Algorithms for Molecular Biology BioMed Central

Research

Open Access

Local sequence alignments statistics: deviations from Gumbel statistics in the rare-event tail

Stefan Wolfsheimer*[1,2], Bernd Burghardt[1] and Alexander K Hartmann[1,2]

Address: [1]Institut für Theoretische Physik, Universität Göttingen, 37077, Göttingen, Friedrich-Hund-Platz 1, Germany and [2]Institut für Physik, Universität Oldenburg, 26111, Oldenburg, Germany

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

* Corresponding author

Published: 11 July 2007

Algorithms for Molecular Biology 2007, **2**:9 doi:10.1186/1748-7188-2-9

Received: 5 October 2006
Accepted: 11 July 2007

This article is available from: http://www.almob.org/content/2/1/9

Abstract

Background: The optimal score for ungapped local alignments of infinitely 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 generating the sequences with a parametrized probability distribution, which is biased with respect to the original biological one, in the framework of Metropolis Coupled Markov Chain Monte Carlo. 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.

Conclusion: 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 l depending score function $g(l)$. Widely used are affine gap costs because for two given sequences of length L and M, because fast algorithms with running time $O(LM)$ are available for this case [7]. Note that for database queries even this is too complex, hence fast heuristics like BLAST [8] are used there.

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 *p-value* for a given score S. This means under a random sequence model one wants to know the probability for the occurrence of at least one hit with a score S greater than or equal to some given threshold value b, i.e. $(S \geq b)$. Often E-values are used instead. They describe the number of expected hits with a score greater than or equal to some threshold value. One possible access to the statistical significance can be achieved under the null model of random sequences. Then the optimal alignment score S becomes a random variable and the probability of occurrence of S under this model $P(s) = (S = s)$ provides estimates for p-values. 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 Poisson 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 E-values. 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, in the present case down to $P(s) = 10^{-40}$. 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 of S 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 state the rare-event approach used here and 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 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 ($k = 1, 2, \ldots, K - 1$, $1 \le i_k < i_{k+1} \le L$ and $1 \le j_k < j_{k+1} \le M$) 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 $\mathbf{x}$ or sequence $\mathbf{y}$ respectively, in detail

$$l_g^x(k) = i_{k+1} - i_k - 1$$
$$l_g^y(k) = j_{k+1} - j_k - 1.$$

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:

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

where $\sigma(a, b)$ $a, b \in \mathcal{A}$ 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

$$S(\mathbf{x}, \mathbf{y}) = \max_{\mathcal{A}} \tilde{S}(\mathbf{x}, \mathbf{y}, \mathcal{A}), \tag{2}$$

which can be obtained in $O(LM)$ time [7].

In the case of gapless optimum local alignments of two random sequences of L and M independent letters from Σ with frequencies $\{f_a\}$ with $a \in \Sigma$ and $\sum_a f_a = 1$, referred as *null model*, 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]

$$(S \ge b) = 1 - \exp[-KLM\, e^{-\lambda b}] \tag{3}$$

or

$$P_{\text{Gumble}}(s) = (S = s) = \lambda\, KLM \exp[-\lambda s - KLM\, e^{-\lambda s}] \tag{4}$$

The parameters λ and K of Eq. (3) can be derived directly from the score matrix $\sigma(a, b)$ 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. (3) 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 $O(LM)$ 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 f for the sum of the k-best nor-

malized scores $T_k = \lambda \sum_i^k (S_i - \frac{\ln KLM}{\lambda})$ (λ and K are the corresponding Gumbel-parameters for the optimal alignment)is given by the integral

$$f(t) = \frac{e^{-t}}{k!(k-2)!} \int_0^\infty y^{k-2} \exp(-e^{(y-t)/k}) dy. \qquad (5)$$

In the tail, i.e. for large t, $f(t)$ is well approximated by

$$f_{\text{tail}}(t) = \frac{e^{-t}}{k!(k-1)!} [t^{k-1} - (k-1)t^{k-2})]. \qquad (6)$$

In the asymptotic theory the score can be seen as a continuous variable and the probabilities Eq. (4) and Eq. (5) become probability densities. Then the probability of finding a normalized score b or larger is given by the integral $\mathbb{P}(S \geq b) = \int_b^\infty f(t)dt$. However in computer simulations the score is a discrete variable and therefore the normalization constants in Eq. (5) differ from continious scoring. Below we will compare the results of our numerical studies to this distribution in the tail of the data for values k = 2, ..., 5.

Sampling of rare-events

Metropolis Hastings Algorithm

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 randomly and independently drawn sequences of finite lengths. The basic idea of our approach is to generate the sequences from different distributions, which are biased towards higher scores.

In order to be more precise let us denote the state space of all possible pairs of sequences $(\mathbf{x}, \mathbf{y})$ as $\mathcal{X}$ and an element in this space as a *configuration*. We write $\mathbf{X} = (\mathbf{x}, \mathbf{y})$.

The probability mass function (pmf) of finding $\mathbf{X}$ under the null model is given by $p(\mathbf{X}) = p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^L f_{x_i} \prod_{j=1}^M f_{y_j}$ and the alignment score as defined in Eq. (2) is a random variable. A direct way to obtain the probability of the occurrence of a certain score s, is to generate n uncorrelated representatives $\mathbf{X}_i \in \mathcal{X}$ according to the null model and then compute the expectation values of the family of indicator functions h_s: $\mathcal{X} \to \mathbb{R}$ with $h_s(\mathbf{X}) = 1$, if $S(\mathbf{X}) = s$ and $h_s(\mathbf{X}) = 0$ otherwise, in other words

$$\mathbb{P}[S(\mathbf{X}) = s] = \mathbb{E}[h_s(\mathbf{X})] = \sum_{\mathbf{X}} h_s(\mathbf{X}) p(\mathbf{X}) \approx \frac{1}{n} \sum_{i=1}^n h_s(\mathbf{X}_i).$$

Since the region of biological interest is located in the rare-event tail a huge amount of samples would be needed to achieve an acceptable accuracy. In practice the rare-event tail becomes inaccessible.

Our method is based on importance sampling of a mixture of chains based on the Metropolis-Hastings algorithm. Before describing the coupling of multiple chains, we introduce the general idea of importance sampling first: The approach is based on sampling from a different distribution, such that the region of interest is sampled with high probability. Since this happens in a controlled manner the true distribution can be obtained afterward, as frequently used in variance reduction techniques. The modified distribution yields a different random variable with a different pmf q. We may write

$$P(s) = \mathbb{P}[S(\mathbf{X}) = s] = \sum_{\mathbf{X}'} h_s(\mathbf{X}') \frac{p(\mathbf{X}')}{q(\mathbf{X}')} q(\mathbf{X}') \approx \frac{1}{n} \sum_{i=1}^n h_s(\mathbf{X}_i') \frac{p(\mathbf{X}_i')}{q(\mathbf{X}_i')}.$$

At least approximately, the distribution of local alignment follows a Gumbel distribution, which exhibits an exponential behavior in the tail. Therefore an obvious choice for the biased distribution is

$$q_T(\mathbf{X}) \equiv \frac{\tilde{q}_T(X)}{Z_T} \equiv \frac{1}{Z_T} p(\mathbf{X}) \cdot \exp[S(\mathbf{X})/T], \qquad (7)$$

where $\tilde{q}_T$ the unnormalized weight of a configuration, Z_T is a (usually unknown) normalization constant and T an adjustable parameter, which we will call "temperature" (In the framework of statistical mechanics, which is closely related to our method, the parameter T describes the temperature of a physical system. The pair of sequences can be seen as a configuration of a physical system and the negative score as the energy function. Then exp [S (**X**)/T] refers to the so called *Gibbs-Boltzmann distribution*.) The close-to Gumbel form of the distribution is also directly related to the so called "large deviation rate function", which basically describes the decay rate of the tail of the distribution. Note that, if the score distribution is an exact Gumbel distribution Eq. (3), i.e. the rate function a known constant λ, then setting $T = 1/\lambda$ in Eq. (7) yields a "flat score histogram" for sufficient large s. Hence, in this case, a simulation at a single carfully chosen value T would be sufficient to obtain the full result. Since $P(s)$ does not follow the Gumbel form exactly, importance sampling has to be applied. Each value of T selects one

region of the distribution around which a high accurracy is obtained.

This importance sampling approach is conceptual related to the method of "measure change" in large deviation theory. For example Siegmund and Yakir [14] approximated the p-value for local sequence alignment by considering the log-likelihood ratio between an alternative measure and the measure of the null model. Under the new measure a rare event occurs more likely than under the original null measure and approximations become possible. Another example can be found in Ref. [29], where techniques from large deviation theory were applied to proof "asymptotic efficiency" of rare-event simulations.

However, since there is no direct method to sample directly according to the modified distribution Eq. (7) we implemented the *Metropolis-Hastings algorithm* [30], which is explained now in detail. It is based on ergodic *Markov chain Monte Carlo (MCMC)* in state space. Ergodic here means, that for a given state in the configuration space $\mathcal{X}$ any other can be achieved by stepwise "local" modifications of configurations in finite time. Note that we work in discrete time steps here. Let $\mathbf{X} \in \mathcal{X}$ a configuration at time t (e.g. at the start of the simulation). To determine the configuration at time $t + 1$, first a *trial configuration* $\mathbf{X}^*$ is selected randomly among its "neighbors". The neighborhood of a configuration depends on the choice of trial steps, which are specified below. For practical reasons we require, that the score within a neighborhood of a given configuration will not change too much. The transition matrix for this trial selection process is denoted by $P(\mathbf{X}, \mathbf{X}^*)$. Now, the trial configuration becomes the configuration at time $t + 1$, i.e. is *accepted*, with probability

$$\tilde{p}(\mathbf{X} \to \mathbf{X}^*) = \max\left\{1, \frac{P(\mathbf{X}^*, \mathbf{X})}{P(\mathbf{X}, \mathbf{X}^*)} \cdot \frac{q_T(\mathbf{X}^*)}{q_T(\mathbf{X})}\right\} = \max\left\{1, \frac{P(\mathbf{X}^*, \mathbf{X})}{P(\mathbf{X}, \mathbf{X}^*)} \exp[\Delta S/T]\right\}, \tag{8}$$

with $\Delta S = S(\mathbf{X}^*) - S(\mathbf{X})$ If the trial configuration is not accepted, the previous configuration $\mathbf{X}$ is kept for the next time step $t + 1$. In this way, the Markov chain fulfills the detailed balance condition $P(\mathbf{X}^*, \mathbf{X})\,\tilde{p}(\mathbf{X}^* \to \mathbf{X}) \cdot q_T(\mathbf{X}^*) = P(\mathbf{X}, \mathbf{X}^*)\,\tilde{p}(\mathbf{X} \to \mathbf{X}^*) \cdot q_T(\mathbf{X})$. In this case it has been proven that an ergodic Markov chain converges to the stationary distribution q_T. Ergodicity means, that there is a non-zero probability for a path between *any* pair$(\mathbf{X}_1, \mathbf{X}_2)$ of configurations.

We used a simple way to define the neighborhood of a configuration and constructed the trial configuration as follows: First a letter a is drawn from the alphabet Σ according to the letter weights f_a and next one of the sequences ($\mathbf{x}$ or $\mathbf{y}$) and a position i is chosen randomly. Finally, the letter at position i is replaced by a.

Given a Monte Carlo chain $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ estimated for a fixed temperature T in principle one may estimate expectation values with respect to any member of the family of distributions q_T by importance reweighting

$$\mathbb{E}_{T'}[g(\mathbf{X})] \approx \frac{1}{n}\sum_{i=1}^{n} \frac{q_{T'}(\mathbf{X}_i)}{q_T(\mathbf{X}_i)} \cdot g(\mathbf{X}_i)$$

Since the normalization of q_T is not trivial, we used a different normalization

$$\mathbb{E}_{T'}[g(\mathbf{X})] \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{q}_{T'}(\mathbf{X}_i)}{\tilde{q}_T(\mathbf{X}_i)} \cdot g(\mathbf{X}_i), \tag{9}$$

and estimate Z from the sample $Z = \sum_{k=1}^{n} \tilde{q}_{T'}(\mathbf{X}_k)/\tilde{q}_T(\mathbf{X}_k)$. A detailed discussion about this issue can be found in Ref. [31,32]. In practice this may work badly as soon as the parameter ranges of the given distribution and the target distribution do not overlap sufficiently. In this case $q_{T'}(\mathbf{X}_i)$ is very small, but the configurations where $q_{T'}(\mathbf{X})/q_T(\mathbf{X})$ is sufficiently large are not generated because $q_T(\mathbf{X})$ is relatively small for those. Therefore we sampled a mixture of many coupled Monte Carlo chains and reweighted the mixture, which is explained in detail in the next section. This allows for large overlap between neighboring distributions and to determine the normalization constants, up to an irrelevant global constant.

Metropolis Coupled MCMC

Metropolis Coupled Markov Chain Monte Carlo (MCMCMC) was first invented by Charles Geyer [33] and then reinvented by Hukushima and Nemoto [34] under the term *exchange Monte Carlo*. In physical literature MCMCMC is often denoted as *parallel tempering*. The method has become a standard tool in disordered systems with a rough (free) energy landscape [35]. These rough energy landscapes are characterized by high energy barriers and can be found for problems like protein folding [36-40], nucleation [41], spin-glasses [42,43] and other models characterized by rare events [19,44]. In the last decade it turned out that MCMCMC accelerates equilibration and mixing remarkably.

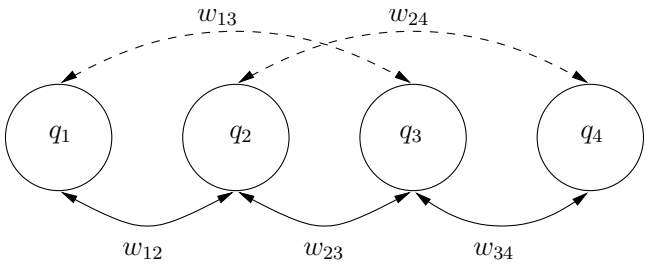

Figure 1
Sketch of the graph of overlapping distributions q_1,..., q_4. Distant distributions have weak overlaps.

In the framework of MCMCMC m copies $\mathbf{X}^{(1)}$, ..., $\mathbf{X}^{(m)}$ of the system held at different temperatures $T_1 < T_2 < ... < T_m$ are simulated in parallel. This means one samples from the product of the state space $\mathcal{X}^m$ weighted with the joint distribution with weights $\prod_{j=1}^{m} q_{T_j}$. Since the different copies are allowed to exchange temperatures during the simulation, let us define the space of all possible mappings from the m configurations to the m temperatures as *temperature space*.

During the simulation, mainly each of the replicated configurations will evolve independently according the underlying MCMC scheme charaterized by the weight Eq. (7) at its current temperature, i.e. according to Eq. (8). In addition to this evolution, every $t_{\text{exchange}}th$ step (for each replicated configuration) a flip between two neighboring replicas k and $k + 1$ is attempted, i.e. for all $k \in \{1, ..., m - 1\}$. If an attempt is successful, the configurations $\mathbf{X}^{(k)}$ and $\mathbf{X}^{(k+1)}$ are exchanged (denoted by $\mathbf{X}^{(k)} \leftrightarrow \mathbf{X}^{(k+1)}$), i.e. the configurations which has previously evolved at temperature T_k will now evolve at temperature T_{k+1} and vice versa. This exchange is accepted with the probability

$$\tilde{p}\left(\mathbf{X}^{(k)} \leftrightarrow \mathbf{X}^{(k+1)}\right) = \max\left\{1, \frac{q_{T_k}(\mathbf{X}^{(k+1)})}{q_{T_k}(\mathbf{X}^{(k)})} \cdot \frac{q_{T_{k+1}}(\mathbf{X}^{(k)})}{q_{T_{k+1}}(\mathbf{X}^{(k+1)})}\right\} = \max\left\{1, \exp[-\Delta\beta_k \Delta S]\right\}, \tag{10}$$

where, $\Delta\beta_k = \frac{1}{T_{k+1}} - \frac{1}{T_k}$, $\Delta S = S(\mathbf{X}^{(k+1)}) - S(\mathbf{X}^{k})$ and all weights are calculated with the configurations before the flip. This leads to a "random walk in temperature space" of the configurations.

Note that another possible approach based on Markov chains to compute p-values of a random model with a random variable X, $[X > b]$ was introduced by Wilbur [45]. The first step is to sample from an unbiased Markov chain based on the model of interest and compute the median of the (high probability) distribution. In the second iteration the random walk is truncated such that only values larger than the median of the first iteration occur. This corresponds to choosing a lower temperaure T in Eq. (7). The third iteration uses the median of the second iteration and so forth. This is repeated until a fraction of 1/4 of all events lay beyond a certain threshold value leading to a non decreasing sequence of splitting intervals defined by the medians of each iteration. This sequence is used in the second stage of the algorithm, where p-values are computed explicitly by multiplying the p-values of the truncated distribution in each iteration.

Although this method is easy to implement and errors can be estimated relatively simply, the MCMCMC approach has the advantage that the different configurations are not subjected to a sequence of decreasing temperatures, but perform a random walk in temperature space, i.e. visit all temperatures several times. Thus, mixing is accelerated and hence fewer Monte Carlo steps are required.

Reweighting the mixture

The production run of MCMCMC yield a set of m different chains of lengths n_j. We denote the *ith* configuration in the chain of *jth* temperature as $\mathbf{X}_i^{(j)}$. Of course this leads to a larger parameter range than simple importance reweighting of a single chain, hence Eq. (9) cannot be applied directly to the mixture. Geyer [46] developed a generalization of the importance reweighting formula to mixtures. His idea is based on Eq. (9), where q_T is replaced by a "mixture weight" q_{mix}, i.e. (using $q_j \equiv \tilde{q}_{T_j}$, i.e. q_j represents the unormalized weights)

$$\mathbb{E}_{T'}[g(\mathbf{X})] \approx \frac{1}{Z} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{q_{T'}(\mathbf{X}_i^{(j)})}{q_{\text{mix}}(\mathbf{X}_i^{(j)})} \cdot g(\mathbf{X}_i^{(j)}). \tag{11}$$

The (global) normalization constant is given by $Z = \sum_{j=1}^{m} \sum_{i=1}^{n_j} q_{T'}(\mathbf{X}_i^{(j)}) / q_{\text{mix}}(\mathbf{X}_i^{(j)})$. The mixture weight function is known up to normalization constants $c_j \equiv Z_{T_j}$:

$$q_{\text{mix}}(\mathbf{X}) = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \frac{q_j(\mathbf{X})}{c_j},$$

with $n = \sum_j n_j$. The unknown constants $\mathbf{c} \equiv (c_1, ..., c_m)$ may be estimated by *reverse logistic regression* introduced by Geyer [46]. Here we used an alternative approach to

obtain the constants **c** developed by Meng and Wong [47], which is explained now.

Since the global normalization constant Z in Eq. (11) is trivial, the problem is reduced to the estimation of $(m - 1)$ ratios of normalization constants to some reference value. One possible choice is to fix the normalization constant of q_1 and estimate the ratios $r_i = c_1/c_i$ $(i = 2, ..., m)$.

Since the support of the mixture distribution is broader than each of the particular distributions, not all pairs of distributions q_i and q_j overlap in general. The overlaps of the empirical data can be measured by the matrix

$$w_{ij} = \frac{1}{n_i n_j} \sum_S \left(\sum_{k=1}^{n_i} h_S(\mathbf{X}_k^{(i)}) \right) \cdot \left(\sum_{l=1}^{n_j} h_S(\mathbf{X}_l^{(j)}) \right)$$

and the set of distributions can be represented by a graph (V, E) with vertices being the weight functions $V = \{q_1, ..., q_m\}$ and the set of all overlaps being the weighted edges $E = \{w_{ij}\}$ with $w_{ij} > 0$(see Fig. 1. We require, that the so constructed graph is connected. In practice one must find paths between each pair of distributions with not too small weights. In this case each distribution has a finite overlap with q_{mix} and reweighting become possible on the full support.

Consider arbitrary weight functions α_{ij} assigned to each edge of the graph and define the following expectation values with respect to q_j

$$b_{ji} = \mathbb{E}_j[q_i(\mathbf{X}) \cdot \alpha_{ij}(\mathbf{X})] = \frac{1}{c_j} \sum_{\mathbf{X}} q_j(\mathbf{X}) \cdot q_i(\mathbf{X}) \cdot \alpha_{ij}(\mathbf{X}) = \frac{c_i}{c_j} b_{ij}. \tag{12}$$

This means, for any given vector **c**, all values $\{b_{ji}\}$ can be calculated using this expression. We require the α_{ij} to be symmetric, i.e. $\alpha_{ij} = \alpha_{ji}$, and a finite overlap with each of the distributions. With $r_1 = 1$ and $r_i b_{ji} = r_j b_{ij}$ it is straight forward to construct a linear system for the remaining $(m - 1)$ ratios, for $i > 1$:

$$b_{i1} = b_{1i} \cdot r_i = \sum_{j \neq i} b_{ji} \cdot r_i - \sum_{j \neq i, j > 1} b_{ij} \cdot r_j \equiv \sum_{j>1} a_{ij} \cdot r_j, \tag{13}$$

with $a_{ii} = \sum_{j \neq i} b_{ij}$ and $a_{ij} = -b_{ij}$ for $i \neq j$. This equations cannot be solved directly, because the coefficients a_{ij} do depend on the unknown ratios. However it is possible to solve Eq. (13) self-consistently. Using $\hat{\mathbf{b}} = (b_{11}, b_{21}, ..., b_{m1})$ and including explicitely the dependence on $\mathbf{r} = (r_1, r_2, ..., r_m)$ we obtain

$$A(\mathbf{r}^{(t)}) \cdot \mathbf{r}^{(t+1)} = \mathbf{b}(\mathbf{r}^{(t)}). \tag{14}$$

This equation can be solved by starting with $\mathbf{r}^{(1)} = (1, 1, ..., 1)$ and iteratively solving for $\mathbf{r}^{(t+1)}$ till convergence. Following the paper of Meng and Wong [47] Eq. (14) with the choice $\alpha_{ij}(\mathbf{X}) = \frac{n_i n_j}{|n|^2} \cdot q_{\text{mix}}(\mathbf{X})$ converges to same estimator as proposed by Geyer [46], which is based on maximization of a quasi-loglikelihood. The desired probability $P(s)$ can be achieved by setting $q_{T'}$ to the unbiased weight $q_\infty = 1$ and estimate the expectation values of the indicator functions h_S in Eq. (11).

Illustration and convergence diagnostics

In order to guarantee start configurations taken from the stationary distribution the first few iterations of the chains have to be discarded. The number of iterations to be discarded is denoted as burning or equilibration period. Usually one starts from a random (i.e. disordered) configuration and equilibrates the system. At the beginning of the simulation the system has a low score and hence it can reach in principle most regions of the score landscape. If the temperature is low, one sees when looking at Eq. (7) that configurations with large score dominate. Hence, typically the score increases or stays the same during the simulation with only few score-decreasing fluctuations.

Note that if "ground states" are also known, i.e. the maxima of the score landscape, the reverse process is possible, i.e. starting from a high maximum 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 barriers in the score landscape. This is the case when the average behavior for two runs, one starting with a disordered configuration and one starting with an "ground-state" configuration, is the same (within fluctuation). If the temperature is too small, this is usually not possible.

It is helpful to consider a simple toy system to illustrate and benchmark the method, in detail consider a 4-letter alphabet of equal weights and sequence lengths $L = M = 10, 20$. The scoring system is defined by the score matrix

$$\sigma(a,b) = \begin{cases} +1 & \text{if } a = b \\ -3 & else \end{cases}. \tag{15}$$

and affine gap costs with $\alpha = 4$ and $\beta = 2$.

An illustration of the equilibration criterion is given in Fig. 2. By "visual inspection" we obtain equilibration times 100 ($T = \infty$),1000 ($T = 1$), 10000 ($T = 0.7$), 15000 ($T = 0.6$) and 20000 ($T = 0.5$), respectively.

A more quantitative method was introduced by Raftery and Lewis [48,49], that estimates equilibration and sample times for a set of quantils. Raftery and Lewis's program, which is available from *StatLib* [50] or in the *CODA* package [51], estimates a *thining interval* n_{thin} as well. That means only every $n_{\text{thin}}th$ step is used for inference in order to avoid correlations between the scores at time t and $t + \Delta t$, that occur in MCMC in constrast to direct generating random sequences. The program requires three parameters: the desired accuracy r, the required probability s of attaining the specified accuracy and a less relevant tolerance parameter ε.

We compared the result of the estimate of the equilibration time with the simple visual approach: For the example given in Fig. 2 we maximized numerical estimate of equilibration time over a set of quantils between 0.1 and 0.95 for $r = 0.0125$, $s = 0.95$, $\varepsilon = 0.001$): The results for the equilibration time obtained by this approach are always much smaller than those obtained by the visual inspection. For example for $L = 20$, the Rafter-Lewis approach gives an equilibration time of 800 steps for the lowest temperature, whereas Fig. 2 suggests 20000 steps. Therefore equilibrium might not be guaranteed with the Rafter-Lewis approach and the visual inspection seems to be more conservative.

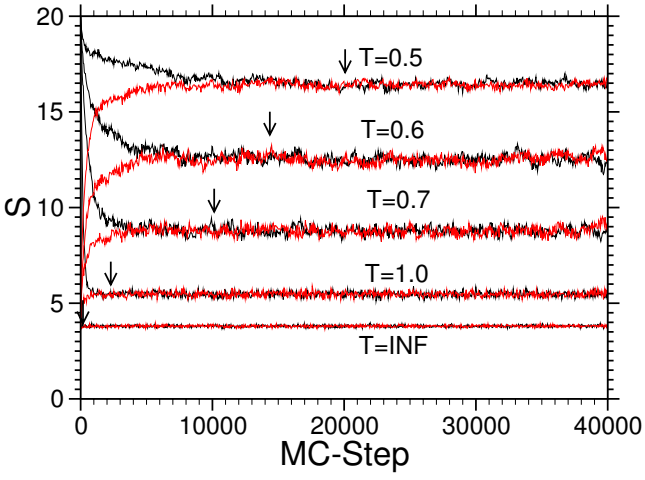

Figure 2
Equilibration of the 4-letter system ($L = M = 20$) with temperatures $T = 0.5, 0.6, 0.7, 1.0, \infty$ Equilibrium is reached after 20000, 15000, 10000, 1000, 100 steps (indicated by arrows) respectively. $S(t)$ is averaged over independent 250 runs.

To estimate the times scales over which the simulation decorrelates, we considered the autocorrelation function

$$\xi(t) = \frac{\langle S(t_0)S(t_0+t)\rangle_{t_0} - \langle S(t_0)\rangle_{t_0}^2}{\langle S(t_0)^2\rangle_{t_0} - \langle S(t_0)\rangle_{t_0}^2}, \qquad (16)$$

$\langle \cdots \rangle_{t_0}$ denoting the average over different times and independent runs. The typical time scale, over which correlation vanish is the correlation time τ defined via $\xi(\tau) = 1/e$. The normalized auto-correlation function for the system of $L = 20$ is shown in Fig. 3. A comparison with Raftery and Lewis diagnostics of n_{thin}, indicated by dots, gives evidence that the two estimates coincide with each other at least in the order of magnitude. 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 when the data are correlated. However the consideration of τ and n_{thin} has some practical issues too: For the application it is only necessary to infere every 100 *th* step, which saves a lot disk space.

Once the equilibration period is estimated one may check the convergence of the remaining parts of the chains to the equilibrium distributions. This was done by computing the Gelman and Rubin shrink factors R [49,52,53]. This diagnostic compares the "within-chain" and the "inter-chain variance" of a set of multiple Monte Carlo chains.

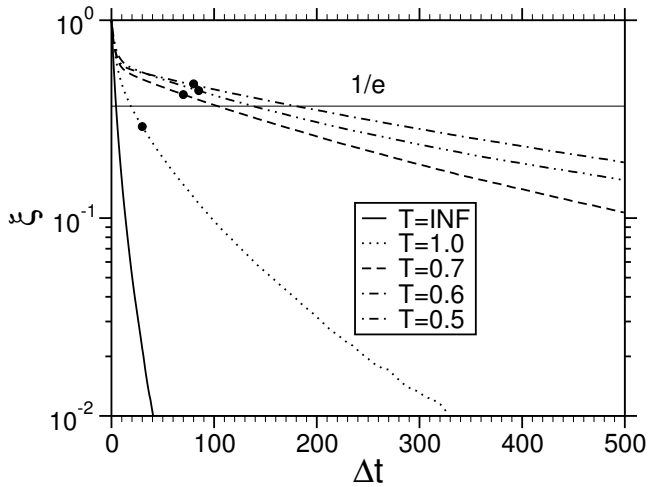

Figure 3
Score auto-correlation function for different temperatures (4 letters, $L = M = 20$). Circles indicate corre-sponding n_{thin} from Raftery and Lewis [48,49].

When the factor R approaches 1 the within-chain variance dominates and the sampler has forgotten its starting point. For the lowest temperature in our toy model $L = 20$ we found $R = 1.03$ for the 99.995% quantile, which appears to be reasonable.

From the equilibrated and converged chains we obtained histograms for different temperatures, which are shown in Fig. 4 for the case $L = 20$.

The empirical overlap matrix of this mixture is estimated by

$$(w_{ij}) \approx \begin{pmatrix} 1 & 0.543 & 0.256 & 0.098 & 0.009 \\ 0.543 & 1 & 0.572 & 0.266 & 0.070 \\ 0.256 & 0.572 & 1 & 0.624 & 0.264 \\ 0.098 & 0.266 & 0.624 & 1 & 0.570 \\ 0.009 & 0.070 & 0.264 & 0.570 & 1 \end{pmatrix}, \quad (17)$$

which has a finite overlap between *all* pairs. Note that in general a weaker condition must be fulfilled, namely that a connected path from the lowest to the hightest temperature must be possible, as outlined before. In more complex models only this condidition might be fulfilled.

Applying the reweighting technique, which was explained in the previous section, we obtain the infinite temperature probability $P(s)$ (see Fig. 5).

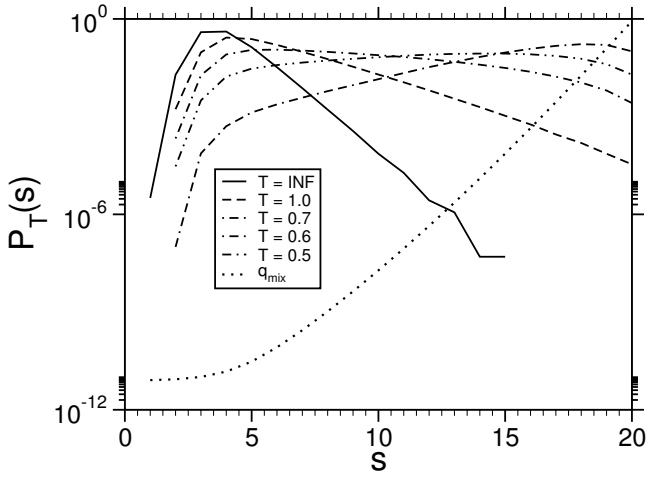

Figure 4
Empirical probabilities for the toy model (4 letters, $L = M = 20$) held at finite temperature. The dottet line showes the normalized mixture weight function $\hat{q}_{\text{mix}}$.

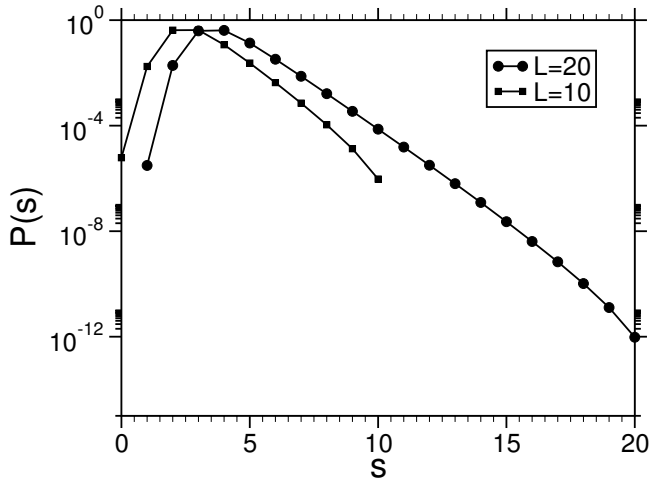

Figure 5
Score probabilities obtained throw the reweighting mixture technique for a 4-letter system with sequence-length $L = 10$, 20 and scoring parameters Eq. (15) using affine gap costs ($\alpha = 4$, $\beta = 2$). For $L = 10$ the $P(s)$ had also been been obtained by exact enumeration of all $4^{2 \times 10}$ configurations. A difference between the empirical curve is not visible in the plot.

Obviously, the toy model has $Z = 4^{2L}$ configurations. The maximum score over the ensemble of all possible configurations is $S_{\max} = L$. This corresponds to a pair of sequences with L equal letters $x_i = y_i$ ($i = 1 \ldots L$). The number of configurations with the highest score is 4^L. Hence, the probability to find a maximum score among all random sequences is $P(S_{\max}) = [S = S_{\max}] = 4^L/4^{2L} = 4^{-L}$. Below, to benchmark the Monte Carlo algorithm, we compare the convergence of the relative error $\varepsilon(S_{\max}) = \dfrac{\left|P_{\text{sample}}(S_{\max}) - 4^{-L}\right|}{4^{-L}}$ for different sequence lengths, $P_{\text{sample}}(s)$ being the corresponding probability obtained from the MC simulation. From Fig. 6, which illustrates convergence of the $\varepsilon(S_{\max})$ as a function of total sample size for all temperatures. In order to get a clear picture we averaged over several blocks of runs.

For small systems one may enumerate all possible configurations and compare the complete distribution with the Monte Carlo data. The empirical probability distribution for $L = 10$ in Fig. 5 coincides with the exact result, such that a the difference is not visible in the plot. However $L = 10$ is a very small system in contrast to real biological sequences, which are considered in section "Results", but exact enumeration is only possible on a modern computer cluster. Hence only for $L = 10$ the relative error

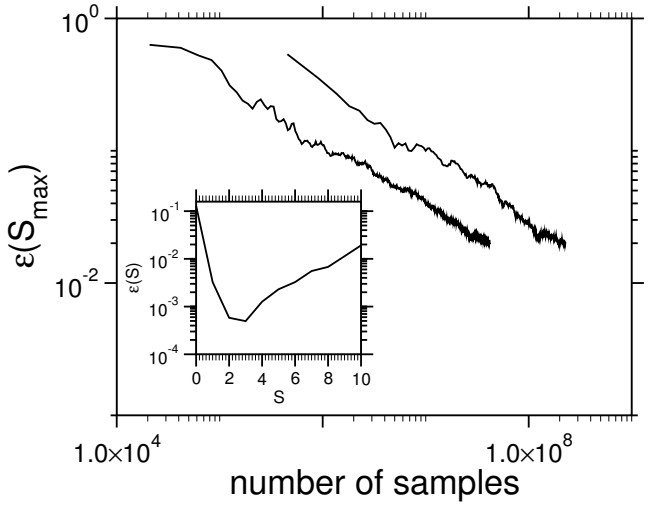

Figure 6
Rate of convergence of the MCMCMC data. The relative error $\varepsilon(S_{max})$ of the ground state for $L = 10$ and $L = 20$ depending on the number $N_{samples}$ of samples is shown. Inset: relative error of the final $P(s)$ incomparison to the exact enumeration of all states for the smallest system $L = 10$.

$\varepsilon(s) = \frac{\left|P_{\text{sample}}(s) - P_{\text{exact}}(s)\right|}{P_{\text{exact}}(s)}$ (see inset of Fig. 6) can be computed on the full support. In principle one is able to reduce variance on the low score end of the distribution by introducing negative temperature values, but this is beyond of the scope of this article.

Error estimation

As mentioned previously, a direct calculation of the errors is hardly possible. The first reason is that the Markov chain data are correlated. Secondly, the iterative estimation of the relative normalization constants is not trivial and contributes also to the overall error. Nevertheless, one can evaluate errors using the jackknife method [54]: First, in order to ensure, that the data are uncorrelated, we took data points which are seperated by at least the correlation time, determined via Eq. (16). Next, the dataset is divided into n_b blocks of equal size (hence, the number should be a multiple of n_b). Quantities of interests g are calculated k times ($k = 1 \ldots n_b$), each time omitting block B_k. These n_b values are averaged over all possibilities of k, in the notation of Eq. (11)

$$\left\langle g(\mathbf{X}_1,\ldots,\mathbf{X}_n)\right\rangle_k^J = \frac{1}{Zn_b}\sum_{k=1}^{n_b}\sum_{j=1}^{m}\sum_{i=1,i\notin B_k}^{n_j}\frac{q_{T'}(\mathbf{X}_i^{(j)})}{q_{\text{mix}}(\mathbf{X}_i^{(j)})}\cdot g(\mathbf{X}_i^{(j)}).$$

The error of g is estimated by

$$\sigma_{g^J} = \sqrt{(n_b-1)\left(\left\langle g^2(\mathbf{X}_1,\ldots,\mathbf{X}_n)\right\rangle_k^J - \left(\left\langle g(\mathbf{X}_1,\ldots,\mathbf{X}_n)\right\rangle_k^J\right)^2\right)}.$$

For example the relative errors $\sigma_{r_j^J}/r_j$ of the normalization constant ratios increase from 8.6×10^{-4} for r_2 to 1.29×10^{-2} for r_5. This indicates that the method is able to capture the error propagation of the relative normalization constants due to weak overlaps of distant distributions (see also Eq. (17)). Similar errors for the probabilities $P(s)$ can be estimated by applying this approach.

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 [55]. We consider different affine gap cost with $10 \le \alpha \le 16$, $\beta = 1$ for the BLOSUM62 matrix and $11 \le \alpha \le 17$, $\beta = 3$ 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 probability 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 MCMCMC 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 Gelman and Rubin shrink factors fell below 1.04 in almost all cases. For BLOSUM62 matrices and $L = 350$, 400 a slightly longer run (10^6) had been required to reduce R. The resulting probabilities were obtained from averaging over 10 ($L = 400$) up to 100 ($L = 40$) runs. The typical overlap matrix for the most complex system ($L = 400$, BLOSUM62) was

$$(w_{ij}) = \begin{pmatrix} 1 & 0.6850 & 0.5017 & 0.2717 & 0.0480 & 0.0015 \\ 0.6850 & 1 & 0.7857 & 0.4624 & 0.0984 & 0.0034 \\ 0.5017 & 0.7857 & 1 & 0.6409 & 0.1607 & 0.0117 \\ 0.2717 & 0.4624 & 0.6409 & 1 & 0.3587 & 0.0549 \\ 0.0480 & 0.0984 & 0.1607 & 0.3587 & 1 & 0.3777 \\ 0.0015 & 0.0034 & 0.0117 & 0.3777 & 0.3777 & 1 \end{pmatrix}.$$

Thus the overlap graph is connected sufficientely. For $L = 40$ we obtained relative errors of the normalization constants between 10^{-4}(highest temperature) and 0.4 (lowest temperature) and similar values for $L = 400$.

The main result is that most of the distributions we obtain deviate strongly from the Gumbel form, which is indicated in Fig. 7 and Fig. 8 by dotted lines. A typical example for the relative error of the results, obtained as explained above, is shown in Fig. 9. Note, that we used normalized scores $s^* = s - s_0$ by subtracting the position of the maximum s_0 of the probability distribution. According to Eq. (3), the form of the Gumbel distribution is independent of the sequence length. In the limit $L = M \to \infty$. In practice this is not the case due to edge effects [17,18] and database applications use adjusted λ's, but the distribution is still assumed to be of Gumbel form. The results in this work suggest that this is only the case for not too small p-values.

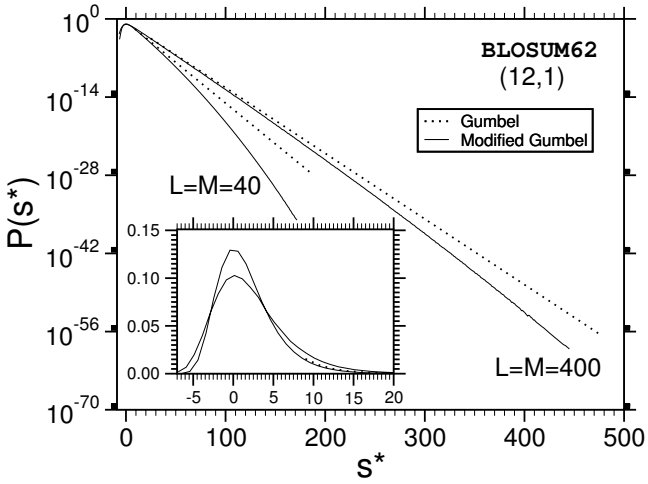

Figure 7
Probability distribution P(s) for gapped sequence alignment using BLOSUM62 matrices and affine gap costs with $\alpha = 12$, $\beta = 1$ for two sequences lengths $L = M = 40$. The results for other lengths are summarized in additional file 1. Strong deviations from the Gumbel distribution become visible in the tail. The dotted lines show the original Gumbel distribution, when fitted to the region of high probability. The inset shows the same data with linear ordinate.

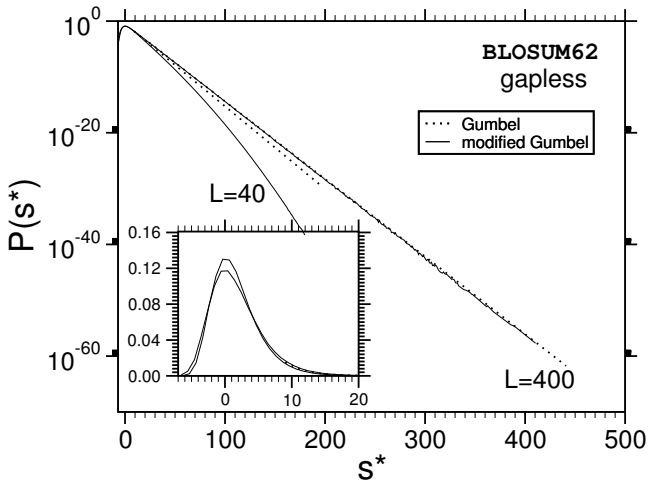

Figure 8
Probability distribution P(s) for ungapped sequence alignment using BLOSUM62-matrices. Deviations form the Gumbel-distribution can only be observed for short sequences ($L < 250$). The inset shows the same data with linear ordinate.

One observes that the discrepancy seems to be stronger for shorter sequences. Also, the case without gaps (Fig. 8) 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

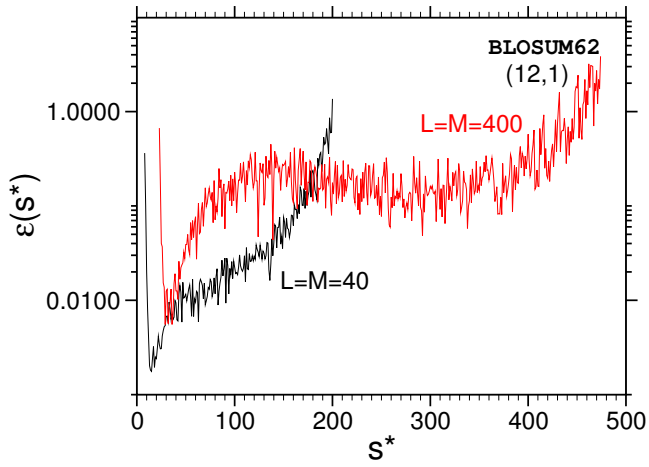

Figure 9
Relative error of the probability estimation using gapped sequence alignment and BLOSUM62 matrices.

Table 1: Fit parameters of the modified Gumbel distribution Eq. (18) using the BLOSUM62 scoring matrix and affine gap costs with α = 10, β = 1 . 10⁴ λ_2^{extra} describes the estimated value of λ_2 using the scaling relation Eq. (19). Fit parameters for other scoring systems are provided as supplementary material to this artilce [see additional file 1].

L, M	λ	$10^4\,\lambda_2$	K	s0	χ^2_*	$10^4\,\lambda_2^{\text{extra}}$
40	0.3272 ± 0.108%	8.6347 ± 0.412%	0.1028 ± 0.65%	15.597 ± 0.0676%	79.05	8.1560 ± 12.485%
60	0.3034 ± 0.086%	6.2007 ± 0.285%	0.0751 ± 0.60%	18.455 ± 0.0645%	49.40	6.1711 ± 12.907%
80	0.2892 ± 0.070%	4.8781 ± 0.222%	0.0612 ± 0.53%	20.644 ± 0.0540%	21.67	5.0458 ± 13.280%
100	0.2747 ± 0.072%	4.3187 ± 0.330%	0.0472 ± 0.58%	22.413 ± 0.0611%	39.42	4.3056 ± 13.627%
150	0.2541 ± 0.083%	3.2974 ± 0.529%	0.0303 ± 0.61%	25.682 ± 0.0422%	39.46	3.2047 ± 14.437%
200	0.2432 ± 0.063%	2.6343 ± 0.344%	0.0241 ± 0.52%	28.257 ± 0.0412%	10.47	2.5806 ± 15.214%
250	0.2359 ± 0.071%	2.1999 ± 0.454%	0.0198 ± 0.60%	30.196 ± 0.0459%	9.40	2.1701 ± 15.984%
300	0.2303 ± 0.061%	1.9101 ± 0.348%	0.0174 ± 0.54%	31.934 ± 0.0408%	2.00	1.8758 ± 16.758%
350	0.2261 ± 0.046%	1.6404 ± 0.239%	0.0153 ± 0.41%	33.334 ± 0.0300%	1.27	1.6525 ± 17.544%
400	0.2224 ± 0.052%	1.4806 ± 0.266%	0.0136 ± 0.49%	34.556 ± 0.0369%	1.36	1.4762 ± 18.347%
600	0.2140 ± 0.062%	1.0206 ± 0.384%	0.0106 ± 0.64%	38.561 ± 0.0472%	2.15	1.0250 ± 21.787%
800	0.2090 ± 0.063%	0.7660 ± 0.419%	0.0088 ± 0.67%	41.320 ± 0.0457%	1.82	0.7691 ± 25.697%

will be given below. Empirically we find that the resulting distribution can be described by a modified Gumbel distribution with a Gaussian correction:

$$P(s) = P_{\text{Gumbel}}(s)\cdot\exp\left[-\lambda_2(s-s_0)^2\right] = \lambda\exp\left[-\lambda(s-s_0)-\lambda_2(s-s_0)^2-e^{-\lambda(s-s_0)}\right], \quad (18)$$

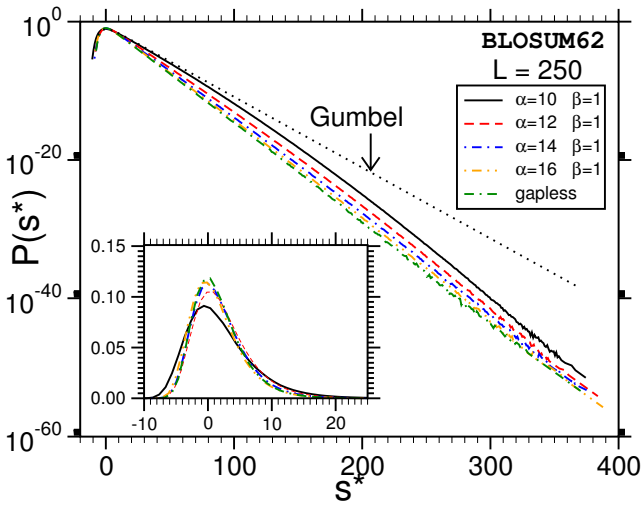

Figure 10
Probability distributions P(s) comparing different gap costs. The dotted line denote the distribution without Gaussian correction (λ_2 = 0). Deviations from the Gumbel distribution become stronger for small gap costs. The inset shows the same data with linear ordinate.

with $s_0 = \log(KLM)/\lambda$. 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. We modeled the data by a minimizing a weighted χ^2 using the program gnuplot [56]. The results including the reduced χ^2 - values ($\chi^2_* = \chi^2$ /degrees of freedom) are documented in Tab. 1 and as an additional CSV-file [see additional file 1].

All estimated standard errors in this paper are written behind the values and separated by "±".

Note that only for not too small sequences χ^2_* is in the order of one. This means that Eq. (18) describes the data better for longer sequences. However biological relevant sequence lengths (L > 200) sit in the range were the fit works fine. Moreover the results for shorter sequences are still several orders of magnitude below the naive Gumbel result, which yield χ^2_* a value of about 10^4 for the L = 40 system.

We also tried smaller gap costs than α < 10 (β = 1, BLOSUM62) and α < 11 (β = 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. 7 and Fig. 8, we expect that λ_2 decreases for $L \to \infty$. Furthermore, when looking at Fig. 10, 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. (18) and compared the resulting fit parameters.

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. 11 and Fig.12. BLOSUM62 and PAM250 behaves qualitatively the same. λ_2 seems to decay with a power law

$$\lambda_2(L) = a\, L^{-b} - \lambda_2^* \qquad (19)$$

for the smallest gap costs and faster than a power law for larger gap costs.

By fitting the limiting cases (two smallest gap costs) to this function an upper bound of the decay could be estimated. The results are summarized in Table 2.

Note that these arguments are purely heuristical attempts to look at the scaling behaviour and its upper bound. It is hard to decide, wether the extrapolation is valid for $L = M \to \infty$. However an important range of biological interesting sequence lengths are governed with this scaling analysis.

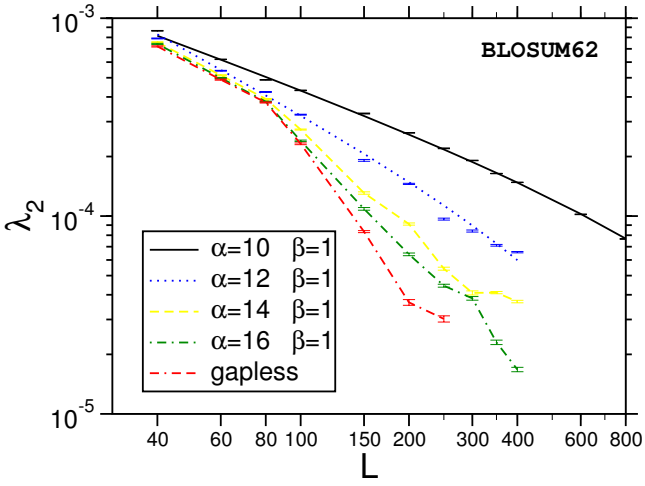

Figure 11
Scaling of the correction parameter λ_2 (BLOSUM62). The decay of λ_2 with system size shows approximately a power law near the logarithm-linear transition (two smallest gap costs). For this cases the fit to Eq. (19) is shown by a line (α = 10) and dots (α = 12). The lines of the remaining cases are guides to the eye conneting the data points.

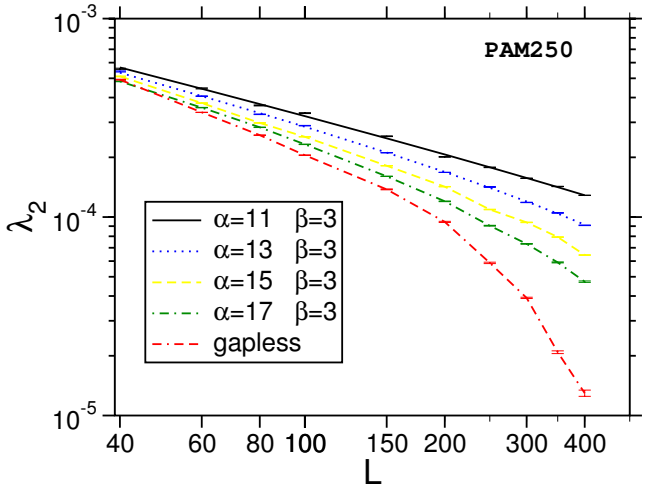

Figure 12
Scaling of the correction parameter λ_2 (PAM250). The decay of λ_2 with system size shows approximately a power law near the logarithm-linear transition (two smallest gap costs). For this cases the fit to Eq. (19) is shown by a line (α = 11) and dots (α = 13). The lines of the remaining cases are guides to the eye conneting the data points.

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 [57], which contains currently $N_{\text{swissprot}} = 210,623$ sequences. In BLAST [58], the E-value, i.e. the expected number of hits exhibiting at certain "cut-off" score b_{cut}, is currently estimated via the cumulative Gumbel distribution

$$E = KLN \cdot e^{-\lambda b_{\text{cut}}}, \qquad (20)$$

Table 2: Fitting parameters of the scaling relation Eq. (19).

Parameter	BLOSUM62 α = 10, β = 1	BLOSUM62 α = 12, β = 1
a	0.00928 ± 0.0001	0.0309 ± 0.01
b	0.643 ± 0.027	0.971 ± 0.08
$10^{-5}\ \lambda_2^*$	4.9 ± 1.2	3.2 ± 2.0
Parameter	**PAM250 α = 11, β = 3**	**PAM250 α = 13, β = 3**
a	0.0049 ± 0.0008	0.0053 ± 0.0005
b	0.575 ± 0.046	0.591 ± 0.023
$10^{-5}\ \lambda_2^*$	3.015 ± 2.0	6.1 ± 1.1

Table 3: Temperature parameters for sum-statistics.

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

where L is the query length and N the total number of amino acids of the entire database, with parameters K = 0.0410 and λ = 0.267. Using the suggested E-value of 10 [58], we find a cut-off of b_{cut} = 64.8 above which a result is considered to be significant, with $[S > b_{cut}] = 4.75 \times 10^{-5}$. Our cumulative distribution achieves this probability at b_{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 precission than when applying the approximation of the Gumbel distribution.

Sum statistics of the k*-best alignments*

The asymptotic distribution of the ungapped sum statistics is well known by Eq. (5). Again, we are interested in the distributions for *finite* sequence lengths. We use the SIM procedure [27] to compute the sum of the k-best alignments (k = 2, ..., 5) within the same type of Markov-chain Monte Carlo simulation as in the previous sections. In this case, we consider only the BLOSUM62 matrix together with affine gap costs α = 12, β = 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 $L \geq 60$ (k = 2) and $L \geq 80$ ($k \geq 3$) have been used for the analysis. The temperature sets varied between {2.75, 3.0, 3.5, 4.0, 7.0, ∞} for L = 100, k = 2 and {6.25, 6.5, 7, 9, 11, ∞} for L = 400, k = 5 (details are shown in Tab. 3).

Note that for k > 3 the systems could not be equilibrated in the very low temperature regime T < 5. Therefore, for theses cases, the tail could only be obtained in an intermediate range of probabilities (~10^{-20}), which is nevertheless low enough to obtain significance figures much better compared to using a simple-sampling approach.

In Fig. 13 we compare different distributions obtained for varying k and fixed sequence length L = 200. Similar to the case of optimal alignment quadratic deviations could be observed which decrease with growing system length for all values of k (not shown).

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. (6) with $t = \lambda\ (s - ks_0)$ we fitted the tail ($p < 10^{-10}$) of the Monte Carlo data to the modified distribution of the sum statistics, where the functional form f_{tail} from Eq. (6) is again modified by a Gaussian factor:

$$P(s) = Cf_{\text{tail}}[\lambda(s - ks_0)] \cdot \exp\left[\lambda_2^{(k)}(S - kS_0)^2\right]. \tag{21}$$

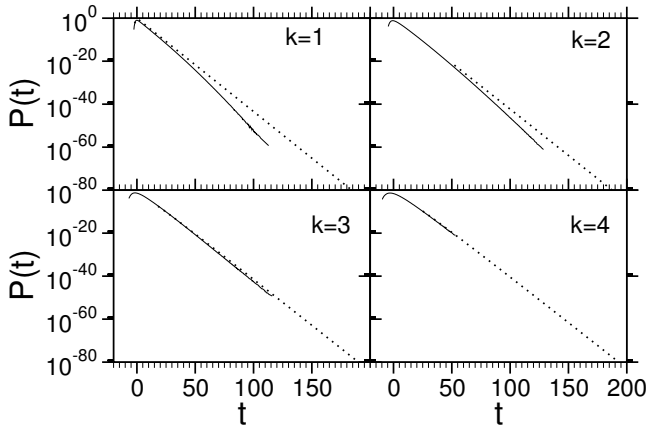

Figure 13
Score probability distributions for sum-statistics of the k-best scores (solid lines) for $L = M$ = 200. The dotted lines denote the distribution without Gaussian correction (λ_2 = 0). Deviations from Eq. (3) or Eq. (6) become only visible in the rare-event tail.

Table 4: Correction parameter λ_2 for the sum statistics $k = 2$ and $k = 3$. λ_2 is estimated by a fit for Eq. (21) 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.

L	$10^4\ \lambda_2^{(k=2)}$	$10^4\ \lambda_2^{(k=3)}$
60	2.692 ± 0.30%	
80	1.631 ± 0.63%	1.074 ± 2.59%
100	1.488 ± 0.23%	0.649 ± 2.06%
150	1.056 ± 0.06%	0.344 ± 1.90%
200	0.749 ± 0.13%	0.280 ± 1.14%
300	0.463 ± 0.15%	0.189 ± 0.70%
400	0.338 ± 0.29%	0.139 ± 0.92%

This was possible for $k = 2$ and $k = 3$. The results are summarized in Tab. 4 and the scaling behaviour of $\lambda_2^{(k)}$ is shown in Fig. 14. 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. 14) 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. 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.

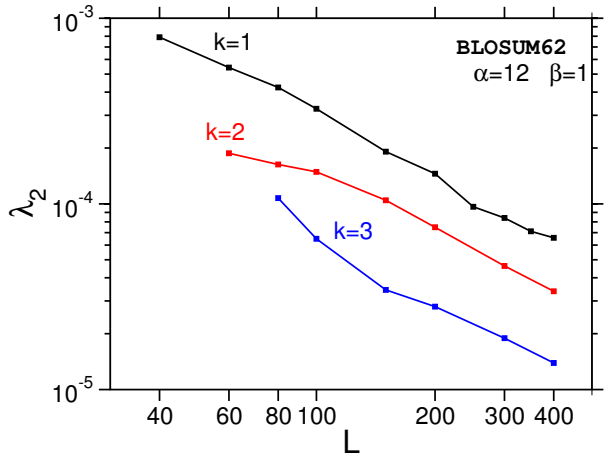

Figure 14
Scaling of the correction parameter for BLOSUM62 sum-statistics ($k = 1, 2, 3$). λ_2 is estimated by a fit for Eq. (21) using optimal the Gumbel-parameters λ and S_0 from optimal score statistics ($k = 1$).

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 of biological relevance, we observed that the distributions can be fitted well by a Gumbel distribution with a Gaussian correction. Therefore, for data bases like BLAST [8,18,58], 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 infinitely long sequences gives good evidence that the ungapped statistical theory describes the gapped case for $L = M \to \infty$ as well.

Additional material

Additional file 1

Fit parameter of the modified Gumbel distribution. *CSV file (tabulator separated) of fit parameters of the modified Gumbel distribution Eq. (18) using different scoring matrices (BLOSUM62 and (PAM250) and gap costs.* $10^4\ \lambda_2^{\text{extra}}$ *describes the estimated value of* λ_2 *using the scaling relation Eq. (19) (for small gap costs only).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1748-7188-2-9-S1.csv]

Acknowledgements

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 Uni-versitäten", and from the European Community via the DYGLAGEMEM program.

References

1. Brown S: *Bioinformatics* Natick (MA): Eaton Publishing; 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:**403-410.
9. Karlin S, Altschul S: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87:**2264.
10. Dembo A, Karlin S, Zeitouni O: **Limit Distribution of Maximal Non-Aligned Two-Sequence Segmental Score.** *Ann Prob* 1994, **22:**2022-2039.
11. Yu Y, Hwa T: **Statistical Significance of Probabilistic Sequence Alignment and Related Local Hidden Markov Models.** *J Comp Biol* 2001, **8(3):**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* edition. 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. Olsen R, Bundschuh R, Hwa T: **Rapid Assessment of Extremal Statistics for Local Alignment with Gaps.** In *Proceedings of the seventh International Conference on Intelligent Systems for Molecular Biology Volume 270.* Edited by: Lengauer T, Schneider R, Bork P, Brutlag D, Glasgow J, Mewes HW, Zimmer R, Menlo Park. CA: AAAI Press; 1999:211-222.
18. Altschul S, Bundschuh R, Olsen R, Hwa T: **The estimation of statistical parameters for local alignment score distributions.** *Nucl Acid Res* 2001, **29(2):**351-361.
19. Hartmann A: **Sampling rare events: Statistics of local sequence alignments.** *Phys Rev E* 2002, **65(5 Pt 2):**056102.
20. Heinkoff S, Heinkoff J: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 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*. Issue 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*. Issue 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:**200-225.
25. Hwa T, Lässig M: **Optimal Detection of Sequence Similarity by Local Alignment.** *Proceedings of the Second Annual International Conference on Computational Molecular Biology (RECOMB98)* 1998:109.
26. Sellers P: **Pattern recognition in genetic sequences by mismatch density.** *Bull Math Biol* 1984, **46:**501-514.
27. Altschul S, Erickson B: **Locally optimal subalignments using nonlinear similartity functions.** *Bull Math Biol* 1986, **48:**633-660.
28. Karlin S, Altschul S: **Applications and statistics for multiple high-scoring segments in molecular sequences.** *Proc Natl Acad Sci USA* 1993, **90:**5873-5877.
29. Dieker A, Mandjes M: **On Asymptotically efficient simulation of large deviation probabilities.** *Adv Appl Prob* 2005, **37:**539-552.
30. Hastings WK: **Monte Carlo Sampling Methods Using Markov Chains and Their Applications.** *Biometrika* 1970, **57:**97-109.
31. Liu J: *Monte Carlo Strategies in Scientific Computing* New York: Springer; 2002.
32. Liu J: **Metropolized independent sampling with comparisons to rejection sampling and importance sampling.** *Statist Comput* 1996, **6:**113-119.
33. Geyer C: **Monte Carlo Maximum Likelihood for Depend Data.** *Proceedings of the 23rd Symposium on the Interface* 1991:156-163.
34. Hukushima K, Nemoto K: **Exchange Monte Carlo Method and Application to Spin Glass Simulations.** *J Phys Soc Jpn* 1996, **65:**1604-1608.
35. Earl D, Deem M: **Parallel tempering: Theory, applications, and new perspectives.** *Phys Chem Chem Phys* 2005, **7:**3910-3916.
36. Zhou R: **Exploring the protein folding free energy landscape: Coupling replica exchange method with P3ME/RESPA algorithm.** *J Molec Graph Mod* 2004, **22(5):**451-463.
37. Zhou R, Berne B: **Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water?** *Proc Natl Acad Sci USA* 2002, **99:**12777-12782.
38. Zhou R, Berne B: **Trp-cage: Folding free energy landscape in explicit water.** *Proc Natl Acad Sci USA* 2002, **100(23):**13280-13285.
39. Garci'a A, Onuchic J: **Folding a protein in a computer: An atomic description of the folding/unfolding of protein.** *Proc Natl Acad Sci USA* 2003, **100:**13898-13903.
40. Zhou R, Berne B, Germain R: **The free energy landscape for β hairpin folding in explicit water.** *Proc Natl Acad Sci USA* 2001, **98:**14931-14936.
41. Auer S, Frenkel D: **Prediction of absolute crystal-nucleation rate in hard-sphere colloids.** *Nature* 2001, **409:**1020-1023.
42. Marinari E, Parisi G, Ruiz-Lorenzo J: **Numerical Simulations of Spin Glass Systems.** In *Spin Glasses and Random Fields, Directions in Condensed Matter Physics Volume 12*. Edited by: Young A. World Scientific; 1998::109.
43. Katzgraber H, Palassini M, Young A: **Monte Carlo simulations of spin glasses at low temperatures.** *Phys Rev B* 2001, **63:**1844221-18442210.
44. 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.
45. Wilbur W: **Accurate Monte Carlo Estimation of Very Small P-Values In Markov Chains.** *Comp Stat* 1998, **13:**153-168.
46. Geyer C: **Estimating Normalization Constants and Reweighting Mixtures in Markov Chain Monte Carlo.** In *Tech Rep 568* School of Statistics, University of Minnesota; 1994.

47. Meng X, Wong W: **Simulating Ratios of Normalization Constants via a Simple Identity: ATheoretical Exploration.** *Statistica Sinica* 1996, **6:**831-860.
48. Raftery A, Lewis S: **How Many Iterations in the Gibbs Sampler.** In *Bayesian Statistics 4* Edited by: Bernardo J, Berger J, Dawid A, Smith A. Oxford University Press; 1992:763-773.
49. Cowles M, Carlin B: **Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review.** *JASA* 1996, **91(434):**883-904.
50. **StatLib** [http://lib.stat.cmu.edu/]
51. **Coda R package** [http://www.r-project.org/]
52. Gelman A, Rubin D: **Inference from iterative simulation using multiple sequences.** *Stat Sci* 1992, **7:**457-472.
53. Brooks S, Gelman A: **General methods for monitoring convergence of iterative simulations.** *J Comput Graph Stat* 1998, **7:**434-455.
54. BEfron: *The Jackknife, the Bootstrap and Other Resampling Plans* New York: SIAM; 1982.
55. Robinson A, Robinson L: **Distribution of glutamine and asparagine residues and their near neighbours in peptides and proteins.** *Proc Natl Acad Sci USA* 1991, **88:**8880-8884.
56. **gnuplot** [http://www.gnuplot.info/]
57. **SWISSPROT** [http://www.expasy.org/]
58. **NCBI BLAST** [http://www.ncbi.nlm.nih.gov/BLAST]