Large-deviation properties of sequence alignment of correlated sequences

Matthias Werner
SFB 1114 Scaling Cascades in Complex Systems,
Free University of Berlin, 14195 Berlin, Germany

Pascal Fieth*
Institute of Physics, University of Oldenburg
26111 Oldenburg, Germany

Alexander Hartmann
Institute of Physics, University of Oldenburg
26111 Oldenburg, Germany

1 Abstract

The significance of alignment scores of optimally aligned DNA sequences can be estimated through the score distribution of pairs of random sequences. It is necessary to obtain statistics for the relevant high-scoring tail of the distribution. For local alignments of iid drawn sequences it has already been shown that the often assumed Gumbel distribution does not hold in the distribution tail, but has to be corrected by a Gaussian factor. Real DNA sequences were observed to show long-range correlations within sequences, which is not correctly modeled by iid random sequences. In this publication the large deviation method that was used in previous studies is applied to local and global alignment of such sequences with long-range correlations. We study the distributions over the full range of the support and obtained probabilities as low as 10^{-55}. We show that again a correction to the Gumbel distribution is necessary and study the dependence of the parameters on the correlation strength. For global alignments the Gamma distribution, which was found heuristically to be a good fit in earlier simple sampling studies, is found to be a poor fit.

2 Introduction

To analyze DNA or amino acid sequences by comparison of new sequences to sets of known ones, one uses sequence alignment, where a vast number of bioinformatics tools exist (Durbin et al., 1998). The alignment of any set of sequences yields an *alignment score* S. To allow conclusions on the significance of the alignment one considers ensembles of randomly drawn sequences, used as null models, and calculates the probability $P(S' \geq S)$ to find a score S' equal to or better than the one observed. This probability is usually called *p-value*. Therefore, these p-values can be calculated from an according score distribution $P(S)$. p-values are widely used, e.g., to estimate the significance of alignments in major public databases like PDB (Berman et al., 2000) or UniProt (The UniProt Consortium, 2017). It is essential to have a good statistical basis to assess the significance of sequence similarity found with alignment algorithms.

The probability distribution for alignment scores of randomly drawn sequences is known analytically only for the theoretical case of gapless local alignments of infinitely long sequences (Karlin et al., 1990). Here a Gumbel distribution was found. Large deviation studies on gapped alignments of *iid* sequences of finite size have shown that, while this distributions fits data well for the high probability region of score distributions, corrections are necessary (Fieth and Hartmann, 2016; Hartmann, 2002; Wolfsheimer et al., 2007) to describe the tails, where the probabilities are very small like 10^{-50}. This region is relevant for biological applications, because nature is shaped by evolution, therefore relevant structures would emerge in random sequence models only with such extremely small probabilities.

For the case of iid sequences, the letters for a sequence are drawn randomly independently and identically (from the DNA alphabet $\Sigma = \{G, T, A, C\}$. Natural DNA sequences often show long-range correlations (Li and Kaneko, 1992) as measured by the correlation function

$$C(r) = \sum_{n \in \Sigma} [P(x_i = x_{i+r} = n) - P(n)^2] \tag{1}$$

for distance r. To use such sequences within computer simulations one can use, e.g., the *CorGen* algorithm (Messer and Arndt, 2006), which generates sequences with $C(r) \propto r^{-\alpha}$ with a decay parameter α. Stronger correlations or lower α values lead to larger scores, shifting the score distribution to higher probabilities.

Messer et al. (2007) analyzed the score distributions of correlated DNA in the high probability region. Nevertheless, to our knowledge, there is no publication which studies the score distributions in the biologically relevant small-probability tail, which cannot be reached by standard sampling.

The aim of this paper is to show results for the application of large deviation studies to the alignment of DNA sequences with correlations. Distributions will be shown to probabilities as low as $P(S) \approx 10^{-55}$ whereas previous studies without the large deviation approach were only able to cover the distribution down to $P(S) \approx 10^{-7}$(Messer et al., 2007). We will show that, again, a Gaussian correction to the Gumbel distribution improves the fit of the function to the whole distribution in case of local alignments. We will further show that the increase of Gumbel tail parameter λ increases with the decay

exponent α, while the correction parameter λ_2 does not show such a clear dependence. Finally we will present the case of global alignment as well for which the in previous studies heuristically found Gamma distribution is shown to be a poor fit (Pang et al., 2005), especially for the alignment of sequences with strong correlation.

3 Methods

3.1 Sequence alignment and score distributions

Pairs of sequences are aligned with the Needleman-Wunsch (Needleman and Wunsch, 1970) (for global alignments) and the Smith-Waterman (Smith and Waterman, 1981) (for local alignments) algorithms. Pairs of letters (a, b) are scored according to

$$s(a,b) = \begin{cases} +1, & a = b \\ -3, & a \neq b. \end{cases} \tag{2}$$

This means, matching letters increase the score by 1, mismatches decrease it by 3. Gaps are penalized with affine gap costs using $\gamma_i = 5$ as gap initiation penalty and $\gamma_e = 2$ as gap extension penalty. $s(a,b)$, γ_i and γ_e are chosen in accordance with Messer et al. (2007).

Score distributions for the gapless local alignment of infinitely long sequences have been analytically found to follow a Gumbel distribution (Karlin et al., 1990). Transferred to pairwise alignment of two sequences of identical length L the distribution is assumed to follow

$$P(S) = \lambda \exp\left(-\lambda(S - S_0) - e^{-\lambda(S-S_0)}\right) \tag{3}$$

with constant λ. Numerical studies have found this to be a good estimate for the high-probability region (Altschul and Gish, 1996). However, such studies with a simple sampling approach can only sample sequences to cover probabilities higher than $P(S) \geq 10^{-10}$. Numerical large deviation analyses of score distributions yielded a deviation from the Gumbel distribution in the low-probability tail (Wolfsheimer et al., 2007; Fieth and Hartmann, 2016). An adjusted form for the distribution can be given by

$$\begin{aligned} P(S) &= P_{\text{Gumbel}}\, e^{-\lambda_2 (S-S_0)^2} \\ &= \lambda \exp\left(-\lambda(S - S_0) - e^{-\lambda(S-S_0)} - \lambda_2 (S - S_0)^2\right), \end{aligned} \tag{4}$$

with a new parameter λ_2 which indicates the strength of a Gaussian correction.

For global alignment, numerical studies of the high-probability region have yielded the Gamma distribution as a heuristic estimate for the score distribution (Pang et al., 2005)

$$P_{\text{gamma}}(S) = \begin{cases} \frac{\lambda^\gamma (S-\mu)^{\gamma-1} e^{-\lambda(S-\mu)}}{\Gamma(\gamma)} & S > \mu \\ 0 & S \leq \mu, \end{cases} \tag{5}$$

with the Gamma function $\Gamma(x)$, and constants λ, γ, μ. Large deviation studies yielded again a deviation for lower probabilities (Fieth and Hartmann, 2016)

$$P_{\text{gc}}(S) = \begin{cases} \frac{\lambda^\gamma (S-\mu)^{\gamma-1} e^{-\lambda(S-\mu)}}{\Gamma(\gamma)} e^{\lambda_2 S^2} & S > \mu \\ 0 & S \leq \mu, \end{cases} \tag{6}$$

with another Gaussian correction indicated by parameter λ_2.

3.2 Large Deviation approach

To obtain the score distributions in the region of low probabilities, the approach for sequence alignment as introduced by AKH in Hartmann (2002) was used. The basic principle is to use the Metropolis algorithm (Metropolis et al., 1953) to sample a Markov Chain of sequence pairs with a bias parameter T. In a Markov Chain parametrized by a "time" t, the system state $\mathcal{C}_t$ (here: the sequence pair) is slightly changed to trial state $\mathcal{C}'$, its new score $\mathcal{C}'(S)$ is calculated and a new step *accepted* with probability $P(\mathcal{C}_{t+1} = \mathcal{C}') = \min[1, \exp(\Delta S/T)]$ with $\Delta S = S(\mathcal{C}') - S(\mathcal{C}_t)$. In case of non-acceptance, the current state is kept, i.e. $\mathcal{C}_{t+1} = \mathcal{C}_t$. This procedure yields when sampling the steady state of the Markov chain a biased score distribution

$$P_T(S) = \frac{\exp(S/T)}{Z_T} P(S) \tag{7}$$

with the partition function Z_T for parameter T as normalization. Note that the unbiased distribution then corresponds to the case $T = \infty$ and we can estimate $Z_{T=\infty} = 1$.

For details see Hartmann (2002). Here we just briefly note that the unbiased distribution can be obtained by successively rescaling the biased distributions by the factor $\exp(S/T)$ and estimating the partition function Z_T comparing overlapping data points with the unbiased distribution. The simulations have to be done for an appropriate range of scaling parameters T, so that rescaled distributions have enough overlap to estimate the respective Z_T. Algorithm performance can then be improved by sampling several Markov chains with different parameters T_i in parallel and systematically switching parameters T_i, T_{i+1}(Geyer, 1991; Hukushima and Nemoto, 1996; Marinari and Parisi, 1992).

3.3 CorGen

The CorGen algorithm as presented by Messer and Arndt (2006) generates sequences with $C(r) \propto r^{-\alpha}$. An iid sequence x of length N_0 is generated initially. A letter x_k in the sequence is chosen at random and either mutated with probability P_{mut} or duplicated, i.e. introduced at position $k+1$ after shifting all letters from position $k+1$ on by one, increasing the sequence length. It is $\alpha = 2P_{\text{mut}}/(1 - P_{\text{mut}})$. Letters are drawn to ensure a GC-content of $g = 0.5$, the initial sequence length is chosen as $N_0 = 6$. In practice, α has to be chosen according to the typical correlation decay observed in assessed genomes. Therefore it will be varied in this study.

Slightly changing a sequence set can be achieved by randomly changing a single letter for iid sequences without correlation. Here, however, we have to maintain the correlation in the sequences. To achieve this we use the fact that sequences are generated by pseudo random numbers. A particular set of sequences showing long-range correlation with CorGen can be replicated by feeding the algorithm the exact same pseudo random numbers as in the first run. However, as the decision for or against mutation changes the number of random numbers needed per generation step, one manipulation of the vector could radically change the generation process. This is not desired in a Markov chain simulation, because small changes in a state should yield also only small changes

in the results. To avoid resulting large "chaotic" changes, three random vectors are used, one to decide to mutate or duplicate, one to decide which element of the sequence this decision is applied to and a third vector that, in case of mutation, chooses the replacing letter from the alphabet. Whilst the i-th element of the first two vectors is used in the i-th step of the sequence generation, the k-th element of the third vector decides the k-th mutation ($k \leq i$) that occurs during the generation process.

4 Results

4.1 Local alignment

Using the large deviation approach, distributions could be obtained down to the maximum possible score ($S = N$ for sequences of length N). Figure 1 shows the distribution as obtained for $N = 100$ and $\alpha = 2.0$. A fit of the Gumbel distribution to the high-probability region $p(S) \geq 10^{-10}$ was attempted. This corresponds roughly to the region obtainable by simple sampling in feasible computation time. For this region only, the fit performs relatively well ($\chi^2/\mathrm{ndf} = 17.6$), but we see that it deviates significantly from the obtained distribution in the tail. Fitting the pure Gumbel distribution to the whole obtained distribution fails ($\chi^2/\mathrm{ndf} = 830$). Using the Gumbel distribution with Gaussian correction (4) for the whole distributions performs better in contrast ($\chi^2/\mathrm{ndf} = 50$). The χ^2 value indicates that this heuristically found correction is still not ideal, but it significantly improves the fit. Anyway, the numerical data display the true distribution, even if the correct function to fit is unknown to us.

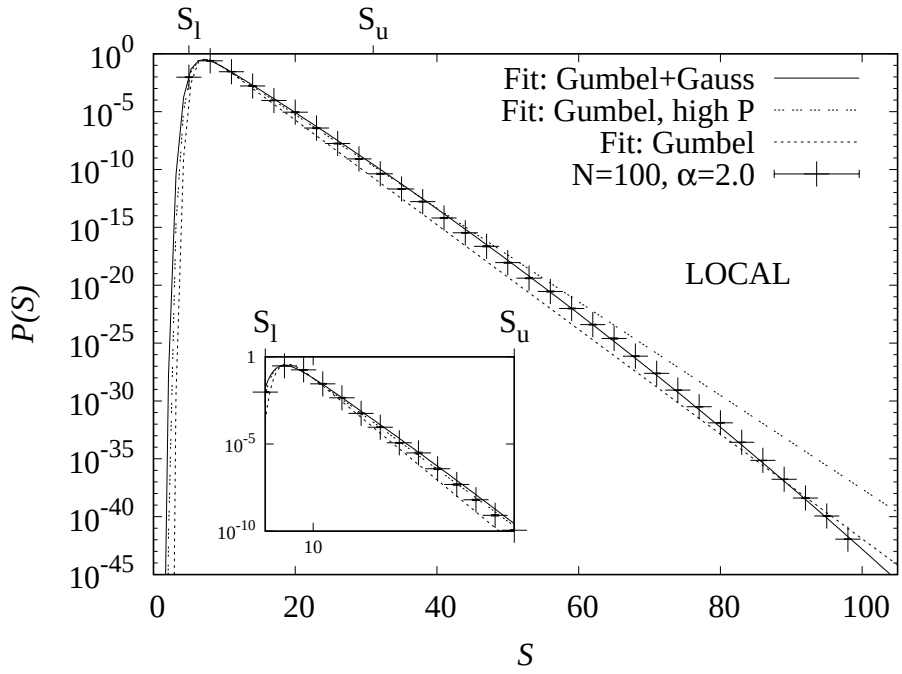

Figure 1: The score distribution from the rare-event simulations for local alignment of sequences of length $N = 100$ and correlation parameter $\alpha = 2.0$. Shown are the fits of the Gumbel distribution to the whole data set as well as the Gumbel distribution restricted to the score range $[S_l = 5, S_u = 31]$, corresponding to high probabilities $P(S) \geq 10^{-10}$. In comparison the Gumbel distribution with Gaussian correction is shown, which fits the data better. The inset shows the high probability region, for which the Gumbel distribution with restricted fit and with Gaussian correction show good results.

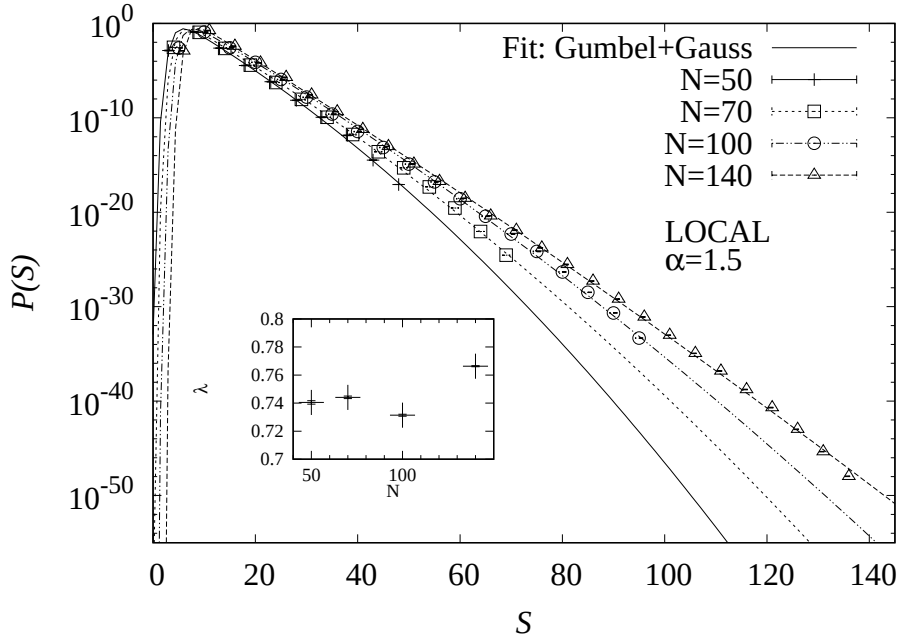

Figure 2: The score distributions for alignments of sequences of different lengths with a fixed correlation parameter of $\alpha = 1.5$. The curvature, fitted by the Gaussian correction parameter λ_2 decreases with increasing sequence length N. The inset shows the parameter λ, indicating little to no influence of the parameter on curve differences.

Figure 2 shows distributions for fixed correlation ($\alpha = 1.5$) and varying sequence lengths. The Gumbel distribution with Gaussian correction was fitted to all distributions. The curvature appears to be increasing with decreasing sequence length. The parameter λ seems to have no conclusive dependence on the sequence length as suggested by the inset of figure 2 which is concordant with the mutual λ for all values of N observed in Messer et al. (2007). But the parameter λ_2, indicating the Gaussian correction, increases with $1/N$ as seen in the inset of figure 3, suggesting it is responsible for describing the increasing curvature with decreasing sequence length. Figure 3 shows the distributions rescaled (Newberg, 2008) by the maximum-possible score $S_{\max} = N$. The distributions coincide for the low-probability region, but deter in the high-probability region.

Note that, we also performed a heuristic rescaling (not shown here) as in Messer et al. (2007). The rescaling yields coinciding distributions in the high-probability region, which show that our results are compatible with the past work. Nevertheless, this rescaling yields a strong deviation in the previously unobserved low-probability region.

For the dependence of the distributions on the correlation we varied parameter α. Figure 4 shows different distributions obtained for $N = 100$. Again the Gumbel distribution with Gaussian correction was fitted to the data. As a general trend, the fit performed worse for distributions of higher correlated sequences, as suggested by the

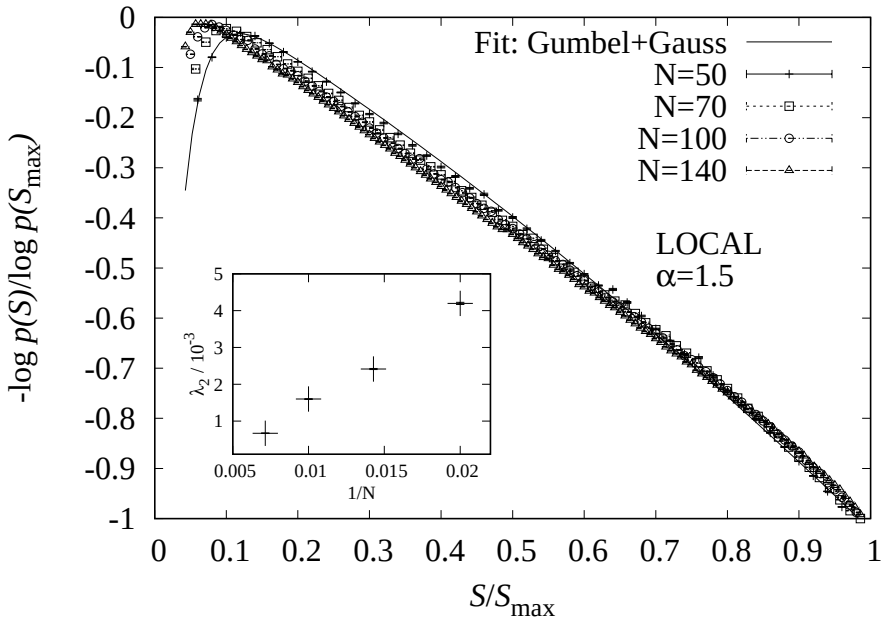

Figure 3: The rescaled score distributions for alignments of sequences of different lengths with a fixed correlation parameter of $\alpha = 1.5$. The distributions coincide better for lower probabilities. The inset shows the Gaussian correction parameter λ_2 over the inverse sequence length $1/N$.

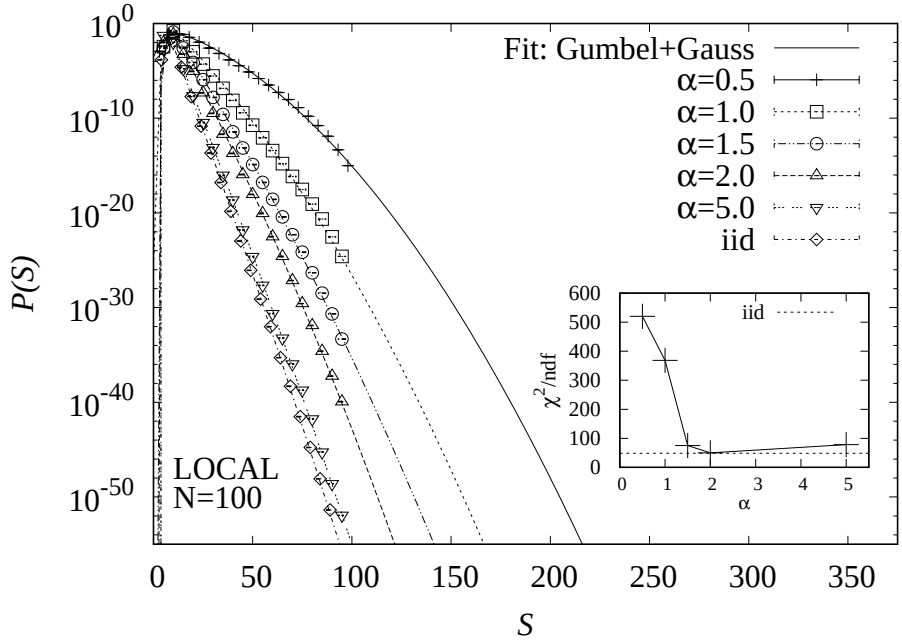

Figure 4: Score distribution for a single sequence length ($N = 100$) and varied values of α. Lines show fits of the Gumbel distribution with Gaussian correction. The inset shows χ^2/ndf over α, the solid line is to guide the eye, the dotted line indicates the value for a randomly sampled sequence without correlation (i.e. $\alpha = \infty$). For increasing correlation the fit of the Gumbel function with Gaussian correction performs increasingly worse.

χ^2 values shown in the inset of figure 4. The parameter λ increases with increasing α and approaches the asymptotic iid case as seen in figure 5. In contrast the Gaussian correction shows no clear trend in its α-dependence. Anyway, the visual inspection of Figure 4 and the comparable large values of λ_2 show that the deviations from the Gumbel distributions are highly significant for small values of α, i.e. strong correlations.

4.2 Global alignment

The same approach was used to to obtain the score distributions for *global* alignment of sequences exhibiting correlation. The distribution for the case $N = 100$ and $\alpha = 2.0$ is shown in figure 6. The whole distribution could be obtained. The Gamma distribution with and without correction was fitted to the data. The plot indicates that neither of the functions fits the data well ($\chi^2 = 224$ and $\chi^2 = 292$). In case of iid sequences both fits perform even worse which is in contrast to former findings in which at least the Gamma distribution with Gaussian correction showed good results (Fieth and Hartmann, 2016). However, in the previous study only part of the distribution was obtained and, indeed, fitting only to a more restricted range improves χ^2 values. Also the previous study dealt with amino acid sequences, not with DNA sequences which might be another

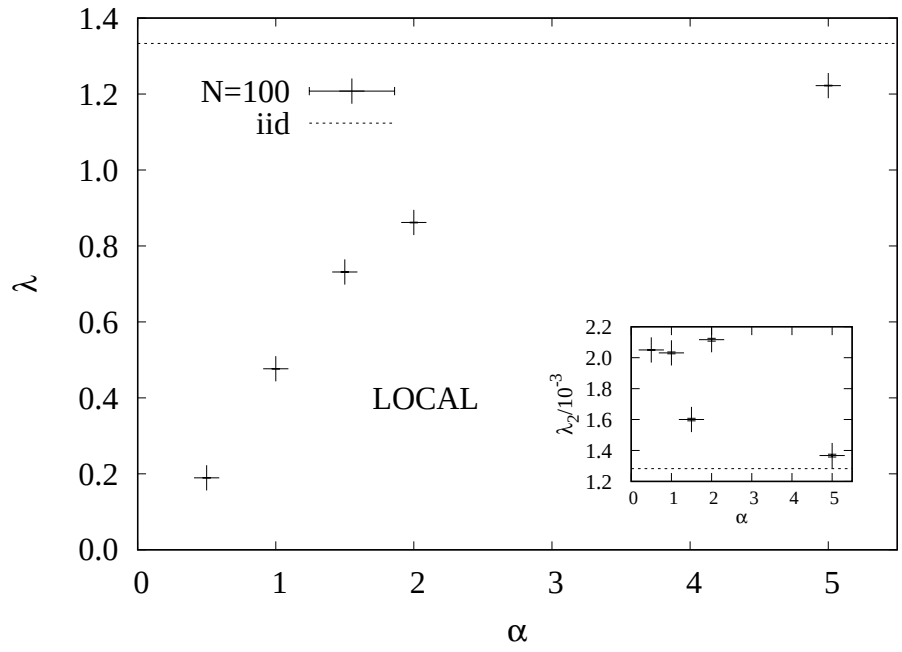

Figure 5: Fit parameter λ over correlation parameter α. The dashed line indicates the value for sequences without correlation ($\alpha = \infty$). The value of λ increases with increasing α or decreasing correlation. The inset show the Gaussian correction parameter λ_2.

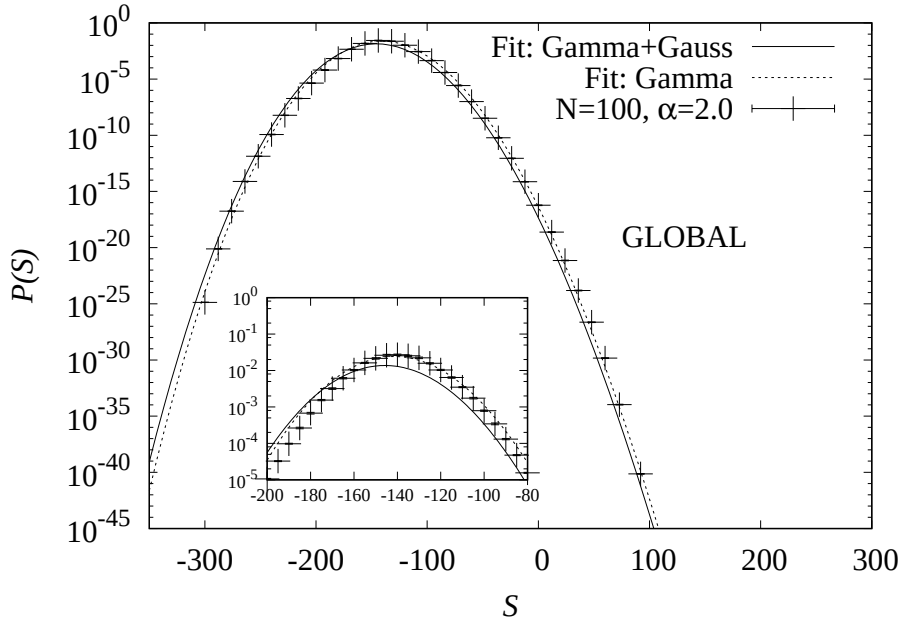

Figure 6: The distribution for global alignment scores as obtained for $N = 100$ and $\alpha = 2.0$. The Gamma function does not match the data well over the whole distribution range. Neither does the Gamma function with Gaussian correction, which was shown to be an in improvement to the sole Gamma function for amino acid sequence alignments in a previous work (Fieth and Hartmann, 2016).

contributing factor. Overall, our results show that the Gamma distribution – with or without corrections – seems not to be a good choice to describe the data. Note that only because we were able to obtain the score distribution over many decades in probability allowed us to reach this conclusion.

Figure 7 shows the distributions obtained for different values of α. It also shows attempts to fit the Gamma distribution (without correction) to the data. The smaller the decay parameter α, the more plateau-like the distribution gets. For $\alpha = 0.5$ a fit of the Gamma distribution was not possible at all. It performed best for intermediate values ($\chi^2 = 292$ for $\alpha = 2.0$) and worse again for less correlated sequences ($\chi^2 = 8781$ for iid).

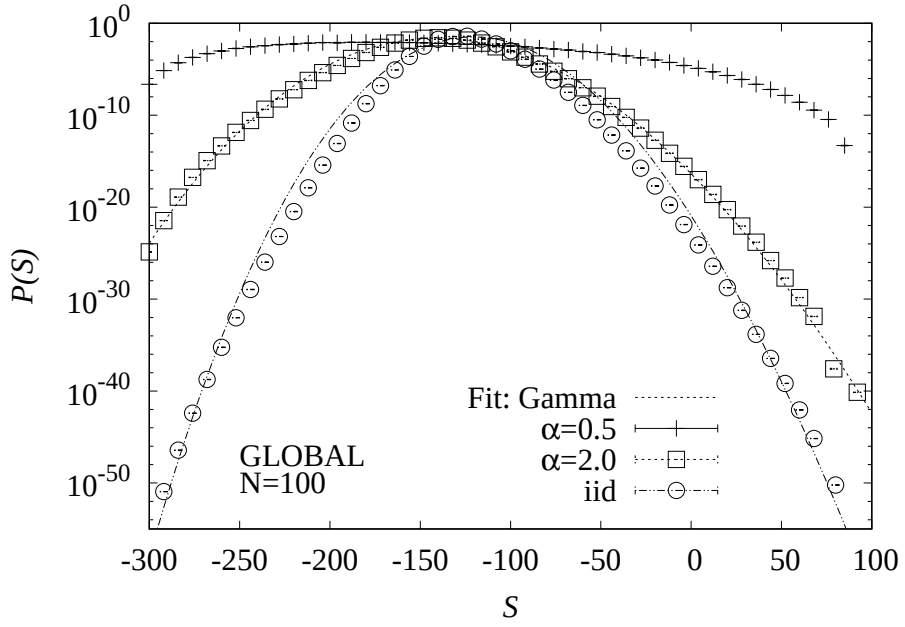

Figure 7: Score distribution for global alignment of sequences of length $N = 100$ and varying α. With increasing correlation, plateaus emerge in the score distributions. The Gamma function performs increasingly worse.

5 Conclusion

With the large-deviation approach we could obtain score distributions of local and global alignment of correlated sequences down to probabilities as low as $p(S) \approx 10^{-55}$. As in previous studies it could be shown that the Gumbel distribution is not a good estimator for local alignments, but can be improved on by a heuristically found Gaussian correction. The strength of this correction depends on the sequence length N, but not very strongly on the decay parameter α. The correlation has a stronger influence on the parameter λ. The fits perform worse for sequences with higher correlation, i.e. lower decay parameter α.

Distributions for global alignment showed that the previously heuristically found Gamma distribution with and without correlation is not a good model of the data. Further work would be necessary to find an alternative distribution to describe the whole numerically found distribution. This is especially true for sequences with low decay parameters that get increasingly plateau-like. As here, only by accessing the distributions over a large range in probability, e.g., by applying similar large-deviations approaches, a final conclusion about the nature of the distribution can be obtained. As long as this is not the case, the numerically obtained data can serve as a good source to obtain high-precision p-values.

6 Disclosure Statement

No competing financial interests exist.

7 Acknowledgements

P.F. acknowledges financial support from the German Science Foundation (DFG) within the Graduiertenkolleg GRK 1885. The simulations were performed at the HPC Cluster CARL, located at the University of Oldenburg (Germany) and funded by the DFG through its Major Research Instrumentation Programme (INST 184/157-1 FUGG) and the Ministry of Science and Culture (MWK) of the Lower Saxony State.

References

Altschul, S. F., and Gish, W., 1996. Local alignment statistics. *In Methods in Enzymology*, Russell F. Doolittle, ed., volume 266 of *Computer Methods for Macromolecular Sequence Analysis*, 460–480. Academic Press.

Berman, H. M., Westbrook, J., Feng, Z., et al., 2000. The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242.

Durbin, R., Eddy, S., Krogh, A., et al., 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Fieth, P., and Hartmann, A. K., 2016. Score distributions of gapped multiple sequence alignments down to the low-probability tail. *Phys. Rev. E* 94, 022127.

Geyer, C. J., 1991. Markov Chain Monte Carlo maximum likelihood. *In 23rd Symposium on the Interface between Computing Science and Statistics*. Interface Foundation of North America.

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

Hukushima, K., and Nemoto, K., 1996. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* 65, 1604–1608.

Karlin, S., Dembo, A., and Kawabata, T., 1990. Statistical composition of high-scoring segments from molecular sequences. *Ann. Statist.* 18, 571–581.

Li, W., and Kaneko, K., 1992. Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding dna sequence. *Europhys. Lett.* 17, 655 – 660.

Marinari, E., and Parisi, G., 1992. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* 19, 451.

Messer, P. W., and Arndt, P. F., 2006. CorGen–measuring and generating long-range correlations for DNA sequence analysis. *Nucleic Acids Res.* 34, W692–695.

Messer, P. W., Bundschuh, R., Vingron, M., et al., 2007. Effects of long-range correlations in DNA on sequence alignment score statistics. *J. Comput. Biol.* 14, 655–668.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., et al., 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092.

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

Newberg, L. A., 2008. Significance of gapped sequence alignments. *J. Comput. Biol.* 15, 1187–1194.

Pang, H., Tang, J., Chen, S.-S., et al., 2005. Statistical distributions of optimal global alignment scores of random protein sequences. *BMC Bioinformatics* 6, 257.

Smith, T. F., and Waterman, M. S., 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169.

Wolfsheimer, S., Burghardt, B., and Hartmann, A. K., 2007. Local sequence alignments statistics: deviations from Gumbel statistics in the rare-event tail. *Algorithms Mol. Biol.* 2, 9.