JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 5, 2007
© Mary Ann Liebert, Inc.
Pp. 655–668
DOI: 10.1089/cmb.2007.R008

Effects of Long-Range Correlations in DNA on Sequence Alignment Score Statistics

PHILIPP W. MESSER,[1] RALF BUNDSCHUH,[2] MARTIN VINGRON,[1]
and PETER F. ARNDT[1]

ABSTRACT

Long-range correlations in genomic base composition are a ubiquitous statistical feature among many eukaryotic genomes. In this article, these correlations are shown to substantially influence the statistics of sequence alignment scores. 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 based on a simple algorithm to efficiently generate long-range correlated random sequences. 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 will therefore change upon incorporation of the correlations in the null model. We discuss the magnitude of this effect in a biological context.

Key words: extreme value distribution, long-range correlations, sequence alignment, statistical significance.

1. INTRODUCTION

Recent years have witnessed an impressive advance of bioinformatics sequence analysis tools, aiming at deeper insight to the functional organization and evolutionary dynamics of genomic DNA sequences. Popular examples include algorithms for genome annotation, homology detection between genomic regions of different organisms, or the prediction of transcription factor binding sites (Waterman, 1995; Durbin et al., 1998).

Bioinformatics methods frequently yield probabilistic statements. Usually the statistical significance of a computational prediction is characterized by a p-value, specifying the likelihood that this prediction could have arisen by chance. The calculation of p-values requires an appropriate null model of DNA, which reflects our assumptions about the "background" statistical features of the sequence under consideration. The challenging task is to decide on the set of statistical features a suitable null model should obey. Ideally, one incorporates those features into the null model which describe the background "noise" of the DNA sequence, but still allow to discern the specific signal the computational analysis tries to detect.

[1] Max Planck Institute for Molecular Genetics, Berlin, Germany.
[2] Department of Physics, Ohio State University, Columbus Ohio.

The simplest DNA background model is an *iid* model, given by a random sequence with letters drawn independently from an identical distribution (Durbin et al., 1998). The iid model can incorporate the length and the average composition of the sequences under consideration, but it lacks any specific structure concerning the arrangement of the nucleotides along the DNA. In particular, it is not capable of incorporating correlations in base composition along the sequences. Up to a certain degree, this additional complexity can be taken into account by an nth order Markov model, specifying the transition probabilities $P(a_{i+1}|a_{i-n+1},\dots,a_i)$ in a genomic sequence $\boldsymbol{a} = a_1,\dots,a_N$ (Durbin et al., 1998). Assuming the sequences to be generated by Markov processes already allows to incorporate a multitude of spatial statistical features into the model, like e.g. the preferential occurrence of DNA motifs, local peculiarities in genomic composition, or specific dinucleotide frequencies. In contrast to iid sequences, where all letters are uncorrelated, Markov processes lead to, so called, *short-range correlations* in the nucleotide composition (Peng et al., 1992). They are characterized by an exponential decay of the correlations between two different bases with increasing distance along the sequence.

A statistical measure of the correlations in genomic base composition is the autocorrelation function $C(r)$. It quantifies the deviations in the joint probability of finding equal bases at a distance of r basepairs along the DNA backbone compared to that in a random sequence of independent letters with the same nucleotide frequencies $p_{a\in\{A,C,T,G\}}$,

$$C(r) \equiv \sum_a \left[P(a_i = a_{i+r} = a) - p_a^2\right]. \tag{1}$$

We have $C(r) = 0$ $(r > 0)$ for iid sequences, while $C(r) \propto \exp(-\beta r)$ for short-range correlated sequences, e.g. those generated by Markov processes.

With the rapidly growing availability of whole-genome sequence data the correlations along genomic DNA can nowadays be studied systematically over a wide range of scales and organisms. A striking observation in this field was the finding of *long-range correlations* in the base composition of genomes more than a decade ago (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992). They are characterized by a power-law decay of the correlation function for large r,

$$C(r) \propto r^{-\alpha}, \tag{2}$$

and therefore decay much slower compared to short-range correlations. By now it is well established that long-range correlations in base composition appear in the genomes of most eukaryotic species (Arneodo et al., 1995; Bernaola-Galvan et al., 2002; Li and Holste, 2005) with two examples shown in Figure 1. Little is known about the origin of genomic long-range correlations, so far. However, their ubiquity among eukaryotic genomes points towards a universal mechanism. A likely dynamical scenario is that they are generated by the stochastic processes of molecular sequence evolution, as has been discussed in Li (1991) and Messer et al. (2005a, 2005b).

The widespread presence of long-range correlations in genomes raises the question if they need to be incorporated into an accurate null model of eukaryotic DNA and how that would change the p-value calculations. In this article, we address this question in the context of sequence alignment, which constitutes the most commonly used computational tool of molecular biology today (Altschul et al., 1990, 1997). We tackle the problem of calculating sequence alignment significance values for null models with long-range sequence composition correlations with both, analytical, as well as numerical methods. On the analytical side, we introduce a novel approach, the Gaussian approximation, which allows us to calculate the corrections to the scale parameter λ of the alignment score distribution for correlated sequences. Long-range correlated sequences cannot be generated by an nth order Markov process with finite n (Peng et al., 1992). The numerical approach therefore only recently has come within reach due to results derived in Messer et al. (2005a, 2005b), where we proposed a biologically motivated algorithm capable of efficiently generating long-range correlated sequences with arbitrary correlation parameters. As the main result of our analysis, it turns out that long-range correlations in the sequences lead to considerable deviations in the score statistics of sequence alignment.

After presenting a short review of sequence alignment in Section 2, we analytically treat the alignment of long-range correlated sequences in Section 3. A numerical evaluation of the approximative analytic results is presented in Section 4. In Section 5, we discuss the relevance of this effect for genomic sequence

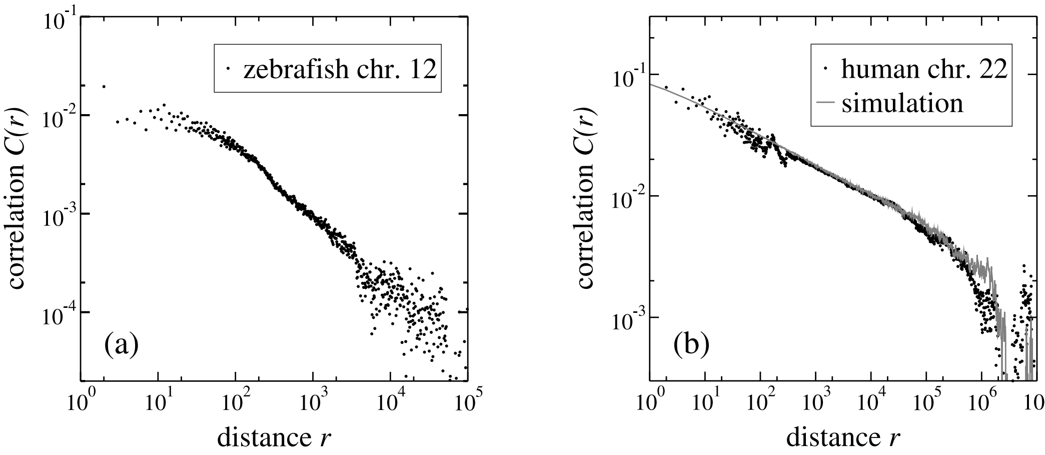

FIG. 1. Long-range correlations in the base composition of two eukaryotic chromosomes. In the double-logarithmic plots, power-law correlations $C(r) \propto r^{-\alpha}$ show up as straight lines with slope α. They extend over distances of several orders of magnitude. We demonstrate our capability of simulating long-range correlated sequences with similar amplitude and correlation exponent $\alpha \approx 0.232$, as measured in Human chromosome 22 **(b)**.

alignment by analyzing the magnitude of the corrections to the score significance values using correlation parameters, measured in eukaryotic genomes. The implications of our findings in a bioinformatics context are discussed at the end of this article.

2. SEQUENCE ALIGNMENT AND SIGNIFICANCE ASSESSMENT

The goal of DNA sequence alignment is to assign to a given pair of genomic sequences $\boldsymbol{a} = a_1, \ldots, a_N$ and $\boldsymbol{b} = b_1, \ldots, b_M$ 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 $a_{i-l+1} \cdots a_i$ of length l of sequence $\boldsymbol{a}$ and a substring $b_{j-l+1} \cdots b_j$ of sequence $\boldsymbol{b}$ 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 (Smith and Waterman, 1981):

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

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

$$S \equiv \max S_{\mathcal{A}}. \tag{4}$$

For the purpose of detecting weak sequence homologies, alignment algorithms can also take into account insertions and deletions in either one of the two sequences during biological evolution (Smith and Waterman, 1981). For such *gapped* alignments, each gap contributes a (negative) gap cost γ to the total score of the alignment. Using affine gap costs, one additionally distinguishes between the gap initiation cost γ_i and the gap extension cost γ_e.

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. 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 in order to be able to distinguish true evolutionary relationship from random similarities. As already mentioned 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 (Karlin and Altschul, 1990, 1993; Karlin and Dembo, 1992); it is a Gumbel or extreme value distribution, with its probability density function given by

$$\mathrm{pdf}(S) = KMN\lambda \exp\left(-\lambda S - KMNe^{-\lambda S}\right). \tag{5}$$

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

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

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 (Smith et al., 1985; Waterman and Vingron, 1994; Altschul and Gish, 1996; Mott, 1999). Using this empirical applicability, it has been shown in Bundschuh (2000, 2002) and Grossmann and Yakir (2004) 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 $\boldsymbol{a}$ and $\boldsymbol{b}$ (we set $M = N$, from now on). One defines a *generating function*

$$Z_N(\lambda) \equiv \langle \exp(\lambda h)\rangle, \tag{7}$$

where the brackets $\langle\cdot\rangle$ denote an average over all possible pairs of random sequences $\boldsymbol{a}$ and $\boldsymbol{b}$ of length N. The *central conjecture* in Bundschuh (2000) then states that λ is determined by the solution of the equation

$$\lim_{N\to\infty} \frac{1}{N} \log Z_N(\lambda) = 0. \tag{8}$$

Following the results of Park et al. (2005) and Chia and Bundschuh (2005), this allows for a very efficient computation of λ for gapped alignment in the iid model.

3. 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. We restrict ourselves to gapless alignment, since we expect qualitatively similar results for the gapped case. This will also be confirmed by the numerical data we present in Section 5. 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 Equation (8). 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 Figure 2. Defining $\boldsymbol{s} = (s_1, \ldots, s_N)$, we have

$$h = \sum_{i=1}^{N} s_i = \mathbf{1}^{\mathrm{t}}\boldsymbol{s}. \tag{9}$$

The ensemble average of Equation (7) over all realizations of the two sequences $\boldsymbol{a}$ and $\boldsymbol{b}$ can therefore be expressed in terms of an average over all score vectors $\boldsymbol{s}$. While the probability of a score vector factorizes in the iid model, $P(\boldsymbol{s}) = \prod_i P(s_i)$, this is no longer the case for correlated sequences. However, approximate values for the probabilities $P(\boldsymbol{s})$ 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

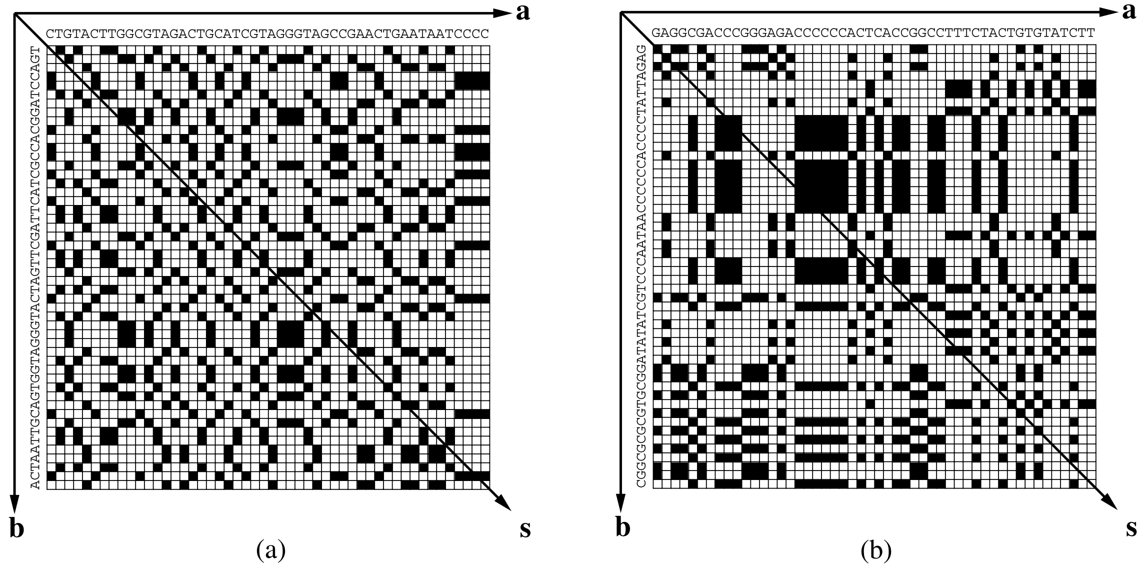
a
CTGTACTTGGCGTAGACTGCATCGTAGGGTAGCCGAACTGAATAATCCCC
ACTAATTGCAGTGGTAGGGTACTAGTTCGATTCATCGCCACGGATCCAGT
b
s
(a)
a
GAGGCGACCCGGGAGACCCCCCACTCACCGGCCTTTCTACTGTGTATCTT
CGGCGCGCGTGCGGATATATCGTCCCAATAACCCCCCACCCCCTATTAGAG
b
s
(b)

FIG. 2. Two realizations of an alignment score lattice: **(a)** The two sequences $\boldsymbol{a}$ and $\boldsymbol{b}$ 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 an expansion-randomization dynamics, as described in Section 4. Cells corresponding to matching nucleotides in $\boldsymbol{a}$ and $\boldsymbol{b}$ are colored black, white cells denote mismatches. The score of an ungapped global alignment is the sum over all elements of the diagonal vector $\boldsymbol{s} = (s_1, \ldots, s_N)$ of the score lattice. Comparison of both figures reveals qualitative differences between the two null models. The alignment lattice of sequences with long-range correlated nucleotide composition shows systematically larger black and white blocks representing exactly matching or mismatching substrings of the two sequences, compared to the iid model.

the alignment lattice will now be allowed to take continuous values, distributed according to a normal distribution

$$\mathrm{pdf}(s_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(s_i - \langle s \rangle)^2}{2\sigma^2}. \tag{10}$$

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

$$P(\boldsymbol{s}) = [(2\pi)^N \det \boldsymbol{\sigma}]^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{s} - \langle \boldsymbol{s} \rangle)^{\mathrm{t}} \boldsymbol{\sigma}^{-1} (\boldsymbol{s} - \langle \boldsymbol{s} \rangle)\right], \tag{11}$$

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

$$\boldsymbol{\sigma}_{ij} = \langle s(i)s(j) \rangle - \langle s(i) \rangle \langle s(j) \rangle. \tag{12}$$

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 $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\boldsymbol{\sigma}_{ij} = \frac{1}{3}(1 + \mu)^2 C^2(|i - j|). \tag{13}$$

In this expression the correlation function $C(r)$ is squared, since (13) describes the correlations of the similarity scores which arise from a comparison of two sequences. The non-diagonal elements vanish for iid sequences.

Using the distribution (11), the calculation of the generating function (7) amounts to the evaluation of an N-dimensional Gaussian integral, which can be solved explicitly,

$$\begin{aligned} Z_N(\lambda) &= \int d\boldsymbol{s}\; P(\boldsymbol{s}) \exp\left(\lambda \mathbf{1}^{\mathrm{t}} \boldsymbol{s}\right) \\ &= \left[(2\pi)^N \det \boldsymbol{\sigma}\right]^{-1/2} \int d\boldsymbol{s}\; e^{-\frac{1}{2}(\boldsymbol{s}-\langle \boldsymbol{s}\rangle)^{\mathrm{t}} \boldsymbol{\sigma}^{-1}(\boldsymbol{s}-\langle \boldsymbol{s}\rangle)+\lambda \mathbf{1}^{\mathrm{t}} \boldsymbol{s}} \\ &= \exp\left(\lambda \mathbf{1}^{\mathrm{t}}\langle \boldsymbol{s}\rangle + \frac{1}{2}\lambda^2 \mathbf{1}^{\mathrm{t}} \boldsymbol{\sigma} \mathbf{1}\right). \end{aligned} \tag{14}$$

The central conjecture (8) then implies

$$0 = \lim_{N\to\infty} \frac{1}{N}\left(\lambda \mathbf{1}^{\mathrm{t}}\langle \boldsymbol{s}\rangle + \frac{1}{2}\lambda^2 \mathbf{1}^{\mathrm{t}} \boldsymbol{\sigma} \mathbf{1}\right). \tag{15}$$

Notice that this expression coincides with the result obtained by applying the central conjecture to the Taylor series approximation of the generating function (7) up to second order. Using Equation (13) yields

$$\lambda = \frac{-2\langle s\rangle}{\sigma^2 + \frac{2}{3}(1+\mu)^2 \lim\limits_{N\to\infty} \sum\limits_{i=1}^{N} C^2(i)}. \tag{16}$$

The first term σ^2 in the denominator of (16) is related to the individual fluctuations of a single score element, irrespective of correlations along the sequences. The second term, on the other hand, 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

$$\lambda = \frac{-2\langle s\rangle}{\sigma^2 + \frac{2}{3}(1+\mu)^2 c^2 \zeta(2\alpha)}, \tag{17}$$

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. Notice that for $\alpha \le 1/2$ the sum $\sum_{i=1}^{\infty} C^2(i)$ diverges, resulting in $\lambda = 0$. This might indicate a transition from local to global alignment in the Gaussian approximation, which will be discussed in Section 4.3.

As a first evaluation of the Gaussian approximation, we investigate its predictions for sequences $\boldsymbol{a} = (a_1, \ldots, a_N)$ 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

$$P(a_{i+1}|a_i) = \begin{cases} p & : \quad a_{i+1} = a_i \\ \frac{1}{3}(1-p) & : \quad a_{i+1} \ne a_i \end{cases}. \tag{18}$$

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 (16) yields

$$\lambda = \frac{-2\langle s\rangle}{\sigma^2 + \frac{2}{3}(1+\mu)^2 c^2/(\exp(2\beta) - 1)}. \tag{19}$$

This can be compared to an exact analytical result for λ obtained by equating the largest eigenvalue of a modified λ-dependent transition matrix of the underlying Markov process to one (Karlin and Dembo,

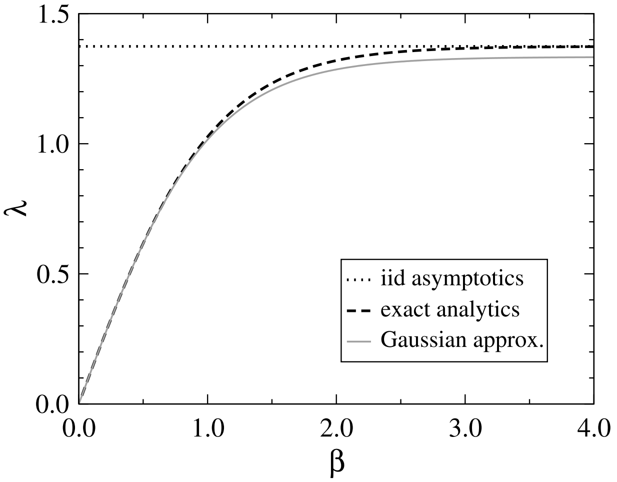

FIG. 3. λ for sequences with short-range correlations generated by a Markov process. The dashed line is the exact result (Karlin and Dembo, 1992) for the Markov process defined in (18), using $\mu = 3$. The solid line is the corresponding result of the Gaussian approximation, as derived in Equation (19). Solving Equation (6) yields the iid asymptotics $\lambda \approx 1.374$ (dotted line).

1992). As is shown in Figure 3, the Gaussian approximation (19) 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 (10) vanish, which they do not for the discrete distribution.

4. NUMERICAL RESULTS

4.1. *Generation of long-range correlated random sequences*

Numerical evaluation of the results obtained in the previous section hinges on the knowledge of the score distribution pdf(S) for local gapless alignment of pairs of long-range correlated random sequences. However, the efficient generation of such sequences is quite intricate. In Messer et al. (2005a), we have proposed a biologically motivated model of sequence evolution which generates sequences with the desired statistical features. Furthermore, it has recently been shown (Messer et al., 2005b) that there exists a much larger class of dynamical processes, so called, *expansion-randomization* processes, which allow for the efficient generation of sequences with arbitrary long-range correlations.

Based on Messer et al. (2005b), we use a single-site duplication-mutation algorithm to generate long-range correlated sequences. We start with a sequence of one random nucleotide a_1, and the dynamics of the model is defined by the following update rules:

1. A random position j of the sequence is chosen.
2. The nucleotide a_j is either mutated to a random but different nucleotide with probability P_{mut}, or duplicated with probability $P_{\text{dup}} = 1 - P_{\text{mut}}$. The duplication process inserts a copy of a_j at position $j + 1$, thereby increasing the sequence length by one.

This process generates sequences of arbitrary length N in a time $O(N)$ with asymptotic long-range correlations in their nucleotide composition. The correlation function of the generated sequences is given in terms of the Euler beta function $B(x, y)$ by Messer et al. (2005a):

$$C(r) = \frac{3}{4}\alpha B(r + 1, \alpha). \tag{20}$$

In the large r limit, this yields $C(r) \propto r^{-\alpha}$. By varying the mutation probability $0 < P_{\mathrm{mut}} < 1$, the decay exponent α of the long-range correlations can be tuned to any desired positive value, as it is determined by

$$\alpha = \frac{8}{3}\frac{P_{\mathrm{mut}}}{1 - P_{\mathrm{mut}}}. \tag{21}$$

The algorithm has recently been implemented in the web server CorGen (Messer and Arndt, 2006), which is publicly available at *http://corgen.molgen.mpg.de*. CorGen can measure long-range correlations in the base composition of DNA and generate random sequences with the same correlation parameters or any other user-defined long-range correlation parameters, using the above described algorithm. If large ensembles of long-range correlated random sequences are needed, as is the case for an accurate measurement of the tail of the score distribution pdf(S), independent realizations of the sequences can directly be obtained from CorGen via non-interactive network clients, e.g., `wget`. 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 (Smith and Waterman, 1981) with scoring matrix (3) and $\mu = 3$.

4.2. Gumbel distribution of alignment scores

Our solution of the Gaussian model is based on the assumption that the alignment score distribution pdf(S) is of Gumbel form for long-range correlated sequences. Consequently, our first numerical analysis aims at a verification of this conjecture. In Figure 4, we show the measured score distributions for two different long-range correlated sequence ensembles with correlation parameters $\alpha = 1.0$ and $\alpha = 0.5$, compared to the Gumbel distribution of the iid model. While 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, and the mean of the distribution is additionally shifted towards larger scores. For large N, the shape of pdf(S) asymptotically approaches a Gumbel form for the correlated ensembles, as can be seen in Figure 5, where we exemplarily show the measured score distribution for long-range correlated sequences with $\alpha = 2.0$ and different

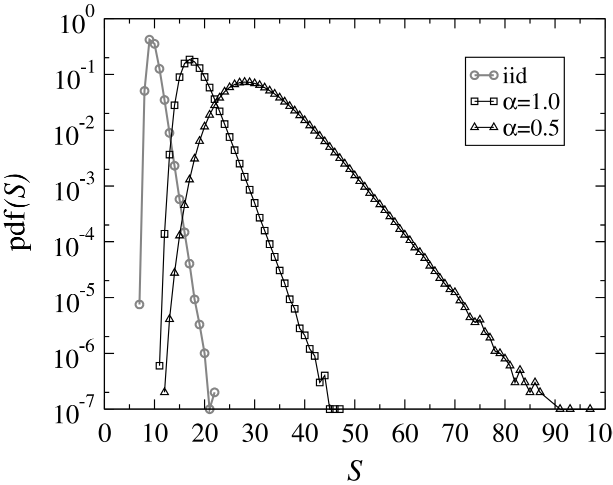

FIG. 4. Numerically measured shape of the alignment score distribution pdf(S) for $N = M = 10^3$ using three different null models: iid sequences, long-range correlated sequences with $\alpha = 1.0$, and such with $\alpha = 0.5$. The distributions were obtained by aligning 10^7 pairs of sequences randomly 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 pdf(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 (indicated by less steeper slopes in the semi-logarithmic plot), and therefore lead to a slower decay of the exponential tail of the alignment score distribution.

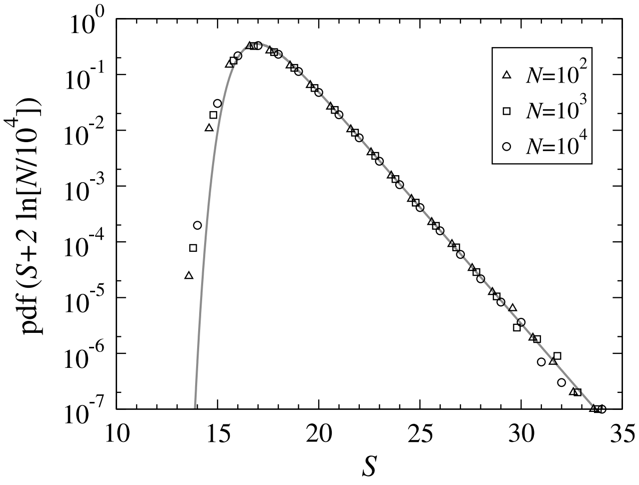

FIG. 5. Convergence of the distribution pdf(S) for long-range correlated sequences with $\alpha = 2.0$ to a Gumbel form. The solid line is a Gumbel distribution, as specified in Equation (5) with $N = M = 10^4$ and fitted parameters $\lambda = 0.9614$ and $K = 0.119$. λ was obtained by fitting a linear function to log[pdf(S)] for $21 < S < 31$, K has then been estimated by fitting the data to (5) in the same interval. In order to be able to compare the shape of pdf(S) for different N, the distributions have to be rescaled by a transformation pdf(S) → pdf($S + 2\ln[N/N_0]$) with reference length $N_0 = 10^4$.

sequence lengths N. As is the case for the iid model, finite-size corrections come into play for small N (Altschul and Gish, 1996; Altschul, 1991; Yu et al., 2002). These deviations primarily show up in the small S regime, while the more relevant large S regime converges fast for increasing sequence lengths.

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 Equation (16), using correlations of the form (20). Results are shown in Figure 6. The Gaussian approximation captures the qualitative behavior of the numerical data. Again,

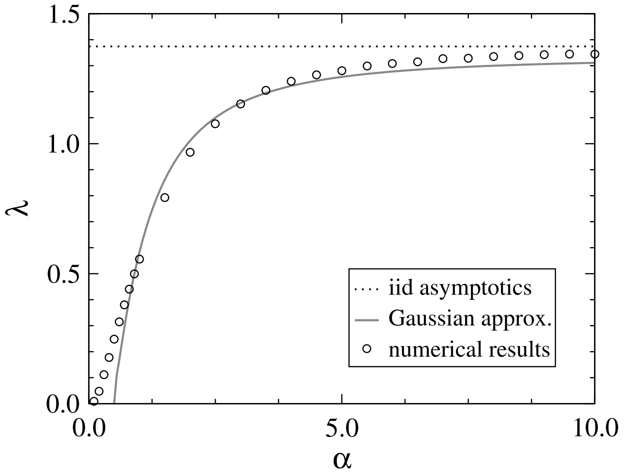

FIG. 6. λ for a null model with long-range correlated sequences in dependence of the correlation exponent α. The solid line is the analytic result of the Gaussian approximation, one obtains by estimating Equation (16) using the correlations (20) of our simulated sequences. Numerically measured values 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.

Table 1. Dependence of $\langle S\rangle$ and K on the Exponent α

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}

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 Equation (22).

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.

So far, our investigations of the alignment score distribution for long-range correlated sequences have focused on the exponential tail of pdf(S). We now turn to the second parameter K. For that purpose, we recall that the mean of a Gumbel distribution (5) is determined by

$$\langle S\rangle = \frac{\Gamma + \log(KN^2)}{\lambda}, \tag{22}$$

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 1 and Figure 7, K is significantly affected by the presence of long-range correlations in the sequences to be aligned; it decreases

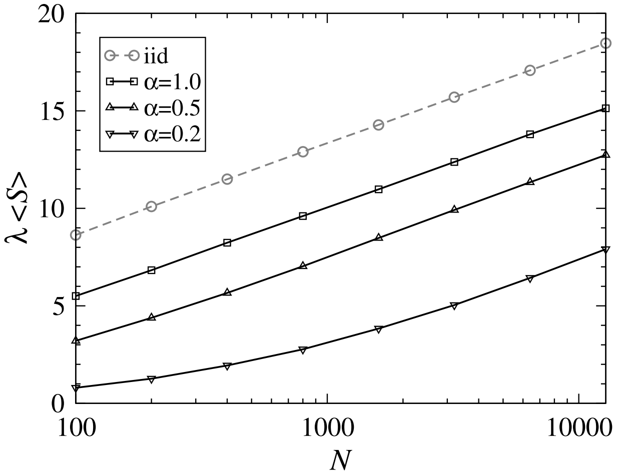

FIG. 7. Mean of the score distribution pdf(S) for different exponents α against increasing sequence length N. Measured values for $\langle S\rangle$ were obtained by averaging over ensembles of 10^4 pairs of sequences. K can be measured using $K = 1/N^2 \exp[\lambda\langle S\rangle - \Gamma]$. The deviations 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 pdf(S) yet significantly deviates from a Gumbel distribution for small N in the strong correlation regime $\alpha \leq 0.5$.

with increasing correlation-strength. However, the mean of the distribution is, as expected, shifted to larger values of S for decreasing values of α, since K contributes only logarithmically in Equation (22) and the change in $\langle S \rangle$ is dominated by the decrease of λ. Figure 7 again reveals the finite-size deviations of the numerically measured score distribution pdf(S) from a Gumbel form (5) in the strong correlation regime $\alpha \leq 1/2$.

4.3. 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 (Fig. 6). The numerical investigation of this regime is complicated by a distinct finite size effect: according to the results derived in Messer et al. (2005b), 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. However, 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 since they lead to systematically higher values of $\langle s \rangle$ and therefore also higher scores S.

As can be seen in Figure 6, 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 this is the case, 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.

5. CONSEQUENCES FOR ALIGNMENTS OF GENOMIC SEQUENCES

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. As an example, we consider the measured correlation function of Human chromosome 22, shown in Figure 1b. Using the simulation algorithm introduced in Section 4.1 we can generate long-range correlated random sequences with the corresponding exponent $\alpha \approx 0.232$. By randomly mutating 85% of the sites after sequence build up, the correlation amplitude is reduced to the genomic value, while the exponent remains unchanged (Messer et al., 2005b). As can be seen in Figure 1b, this procedure allows us to generate random sequences featuring comparable correlations as Human chromosome 22.

We perform ungapped, as well as gapped alignment with affine gap costs for 10^7 pairs of random sequences with 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 $\gamma_e = 2$ (NCBI, 2007). In Figure 8, we show the measured score distributions for the simulated chromosome 22 sequences compared to iid sequences. The resulting parameters λ and $\langle S \rangle$ are presented in Table 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 higher. In combination, both effects substantially change the p-values of high scores compared to the iid model, as can be seen in Table 2. The p-value of a specific score S' is thereby defined by the integral $P(S \geq S') = \int_{S'}^{\infty} \text{pdf}(S)dS$. 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.

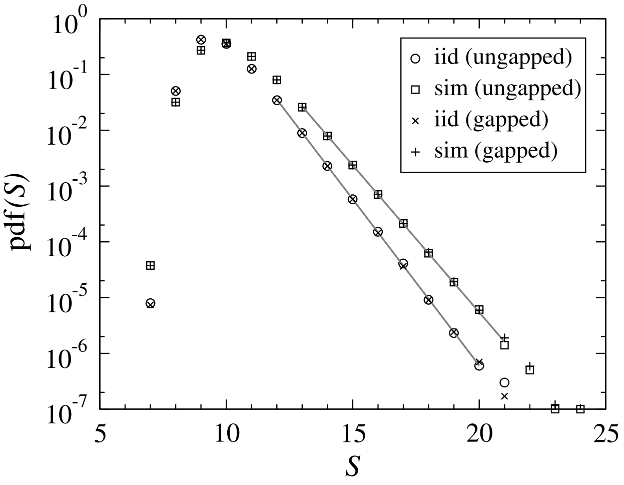

FIG. 8. The score distribution for ungapped and gapped alignment of simulated sequences with correlations comparable to those of Human chromosome 22. The straight lines are the fits to the exponential tails of the score distributions, obtained by fitting a linear function to log[pdf(S)] in the depicted intervals.

6. DISCUSSION

Long-range correlations are a widespread statistical feature of eukaryotic DNA. In this article, 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 significantly 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. 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. 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 analysis tools.

TABLE 2. FITTED PARAMETERS λ AND $\langle S\rangle$ FOR THE iid ENSEMBLE AND SIMULATED HUMAN CHROMOSOME 22 SEQUENCES OF LENGTH $N = 10^3$

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}

In the last column, exemplary p-values of a score $S' = 18$ are shown.

ACKNOWLEDGMENT

This work has been partially funded by the U.S. National Science Foundation (grant DBI-0317335 to R.B.).

REFERENCES

Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219, 555–565.

Altschul, S.F., and Gish, W. 1996. Local alignment statistics. *Methods Enzymol.* 266, 460–480.

Altschul, S.F., Gish, W., Miller, W., et al. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Arneodo, A., Bacry, E., Graves, P.V., et al. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.* 74, 3293–3296.

Bernaola-Galvan, P., Carpena, P., Roman-Roldan, R., et al. 2002. Study of statistical correlations in DNA sequences. *Gene* 300, 105–115.

Bundschuh, R. 2000. An analytic approach to significance assessment in local sequence alignment with gaps. *RECOMB 2000*, 86–95.

Bundschuh, R. 2002. Asymmetric exclusion process and extremal statistics of random sequences. *Phys. Rev. E* 65, 031911, 1–19.

Chia, N., and Bundschuh, R. 2005. A practical approach to significance assessment in alignment with gaps. *RECOMB 2005*, 474–488.

Durbin, R., Eddy, S., Krogh, A., et al. 1998. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Grossmann, S., and Yakir, B. 2004. Large deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignments. *Bernoulli* 10, 829–845.

Karlin, S., and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87, 2264–2268.

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

Karlin, S., and Dembo, A. 1992. Limit distribution of the maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Prob.* 24, 113–140.

Li, W. 1991. Expansion-modification systems: a model for spatial $1/f$ spectra. *Phys. Rev. A* 43, 5240–5260.

Li, W., and Holste, D. 2005. Universal $1/f$ noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome. *Phys. Rev. E* 71, 041910, 1–9.

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.

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

Messer, P.W., Arndt, P.F., Lässig, M. 2005a. Solvable sequence evolution models and genomic correlations. *Phys. Rev. Lett.* 94, 138103.

Messer, P.W., Lässig, M., Arndt, P.F. 2005b. Universality of long-range correlations in expansion-randomization systems. *J. Stat. Mech.*, P10004, 1–23.

Mott, R. 1999. Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* 54, 59–75.

NCBI. 2007. Available at: *http://www.ncbi.nlm.nih.gov/BLAST*. Accessed May 1, 2007.

Park, Y., Sheetlin, S., Spouge, J.L. 2005. Accelerated convergence and robust asymptotic regression of the Gumbel scale parameter for gapped sequence alignment. *J. Physics A* 38, 97–108.

Peng, C.K., Buldyrev, S.V., Goldberger, A.L., et al. 1992. Long-range correlations in nucleotide sequences. *Nature* 356, 168–170.

Smith, S.F., and Waterman, M.S. 1981. Comparison of biosequences. *Adv. Appl. Math.* 2, 482–489.

Smith, T.F., Waterman, M.S., Burks, C. 1985. The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.* 13, 645–656.

Voss, R.F. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* 68, 3805–3808.

Waterman, M.S. 1995. *Introduction to Computational Biology: Maps, Sequences, and Genomes*. CRC Press, Boca Raton, FL.

Waterman, M.S., and Vingron, M. 1994. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci. USA* 91, 4625–4628.

Yu, Y.K., Bundschuh, R., Hwa, T. 2002. Statistical significance and extremal ensemble of gapped local hybrid alignment. *LNP* 585, 3–21.

Address reprint requests to:
Philipp W. Messer
Max Planck Institute for Molecular Genetics
Ihnestr. 73
14195 Berlin, Germany

E-mail: philipp.messer@molgen.mpg.de