Entropies of biosequences: The role of repeats

Hanspeter Herzel*
Institute of Theoretical Physics, Technical University, Hardenbergstrasse 36, D-10623 Berlin, Germany

Werner Ebeling and Armin O. Schmitt
Institute of Physics, Humboldt University, Invalidenstrasse 110, D-10115 Berlin, Germany
(Received 23 May 1994)

DNA sequences of higher organisms contain thousands of nearly identical dispersed repetitive sequences. In order to understand the effect of such repeats on word entropies, we construct a model that can be analyzed analytically. The hypothetical model sequences consist of independent equidistributed symbols with randomly interspersed repeats. As a conclusion, we predict that the entropy of DNA sequences measuring the information content is much lower than suggested by earlier empirical studies.

PACS number(s): 87.10.+e, 05.40.+j, 02.50.Ey

I. INTRODUCTION

The immense progress of molecular biology revealed that genomes are of extraordinary complexity [1–7]. Despite the complicated mechanisms of recombination and gene expression, a first inspection of frequencies of nucleotides, dinucleotides, etc. indicates that letters and short words are approximately equidistributed and hence the entropy of the source was estimated to be higher than 1.9 bits per nucleotide [8–12]. However, available data bases are strongly biased towards genes since these fractions of the genome are naturally of particular interest. But in order to discuss the overall information content, the repetitive character of large parts of the DNA has to be taken into account.

Repeated nucleotide sequences are a characteristic feature of eukaryotic genomes. Reassociation experiments from single to double strand DNA revealed that up to 50% of the DNA consists of repeats. These subsequences include repetitions of segments coding for RNAs and histones, of pseudogenes, of "satellite DNA" (long runs of tandemly repeated short words), and of randomly interspersed repeats. The latter are divided into "long interspersed repeat sequences" (LINES), which are typically 6000 bases long, and "short interspersed repeat sequences" (SINES) of a few hundred bases. These seemingly randomly dispersed repeats have presumably been distributed by replicative transpositions [3,7]. A prominent example of SINES is the family of Alu repeats with an approximate length of 300 bases. The name Alu derives from the fact that nucleotide sequences occur in these repeats that can be cut by a restriction enzyme, AluI. It has been estimated that the human genome contains about 900 000 copies of this repeat corresponding to about 9% of the whole genome [5]. Hence this family alone makes up more DNA than protein coding regions.

The biological role of repeats is now recognized as an extremely important subject. For multigene families, which represent a specific class of repeats, the function is known. They are needed for the synthesis of proteins with a similar, not identical, function in differentiated cells and tissues [7]. However, the function of most repetitive sequences is, so far, obscure. There are indications that repeats may help to ensure the complex hierarchical system of genome regulation, i.e., folding of DNA into chromatin, its reproduction, storage, and differential gene expression. For instance, Alu repeats are assumed to affect the expression of adjacent genes because of the functional sites contained in these repeats [7].

Obviously, repetitive nucleotide sequences imply redundancy and reduce the information content of DNA. This effect is quantified in our paper with the aid of word entropies. However, the combinatorial explosion of the number of possible words with their lengths prohibits the direct estimation of entropies of long words even from the longest DNA sequences available. Although sophisticated finite sample corrections have been developed [12–16], no reliable estimations are possible so far for word lengths $n > 12$ [17]. In this paper we circumvent finite sample effects by studying hypothetical processes corresponding to (in principle) infinite strings. Our model, which will be formulated in Sec. III, consists of a "sea" of random symbols with interspersed repeats that are characterized by their lengths and probability of occurrence.

In Sec. IX we discuss also the role of repeats for long-range correlations which have been detected in DNA sequences with the aid of mutual information [12,16,18], correlation functions [19,20], and spectra [21,22].

II. THE ENTROPY CONCEPT

In sequence analysis, the study of correlation functions and spectra implies a somewhat arbitrary identification of symbols with real numbers (see [23] for a discussion of this point). Moreover, correlation coefficients mea-

*Electronic address: herzel@itp1.physik.tu-berlin.de

sure only linear dependences. By contrast, information-theoretical quantities such as those discussed below detect any statistical dependences. In this section we introduce information-theoretical measures that are widely used in linguistics and nonlinear dynamics [24–29].

Symbol sequences are composed of λ "letters" $A_1, A_2, ..., A_\lambda$. Their corresponding probabilities of occurrence are denoted by p_i $(i = 1, 2, ..., \lambda)$. Then, the entropy

$$H_1 = \sum_{i=1}^{\lambda} -p_i \log_2 p_i \tag{1}$$

gives the (average) information of a single symbol. Analogously, the n-word entropy H_n is defined for the probabilities $p_i^{(n)}$ of "words" $\mathcal{S}_i^{(n)}$:

$$H_n = \sum_i -p_i^{(n)} \log_2 p_i^{(n)}. \tag{2}$$

The summation has to be carried out over all words with nonvanishing probability. The maximum number of possible n-words is λ^n. In our paper, we always choose $\lambda = 4$ referring to the four nucleotides A, C, G, and T. All logarithms are taken to base 2 and thus the entropies are measured in bits. Consequently, a sequence of four independent equidistributed letters gives word entropies $H_n = 2n$.

The differential entropies h_n are a closely related measure:

$$h_n = H_{n+1} - H_n. \tag{3}$$

They indicate the information contained in the $(n+1)$th letter, presuming the n previous letters are known. We note that h_n can be rewritten as the average logarithm of the conditional probability that a certain letter A_j follows a word $\mathcal{S}_i^{(n)}$:

$$h_n = \langle -\log_2 p(A_j|\mathcal{S}_i^{(n)}) \rangle. \tag{4}$$

Here $\langle\,\rangle$ denotes the average over all $(n+1)$-tuples $(\mathcal{S}_i^{(n)}, A_j)$.

The asymptotic information gain was termed "entropy of the source" [24]

$$h = \lim_{n\to\infty} h_n. \tag{5}$$

It plays a central role in coding theory [30,31] and is intimately related to the Kolmogorov entropy in dynamical systems theory [27].

Besides the limit h, the convergence of the h_n contains valuable information [8,28,32] since it quantifies memory effects within the string. Grassberger introduced, for example, the sum

$$S_{\mathrm{EMC}} = \sum_{n=0}^{\infty} (h_n - h) \tag{6}$$

as the "effective measure complexity" [28] and some authors study the corresponding decay rate for chaotic maps [32].

For Markov chains with memory m, the asymptotic value h is already reached for $n = m$, i.e.,

$$h_m = h_{m+1} = \cdots = h. \tag{7}$$

In summary, the entropy h quantifies the information content per symbol and the decay of the h_n measures correlations within the sequence.

III. MODEL

Frequencies and lengths of repeats vary in a wide range. There are rather short words of a few bases which appear millions of times (simple-sequence DNA) and very long strings of several thousand nucleotides may appear only twice (e.g., duplicated genes) [3–6]. Our model is sufficiently general to allow adaptation to observed frequency and length distributions of repeats.

The basic assumption of our model is to distinguish between nonrepetitive DNA, which is close to a random succession of letters, and interspersed repeats with a certain length and probability of occurrence. It may sound strange to classify nonrepetitive DNA as quasirandom since there are many structures detectable [2] such as an alternation of protein coding sequences (exons) and intervening sequences (introns), promotor sequences, ribosome binding sites, etc. Moreover, characteristic distributions of 3-words (codons) can be found within exons [33]. However, the effect of these structures on entropies is relatively small. As an example we discuss the codon usage table of protein coding sequences [34]. The characteristic uneven distribution of codons in exons is one of the most striking signals in DNA sequences [35,12] and is widely used to identify protein coding regions [33]. Even though there are remarkable deviations from an equidistribution [three stop codons are absent and $p(\mathrm{CGA}) = 0.004$ is, e.g., much smaller than $p(\mathrm{CTG}) = 0.033$], the corresponding entropy is close to its maximum value of 6.0 bits:

$$H_{\mathrm{codon}} = 5.80. \tag{8}$$

Several authors estimated entropies up to H_5 from various mostly nonrepetitive DNA sequences. Their estimations of $h_4 = H_5 - H_4$ are close to 2 bits as well:

$$\begin{aligned}
&h_4 = 1.94 && \text{(rabbit liver [8])},\\
&h_4 = 1.93 && \text{(viral DNA [9])},\\
&h_4 = 1.92 && \text{(mammalian genes [10])},\\
&h_4 = 1.92 && \text{(bacteria [10])},\\
&h_4 = 1.97 && \text{(Rous Sarcoma virus [12])},\\
&h_4 = 1.95 && \text{(yeast chromosome III [14])}.
\end{aligned} \tag{9}$$

In these cases, the analyzed sequences contain a large amount of genes and only a few repeats. For example, words with a length of $n \geq 20$ that appear at least

twice constitute less than 4% of the yeast chromosome III. These examples illustrate that the base composition of DNA is close to a random one and that short-range correlations only weakly reduce the entropy.

In order to keep our model simple, we approximate nonrepetitive DNA by a Bernoulli process of independent equidistributed letters. Hence we concentrate on the lowering of the entropy due to interspersed repeats. As mentioned in the Introduction, repeats may constitute more than 50% of the DNA of higher organisms [3–6]. The various families of SINES and LINES have characteristic lengths ranging from about a hundred to several thousand bases.

In our model, we introduce repeats as follows. A repeat $\mathcal{R}_j(\rho_j, l_j)$ is characterized by its relative frequency of occurrence ρ_j and its length l_j. Here ρ_j is the probability to find the first symbol of the repeat $\mathcal{R}_j$ at an arbitrary site of the sequence. In other words, ρ_j is the expected number of copies divided by the length of the whole sequence. Consequently, the product $\rho_j l_j$ is the fraction of the total sequence covered by copies of the repeat $\mathcal{R}_j$. The copies of the repeats are assumed to be randomly distributed within the sequence. Then the mean distance between the start of two copies of $\mathcal{R}_j$ is $d_j = \frac{1}{\rho_j}$. Furthermore, we assume that there are no specific statistical features within the repeating sequence, i.e., the strings $\mathcal{R}_j$ have a random composition of independent letters. Calculations of the entropy H_1 for repeats gave the values [36]

$$H_1 = 1.97 \quad \text{(Alu repeats)}$$
$$H_1 = 1.91 \quad \text{(non-Alu repeats)}, \tag{10}$$

indicating that the symbol composition is not too far from being random in many cases.

In this way, a process is defined that contains interspersed repeats characterized by their frequency and length in a "random sea" of letters. Such a process can be considered as a generator of infinitely long sequences and hence finite sample effects do not have to be considered. However, in order to compare analytical results with numerical simulations, we construct also finite realizations of the above process in the following way. First, a random string of length $L(1-\rho_j l_j)$ is generated. Then a repeat of length l_j is inserted at $\rho_j L$ random positions. This procedure has some analogy to the "transpositions of sequences," which is assumed to be a relevant mechanism for the spreading of repeats.

Despite drastic simplifications the constructed process exhibits statistically relevant features of eukaryotic DNA sequences. It will be shown below that the effect of repeats on entropies can be understood very clearly with the aid of our hypothetical model.

IV. CALCULATION OF WORD DISTRIBUTIONS

In this section we discuss, for simplicity, processes with but a single type of repeat $\mathcal{R}(\rho, l)$. The corresponding results for the case of several repeats can be obtained through superposition afterwards.

As discussed in Sec. II, the calculation of entropies H_n is based on the probability distributions of n-words. In order to obtain these distributions we distinguish between three possible cases: (a) an n-word is located within the random sea, (b) the n-word lies partially in the repeat and has an overlap of $k < n$ letters with a repeat, and (c) the word lies completely within the repeat. The task now is to find out how many of the λ^n words belong to each of these classes. Moreover, the probabilities of words with an overlap k have to be estimated. We introduce the following notations: $Z(k)$ denotes the number of words with an overlap k. Note that $Z(0)$ and $Z(n)$ correspond to the classes (a) and (c), respectively. Furthermore, we calculate the probability $P_k(\mathcal{S}_i^{(n)})$ of an n-word having an overlap k.

In order to derive expressions for $Z(k)$ and $P_k(\mathcal{S}_i^{(n)})$, we first consider a part of our hypothetical string consisting of just one repeat (small letters in the sketch below) and a following random part (capital letters). The length of this part is denoted by $\hat{d}$ and hence the number of random symbols is $\hat{d}-l$. The notations are visualized in the following sketch:

$$\ldots \text{AGAT } \overbrace{\text{CTcgg}}^{n=5,k=3} \text{ acttaTATG } \overbrace{\text{cggactta}}^{l=8} \underbrace{\text{CTCAGGA}}_{\hat{d}} \text{cgg} \ldots .$$

Considering all words of length n starting in this chosen part, we find $\hat{d}-l-n+1$ words completely within the random string [i.e., with $k=0$, class (a)], $l-n+1$ words within the repeat [$k=n$, class(c)], and two words overlapping by k ($0<k<n$) letters (note that words starting near the end of the random subsequence have an overlap with the next repeat).

From these preliminaries, we can now deduce the overall fraction of n words with overlap k. It was argued in Sec. III that a given probability ρ of a repeat corresponds to a mean distance $d = \frac{1}{\rho}$ between the start of two copies. Hence the average length of the random subsequences is $d-l$. Replacing the specific length $\hat{d}$ in the above example by the average value d, we obtain the fraction of words within the random sea

$$F(0) = \frac{d-l-n+1}{d}. \tag{11}$$

The fraction of words with overlap k is

$$F(k) = \frac{2}{d} \qquad (k = 1, 2, \ldots, n-1) \tag{12}$$

and of those completely within the repeat

$$F(n) = \frac{l-n+1}{d}. \tag{13}$$

Per construction, the fractions are normalized

$$\sum_{k=0}^{n} F(k) = 1. \tag{14}$$

We have learned how the probability is distributed among the different classes with overlaps $k \in [0, n]$.

In order to obtain the probabilities of specific words, we have to study how many words share the corresponding amount of probability $F(k)$. Obviously, within the random sea, all λ^n words appear. Hence any word can be found with a basic rate of

$$P_0(\mathcal{S}_i^{(n)}) = \frac{F(0)}{\lambda^n} = \frac{d-l-n+1}{d\ \lambda^n}. \tag{15}$$

n-words occurring in the repeat once have an additional probability ρ to appear

$$P_n(\mathcal{S}_i^{(n)}) = \rho + \frac{F(0)}{\lambda^n}. \tag{16}$$

Considering overlapping words, there are λ^{n-k} different words that can have an overlap of $n-k$ letters with the random sea (the other k letters are fixed as part of the repeat $\mathcal{R}$) and therefore these λ^{n-k} words share the probability ρ

$$P_k(\mathcal{S}_i^{(n)}) = \frac{\rho}{\lambda^{n-k}} + \frac{F(0)}{\lambda^n}. \tag{17}$$

Now we know the probabilities of words with a certain overlap. It remains to count the number of words with these probabilities $P_k(\mathcal{S}_i^{(n)})$. There are obviously $l-n+1$ words of length n within a repeat, i.e.,

$$Z(n) = l - n + 1. \tag{18}$$

As already discussed, the number of words with overlap k is

$$Z(k) = 2\lambda^{n-k}. \tag{19}$$

The factor 2 reflects the fact that each repeat has overlapping words at its beginning and at its end. The number of the remaining words that appear only in the random sea can be obtained from normalization

$$\sum_{k=0}^{n} Z(k) = \lambda^n. \tag{20}$$

Inserting the results above, we obtain

$$Z(0) = \lambda^n - (l-n+1) - \frac{2(\lambda^n - \lambda)}{\lambda - 1}. \tag{21}$$

Using these expressions, the entropy H_n can be calculated

$$H_n = \sum_{k=0}^{n} -Z(k)P_k(\mathcal{S}_i^{(n)}) \log_2 P_k(\mathcal{S}_i^{(n)}). \tag{22}$$

At this point, we want to emphasize that the above expression is not exact, but it is a reasonable approximation for sufficiently large n. Our implicit assumption was that words $\mathcal{S}_i^{(n)}$ within the repeat and with a certain overlap $k < n$ are all different. This ansatz would fail, for example, if a word appeared twice within the repeat. Then its probability would be higher than $P_n(\mathcal{S}_i^{(n)})$ and, moreover, the numbers $Z(k)$ would be affected. For $\lambda^n \gg l$, this observation is not very likely. However, certain kinds of repeats such as retroposons contain such "repeats within repeats" (e.g., "long terminal repeats"). In some cases even hierarchies of words can be found [3,23]. Our model should be refined in such a situation.

Moreover, we assume that copies of repeats are sufficiently far apart so that n-words have an overlap only with one single copy. It will be shown in the following section that our theoretical predictions are in excellent agreement with simulations despite these approximations.

V. COMPARISON WITH NUMERICAL SIMULATIONS

As described in Sec. III, we first generate a long random string of independent equidistributed symbols with length $L(1-\rho l)$. Then repetitive sequences of length l are inserted at ρL randomly chosen positions.

As a first feature of such a process, we compare rank-ordered distributions of n-words. For this purpose the frequencies of all λ^n words are counted and the words are arranged according to their frequency of occurrence. Such "ordered histograms" are widely used in linguistics [37]. Our theoretical consideration in the preceding section predicts a staircaselike rank-ordered distribution. There should be a pedestal at a level of $P(0)L$ and a plateau of a height $P(n)L$ due to the $Z(n)$ words within the repeats. Words with overlaps k form the steps of the staircase from the pedestal to the plateau.

Figure 1 shows fairly good agreement between the the-

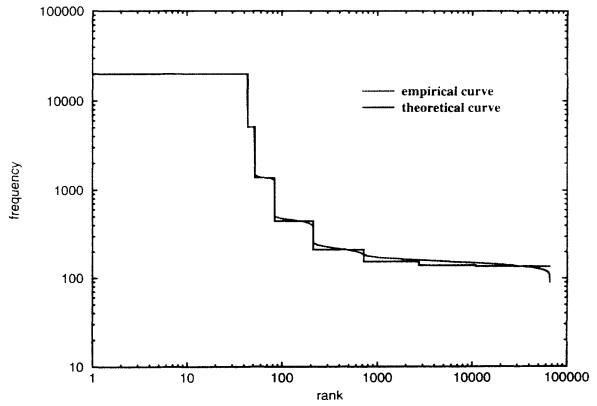

FIG. 1. Rank-ordered statistics of 8-words from a sequence of $L = 10^7$ letters with 20 000 interspersed repeats of length $l = 50$. Full line (staircase), theoretical predictions according to $Z(k)$ and $P(k)$ in Sec. IV; dotted line, rank-ordered histogram from a realization.

oretical curve and the distribution from the simulations. The rounding of the steps is caused by finite sample effects [16]. The rank-ordered histogram exhibits an interesting feature: The decay from the plateau to the pedestal obeys a power law i^{-1} ("Zipf's law"), i being the rank of the word. This is due to the fact that the height of steps decays proportional to λ^{k-n}, whereas the number of words per step increases as λ^{n-k}. Thus, as also discussed in [38] for another random process, fairly simple mechanisms can generate Zipf-like word distributions.

Now we turn to the comparison of word entropies H_n from Eq. (22) with estimations from simulations. Even for long realizations, we have to care about systematic errors due to finite sample effects. It was shown in [12,16] that

$$\Delta H_n^{\mathrm{sys}} = \frac{\lambda^n}{2L \ln 2} \tag{23}$$

gives a first approximation of the systematic underestimation of entropies. Hence, for $L = 10^7$, we can trust the estimations of H_n for $n \leq 9$. In Fig. 2, we corrected the entropy estimations by the term in Eq. (23). Figure 2 visualizes the excellent agreement of the theoretical entropies and those from a simulated process.

The characteristic features of the curve h_n versus n is a decrease from $h_n = 2$ to $h_n \approx 1.8$ at about $n = 5$. It will be discussed in the following section how the asymptotic value (about 1.8 bit in the above case) and the crossover to this value can be predicted by rather simple calculations.

VI. DERIVATION OF A SIMPLIFIED ENTROPY FORMULA

In this section, we derive simplified expressions for the differential entropies h_n, allowing an easy interpretation and generalization. We recall that h_n measures the mean uncertainty about the $(n+1)$th symbol if the preceding n letters are known. Quantitatively, h_n can be obtained from the mean logarithm of the conditional probability $p(A_j|\mathcal{S}_i^{(n)})$ [see Eq. (4)]. The average has to be carried out over all words $\mathcal{S}_i^{(n)}$ $(i = 1, 2, ..., \lambda^n)$ and continuations by the $(n+1)$th letter A_j $(j = 1, 2, ..., \lambda)$. The conditional probabilities are just a ratio of the corresponding word probabilities:

$$p(A_j|\mathcal{S}_i^{(n)}) = \frac{p_{ij}(\mathcal{S}_{ij}^{(n+1)})}{p_i(\mathcal{S}_i^{(n)})}. \tag{24}$$

Here the double index ij refers to the corresponding $(n+1)$-word composed of the n-word $\mathcal{S}_i^{(n)}$ and the letter A_j. Explicit expressions for these probabilities are available from the considerations in Sec. IV: If a word $\mathcal{S}_{ij}^{(n+1)}$ appears within the random sea, we obtain

$$p(A_j|\mathcal{S}_i^{(n)}) = \frac{(d-l-n)\quad \lambda^n}{(d-l-n+1)\quad \lambda^{n+1}} \approx \frac{1}{\lambda}. \tag{25}$$

That means that all continuations A_j are equally probable and hence no prediction better than guessing is possible. The corresponding uncertainty thereupon is 2 bits. For words within the repeat, we obtain the following formula for the "correct" continuation with $\rho \gg \frac{1}{\lambda^n}$:

$$p(A_j \mid \mathcal{S}_i^{(n)}) = \frac{\rho + \frac{(d-l-n)}{\lambda^{n+1}}}{\rho + \frac{(d-l-n+1)}{\lambda^n}} \approx 1. \tag{26}$$

In this case, the next letter is almost surely predictable. Our simulation in Fig. 3 visualizes the switch of the conditional probabilities from $\frac{1}{4}$ to 1 when a repeat is detected.

We note in passing that such plots may be exploited to identify repeats in sequences. Figure 3 is simply obtained by counting words of length $n+1$ without *a priori* assumptions. All words that are much more frequent than random words of the same length become visible

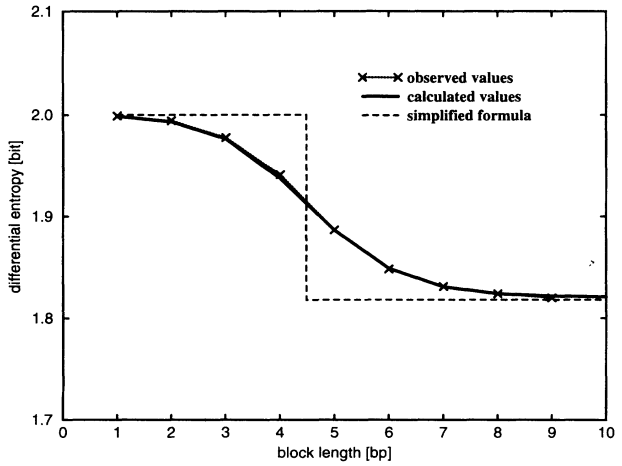

FIG. 2. Differential entropies h_n from a realization (×) ($L = 10^7$, $\lambda = 4$, $\rho = 0.002$, and $l = 50$) and from Eq. (22) (full line). The dashed line corresponds to Eq. (33) in Sec. VI.

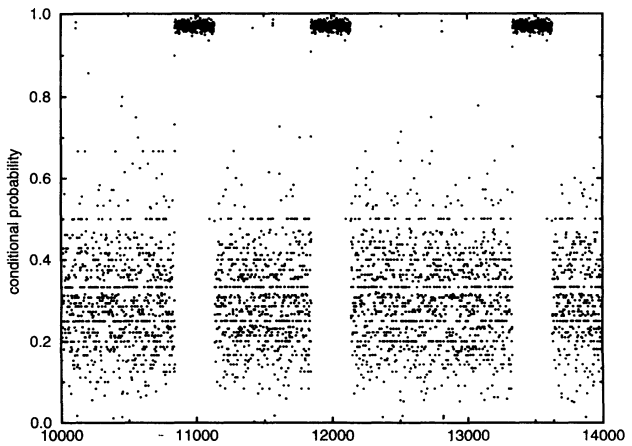

FIG. 3. Part of a simulated process with $L = 10^6$, $\rho = \frac{1}{3000}$, and $l = 300$. The estimated conditional probabilities from the preceding 8-words to the next letter are plotted vertically.

by such plots. Moreover, the method is robust against "mutations" as demonstrated in Fig. 4.

Figure 4 illustrates also the n dependence of the transition probabilities. For small n, even within a repeat, no definite prediction about the next letter is possible since ρ is not larger than $\frac{1}{\lambda^n}$. Consequently, there should be a crossover at n_c given by

$$\rho = \frac{1}{\lambda^{n_c}}, \tag{27}$$

leading to

$$n_c = \frac{\log_2 \frac{1}{\rho}}{\log_2 \lambda}. \tag{28}$$

Hence we have a critical word length n_c for a given probability ρ which can be interpreted as the number of letters necessary to identify the repeat. For $n \ll n_c$, the repeat looks like a random text, whereas for $n \gg n_c$ the repeat can be easily recognized with the aid of conditional probabilities. For $\rho = \frac{1}{3000}$, we find, for example, $n_c \approx 5.8$, which explains that the repeats are hardly visible for $n = 3$ in Fig. 4.

Similar considerations are valid for the beginning of a repetitive sequence. It is intuitively clear that the first letter cannot be predicted from the preceding n-word. For words with overlap k with the repeat, we found

$$p(A_j \mid \mathcal{S}_i^{(n)}) = \frac{\frac{\rho}{\lambda^{n-k}} + \frac{d-l-n}{\lambda^{n+1}}}{\frac{\rho}{\lambda^{n-k}} + \frac{d-l-n+1}{\lambda^n}}. \tag{29}$$

The k dependence of this expression resembles a Fermi function: There is a stepwise increase of the conditional probabilities from $\frac{1}{\lambda}$ to 1 for increasing k. For $\frac{\rho}{\lambda^{n-k}} \gg \frac{1}{\lambda^n}$, the conditional probability approaches 1. Thus we can define (somewhat arbitrarily) a critical k_c from

$$\frac{\rho}{\lambda^{n-k_c}} = \frac{1}{\lambda^n}. \tag{30}$$

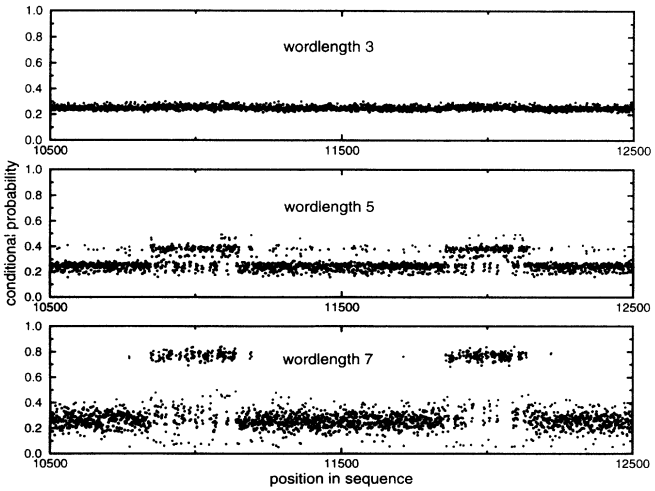

FIG. 4. Same as in Fig. 3, but all copies of the repeat are mutated at 30 randomly chosen sites. $n = 3$, 5, and 7 (from top to bottom).

This threshold means that the repeat has been recognized for

$$k > k_c = \frac{\log_2 \frac{1}{\rho}}{\log_2 \lambda} \equiv n_c. \tag{31}$$

Consequently, the first k_c letters of a repeat are taken as unpredictable. Again, the quantity $k_c \equiv n_c$ appears since it represents the critical number of letters required to identify a repeat with a given frequency of occurrence ρ.

Using this threshold k_c, we can divide all letters of the process under consideration into two classes: All letters in the random sea together with the first k_c letters of each copy of a repeat $\mathcal{R}(\rho, l)$ are assumed to be unpredictable. The other letters within the repetitive sequences are predictable, i.e., we set

$$\log_2 p(A_j \mid \mathcal{S}_i^{(n)}) = 0. \tag{32}$$

Since the differential entropy h_n can easily be obtained by averaging the logarithm of these conditional probabilities, the values h_n can be approximated as weighted sums of 2 bits in the case of unpredictable and 0 bits for predictable symbols:

$$h_n = 2\,[1 - \rho(l - k_c)\Theta(n - k_c),]$$

$$\Theta(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x \le 0, \end{cases} \tag{33}$$

$$k_c = \frac{1}{2} \log_2 \frac{1}{\rho} \quad \text{for } \lambda = 4.$$

The value $\rho(l - k_c)$ is just the fraction of predictable letters which approaches the total fraction of repeats ρl for $l \gg k_c$. The Θ function reflects the fact that the repetitive sequences are "recognizable" only for $n > k_c$.

Formula (33) gives the lowering of the h_n for large n and, moreover, the approximate value k_c of a decrease of the entropies h_n. The dashed line in Fig. 2 derived from the above formula is indeed a reasonable approximation of the actual decay of the entropies.

In the Appendix, we give another derivation of the entropy of the source $h = \lim_{n\to\infty} h_n$, which is in accordance with Eq. (33). Equation (33) essentially predicts the following: For a given repeat $\mathcal{R}(\rho, l)$, the differential entropies h_n decay at $n = \frac{1}{2} \log_2 \frac{1}{\rho}$ from 2 to about $2 - \rho l$ bits. In order to demonstrate the relevance of these rules, we present several results from simulations and analytical calculations in Figs. 5 and 6. Figure 5 shows the entropy decay for fixed probability ρ, but with fractions of repeats ρl from 1% to 40%. There is indeed a decay of the entropies at around $n = k_c \approx 5$ to values somewhat above $2 - \rho l$. Figure 6 illustrates that the crossover n_c depends logarithmically on the probability ρ. The predicted critical values k_c for the curves in Fig. 6 are at about $n = 3.3$, 5.0, and 6.6. As expected, the decay of h_n is shifted due to the variation of ρ. We underscore that Figs. 5 and 6 demonstrate again the excellent agreement of estimated and calculated entropies for $n \le 8$.

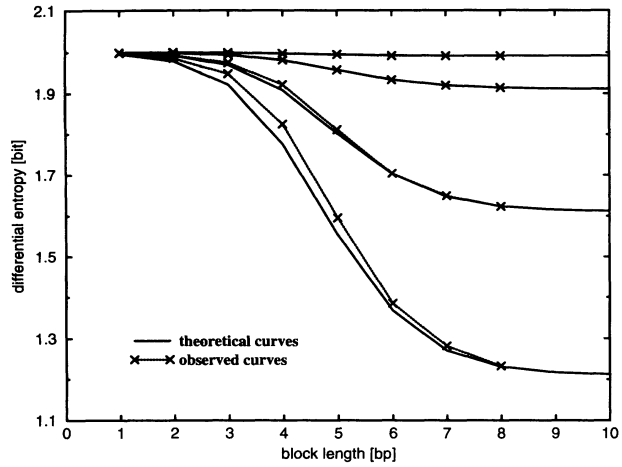

FIG. 5. Entropies from realizations ($L = 10^7$) of processes with dispersed repeats $\mathcal{R}$ ($\rho = 0.001$ and $l = 10, 50, 200, 400$) ($\times$) and from Eq. (22) (full lines). As in Fig. 2, the length correction (23) was applied.

VII. GENERALIZATIONS

So far, we mostly studied processes with dispersed identical copies of a single type of repeat. In this section, we will argue that most of our results also apply to more general situations such as "mutated" repeats and ensembles of several repeats.

Up to now, we have regarded the copies of a repeat as identical. However, in reality, they are quite similar, but not identical. For example, Alu repeats exhibit 87% homology to a consensus sequence [3]. In Fig. 4 we have demonstrated that 10% mutations have only minor effects on conditional probabilities. The modifications of our entropy calculations can be treated as follows. Let us assume that a fraction ϵ of a repeat is mutated at random to another symbol. Then we have a probability $1 - \epsilon$ to find the original symbol of the consensus repetitive sequence at any site of the repeat. Consequently, the probability to find the "right" word $\mathcal{S}_i^{(n)}$ is modified to

$$\tilde{P}(n) = \rho(1-\epsilon)^n + \frac{F(0)}{\lambda^n}. \tag{34}$$

Now the related conditional probability reads, for sufficiently large n,

$$\tilde{p}(A_j \mid \mathcal{S}_i^{(n)}) = \frac{\rho(1-\epsilon)^{n+1} + \frac{(d-l-n)}{\lambda^{n+1}}}{\rho(1-\epsilon)^n + \frac{(d-l-n+1)}{\lambda^n}} \approx 1 - \epsilon. \tag{35}$$

Hence the probability to find the right continuation is $1 - \epsilon$ (as intuitively expected) and any of the other continuations is found with probability $\frac{\epsilon}{3}$. Hence the entropy contribution (uncertainty) of a letter within the repeat is not zero, but

$$\Delta\tilde{h}_{\text{rep}} = -(1-\epsilon)\log_2(1-\epsilon) - \epsilon\log_2\frac{\epsilon}{3}. \tag{36}$$

For $\epsilon = 0.1$, this yields, for example, about 0.6 bit. In this way, the entropy decay is somewhat reduced due to mutations. Figure 7 shows the effect of random mutations on differential entropies.

The model of randomly dispersed copies of repeats is certainly valid for LINES and SINES (long and short interspersed elements). Besides the dispersed repeats, diverse tandemly repeated sequences are also known [1–6]. For example, human DNA exhibits very long series of CCCTAACCCTAACCCTAA... in the telomere region of chromosomes [6]. Another type of tandem repeats are clustered segments encoding histones or RNAs. Sea urchin DNA contains, e.g., nearly 1000 tandemly repeated copies of histone genes with a length of about 6300 bases [6]. Our model could easily be adapted to

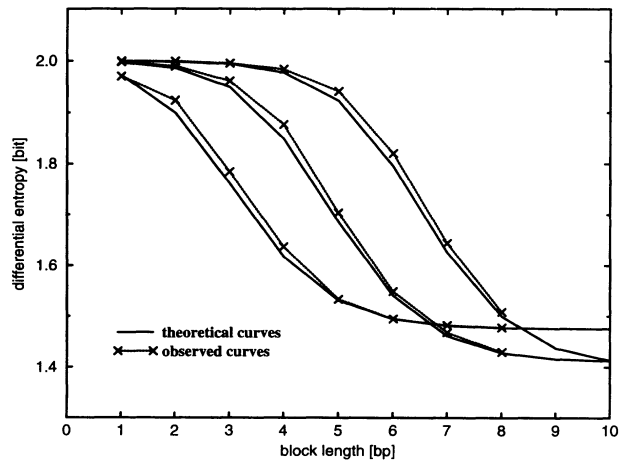

FIG. 6. Entropies from simulated processes ($\times$), ($L = 10^7$) and from Eq. (22) (full lines) with a fixed fraction of repeats $\rho l = 30\%$; from left to right $\rho = 0.01$, $\rho = 0.001$, and $\rho = 0.0001$.

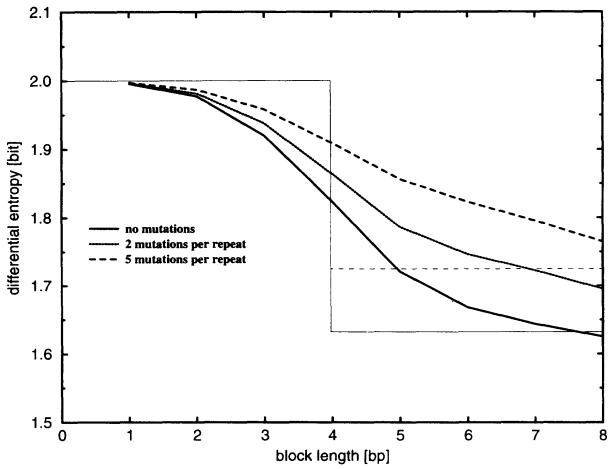

FIG. 7. Entropies of simulated processes with $L = 10^7$, $\rho = 0.004$, and $l = 50$. Each copy of the repeat was randomly "mutated" at 5, 2, and 0 sites (from top to bottom). Since "neutral mutations" were allowed, i.e., a symbol could be "mutated" to itself, the effective mutation rates ϵ were 7.5% and 3%, respectively. The thin lines mark the approximated theoretical values h_n according to Eqs. (33) and (36).

such situations of tandem repeats. The number of overlapping words to nonrepetitive DNA would decrease in such cases. For short repeats such as CCCTAA, the assumption of a random base composition of repeats has to be modified. However, such specific models are not the aim of this paper.

As a generalization of single repeats $\mathcal{R}(\rho, l)$ we will now discuss ensembles $\mathcal{R}_j(\rho_j, l_j)$. As long as overlaps and insertions of repeats into repeats can be neglected, the decay of the entropies can be estimated by superposition from Eq. (33)

$$h_n = 2\,[1 - \sum_j \rho_j (l_j - k_{cj})\Theta(n - k_{cj})],$$
$$k_{cj} = \frac{1}{2}\log_2 \frac{1}{\rho_j} \qquad \text{with} \quad \lambda = 4. \tag{37}$$

Figure 8 is obtained from a process with a short, but frequent repeat $\mathcal{R}_1(0.02, 10)$ and another rare, but long repeat $\mathcal{R}_2(0.001, 200)$. The entropies display a superposition of a decay at around $k_{c1} \approx 2.8$ and $k_{c2} \approx 5$. The full line is again derived from Eq. (22). The dashed line refers to the simplified formula (33).

The total fraction of repeats provides again an upper bound of the sum in Eq. (37) giving

$$h_n > 2\left[1 - \sum_j \rho_j l_j\right]. \tag{38}$$

The n dependence, i.e., the decay to the entropy of the source h, is encoded in the Θ function. Very frequent repeats are taken into account even for small n, whereas seldom repeats govern the decay of the h_n for larger n.

Since $k_{cj} \equiv n_{cj}$ depends logarithmically on the frequency of a certain repeat $\mathcal{R}_j$, a relatively fast decay of the differential entropies is observed. Hence long repeats do not induce automatically long-range correlations. This point will be discussed further in Sec. IX.

VIII. APPLICATION TO SPECIFIC DNA SEQUENCES

In this section, we present some preliminary results concerning the role of repeats in real DNA sequences. As a first example, we discuss sea urchin DNA. Measurements of the reassociation kinetics of single strand to double strand DNA indicate that repeats constitute about 50% of the whole DNA [3]. A reasonable fit of the experimental curves is obtained by the assumption of two kinds of repeats: 19% of the DNA appear about 160 times and 27% about 10 times [3]. The whole genome contains about 8.6×10^8 base pairs (bp), i.e., $\rho_1 \approx 1.9 \times 10^{-7}$ and $\rho_2 \approx 1.2 \times 10^{-8}$. From our approach, we would expect a stepwise decay of the entropies h_n at around $n = k_{c1} \approx 11$ and $n = k_{c2} \approx 13$ (comparable to the two steps in Fig. 8). The asymptotic value h can be as low as 1 bit if the repeats are relatively long [see Eq. (38)]. This example illustrates that an understanding of the role of repeats allows one to draw several conclusions from rather limited experimental knowledge.

In Fig. 9 the detection of repeats with the aid of conditional probabilities is exemplified for the DNA of the Epstein-Barr virus. For the displayed part of the DNA sequence, transitions towards increased probabilities can be seen at 7421 and 12 001 bp, where repetitive regions start according to the documentation. The detection of such peculiarities of conditional probabilities implies changes of the entropy of the source. In the following, we discuss entropy estimations of real DNA sequences even though dramatic finite sample effects have to be taken into account for $n > 6$. Figure 10 shows entropies of two long DNA sequences: the yeast chromosome III (315 338 bp) and the genome of the Epstein-Barr virus (17 2281 bp). The yeast DNA contains relatively few repeats (the words with length $l \geq 20$ appearing twice constitute about 4% of the DNA). In contrast, the documented repetitive regions in the viral DNA occupy 25.3% of the whole sequence (there are, for example, 12 copies of a subsequence of a length $l = 3072$).

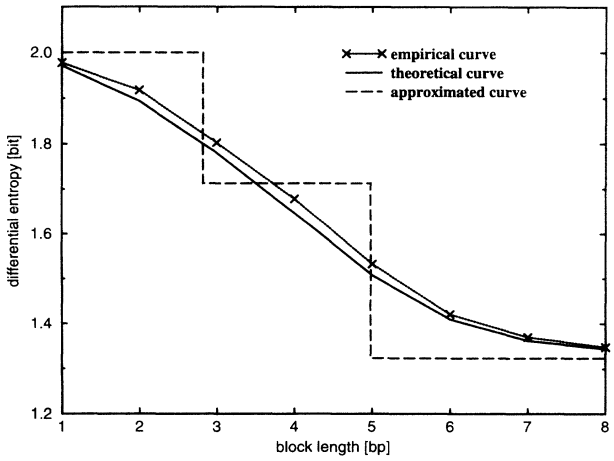

FIG. 8. Differential entropies for a superposition of two types of repeats.

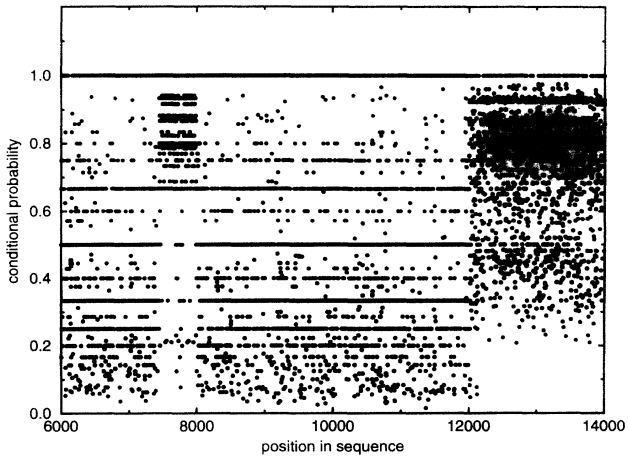

FIG. 9. Conditional probabilities for a part of the Epstein-Barr virus DNA indicating repeat regions.

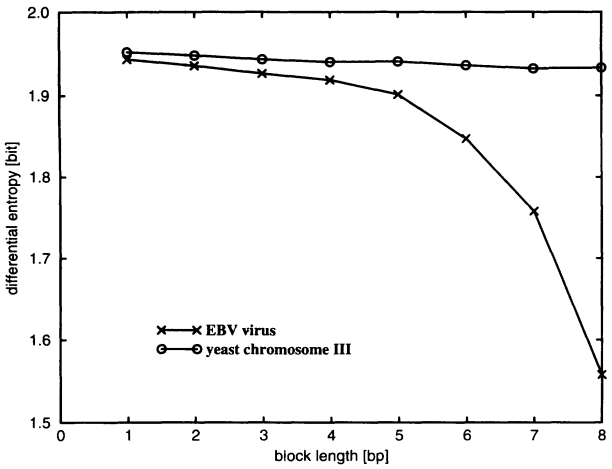

FIG. 10. Estimated entropies for the yeast chromosome III (×) and the Epstein-Barr virus (○) using finite sample corrections discussed in Refs. [14,16,17].

From our theoretical considerations, we would predict a moderate decay for the yeast DNA due to repeats (at most 0.08 bit) at $n = k_c = -\frac{1}{2}\log_2 \frac{2}{315\,338} \approx 8.6$. A drastic decay by about 0.5 bit at $n = -\frac{1}{2}\log_2 \frac{12}{172\,281} \approx 6.9$ is expected for the virus DNA.

Direct estimations of the entropies h_n from word frequencies are seriously affected by finite sample effects (note that $4^{10} \approx 10^6$ combinations of length 10 are possible). However, due to sophisticated finite length corrections [14,16,17] some estimations are available (see [17] for details). The corrected values are consistent with our predictions: There is nearly no decay of the yeast entropies and a sharp decay of the entropies for the Epstein-Barr virus at around $n = 7$.

Our models can be easily updated to reflect the statistical properties of DNA sequences more accurately. For example, the real base, codon, and dicodon usage may be approximated by Markov processes of fifth order. Moreover, specific properties of repeats (e.g., enlarged AT content or tandemly repeated copies) may be included in further studies.

IX. SUMMARY AND DISCUSSION

The information content of biosequences can be quantified by word entropies. However, the direct calculation of higher-order entropies is limited by the finite length of available sequences. Therefore, we constructed stochastic processes that reflect essential features of DNA. In a similar spirit, Young and Crutchfield [39] recently constructed appropriate "ϵ-machines" for the estimation of entropies and Li developed "expansion-modification models" as generators of long-range correlations [18].

The basic assumptions of our model are that (a) DNA outside repeat regions is close to a Bernoulli process, which is substantiated in Sec. III by precise estimations of low-order entropies [8–10,12,14], and (b) in addition to nonrepetitive DNA, relatively long copies of "repeats" $\mathcal{R}_j$ constitute a significant fraction of eukaryotic DNA. Several classes of repetitive sequences such as tandem repeats, SINES, and LINES are well documented [1–6]. According to our statistical approach, we may define repeats fairly generally: Any subsequence appearing much more frequently than expected from the base composition may be called "repeat" if it is not merely part of a longer repeat. In the framework of our model, one has to require for a repeat

$$\rho \gg \frac{1}{\lambda^l}. \tag{39}$$

For $\lambda = 4$ this condition leads to

$$l \gg \frac{1}{2}\log_2 \frac{1}{\rho} = k_c. \tag{40}$$

If Eq. (39) holds, the discussed conditional probabilities tend toward one (see Fig. 3) and consequently entropies are reduced due to the repeat.

We note that words of a length of $l = 20$ that appear but twice in the whole human genome (about 3.5×10^9 nucleotides) already fulfill this requirement:

$$\rho = \frac{2}{3.5 \times 10^9} \gg \left(\frac{1}{4}\right)^{20}. \tag{41}$$

The above general definition of repeats also includes, for example, multiple copies of genes and long repetitive words within pseudogenes. This appears reasonable since such repetitions reduce the information content h as well.

It was shown that for our model, introduced in Sec. III, fairly accurate formulas could be derived for the word distributions and the entropies H_n. Such constructed processes with nontrivial structure (the rank order statistics exhibits a plateau and a power-law decay to a pedestal) may also serve as test cases for finite sample corrections [12–14,23].

A simplified formula for the differential entropies h_n has been derived by analyzing conditional probabilities. This approach revealed much insight into the decay properties of the entropies h_n. The main conclusions are the following.

(a) As intuitively expected, the reduction of the entropy of the source h is intimately related to the total fraction of all repeats [see Eq. (38)].

(b) Since the first $k_{cj} = \frac{1}{2}\log_2 \frac{1}{\rho_j}$ letters of a repeat are hardly predictable, we can improve the lower bound in Eq. (38). For repeats of the Alu family $\mathcal{R}_{\mathrm{Alu}}(\rho \approx \frac{1}{4000}, l \approx 300)$ we obtain, for example, $k_{c\mathrm{Alu}} \approx 6$, i.e., the rest can be considered as predictable.

(c) Another result concerns the decay of the entropies h_n to their limit h. Repetitive subsequences $\mathcal{R}_j(\rho_j, l_j)$ can be detected for $n > k_{cj} = \frac{1}{2}\log_2 \frac{1}{\rho_j}$ and hence the corresponding decay of the h_n is found at $n = k_{cj}$.

It was argued in Sec. VII that repeats do not lead to long tails of the h_n decay. This can be illustrated as follows: Even if a word appears only twice in the whole human genome ($\rho \approx 6 \times 10^{-8}$) the corresponding decay is located around $k_c = 13.7$. Hence the asymptotic value h is reached for relatively small n.

Consequently, randomly distributed repeats do not seem to be candidates to introduce long correlations in biosequences. However, it has been reported that repeats might be distributed in a nonrandom manner [7]. In the case of "short-period interspersions" repeats of length 300 alternate with intervening nonrepetitive sequences of about 1000 bases [3]. In this way repeats might induce long correlations.

Moreover, repeats may contribute to long correlations in DNA sequences due to a substructure. LINES often contain "long terminal repeats" at their ends. Keeping in mind that more than 20 000 LINES extending over several kilobases exist in genomes, such "repeats within repeats" may induce detectable long correlations over thousands of bases as found empirically [19,21,22,16].

Summarizing, we emphasize that the direct statistical analysis of biosequences can profit from careful studies of appropriate model processes as demonstrated in this paper.

ACKNOWLEDGMENTS

We acknowledge support by the Deutsche Forschungsgemeinschaft (H.H.) and the Stiftung Volkswagenwerk (A.O.S.).

APPENDIX: AN ALTERNATIVE CALCULATION OF THE ENTROPY

An approximate formula for the entropy of the source h is derived for a process with independent equidistributed letters A, C, G, and T and randomly dispersed repeats $\mathcal{R}(\rho, l)$. Overlaps of copies of $\mathcal{R}(\rho, l)$ are neglected. Such a process can be considered as a Bernoulli process with the $\lambda = 5$ symbols A, C, G, T, and $\mathcal{R}$ and their probabilities

$$p_i = \frac{1-\rho}{4} \qquad (i = 1, 2, 3, 4),$$
$$p_5 = \rho. \tag{A1}$$

Consequently, the entropy of the source is given by

$$h^{\lambda=5} = \sum_{i=1}^{4} -\frac{1-\rho}{4} \log_2 \frac{1-\rho}{4} - \rho \log_2 \rho \approx 2 + \rho \log_2 \frac{1}{\rho}. \tag{A2}$$

In order to get the mean information per letter of the original string with $\lambda = 4$ letters we have to renormalize the above entropy by dividing it by the mean number of nucleotides per symbol of the process with $\lambda = 5$ [40]:

$$h = \frac{h^{\lambda=5}}{1-\rho+\rho l} \approx 2 \left[1 - \rho \left(l - \frac{1}{2} \log_2 \frac{1}{\rho} \right) \right]. \tag{A3}$$

In this way, an alternative estimation of the decrease of the entropy of the source due to repeats is calculated, which is in perfect harmony with Eq. (33).

[1] E. N. Trifonov and V. Brendel, *Gnomic — A Dictionary of Genetic Codes* (Balaban, Rehovot, 1986).
[2] E. N. Trifonov, Bull. Math. Biol. 51, 417 (1989).
[3] B. Lewin, *Genes V* (Oxford University Press, New York, 1994).
[4] R. Knippers, P. Philippsen, K. P. Schaefer, and E. Fanning, *Molekulare Genetik* (Georg Thieme Verlag, Stuttgart, 1990).
[5] J. D. Watson, M. Gilman, J. Witkowski, and H. Zoller, *Recombinant DNA* (Freeman, New York, 1992).
[6] P. Berg and M. Singer, *Dealing with Genes* (University Science Books, Mill Valley, CA, 1992).
[7] N. A. Kolchanov and H. A. Lim, *Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution* (World Scientific, Singapore, 1994).
[8] L. L. Gatlin, *Information Theory and the Living System* (Columbia University Press, New York, 1972).
[9] G. W. Rowe and L. E. H. Trainor, J. Theor. Biol. 101, 151 (1983).
[10] M. S. Koref, Ph.D. thesis, Humboldt-University, Berlin, 1987.
[11] W. Ebeling, R. Feistel, and H. Herzel, Phys. Scr. **35**, 761 (1987).
[12] H. Herzel, Syst. Anal. Mod. Sim. **5**, 435 (1988).
[13] P. Grassberger, Phys. Lett. A **128**, 369 (1988).
[14] A. O. Schmitt, H. Herzel, and W. Ebeling, Europhys. Lett. **23**, 303 (1993).
[15] D. Wolpert and D. Wolf, Santa Fe Institute Report No. TR-93-07-046, 1993 (unpublished).
[16] H. Herzel, A. O. Schmitt, and W. Ebeling, Chaos, Solit. Fractals **4**, 97 (1994).
[17] A. O. Schmitt, H. Herzel, and W. Ebeling, Institute of Physics, Humboldt University Berlin Report No. ITP-349, 1994 (unpublished).
[18] W. Li, Int. J. Bif. Chaos **2**, 137 (1992).
[19] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 186 (1992).
[20] B. Borštnik, D. Pumpernik, and D. Lukman, Europhys. Lett. **23**, 389 (1993).
[21] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[22] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[23] H. Herzel, W. Ebeling, A. O. Schmitt, and M. A. Jiménez-Montaño, in *From Simplicity to Complexity in Chemistry*, edited by A. Müller, A. Dress, and F. Vögtle (Vieweg, Braunschweig, in press).
[24] C. E. Shannon, Bell Syst. Techn. J. **27**, 379 (1948).
[25] A. M. Jaglom and I. M. Jaglom, *Wahrscheinlichkeit und Information* (Verlag der Wissenschaften, Berlin, 1965).
[26] J. D. Farmer, Z. Naturforsch. A **37**, 1304 (1982).
[27] J. P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).
[28] P. Grassberger, Int. J. Theor. Phys. **25**, 907 (1986).
[29] W. Ebeling and G. Nicolis, Europhys. Lett. **14**, 191 (1991); Chaos, Solit. Fractals **2**, 635 (1992).
[30] B. McMillan, Ann. Math. Statist. **24**, 196 (1953).

[31] A. Khinchin, *Mathematical Foundations of Information Theory* (Dover, New York, 1967).
[32] A. Csordás, G. Győrgyi, P. Szépfalusy, and T. Tél, Chaos **3**, 31 (1993).
[33] J. W. Fickett and Chang-Shung Tung, Nucl. Acid Res. **20**, 6441 (1992).
[34] R. Staden, Nucl. Acid Res. **12**, 551 (1984).
[35] J. W. Fickett, Nucl. Acid Res. **10**, 5303 (1982).
[36] A. K. Konopka and J. Owens, Gene Anal. Technol. Appl. **7**, 35 (1990).
[37] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Reading, MA, 1963).
[38] W. Li, IEEE Trans. Inf. Theory **38**, 1842 (1992).
[39] K. Young and J. P. Crutchfield, Chaos, Solit. Fractals **4**, 5 (1994).
[40] J. Freund and H. Herzel, Chaos, Solit. Fractals (to be published).