds of runs, one starting in ordered configurations, e.g. pairs of equal sequences, and another one in disordered configurations (two completely uncorrelated random sequences). Equilibrium is reached, when both starting configurations lead to the same score within statistical fluctuations. We denote the number of Monte Carlo sweeps to reach equilibrium by $t_{\rm eq}$ (see Fig. 2). Since the score $S(t)$ fluctuates strongly, the result of $S(t)$ shown in Fig. 2 is an average over 1000 independent runs. For systems $L > 10$ the equilibration time of the smallest temperatures ($T < 0.5$ for $L = 20$) seems to diverge. For this reason the data could not be used for data production. An interpretation of this effect from the statistical physics point of view has been given in a previous work [19].

In contrast to direct generating random sequences, the Monte Carlo scheme induces correlations between

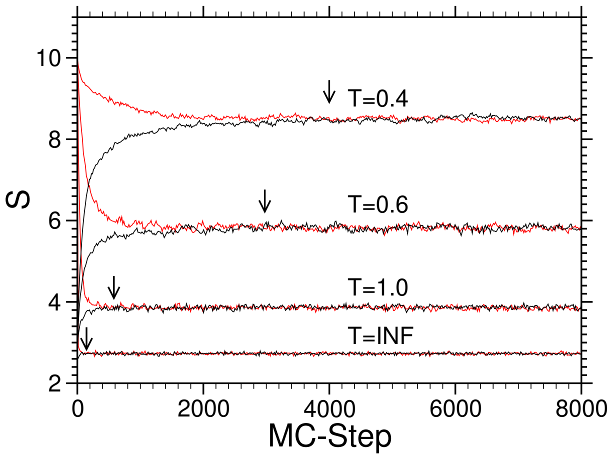

Figure 2: Equilibration of the 4-letter system ($L = M = 10$) with temperatures $T = 0.4, 0.6, 1.0, \infty$ Equilibrium is reached after 150,600,3000,4000 steps (indicated by arrows) respectively. $S(t)$ is averaged over independent 1000 runs.

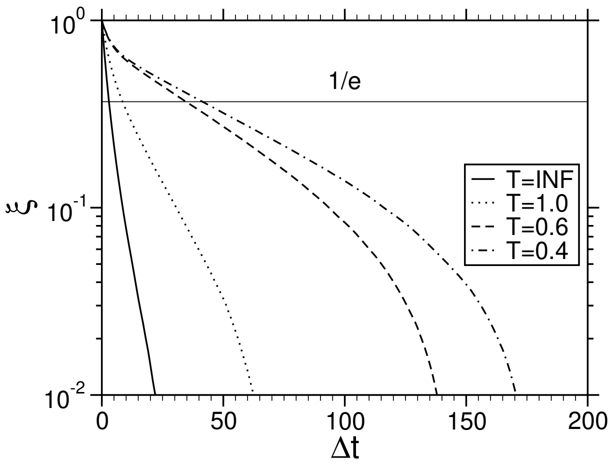

Figure 3: Score auto-correlation function for different temperatures (4 letters, $L = M = 10$).

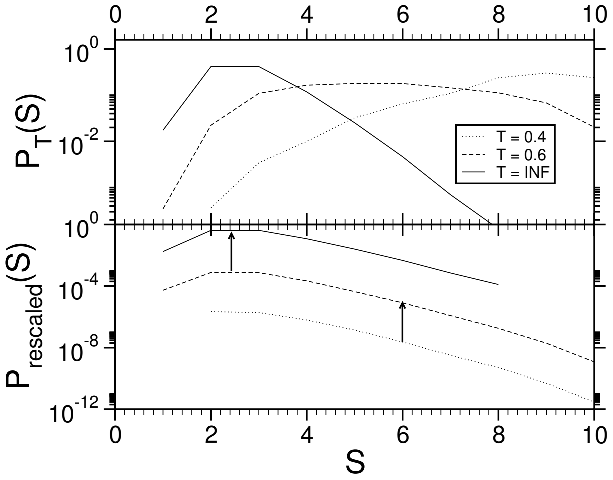

Figure 4: Illustration of the rescaling procedure. Upper: Unscaled finite temperature histograms (4 letters, $L = M = 10$) for temperatures 0.4, 0.6 and ∞. Bottom: rescaled histograms. The arrows indicate the offset between the curves (see text).

the scores at time t and $t+\Delta t$. This correlation can be quantified by the normalized auto-correlation function

$$\xi(t) = \frac{\langle S(t_0)S(t_0+t)\rangle_{t_0} - \langle S(t_0)\rangle_{t_0}^2}{\langle S(t_0)^2\rangle_{t_0} - \langle S(t_0)\rangle_{t_0}^2}\,, \tag{10}$$

$\langle\cdots\rangle_{t_0}$ denoting the average over different times and many independent runs. The typical time scale, over which the correlations vanishes is the correlation time τ defined via $\xi(\tau) = 1/e$. The normalized auto-correlation function for the system of $L = 10$ is shown in Fig. 3. The correlation time increases with decreasing temperature, which corresponds to a growth of the equilibration time with decreasing temperature in Fig. 2. However by the generation of the histograms the correlations will average out, but estimates of the errors are more complicated. Therefore we used resampling methods [51] in the data production for the relevant biological systems, see section "Results".

For each temperature T, the configurations from simulation times before the equilibration time $t_{eq}(T)$ being ignored, a normalized histogram $P_T(S) = \rho_T(-S)$ of scores/energies is sampled. The result is displayed in the upper part of Fig. 4. According to Eq. (8), to obtain $P(S)$, the first step is to multiply $P_T(S)$ with $e^{-S/T}$. The result is shown in the bottom part of Fig. 4.

In order to match the data points between the curves obtained at different temperatures, i.e. when calculating the partition functions Z_T up to a global normalization constant, we use the following procedure: First the multiplicative factors $e^{\Delta_{i-1,i}}$, or in a logarithmic picture the offsets $\Delta_{i-1,i}$, are determined such that the curves overlap, as indicated in Fig. 4. Each time, one considers the sum of the squared differences of the rescaled distributions of two neighboring temperatures $\log P_{T_{i-1}}(S)$ and $\log P_{T_i}(S)$

$$\begin{aligned} f(\Delta_{i-1,i}) = \quad {\textstyle\sum_S}'[\quad & (\log P_{T_i}(S) - S/T_i + \Delta_{i-1,i}) - \\ & \big(\log P_{T_{i-1}}(S) - S/T_{i-1} + \Delta_{i-2,i-1}\big)]^2 \end{aligned}$$

(the sum runs over all values of S, where the two histograms $P_{T_{i-1}}(S)$ and $P_{T_i}(S)$ are non-zero). Minimizing this functions for each pair $(i, i+1)$ of neighboring temperatures for $i > 1$ yields the wanted normalization factors. The temperatures are arranged in descending order with $T_1 = \infty$ and $\Delta_{0,1}$ is set to 0, since for the case of unbiased sampling, the histogram $P_{\infty(S)}$ is almost identical to $P(S)$, i.e. the normalization factor is within reasonable accuracy equal to the global normalization factor. Next, $\Delta_{1,2}$ is obtained from

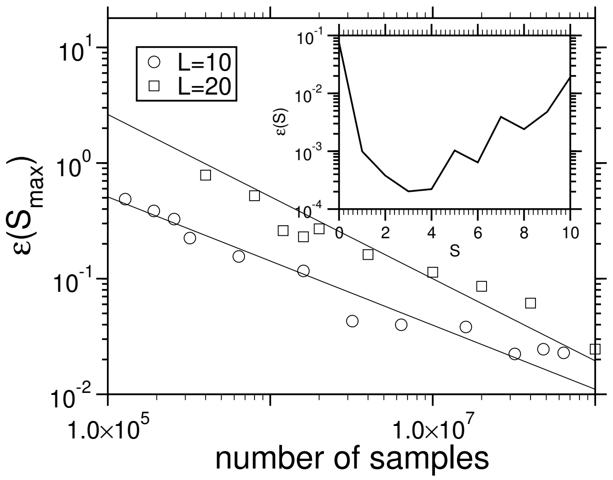

Figure 5: Convergence of the parallel tempering data. The relative error $\epsilon(S_{max})$ of the ground state for $L = 10$ and $L = 20$ depending on the number $N_{\rm samples}$ of samples is shown. The lines denote power laws $\sim N^\tau$ with $\tau = -0.55$ ($L = M = 10$) and $\tau = -0.75$ ($L = M = 20$). Inset: relative error of the distribution for the smallest system $L = 10$.

minimizing $f(\Delta_{1,2})$. Next $f(\Delta_{2,3})$ is minimized and so on. This results in rescaled and matched distributions $\tilde{P}_{T_i}(S) = P_{T_i}(S)e^{-S/T_i+\Delta_{i-1,i}}$, which are then averaged over all temperatures:

$$P(S) = \frac{1}{\sum_{i=1}^{N_T} w_i(S)} \sum_{i=1}^{N_T} w_i(S)\tilde{P}_{T_i}(S). \tag{11}$$

The weights w_i are chosen such that the high statistical values contribute most.

We used the inverse relative error of the unscaled distributions $w_i(S) = \sqrt{P_{T_i}(S)(1 - P_{T_i}(S))/N_i}$, where N_i is the number of independet samples at temperature T_i used to calculate the histogram.[3]

Accuracy of the algorithm

In this section we discuss the error bars and efficiency of the algorithm. Therefore we consider the relative difference $\epsilon(S) = \frac{|P_{\rm sample}(S)-P_{\rm exact}(S)|}{P_{\rm exact}(S)}$ (see inset of Fig. 5) between the exact distribution and the parallel tempering data. To evaluate the convergence rate of the algorithm, we evaluate the relative error $\epsilon(S_{max})$ at the final point of the distribution as a function of the number $N_{\rm samples}$ of samples, see Fig. 5. The number of samples is the total amount of alignments to be computed including equilibration. One clearly observes a strong decrease, meaning that the algorithm is able to converge to the exact solution even at the tail, which is hardest to sample. The convergence towards the exact solution, i.e. the efficiency, can be well described by a power-law $\epsilon \sim N^\tau_{\rm sample}$, with $\tau = -0.55$ for $L = 10$ and $\tau = -0.75$ for $L = 20$. These values of $\tau < -0.5$, hence seemingly becoming more negative with growing sequence length, show that the algorithm works very efficiently.

[3] In our simulations, we assured that for all temperatures N_i is almost the same, hence this drops out in Eq. (11).

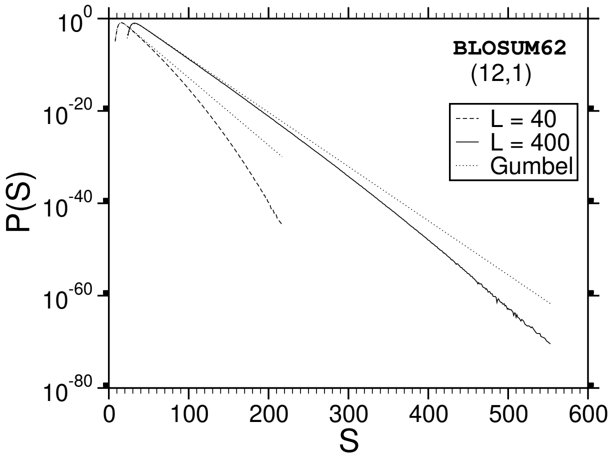

Figure 6: Density distribution of the score for gapped sequence alignment using BLOSUM62 matrices and affine gap costs with $\alpha = 12, \beta = 1$ for two sequences lengths $L = M = 40$. The results for other lengths are summarized in tables 1-4. Strong deviations from the Gumbel distribution become visible in the tail. The dotted lines show the original Gumbel distribution, when fitted to the region of high probability.

Results

Optimal alignment statistics

Next, we show the results from the application of the method to biologically relevant systems: local sequence alignment of protein sequences using BLOSUM62 [20] and PAM250 [21, 22] matrices. We apply amino acid background frequencies by Robinson and Robinson [52]. We consider different affine gap cost with $10 \leq \alpha \leq 16$, $\beta = 1$ for the BLOSUM62 matrix and $11 \leq \alpha \leq 17$, $\beta = 3$ when using the PAM250 matrix, as well as infinite gap costs. We study ten different sequence lengths between $M = L = 40$ and $M = L = 400$, in detail $L = 40, 60, 80, 100, 150, 200, 250, 300, 350, 400$.

Since the complexity of this system is much larger than the simple 4-letter system, the ground states could not be reached. Only temperatures where equilibration was guaranteed within a reasonable computation time were used for the calculation of $P(S)$. This means that we cannot resolve the score density distribution over its full support. But the range of temperatures is large enough to evaluate the distributions down to values $P(S) \sim 10^{-60}$. The temperature sets we have used in the parallel tempering technique were varied between $\{2.00, 2.25, 2.50, 3.00, 5.00, 7.00, \infty\}$ ($L = 40$) and $\{3.25, 3.50, 4.00, 5.00, 7.00, \infty\}$ ($L = 400$) for BLOSUM62 matrices and between $\{2.75, 3.00, 3.25, 4.00, 5.00, 7.00, \infty\}$ and $\{4.00, 4.25, 4.50, 5.00, 8.00, \infty\}$ for the PAM250 matrices. For each run we performed 8×10^5 Monte Carlo steps. The resulting histograms were obtained from averaging over 10 ($L = 400$) up to 100 ($L = 40$) runs.

The main result is that most of the distributions we obtain deviate strongly from the Gumbel form, as indicated in Fig. 6 and Fig. 7 by dotted lines. One observes that the discrepancy seems to be stronger for shorter sequences. Also, the case without gaps (Fig. 7) deviates, at least for $L = M = 400$, only weakly from the Gumbel distribution. This might be expected due to the previous analytical work [9, 10]. Qualitatively the behavior of the PAM250-matrices is the same and therefore the plots are not shown. A quantitative analysis of all results will be given below. Empirically we find that the resulting distribution can be described by a modified Gumbel distribution with a Gaussian correction:

$$P(S) = \lambda \exp\left(-\lambda\,(S - S_0) - \lambda_2\,(S - S_0)^2 - e^{-\lambda(S-S_0)}\right). \tag{12}$$

Note that we would have to use a different normalization constant here, but since the correction dominates the tail of the distribution, the real normalization constant is numerically indistinguishable from λ.

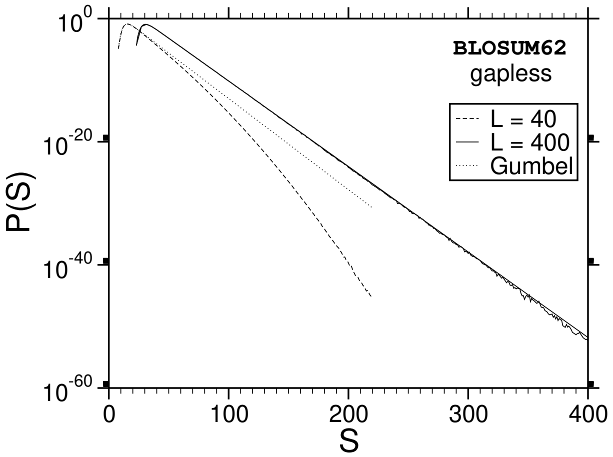

Figure 7: Density distribution of the score for ungapped sequence alignment using BLOSUM62-matrices. Deviations form the Gumbel-distribution can only be observed for short sequences ($L < 250$).

We also tried smaller gap costs than $\alpha < 10$ ($\beta = 1$,BLOSUM62) and $\alpha < 11$ ($\beta = 3$, PAM250 matrices), but in this case the distributions deviate from Gumbel not only in the tail but even in the high-probability region. The reason is presumably that the values of the parameters are close to the critical value of the linear-logarithmic phase transition [24], i.e. the alignment is not really local any more.

Next, we study the scaling behavior of the correction parameter λ_2. Since the distributions seem to approach the Gumbel distribution with increasing sequence length, as can be seen in Fig. 6 and Fig. 7, we expect that λ_2 decreases for $L \to \infty$. Furthermore, when looking at Fig. 8, where $P(S)$ is shown for one sequence length $L = M = 250$ but for different gap-opening costs α, we expect a weak dependence of λ_2 on α. In order to provide more quantitative evidence, we fitted all distributions by Eq. (12) and compared the resulting fit parameters. The results are documented in Tab.(1) - Tab.(5). In the gapless case no deviations from Gumbel could be detected for sequence lengths $L > 200$. For the other cases, the dependence of the scaling behavior λ_2 on the sequence length is plotted in Fig. 9 and Fig. 10. BLOSUM62 and PAM250 behaves qualitatively the same. λ_2 seems to decay with a power law

$$\lambda_2\,(L) = a\,L^{-b} - \lambda_2^\star \tag{13}$$

for the smallest gap costs and faster than a power law for larger gap costs. The case of infinite gap costs could be described roughly by an exponential decay

$$\lambda_2^\infty\,(L) = c\exp\left(-\frac{L}{L_0}\right). \tag{14}$$

For all other gap costs the decay behaves in between these extremes. By fitting the limiting cases to these functions the bounds of the decay could be estimated. For BLOSUM62-matrices and gap costs with $\alpha = 10, \beta = 1$, we found a scaling exponent of $a = 0.02(1)$ and $b = 0.8(8)$ ($a = 0.004(6)$ and $b = 0.5(5)$ for PAM250-matrices and $\alpha = 13, \beta = 3$ gap costs). $\lambda_2^\star$ vanished within error bars. The exponential decay of the gappless case could be quantified by $L_0 = 63 \pm 4$ and $c = 1.(34) \times 10^{-3}$ ($L_0 = 93 \pm 7$ and $c = 0.6(73) \times 10^{-3}$ for PAM250-matrices).

In order to see the relevance of our result we consider a simple example, the E-value of a pair of sequences of length $L = 100$ using $\alpha = 12, \beta = 1$ gap costs, the BLOSUM62-matrix and the SWISSPROT database [53], which contains currently $N_{\rm swissprot} = 210,623$ sequences. In BLAST [54], the E-value, i.e. the expected number of hits exhibting at certain “cut-off” score $S_{\rm cut}$, is currently estimated via the cumulative Gumbel

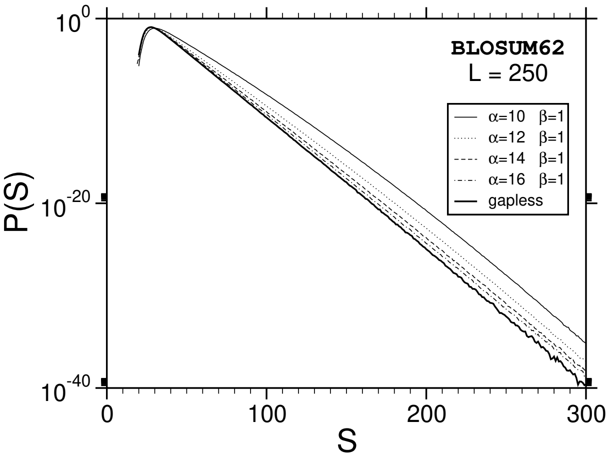

Figure 8: Density distribution of the score comparing different gap costs. Deviations from the Gumbel distribution become stronger for small gap costs.

	λ	$10^4\lambda_2$	S_0
40	0.325(1)	8.8(08)	15.5(26)
60	0.303(2)	6.21(5)	18.4(37)
80	0.289(4)	4.85(6)	20.6(61)
100	0.276(0)	4.2(23)	22.4(11)
150	0.256(6)	3.1(99)	25.7(55)
200	0.246(4)	2.5(51)	28.2(38)
250	0.235(6)	2.21(8)	30.1(84)
300	0.229(8)	1.93(2)	31.8(98)
350	0.225(7)	1.64(9)	33.2(56)
400	0.221(7)	1.50(0)	34.5(27)

Table 1: Fit parameters of the modified Gumbel distribution Eq. (12) using the BLOSUM62 scoring matrix and affine gap costs with $\alpha = 10, \beta = 1$.

distribution

$$E = kLNe^{-\lambda S}, \tag{15}$$

with $S_0 = \frac{\ln(k\,L\,N)}{\lambda}$ in our notation and N the number of amino acids of the entire database. The basic feature of this notation is that the search space is independent of the parameters $k = 0.0410$ and $\lambda = 0.267$. Using the suggested E-value of 10 [54], we find a cut-off of $S_{\text{cut}} = 64.8$ above which a result is considered to be significant, with $P(S > S_{\text{cut}}) = 4.75 \times 10^{-5}$. Our cumulative distribution achieves this probability at $S_{\text{cut}} = 54$, i.e. significantly below the BLAST value. Hence, using the true distributions of the scores, a considerable amount of queries, those which have a score between 54 and 64, are significant in contrast to the result of the significance estimation within the Gumbel approximation. Hence, using the data provided in this work, one is able to estimate the significance of protein-data-base queries for the most commonly used parameter sets with much higher precison than when applying the approximation of the Gumbel distribution.

	λ	$10^4\lambda_2$	S_0
40	0.336(3)	7.9(02)	15.4(06)
60	0.314(2)	5.6(58)	18.0(76)
80	0.302(3)	4.33(8)	20.0(56)
100	0.293(2)	3.55(2)	21.6(48)
150	0.287(0)	2.1(53)	24.6(09)
200	0.278(7)	1.7(07)	26.7(71)
250	0.277(4)	1.17(6)	28.3(54)
300	0.274(1)	0.93(0)	29.7(83)
350	0.272(9)	0.81(1)	30.8(59)
400	0.270(2)	0.70(2)	31.8(47)

Table 2: Fit parameters of the modified Gumbel distribution Eq. (12) using the `BLOSUM62` scoring matrix and affine gap costs with $\alpha = 12, \beta = 1$.

	λ	$10^4\lambda_2$	S_0
40	0.337(6)	7.8(53)	15.3(16)
60	0.318(2)	5.4(46)	17.9(08)
80	0.307(8)	4.1(26)	19.7(91)
100	0.307(2)	2.9(78)	21.4(72)
150	0.304(6)	1.5(81)	24.3(39)
200	0.298(5)	1.20(2)	26.2(76)
250	0.299(7)	0.66(12)	27.8(03)
300	0.296(7)	0.54(05)	29.0(39)
350	0.297(2)	0.46(46)	30.1(46)
400	0.294(7)	0.43(30)	31.0(50)

Table 3: Fit parameters of the modified Gumbel distribution Eq. (12) using the `BLOSUM62` scoring matrix and affine gap costs $\alpha = 14, \beta = 1$.

	λ	$10^4\lambda_2$	S_0
40	0.338(1)	7.7(82)	15.2(63)
60	0.322(1)	5.2(46)	17.9(46)
80	0.310(8)	3.9(98)	19.7(37)
100	0.314(3)	2.6(18)	21.4(56)
150	0.312(1)	1.3(71)	24.1(65)
200	0.310(2)	0.8(94)	26.1(00)
250	0.312(3)	0.34(9)	27.63(1)
300	0.311(0)	0.37(6)	28.84(9)
350	0.308(4)	0.29(2)	29.8(75)
400	0.307(6)	0.21(7)	30.79(5)

Table 4: Fit parameters of the modified Gumbel distribution Eq. (12) using the `BLOSUM62` scoring matrix and affine gap costs with $\alpha = 16, \beta = 1$.

	λ	$10^4\lambda_2$	S_0
40	0.340(6)	7.6(48)	15.3(28)
60	0.322(5)	5.1(81)	17.8(45)
80	0.313(7)	3.8(70)	19.7(73)
100	0.319(2)	2.4(98)	21.4(13)
150	0.322(7)	1.0(31)	24.1(12)
200	0.320(9)	0.63(2)	26.0(41)
250	0.322(6)		27.50(9)
300	0.322(9)		28.70(3)
350	0.321(0)		29.72(5)
400	0.319(7)		30.56(8)

Table 5: Fit parameters of the modified Gumbel distribution Eq. (12) using the `BLOSUM62` scoring matrix and gapless alignment.

	λ	$10^4\lambda_2$	S_0
40	0.262(2)	5.6(57)	19.4(4)
60	0.242(6)	4.1(72)	23.6(6)
80	0.228(5)	3.3(96)	26.9(1)
100	0.209(9)	3.4(44)	29.4(2)
150	0.194(2)	2.4(98)	34.5(9)
200	0.185(9)	1.90(0)	38.5(1)
250	0.172(2)	1.82(8)	41.3(8)
300	0.168(1)	1.53(8)	44.0(9)
350	0.162(2)	1.37(3)	46.3(7)
400	0.159(2)	1.26(5)	48.2(7)

Table 6: Fit parameters of the modified Gumbel distribution Eq. (12) using the `PAM250` scoring matrix and affine gap costs with $\alpha = 11, \beta = 3$.

	λ	$10^4\lambda_2$	S_0
40	0.269(86)	5.1(90)	19.3(1)
60	0.247(01)	4.0(51)	23.1(5)
80	0.235(39)	3.2(10)	26.1(3)
100	0.222(83)	2.92(1)	28.4(2)
150	0.208(35)	2.11(6)	32.8(3)
200	0.1995(4)	1.69(7)	36.0(6)
250	0.1924(7)	1.43(6)	38.5(7)
300	0.1888(3)	1.18(9)	40.7(0)
350	0.1837(2)	1.08(3)	42.3(6)
400	0.1823(9)	0.93(4)	43.9(0)

Table 7: Fit parameters of the modified Gumbel distribution Eq. (12) using the `PAM250` scoring matrix and affine gap costs with $\alpha = 13, \beta = 3$.

	λ	$10^4\lambda_2$	S_0
40	0.273(39)	4.8(92)	19.2(2)
60	0.254(59)	3.6(09)	23.0(4)
80	0.240(57)	2.97(6)	25.7(7)
100	0.231(09)	2.57(2)	27.9(3)
150	0.217(92)	1.84(2)	32.0(2)
200	0.209(97)	1.44(2)	34.8(7)
250	0.206(10)	1.12(9)	37.1(9)
300	0.203(04)	0.95(6)	39.1(0)
350	0.199(64)	0.83(9)	40.5(4)
400	0.199(66)	0.68(2)	41.9(2)

Table 8: Fit parameters of the modified Gumbel distribution Eq. (12) using the PAM250 scoring matrix and affine gap costs with $\alpha = 15, \beta = 3$.

	λ	$10^4\lambda_2$	S_0
40	0.274(42)	4.8(66)	19.2(0)
60	0.256(72)	3.50(3)	22.9(5)
80	0.243(64)	2.87(2)	25.6(0)
100	0.236(60)	2.35(7)	27.7(0)
150	0.224(23)	1.65(1)	31.5(7)
200	0.217(88)	1.24(3)	34.4(0)
250	0.215(36)	0.95(5)	36.5(6)
300	0.212(41)	0.78(1)	38.2(0)
350	0.212(35)	0.61(6)	39.9(2)
400	0.210(93)	0.52(5)	41.0(9)

Table 9: Fit parameters of the modified Gumbel distribution Eq. (12) using the PAM250 scoring matrix and affine gap costs with $\alpha = 17, \beta = 3$.

	λ	$10^4\lambda_2$	S_0
40	0.276(7)	4.6(9)	19.1(9)
60	0.259(1)	3.34(6)	22.8(4)
80	0.249(4)	2.61(5)	25.5(2)
100	0.242(2)	2.14(1)	27.5(2)
150	0.231(4)	1.46(1)	31.1(3)
200	0.230(5)	0.99(7)	33.8(6)
250	0.232(7)	0.53(8)	36.2(0)
300	0.232(5)	0.38(7)	37.8(1)
350	0.232(6)	0.22(7)	39.3(3)
400	0.232(3)	0.13(5)	40.6(2)

Table 10: Fit parameters of the modified Gumbel distribution Eq. (12) using the PAM250 scoring matrix and gapless alignment.

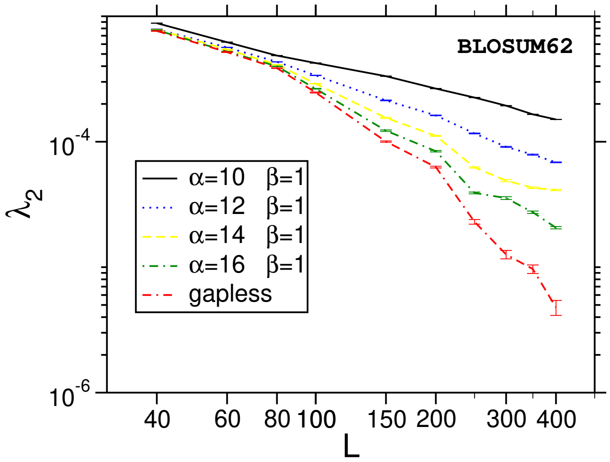

Figure 9: Scaling of the correction parameter λ_2 (BLOSUM62). The decay of λ_2 with system size varies between a power law near the logarithm-linear transition and a exponential behavior for the gapless case.

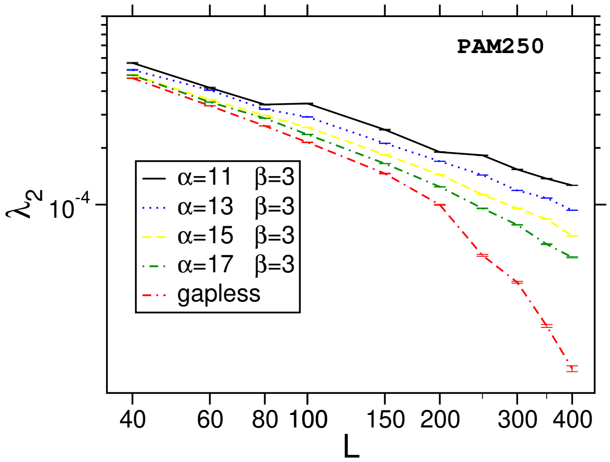

Figure 10: Scaling of the correction parameter λ_2 (PAM250). The decay of λ_2 with system size varies between a power law near the logarithm-linear transition and a exponential behavior for the gapless case.

L	$k=2$	$k=3$	$k=4$	$k=5$
40	$2.75, 3, 3.5, 4, 7, \infty$			
60	$2.75, 3, 3.5, 4, 7, \infty$			
80	$2.75, 3, 3.5, 4, 7, \infty$	$3.75, 4, 4.5, 5, 8, \infty$	$5.25, 5.5, 6, 8, \infty$	$6, 6.25, 6.5, 7, 8, 12, \infty$
100	$2.75, 3, 3.5, 4, 7, \infty$	$3.75, 4, 4.5, 5, 8, \infty$	$5.25, 5.5, 6, 8, \infty$	$6, 6.25, 6.5, 7, 8, 12, \infty$
150	$2.75, 3, 3.5, 4, 7, \infty$	$3.75, 4, 4.5, 5, 8, \infty$	$5.25, 5.5, 6, 8, \infty$	$6, 6.25, 6.5, 7, 8, 12, \infty$
200	$3.25.3.5, 4, 7, \infty$	$3.75, 4, 4.25, 4.5, 5, 8, \infty$	$4.75, 5, 5.25, 5.5, 6, 8, \infty$	$5.75, 6, 6.25, 6.5, 7, 8, 12, \infty$
300	$3.25.3.5, 4, 7, \infty$	$3.75, 4, 4.25, 4.5, 5, 8, \infty$	$4.75, 5, 5.25, 5.5, 6, 8, \infty$	$5.75, 6, 6.25, 6.5, 7, 8, 12, \infty$
400	$3.25.3.5, 3.75,$ $4, 4.25, 5, 8, \infty$	$3.75, 4, 4.25, 4.5, 5, 8, \infty$	$5.25, 5, 5.75, 6, 8, 10, \infty$	$6, 6.25, 6.5, 7, 9, 11, \infty$

Table 11: Temperature parameters for sum-statistics.

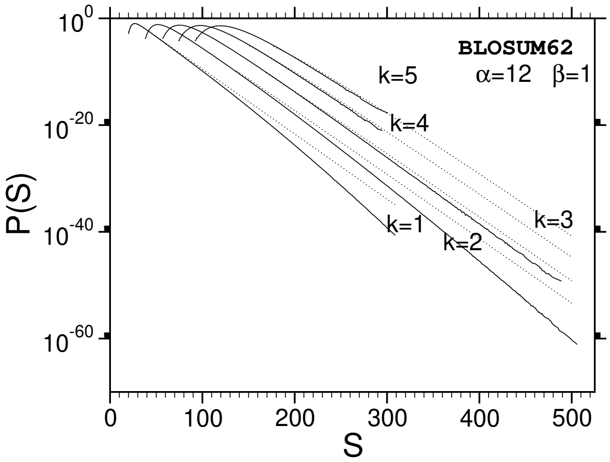

Figure 11: Score distributions for sum-statistics of the k-best scores (solid lines) for $L = M = 200$. The dotted lines denote the distribution without Gaussian correction ($\lambda_2 = 0$). Deviations from Eq. (2) or Eq. (4) become only visible in the rare-event tail.

Sum statistics of the k-best alignments

The densitiy distribution of the ungapped sum statistics asymptotic distribution is well known by Eq. (3). We use the `SIM` procedure [27] to compute the sum of the two and three best alignments for the same Markov-chain Monte Carlo simulation as in the previous sections. In this case, we consider only the `BLOSUM62` matrix together with affine gap costs $\alpha = 12, \beta = 1$, a commonly used scoring system. We observed large fluctuations for short sequences ($L < 100$) and equilibration turned out to be harder for this case. Thus only sequences with $L \geq 100$ had been used for the analysis. The temperature sets varied between $\{2.75, 3.0, 3.5, 4.0, 7.0, \infty\}$ for $L = 100, k = 2$ and $\{6.25, 6.5, 7, 9, 11, \infty\}$ for $L = 400, k = 5$ (details are shown in Tab. 11).

Note that the systems could not be equilibrated for $k > 3$ in the very low temperature regime $T < 5$. Therefore, for theses cases, the tail could only be obtained in an intermediate range of probabilities ($\sim 10^{-20}$), which is nevertheless low enough to obtain precise significance figures.

In Fig. 11 we compare different distributions obtained for varying k and fixed sequence length $L = 200$. In order to quantitatively compare the distribution with theoretical predictions from Karlin-Altschul statistics [28], we used the estimated Gumbel parameters λ and S_0 from the optimal score distributions. Corresponding to substituting the normalized score in Eq. (4) with $t = \lambda(S - kS_0)$ we fitted the the tail ($p < 10^{-10}$) of the

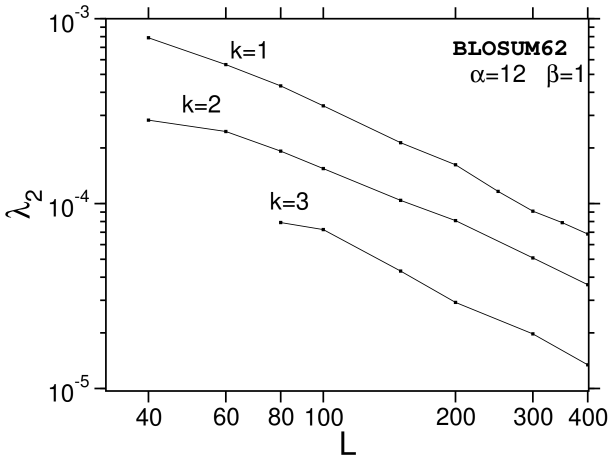

Figure 12: Scaling of the correction parameter for BLOSUM62 sum-statistics ($k = 1, 2, 3$). λ_2 is estimated by a fit for Eq. (16) using optimal the Gumbel-parameters λ and S_0 from optimal score statistics ($k = 1$).

	$10^4\lambda_2^{(k=2)}$	$10^4\lambda_2^{(k=3)}$
60	2.4(54)	
80	1.92(2)	0.7(90)
100	1.54(7)	0.7(24)
150	1.04(1)	0.43(2)
200	0.80(9)	0.29(2)
300	0.50(8)	0.19(7)
400	0.36(4)	0.13(4)

Table 12: Correction parameter λ_2 for the sum statistics $k = 2$ and $k = 3$. λ_2 is estimated by a fit for Eq. (16) using optimal the Gumbel-parameters λ and S_0 from optimal score statistics ($k = 1$). BLOSUM62 with affine gap costs ($\alpha = 12, \beta = 1$) was used as scoring system. .

Monte Carlo data to the modified density of the sum statistics:

$$P(S) = Cf_{tail}(\lambda(S - kS_0)) \exp\left[\lambda_2^{(k)}(S - kS_0)^2\right] . \tag{16}$$

Again quadratic deviations could be observed which decrease with growing system length for all values of k, see Fig. 12 and Tab.(12).

As in the case of the optimal score ($k = 1$), deviations from the theoretical form are significant only in the regime of small probabilities, which is not accessible with naive sampling methods. The data for $k = 1$ to $k = 3$ (Fig. 12) give evidence that the edge effect is reduced by increasing k. Note that in Ref. [16], best agreement with theory was achieved with $k = 6$.

Discussion and summary

We have studied the distribution of optimum alignment scores over a wide range using a rare event sampling method which is based on notions from statistical physics. First, by comparing the results for a small 4-letter test system, we illustrated how the method works and provided some evidence for its convergence. In the main part, we considered protein alignment for two types of substitution matrices, i.e. BLOSUM and PAM matrices. We also studied many different sets of biologically relevant parameters by varying gap costs and sequence lengths.

For large enough gap costs it was previously assumed that the distribution follows the Gumbel extreme-value distribution, even when aligning finite sequences and allowing for gaps. Hence, the Gumbel distribution is used for calculating p-values in protein data bases so far . We observe clear deviations from the Gumbel distribution in the biologically relevant rare-event-tail, which is out of reach of simple sampling methods used so far.

An analysis of the scaling behavior of the correction parameter λ_2 gives evidence that the Gumbel distribution correctly describes the data only in the limit of infinite sequence lengths, even for gapped sequence alignments. For finite protein lengths, which have the highest biological relevance, we observed that the distributions can be fitted well by a Gumbel distribution with a Gaussian correction. Nevertheless, in data bases like BLAST [8,18,54], we recommend to use distribution functions determined by the empirical fitting parameters provided in this work because the critical value S_{cut}, above which a result is considered to be significant, changes considerably, as we have seen.

We have also studied the sum-statistics of the k-best alignments. Again a Gaussian correction to the assumed form of the distribution was found empirically. Extrapolation to infinitly long sequences gives good evidence that the ungapped statistical theory describes the gapped case for $L = M \to \infty$ as well.

Acknowledgments

We thank B. Morgenstern and P. Müller for critically reading the manuscript. The authors have received financial support from the *VolkswagenStiftung* (Germany) within the program "Nachwuchsgruppen an Universitäten", and from the European Community via the DYGLAGEMEM program.

References

1. Brown S: *Bioinformatics*. Natick (MA): Eaton Pblishing 2000.
2. Rashidi S, Buehler L: *Bioinformatics Basics*. Boca Raton (FL): CRC Press 2000.
3. **The Protein Data Bank**[http://www.pdb.org].
4. Fraser C, Gocayne J: **The minimal gene complement of Mycoplasma genitalium.** *Science* 1995, **270**:397.
5. Needleman SB, Wunsch CD: **A general method applicabel to search for similarities in the amino acid sequence of two proteins**. *J.mol.Biol.* 1970, **48**:443–453.
6. Smith TF, Waterman MS: **Identification of Common Molecular Subsequences**. *J.mol.Biol.* 1981, **147**:195–197.
7. Gotoh O: **An improved algorithm for matching biological sequences**. *J. Mol. Biol.* 1982, **162**:705.
8. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool**. *J. Mol. Biol.* 1990, **215**:403.
9. Karlin S, Altschul S: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes**. *Proc.Natl.Acad.Sci.U.S.A.* 1990, **87**:2264.
10. Dembo A, Karlin S, Zeitouni O: **Limit distribution of maximal non-aligned two-sequence seqmental score**. *Ann. Prob.* 1994, **22**:2022.
11. Yu Y, Hwa T: **Statistical Significance of Probabilistic Sequence Alignment and Related Local Hidden Markov Models**. *J. Comp. Biol.* 2001, **8**:249–282.
12. Yu Y, Bundschuh R, Hwa T: **Statistical Significance and Extreme Ensemble of Gapped Local Hybrid Alignment**. In *Biological Evolution and Statistical Physics.* Edited by Lässig M, Valeriani A, Berlin: Springer-Verlag 2002:3–22.
13. Kschischo M, Lässig M, Yu Y: **Toward an Accurate Statistics of Gapped Alignments**. *Bull. Math. Biol.* 2004, **67**:169–191.

14. Siegmund D, Yakir B: **Approximate p-Values for Local Sequence Alignments**. *Annals of Statistics* 2000, **28**:657–680.

15. Metzler D, Grossmann S, Wakolbinger A: **A Poisson Model for Gapped Local Alignments**. *Stat. Prob. Letters.* 2002, **60**:91–100.

16. Altschul S, Gish W: **Local Alignment Statistics**. *Meth. Enzym.* 1996, **266**:460.

17. Lengauer T, Schneider R, Bork P, Brutlag D, Glasgow J, Mewes HW, Zimmer R (Eds): *Rapid assessment of extremal statistics for gapped local alignment*, *Volume 270*, Menlo Park, CA: AAAI Press 1999.

18. Altschul S, R Bundschuh TH R Olsen: **The estimation of statistical parameters for local alignment score distributions**. *Nucl.Acid.Res.* 2001, **29**:351.

19. Hartmann A: **Sampling rare events: statistics of local sequence alignments**. *Phys. Rev. E* 2002, **65**:056102.

20. Heinkoff S, Heinkoff J: **Amino acid substitution matrices from protein blocks**. *Proc.Natl.Acad.Sci.U.S.A.* 1992, **89**:10915–10919.

21. Dayhoff M, Schwartz R, Orcutt B: **A model of evolutionary change in proteins**. In *Atlas of Protein Sequence and Structure*, *Volume 5,Suppl.3*. Edited by Dayhoff M, Washington,D.C.: National Biomedical Research Foundation 1978:345–352.

22. Schwartz R, Dayhoff M: **Matrices for detecting distant relationships**. In *Atlas of Protein Sequence and Structure*, *Volume 5,Suppl.3*. Edited by Dayhoff M, Washington,D.C.: National Biomedical Research Foundation 1978:353–358.

23. Gumbel E: *Statistics of Extremes*. New York: Columbia University Press 1958.

24. Arratia R, Waterman M: **A Phase Transition for the Score in Matching Random Sequences Allowing Deletions**. *Ann.Appl. Prob.* 1994, **4**:200.

25. Istrail S, Pevzner P, Waterman M (Eds): *Optimal Detection of Sequence Similarity by Local Alignment* 1998.

26. Sellers P: **Pattern recognition in genetic sequences by mismatch density**. *Bull.Math.Biol.* 1984, **46**:501–514.

27. Altschul S, Erickson B: **Optimal Sequence Alignment using Affine Gap Costs**. *Bull.Math.Biol.* 1986, **48**:633.

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

29. Mantegna RN, Stanley HE: *Introduction to Econophysics. Correlations and Complexity in Finance*. Cambridge: Cambridge University Press 1999.

30. Hartmann AK, Weigt M: *Phase Transitions in Combinatorial Optimization Problems*. Weinheim: Wiley-VCH 2005.

31. Müller B, Reinhardt J: *Neural Networks*. Berlin: Springer-Verlag 1990.

32. Lässig M, Valleriani A (Eds): *Biological Evolution and Statistical Physics*, Berlin: Springer-Verlag 2002.

33. Sardiu M, Alves G, Yu Y: **Score statistics of global sequence alignment from the energy distribution of a modified directed polymer and directed percolation problem**. *Phys.Rev.E.* 2005, **72**:061917.

34. Zhang M, Marr T: **Alignment of molecular sequences seen as random path analysis**. *J. Theor.Biol.* 1995, **174**:119–129.

35. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E: **Equation of state calculations by fast computing machines**. *J.Chem.Phys. 21* 1953, **21**:1087.

36. Earl D, Deem M: **Parallel Tempering: Theory, Applications, and New Perspectives**. *Phys. Chem. Chem. Phys.* 2005, **7**:3910–3916.

37. Zhou R: **Exploring the protein folding free energy landscape: Coupling replica exchange method with P3ME/RESPA algorithm**. *J. Molec.Graph.Mod.* 2004, **22**:451–463.

38. Zhou R, Berne B: **Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water?** *Proc.Nat.Acad.Sci. USA* 2002, **99**:12777–12782.

39. Zhou R, Berne B: **Trp-cage: Folding free energy landscape in explicit water**. *Proc.Nat.Acad.Sci. USA* 2002, **100**:13280–13285.

40. Garci'a A, Onuchic J: **Folding a protein in a computer: An atomic description of the folding/unfolding of protein**. *Proc.Nat.Acad.Sci. USA* 2003, **100**:13898–13903.

41. Zhou R, Berne B, Germain R: **The free energy landscape for β hairpin folding in explicit water**. *Proc.Nat.Acad.Sci. USA* 2001, **98**:14931–14936.

42. Auer S, Frenkel D: **Prediction of absolute crystal-nucleation rate in hard-sphere colloids**. *Nature* 2001, **409**:1020–1023.

43. Marinari E, Parisi G, Ruiz-Lorenzo J: **Numerical Simulations of Spin Glass Systems**. In *Spin Glasses and Random Fields, Directions in Condensed Matter Physics Vol. 12*. Edited by Young A, World Scientific 1998:109.

44. Katzgraber H, Palassini M, Young A: **Monte Carlo simulations of spin glasses at low temperatures**. *Phys.Rev.B* 2001, **63**:1844221–18442210.

45. Körner M, Katzgraber H, Hartmann A: **Probing tails of energy distributions using importance-sampling in the disorder with a guiding function**. *Stat.Mech.* 2006, :P04005.

46. Marinari E, Parisi G: **Simulated Tempering: a New Monte Carlo Scheme**. *Europhys.Lett.* 1992, **19**:451.

47. Hukushima K, Nemoto K: **Exchange Monte Carlo Method and Application to Spin Glass Simulations**. *J. Phys.Soc.Jpn.* 1996, **65**:1604.

48. Berg B, Neuhaus T: **Multicanonical ensemble: A new approach to simulate first-order phase transitions**. *Phys.Rev.Lett.* 1992, **68**:9.

49. Wang F, Landau D: **Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram**. *Phys. Rev. E* 2001, **64**:056101.

50. Wang F, Landau D: **Efficient, multiple-range random walk algorithm to calculate the density of states**. *Phys. Rev. Lett.* 2001, **86**:2050.

51. Efron B: **Bootstrap methods: Another look at the jackknife**. *Annals of Statistics* 1979, **7**:1.

52. Robinson A, Robinson L: **Distribution of glutamine and asparagine residues and their near neighbours in peptides and proteins**. *Proc. Natl. Acad. Sci. U.S.A.* 1991, **88**:8880–8884.

53. **SWISSPROT**[http://www.expasy.org/].

54. **NCBI BLAST**[http://www.ncbi.nlm.nih.gov/BLAST].