Chapter 4

Genomic correlations and sequence alignment statistics

Long-range correlations in genomic base composition are a ubiquitous statistical feature among many eukaryotic genomes. As we have shown in the Chapter 3, the emergence of such correlations is a natural consequence of the elementary local mutation processes in genome evolution. In this chapter, we show that long-range correlations substantially influence the statistics of sequence alignment scores if they are incorporated in our null model of DNA sequences. Using a Gaussian approximation to model the correlated score landscape, we calculate the corrections to the scale parameter λ of the extreme value distribution of alignment scores. Our approximate analytic results are supported by a detailed numerical study. We find both, mean and exponential tail of the score distribution for long-range correlated sequences to be substantially shifted compared to random sequences with independent nucleotides. The significance of measured alignment scores changes upon incorporation of the correlations in the null model. We discuss the magnitude of this effect in a biological context at the end of this chapter.

4.1 Sequence alignment and significance assessment

The goal of DNA sequence alignment is to assign to a given pair of genomic sequences $\vec{a} = (a_1, \ldots, a_N)$ and $\vec{b} = (b_1, \ldots, b_M)$ a measure of their similarity. The simplest version of sequence alignment is *gapless* alignment. A local gapless alignment $\mathcal{A}$ of the two sequences consists of a substring $(a_{i-l+1}, \ldots, a_i)$ of length l of sequence $\vec{a}$ and a substring $(b_{j-l+1}, \ldots, b_j)$ of sequence $\vec{b}$ of the same length. Each such alignment is assigned a score $S_{\mathcal{A}} = \sum_{k=0}^{l-1} s(a_{i-k}, b_{j-k})$, where $s(a,b)$ is some given scoring matrix measuring the mutual degree of similarity of the different letters of the alphabet. For DNA sequence comparison, one often uses the simple match-mismatch matrix [148]

$$s(a,b) = \left\{ \begin{array}{lcl} 1 & : & a = b \\ -\mu & : & a \neq b \end{array} \right. . \tag{4.1}$$

The computational task is to find the alignment $\mathcal{A}$, which gives the highest score

$$S \equiv \max S_{\mathcal{A}}. \tag{4.2}$$

For the purpose of detecting weak sequence similarities, alignment algorithms can also take into account insertions and deletions in either one of the two sequences during biological evolution [148]. For such *gapped* alignments, each gap contributes a gap cost γ to the total score of the alignment. Using affine gap costs, one additionally distinguishes between gap initiation cost γ_i and gap extension cost γ_e.

DNA sequence alignment is generally used to identify homology between sequences, i. e. common evolutionary origin from an ancestral state. Since an alignment score S is assigned to any pair of sequences, also to biologically completely unrelated ones, it is helpful to know the distribution of S in an appropriate null model in order to be able to distinguish true evolutionary relationship from random similarities. The knowledge of this distribution gives the possibility to assign p-values to alignment results. They specify the probability that a high score could have arisen by chance.

As already discussed in the introduction, a frequently used null model for that purpose is the iid model. For ungapped alignment of long sequences $(M, N \gg 1)$, the distribution of S for the iid model has been worked out rigorously [73, 76, 74]; it is a Gumbel or extreme value distribution, with its probability density function given by

$$\mathrm{pdf}(S) = KMN\lambda \exp\left(-\lambda S - KMNe^{-\lambda S}\right). \tag{4.3}$$

The distribution is characterized by the two parameters λ and K. In the iid case, the scale parameter λ is the unique positive solution of the equation

$$\langle \exp(\lambda s) \rangle = \sum_{a,b} \rho_a \rho_b \exp\left[\lambda s(a,b)\right] = 1, \tag{4.4}$$

where ρ_x is the frequency of nucleotide $x \in \{A, C, G, T\}$ in the model. The other parameter K then determines the mean of the distribution.

For gapped alignment, no rigorous theory for the distribution of S exists, so far. However, numerical evidence strongly suggests that the distribution is still of Gumbel form [149, 168, 4, 118]. Using this empirical applicability, it has been shown in [27, 28, 62] that λ for local gapped alignment in the iid model can be derived solely from studying the much simpler global alignment, where one is interested in the path with the highest score $h \equiv \max h_{\mathcal{A}}$, connecting the beginning (a_1, b_1) to the end (a_N, b_N) of a given pair of sequences $\vec{a}$ and $\vec{b}$ (we set $M = N$, from now on). If we denote the average over all possible pairs of random sequences $\vec{a}$ and $\vec{b}$ of length N by brackets $\langle\cdot\rangle$, we can define a *generating function*

$$Z_N(\lambda) \equiv \langle \exp(\lambda h) \rangle. \tag{4.5}$$

The *central conjecture* in [27] then states that λ is determined by the solution of

$$\lim_{N\to\infty} \frac{1}{N} \log Z_N(\lambda) = 0. \tag{4.6}$$

Following the results of [128, 34], this allows for a very efficient computation of λ for gapped alignment in the iid model.

The iid model is the simplest feasible DNA background model at hand. It allows to incorporate length and average nucleotide composition of a sequence, but lacks any specific structure concerning the arrangement of nucleotides along the DNA sequence. The use of an iid null model for DNA is conclusively justified if nucleotides evolve independent of each other in the sequence, and insertions are comprised of randomly drawn nucleotides. However, we have shown in Chapter 2 that the latter is not the case for genomic evolution. In human, for example, the majority of recent short DNA insertions were found to result from tandem duplication events. It has further been shown in Chapter 3 that tandem duplication insertions lead to long-range correlations in genomic base composition. These findings provide a likely explanation for the widespread presence of such correlations among the genomes of many eukaryotic species, but they also raise the question to what degree iid sequences are still a suitable null model for genomic DNA. It is the aim of the following analysis to investigate the effects on significance estimation of sequence alignment scores that result when replacing the iid model by a more accurate DNA null model, which incorporates long-range correlations in sequence composition.

4.2 The Gaussian approximation

In this section, we derive approximate analytical results for the parameter λ of the score distribution one obtains for alignment of random sequences with long-range correlations in nucleotide composition. We restrict ourselves to gapless alignment, as we expect qualitatively similar results for the gapped case. This will also be confirmed by the numerical data we present in Section 4.4. For simplicity, we furthermore assume a uniform distribution of the four nucleotides; a generalization to sequences with biased composition is straightforward.

The approach employed in the following is based on the assumption that for local gapless alignment of correlated sequences the distribution of the maximal scores obeys Gumbel form, and λ is still determined by Eq. (4.6). The score of the global alignment is given by the sum over all elementary scores $s_i = s(a_i, b_i)$ along the diagonal of the alignment-lattice. Two exemplary realizations of an alignment score lattice are shown in Fig. 4.1. Defining $\vec{s} = (s_1, \ldots, s_N)$, we have

$$h = \sum_{i=1}^{N} s_i = \vec{1}^{\mathrm{t}} \vec{s}. \tag{4.7}$$

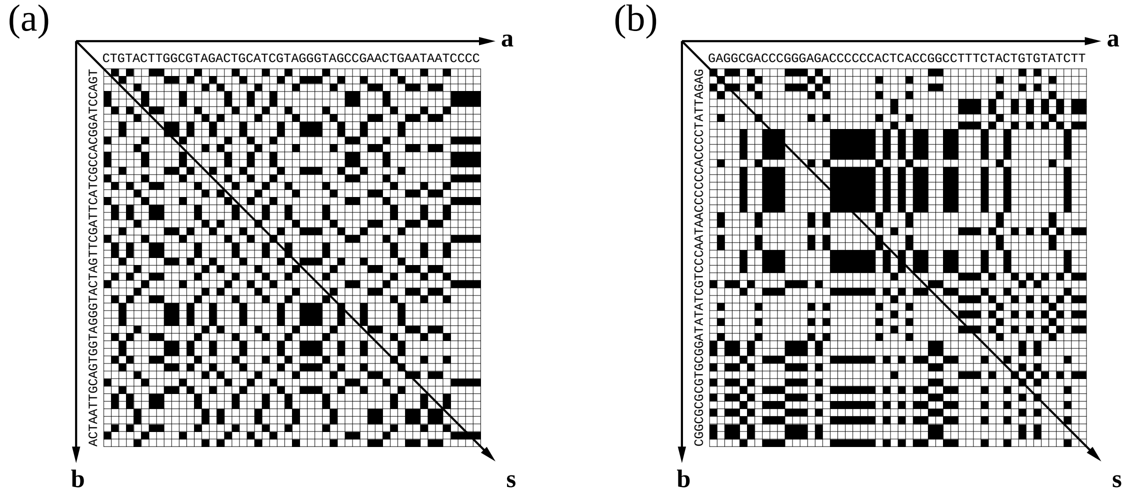
(a)
a
CTGTACTTGGCGTAGACTGCATCGTAGGGTAGCCGAACTGAATAATCCCC
ACTAATTGCAGTGGTAGGGTACTAGTTCGATTCATCGCCACGGATCCAGT
b
s
(b)
a
GAGGCGACCCGGGAGACCCCCCACTCACCGGCCTTTCTACTGTGTATCTT
CGGCGCGCGTGCGGATATATCGTCCCAATAACCCCCCACCCCTATTAGAG
b
s

Figure 4.1: Two realizations of an alignment score lattice: (a) The sequences $\vec{a}$ and $\vec{b}$ are drawn from the iid model. (b) Both sequences are random sequences featuring long-range correlations in their base composition with decay exponent $\alpha = 0.5$. They are generated by a duplication-mutation dynamics, as described in Section 4.3. Cells corresponding to matching nucleotides in $\vec{a}$ and $\vec{b}$ are colored black, white cells denote mismatches. The score of an ungapped global alignment is the sum over all elements of the diagonal vector $\vec{s} = (s_1, \ldots, s_N)$ of the score lattice. Comparison of both figures reveals qualitative differences between the two null models. The alignment lattice of long-range correlated sequences shows systematically larger black and white blocks representing exactly matching or mismatching substrings of the two sequences, compared to the iid model.

The ensemble average of Eq. (4.5) over all realizations of the two sequences $\vec{a}$ and $\vec{b}$ can therefore be expressed in terms of an average over all score vectors $\vec{s}$. While the probability of a score vector factorizes in the iid model, $P(\vec{s}) = \prod_i P(s_i)$, this is no longer the case for correlated sequences. However, approximate values for the probabilities $P(\vec{s})$ in the correlated case can still be derived by a Gaussian approximation. The idea of this approach is to replace the discrete variables s_i by continuous Gaussian variables. More precisely, an individual discrete score $s_i = \{1, -\mu\}$ at position i along the diagonal of the alignment lattice will now be allowed to take continuous values, distributed according to a normal distribution

$$\mathrm{pdf}(s_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(s_i - \langle s \rangle)^2}{2\sigma^2}. \tag{4.8}$$

Mean and variance are chosen in accordance with the original discrete score distribution, i.e., $\langle s \rangle = 1/4 - 3\mu/4$, and $\sigma^2 = 3(1+\mu)^2/16$.

The probability $P(\vec{s})$ of a score vector $\vec{s}$ is then determined by an N-dimensional Gaussian distribution

$$P(\vec{s}) = [(2\pi)^N \det \boldsymbol{\sigma}]^{-1/2} \exp[-\frac{1}{2}(\vec{s} - \langle \vec{s} \rangle)^{\mathrm{t}} \boldsymbol{\sigma}^{-1} (\vec{s} - \langle \vec{s} \rangle)], \tag{4.9}$$

with $\langle \vec{s} \rangle = (\langle s \rangle, \ldots, \langle s \rangle)$ and the covariance matrix $\boldsymbol{\sigma}$, defined by

$$\boldsymbol{\sigma}_{ij} = \langle s(i)s(j) \rangle - \langle s(i) \rangle \langle s(j) \rangle. \tag{4.10}$$

The diagonal elements of $\boldsymbol{\sigma}$ are given by the variance of an individual score, $\boldsymbol{\sigma}_{ii} = \sigma^2$. The non-diagonal elements $\boldsymbol{\sigma}_{i\neq j}$ can be expressed in terms of the correlation function $C(r)$ of the sequences $\vec{a}$ and $\vec{b}$, where $C(r)$ is defined by Eq. (1.5). We have

$$\boldsymbol{\sigma}_{ij} = \frac{1}{3}(1+\mu)^2 C^2(|i-j|). \tag{4.11}$$

In this expression the correlation function $C(r)$ is squared because (4.11) describes the correlations of the similarity scores which arise from a comparison of two sequences. The non-diagonal elements vanish for iid sequences.

Using distribution (4.9), the calculation of the generating function (4.5) amounts to the evaluation of an N-dimensional Gaussian integral, which can be solved explicitly,

$$\begin{aligned} Z_N(\lambda) &= \int d\vec{s}\, P(\vec{s}) \exp(\lambda \vec{1}^{\mathrm{t}} \vec{s}) \\ &= [(2\pi)^N \det \boldsymbol{\sigma}]^{-1/2} \int d\vec{s}\, e^{-\frac{1}{2}(\vec{s}-\langle\vec{s}\rangle)^{\mathrm{t}} \boldsymbol{\sigma}^{-1} (\vec{s}-\langle\vec{s}\rangle) + \lambda \vec{1}^{\mathrm{t}} \vec{s}} \\ &= \exp(\lambda \vec{1}^{\mathrm{t}} \langle\vec{s}\rangle + \frac{1}{2}\lambda^2 \vec{1}^{\mathrm{t}} \boldsymbol{\sigma} \vec{1}). \end{aligned} \tag{4.12}$$

The central conjecture (4.6) then implies

$$0 = \lim_{N\to\infty} \frac{1}{N} (\lambda \vec{1}^{\mathrm{t}} \langle\vec{s}\rangle + \frac{1}{2}\lambda^2 \vec{1}^{\mathrm{t}} \boldsymbol{\sigma} \vec{1}). \tag{4.13}$$

Notice that this expression coincides with the result obtained by applying the central conjecture to the Taylor series approximation of the generating function (4.5) up to second order. Using Eq. (4.11) yields

$$\lambda = \frac{-2\langle s\rangle}{\sigma^2 + \frac{2}{3}(1+\mu)^2 \lim_{N\to\infty} \sum_{i=1}^{N} C^2(i)}. \tag{4.14}$$

The first term σ^2 in the denominator is related to the individual fluctuations of a single score element, irrespective of correlations along the sequences. The second term vanishes for iid sequences and determines the corrections to λ due to correlations.

In case of long-range correlations, i.e., $C(r) = cr^{-\alpha}$, and assuming $\alpha > 1/2$, we obtain

$$\lambda = \frac{-2\langle s\rangle}{\sigma^2 + \frac{2}{3}(1+\mu)^2 c^2 \zeta(2\alpha)}, \tag{4.15}$$

where $\zeta(x)$ is the Riemann zeta function. Consequently, the Gaussian approximation predicts deviations in λ for the alignment of long-range correlated sequences compared to iid sequences. A detailed numerical analysis of this analytic result will be performed in Section 4.3. Notice that for $\alpha \leq 1/2$ the sum $\sum_{i=1}^{\infty} C^2(i)$ diverges, resulting in

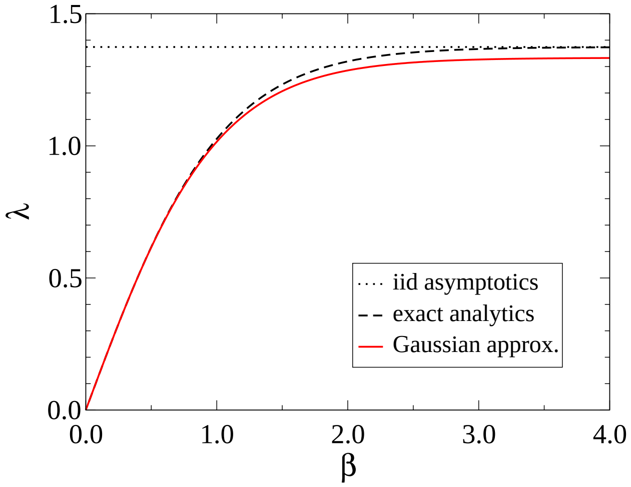

Figure 4.2: λ for sequences with short-range correlations generated by a Markov process. The dashed line is the exact result [76] for the Markov process defined in (4.16), using $\mu = 3$. The solid line is the corresponding result of the Gaussian approximation, as derived in Eq. (4.17). Solving Eq. (4.4) yields the iid asymptotics $\lambda \approx 1.374$.

$\lambda = 0$. This might indicate a transition from local to global alignment in the Gaussian approximation, which will be discussed at the end of Section 4.3.

As a first evaluation of the Gaussian approximation, we investigate its predictions for sequences $\vec{a} = (a_1, \ldots, a_N)$ generated by a Markov process. We consider a first order process with four different states $A_i \in \{A, C, T, G\}$. Starting with a random nucleotide a_1, the transition probabilities are defined by

$$P(a_{i+1}|a_i) = \left\{ \begin{array}{ccc} p & : & a_{i+1} = a_i \\ \frac{1}{3}(1-p) & : & a_{i+1} \neq a_i \end{array} \right. . \tag{4.16}$$

This process generates short-range correlations in the sequences of the form $C(r) = c\exp(-\beta r)$ with $\beta = -\log(4p/3 - 1/3)$ and $c = 3/4$. For this case, the Gaussian approximation (4.14) yields

$$\lambda = \frac{-2\langle s\rangle}{\sigma^2 + \frac{2}{3}(1+\mu)^2 c^2/(\exp(2\beta) - 1)}. \tag{4.17}$$

This can be compared with the exact analytic result for λ obtained by equating the largest eigenvalue of a modified λ-dependent transition matrix of the underlying Markov process to one [76]. As is shown in Fig. 4.2, the Gaussian approximation (4.17) fits well to the exact results; deviations for large β vanish for decreasing β. Notice that the limit $\beta \to \infty$ corresponds to $p \to 1/4$, describing the asymptotics of an uncorrelated iid sequence. The deviations of the Gaussian approximation for this regime result from the fact that the third and all higher cumulants of the distribution (4.8) vanish, which they do not for the discrete distribution.

4.3 Numerical results

Generating long-range correlated random sequences Numerical evaluation of the results obtained in the previous section hinges on the knowledge of the score

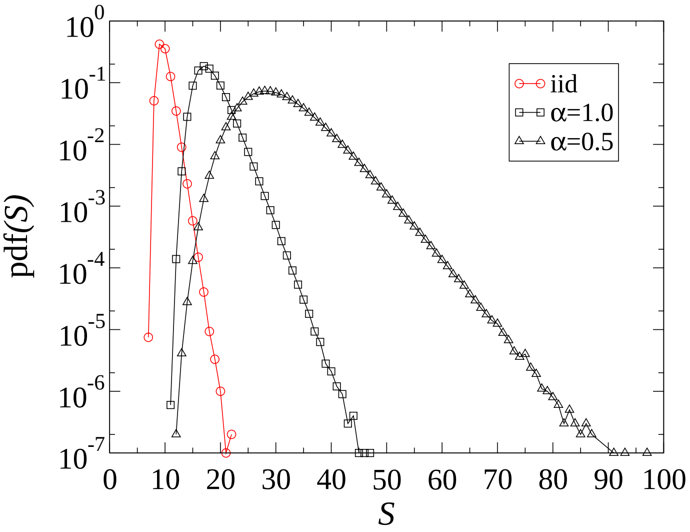

Figure 4.3: Numerically measured shape of the alignment score distribution $\mathrm{pdf}(\mathrm{S})$ for $N = M = 10^3$ using three different null models: iid sequences, long-range correlated sequences with $\alpha = 1.0$, and such with $\alpha = 0.5$. The distributions were obtained by aligning 10^7 pairs of sequences randomly drawn from the particular null model ensemble. As is the case for the iid ensemble, the tail of the score distributions for long-range correlated sequences features the Gumbel-typical exponential decay characterized by a decay exponent λ. However, the mean of $\mathrm{pdf}(\mathrm{S})$ is systematically shifted towards larger scores for increasing correlation strength, i.e. smaller values of α, compared to the iid model. Moreover, long-range correlations decrease the decay exponent λ and therefore lead to a slower decay of the exponential tail of the alignment score distribution.

distribution $\mathrm{pdf}(S)$ for local gapless alignment of pairs of long-range correlated random sequences. Based on our results derived in Chapter 3, we use a simple single-site duplication-mutation algorithm to generate long-range correlated sequences. We start with a sequence of one random nucleotide a_1. The dynamics is defined by a single-site duplication process occurring at rate $\gamma_1^+ = 1.0$, and mutations specified by a rate matrix (3.58) with tunable GC-content g and rate-parameter μ. No other processes are acting on the sequence. This dynamics generates random sequences with $C(r)$ determined by Eq. (3.66). Asymptotically we have long-range correlations $C(r) \propto r^{-\alpha}$ for large r. The decay exponent is determined by $\alpha = 2\mu$, as derived in Eq. (3.65) using $\gamma_1^+ = 1$. By varying the mutation parameter μ we can hence tune α to any desired positive value. Due to the algorithm's fast runtime of $O(N)$ we can efficiently generate the large ensembles of long-range correlated sequences needed for our analysis. For the alignment, we use the standard Smith-Waterman dynamic programming algorithm [148] with scoring matrix (4.1) and $\mu = 3$.

The Gumbel distribution of alignment scores Our solution of the Gaussian model is based on the assumption that the alignment score distribution $\mathrm{pdf}(S)$ is of Gumbel form for long-range correlated sequences. Consequently, our first numerical analysis aims at a verification of this conjecture. In Fig. 4.3 we show the measured score distributions for two different long-range correlated sequence ensembles with

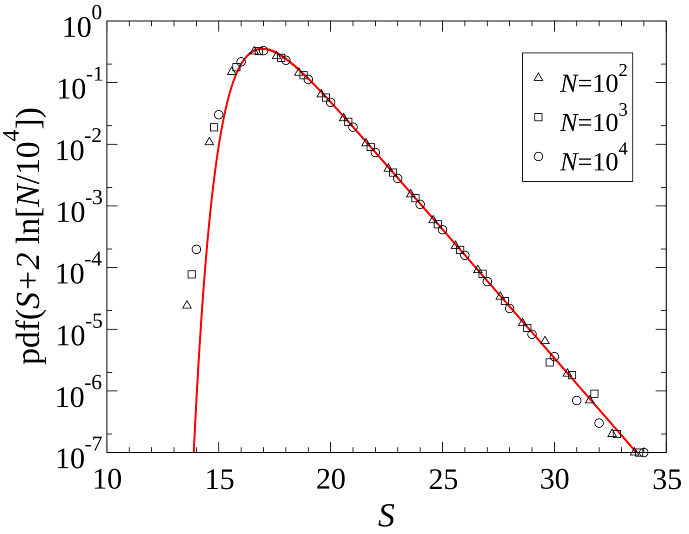

Figure 4.4: Convergence of the distribution $\mathrm{pdf}(S)$ for long-range correlated sequences with $\alpha = 2.0$ to a Gumbel form. The solid line is a Gumbel distribution, as specified in Eq. (4.3) with $N = M = 10^4$ and fitted parameters $\lambda = 0.9614$ and $K = 0.119$. λ was obtained by fitting a linear function to $\log[\mathrm{pdf(S)}]$ for $21 < S < 31$, K has then been estimated by fitting the data to Eq. (4.3) in the same interval. In order to be able to compare the shape of $\mathrm{pdf}(S)$ for different N, the distributions have to be rescaled by a transformation $\mathrm{pdf}(S) \to \mathrm{pdf}(S + 2\ln[N/N_0])$ with reference length $N_0 = 10^4$.

correlation parameters $\alpha = 1.0$ and $\alpha = 0.5$, compared to the Gumbel distribution of the iid model. Whereas the tail of the score distributions for long-range correlated sequences still features the Gumbel-typical exponential decay, its decay parameter λ systematically decreases with increasing correlation strength, i. e. smaller values of α. In addition, the mean of the distribution is shifted towards larger scores. For large N, the shape of $\mathrm{pdf}(S)$ asymptotically approaches a Gumbel form for the correlated ensembles. This can be seen in Fig. 4.4, where we exemplarily show the measured score distribution for long-range correlated sequences with $\alpha = 2.0$ and different sequence lengths N. As is the case for the iid model, finite-size corrections come into play for small N [4, 3, 176]. These deviations primarily show up in the small S regime. The more relevant large S regime converges fast for increasing N.

Now, that we have verified the shape of the score distribution to be of Gumbel form, we can test the accuracy of the analytic predictions for λ derived by the Gaussian approximation. Here we restrict ourselves to the discussion of the regime $\alpha > 1/2$, where the Gaussian approximation predicts finite values of λ. The regime $\alpha \leq 1/2$ will be investigated below. We compare our numerical data to Eq.(4.14), using correlations of the form (3.66). Results are shown in Fig. 4.5 (a). The Gaussian approximation captures the qualitative behavior of the numerical data. Again, the right hand side of the plot reveals the deviations of the Gaussian approximation concerning its iid asymptotics given by $\alpha \to \infty$. With increasing correlation strength, i. e. smaller values of α, λ decreases, confirming that long-range correlations systematically raise the probability of measuring high alignment scores.

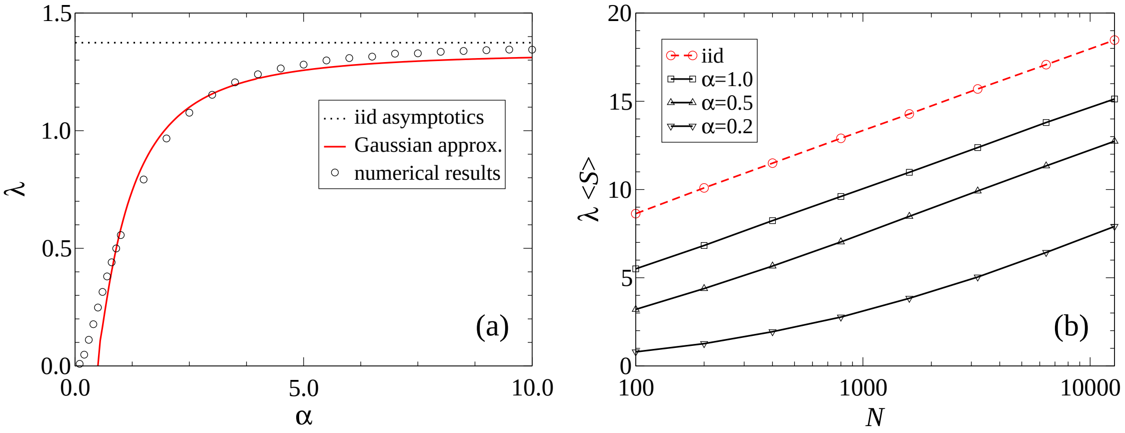

Figure 4.5: (a) λ for a null model with long-range correlated sequences in dependence of the correlation exponent α. The solid line is the analytic result of the Gaussian approximation one obtains by estimating Eq. (4.14) using the correlations (3.66) of our simulated sequences. Numerically measured values of λ for different correlation parameters α are denoted by symbols. For our simulation, we use sequences of length $N = 10^3$ and average over ensembles of 10^8 pairs of sequences. (b) Mean of the score distribution $\mathrm{pdf}(S)$ for different exponents α against increasing sequence length N. Measured values for $\langle S \rangle$ were obtained by averaging over ensembles of 10^4 pairs of sequences. K can be inferred from $K = 1/N^2 \exp\left[\lambda\langle S\rangle - \Gamma\right]$. The deviations from the form $\lambda\langle S\rangle = \Gamma + \log(K) + 2\log(N)$ for the strong correlations $\alpha = 0.2$ are due to the fact that $\mathrm{pdf}(S)$ deviates from a Gumbel distribution for small N in the strong correlation regime $\alpha \leq 0.5$.

So far, our investigations of the alignment score distribution for long-range correlated sequences have focused on the exponential tail of $\mathrm{pdf}(S)$. We now turn to the second parameter K. For that purpose, we recall that the mean of a Gumbel distribution (4.3) is determined by

$$\langle S \rangle = \frac{\Gamma + \log(KN^2)}{\lambda}, \tag{4.18}$$

where $\Gamma \approx 0.5772$ is the Euler-Mascheroni constant. Thus, knowing λ, the parameter K can easily be calculated by measuring the mean $\langle S \rangle$ of the score distribution. As shown in Table 4.1 and Fig. 4.5 (b), K is significantly affected by the presence of long-range correlations in the sequences to be aligned; it decreases with increasing correlation-strength. However, the mean of the distribution is, as expected, shifted to larger values of S for decreasing values of α because K contributes only logarithmically in Eq. (4.18) and the change in $\langle S \rangle$ is dominated by the decrease of λ. Fig. 4.5 (b) again reveals the finite-size deviations of the numerically measured score distribution $\mathrm{pdf}(S)$ from a Gumbel form (4.3) in the strong correlation regime $\alpha \leq 1/2$.

The score distribution for $\alpha \leq 1/2$ In the regime $\alpha > 1/2$, the score distribution is of Gumbel form and the Gaussian approximation suitably fits the numerical values of λ. For values of $\alpha \leq 1/2$, the Gaussian approximation yields $\lambda = 0$, which might indicate a transition from local to global alignment. For simulated sequences of finite length, on the other hand, one still measures finite values of λ, as shown in Fig. 4.5 (a).

α	λ	$\langle S\rangle$	K
(iid)	1.374	9.71	3.50×10^{-1}
4.0	1.240	10.61	2.90×10^{-1}
2.0	0.967	12.65	1.15×10^{-1}
1.0	0.556	18.07	1.30×10^{-2}
0.5	0.248	51.30	1.15×10^{-3}
0.2	0.048	165.08	9.47×10^{-6}

Table 4.1: Dependence of $\langle S\rangle$ and K on the exponent α. We use simulated sequences of length $N = 10^3$ and average over ensembles of 10^8 pairs of sequences for each value of α to obtain numerical values of λ and $\langle S\rangle$. The values of K have been calculated using Eq. (4.18).

The numerical investigation of this regime is complicated by a distinct finite size effect: according to the results derived in Section 3.4, an individual alignment of two finite sequences will have a systematic bias of $\langle s\rangle$ towards either $\langle s\rangle = 1$, or $\langle s\rangle = -\mu$, depending on whether by chance the two initial random letters a_1 and b_1 of our sequence generation algorithm were equal for the two sequences to be aligned, or not. This effect causes strong deviations of pdf(S) from a Gumbel form for small S. The tail of the distribution is still exponential for finite sequences and therefore allows for a measurement of λ. It is dominated by those realizations of the ensemble, where both sequences started with the same letter as they lead to systematically higher values of $\langle s\rangle$ and therefore also higher scores S.

As can be seen in Fig. 4.5 (a), λ approaches zero for finite sequences not until the "infinite" correlation strength limit $\alpha \to 0$. Further analysis is needed to decide on whether there actually is a transition to global alignment for a particular $\alpha > 0$ in the limit $N \to \infty$, or not. If there is, then the rate of convergence for $\lambda \to 0$ is at most logarithmically. However, for practical applications this transition is irrelevant. Finite sequences always have a positive λ, also in the regime $\alpha \leq 1/2$. For these particular choices of parameters, λ needs to be measured numerically.

4.4 Consequences for genomic alignments

It has been shown that long-range correlations in base composition increase the probability of measuring high scores for pairwise sequence alignment. In a biological context, this raises the question whether the effect causes a significant change of the p-values for DNA alignment? In order to address this issue, we investigate the deviations of the score distribution for correlation parameters of genomic magnitude compared to iid sequences. For this purpose, we need to generate ensembles of long-range correlated random sequences with correlation parameters of genomic scale. This can conveniently be achieved by our web server CorGen, described in Section 3.7.

To investigate the magnitude of p-value changes, we consider as an example the measured correlation function of human chromosome 22. As shown in Fig. 4.6 (a), human chromosome 22 shows clear long-range correlations in its base composition, and we

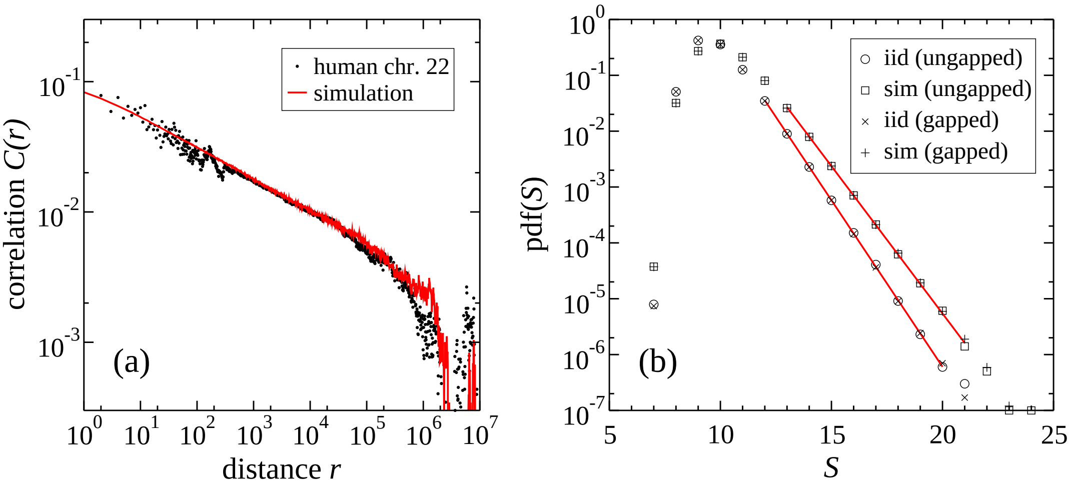

Figure 4.6: (a) Long-range correlations in the base composition of human chromosome 22 (symbols) and demonstration of our capability to produce random sequences with comparable correlation parameters (red line). (b) The score distribution for ungapped and gapped alignment of simulated sequences with correlations comparable to those of human chromosome 22. The straight lines are the fits to the exponential tails of the score distributions, obtained by fitting a linear function to $\log[\mathrm{pdf(S)}]$ in the depicted intervals.

can accurately generate long-range correlated random sequences with comparable exponent $\alpha \approx 0.232$ and amplitude obtained from fitting a power-law to the measured $C(r)$ in the interval $10^2 < r < 10^4$.

We perform ungapped, as well as gapped alignment with affine gap costs for 10^7 pairs of random sequences of length $N = 10^3$ from the above specified ensemble. Alignment parameters are chosen in accordance with the NCBI default values $\mu = 3$, gap initiation cost $\gamma_i = 5$, and gap extension cost $\gamma_e = 2$ [120]. In Fig. 4.6 (b) we show the measured score distributions for the simulated chr. 22 sequences compared to iid sequences. The resulting parameters λ and $\langle S \rangle$ are presented in Table 4.2.

It turns out that the difference in the score distributions between ungapped and gapped alignment is negligible for the parameters used. The deviations in λ between the iid ensemble and the simulated human chromosome 22 sequences are approximately 15% in both cases, and the mean of the score distributions for the correlated sequences is significantly larger. In combination, both effects substantially change the p-values of high scores compared to the iid model, as can be seen in Table 4.2. The p-value of a specific score S' is thereby defined by the integral $P(S \geq S') = \int_{S'}^{\infty} \mathrm{pdf}(S) dS$. For an exemplary score $S' = 18$, this p-value will be increased by almost one order of magnitude if one incorporates the genomic correlations into the null model.

Long-range correlations are a widespread statistical feature of eukaryotic DNA. In this chapter, it has been shown that incorporation of this feature into the null model substantially influences the score statistics of sequence alignment. While the p-values of the scores are systematically increased, the ranking of hits will not be signifi-

ensemble	λ	$\langle S \rangle$	$P(S \geq 18)$
iid (ungapped)	1.374	9.714	3.3×10^{-6}
sim. chr. 22 (ungapped)	1.191	10.164	2.8×10^{-5}
iid (gapped)	1.373	9.714	3.2×10^{-6}
sim. chr. 22 (gapped)	1.215	10.163	2.7×10^{-5}

Table 4.2: Fitted parameters λ and $\langle S \rangle$ for the iid ensemble and simulated human chr. 22 sequences of length $N = 10^3$. In the last column, exemplary p-values of $S' = 18$ are shown.

cantly changed. The effect is therefore relevant whenever one is actually interested in p-values, e.g., when specifying a cutoff in order to distinguish true evolutionary relationship from random similarities.

One has to keep in mind that genomic DNA is a highly heterogeneous environment: it consists of genes, noncoding regions, repetitive elements etc., and all of these substructures may imprint their signature on the amount of correlations found in a particular genomic region [75]. Long-range correlations are by definition a feature on larger scales. Our findings are therefore naturally applicable to the alignment of larger genomic regions. This includes the identification of duplicated regions, or conserved syntenic segments between chromosomes of different species, which often extend over many kilobases up to several megabases. However, long-range correlations will also influence the statistics of search algorithms for short DNA motifs if the query sequences are large enough for long-range correlations to be measured.

Moreover, it will be interesting to analyze possible effects of long-range correlations on the statistics of other widely used sequence analysis tools, e. g. the prediction of transcription factor binding sites [157]. Further investigation is needed to assess the relevance of long-range correlations for other statistical predictions. Finally, more accurate null models of DNA sequences utilizing quantitative correlation features will help to reduce the often encountered high false-positive rate of bioinformatics tools.