Statistical analysis of the DNA sequence of human chromosome 22

Dirk Holste,[1] Ivo Grosse,[2] and Hanspeter Herzel[3]
[1]*Department of Theoretical Biophysics, Humboldt University Berlin, Invalidenstrasse 42, D-10115, Berlin, Germany*
[2]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724*
[3]*Institute for Theoretical Biology, Humboldt University Berlin, Invalidenstrasse 43, D-10115, Berlin, Germany*
(Received 19 April 2001; published 26 September 2001)

We study statistical patterns in the DNA sequence of human chromosome 22, the first completely sequenced human chromosome. We find that (i) the 33.4×10^6 nucleotide long human chromosome exhibits long-range power-law correlations over more than four orders of magnitude, (ii) the entropies H_n of the frequency distribution of oligonucleotides of length n (n-mers) grow sublinearly with increasing n, indicating the presence of higher-order correlations for all of the studied lengths $1\leqslant n\leqslant 10$, and (iii) the generalized entropies $H_n(q)$ of n-mers decrease monotonically with increasing q and the decay of $H_n(q)$ with q becomes steeper with increasing $n\leqslant 10$, indicating that the frequency distribution of oligonucleotides becomes increasingly nonuniform as the length n increases. We investigate to what degree known biological features may explain the observed statistical patterns. We find that (iv) the presence of interspersed repeats may cause the sublinear increase of H_n with n, and that (v) the presence of monomeric tandem repeats as well as the suppression of CG dinucleotides may cause the observed decay of $H_n(q)$ with q.

DOI: 10.1103/PhysRevE.64.041917 PACS number(s): 87.14.Gg, 87.10.+e, 02.50.−r, 05.40.−a

I. INTRODUCTION

The study of statistical patterns in DNA sequences is important as it may improve our understanding of the organization and evolution of life on the genomic level [1–7]. As the DNA sequences of the first human chromosomes have become available [8–11], statistical patterns can be comprehensively analyzed on a chromosomal scale [8–15]. The findings of long-range correlations in DNA sequences have attracted much attention [16–19], and attempts have been made to relate those findings to known biological features such as the presence of triplet periodicities in protein-coding DNA sequences [20,21], the evolution of DNA sequences [22], the length distribution of protein-coding regions [23], or the expansion of simple sequence repeats [24]. While it is possible to statistically explain most of the observed long-range correlations in chromosomes with a large fraction of coding sequences [25], this explanation fails in the human genome, where less than 2% of the DNA is used for encoding proteins [10,11]. On the other hand, about 50% of the human genome contains intermediate and highly repetitive DNA [10]. Hence we may expect that many statistical patterns in human DNA sequences are due to the presence of repeats.

This paper is divided into two parts. In the first part (Secs. II and III) we present an empirical study of statistical patterns in the 33.4×10^6 nucleotide long DNA sequence of human chromosome 22, and in the second part (Secs. IV–VI) we investigate to which degree those statistical patterns might be explained by known biological features, such as the presence of Alu repeats and monomeric tandem repeats or the suppression of CG dinucleotides. Specifically, we study in the first part (i) long-range nucleotide-nucleotide correlations by computing the mutual information function as well as three binary autocorrelation functions, (ii) higher-order (n-point) correlations by computing the entropies H_n as a function of the oligonucleotide length n, and (iii) the nonuniformity of the frequency distribution of oligonucleotides of length n by computing the generalized entropies $H_n(q)$. In the second part we introduce a simple stochastic model that incorporates the presence of interspersed repeats. We study how the presence of repeats influences (i) the mutual information function $I(k)$ and the autocorrelation function $C(k)$, (ii) the increase of the entropy H_n with n, and (iii) the decay of the generalized entropies $H_n(q)$ as a function of q. We find that the presence of interspersed repeats alone cannot reproduce the observed entropies $H_n(q)$, so we study two extensions of the model by incorporating the presence of monomeric tandem repeats and the suppression of CG dinucleotides.

II. LONG-RANGE TWO-POINT CORRELATIONS

DNA can be considered as a sequence composed of $\lambda=4$ nucleotides $\{A_1,A_2,A_3,A_4\}\equiv\{\mathrm{A,C,G,T}\}$, in which A refers to adenine, C refers to cytosine, G refers to guanine, and T refers to thymine. We denote by p_i $(i=1,2,\ldots,\lambda)$ the relative frequency of A_i in human chromosome 22 and by $p_{ij}(k)$ the relative frequency of the pair of nucleotides A_i and A_j that are spaced by $k-1$ nucleotides. Under the assumption that the DNA sequence of human chromosome 22 can be considered as a realization of a stationary and ergodic stochastic process, one may associate with p_i the probability of finding at any given sequence position the nucleotide A_i, and one may associate with $p_{ij}(k)$ the joint probability of finding at any given positions the pair of nucleotides A_i and A_j that are spaced by $k-1$ nucleotides. Two nucleotides spaced by a distance $k-1$ are defined to be statistically independent if, and only if, $p_{ij}(k)=p_ip_j$ for all i and j. One natural and intuitive quantity that measures any deviation from statistical independence is the mutual information function [26]

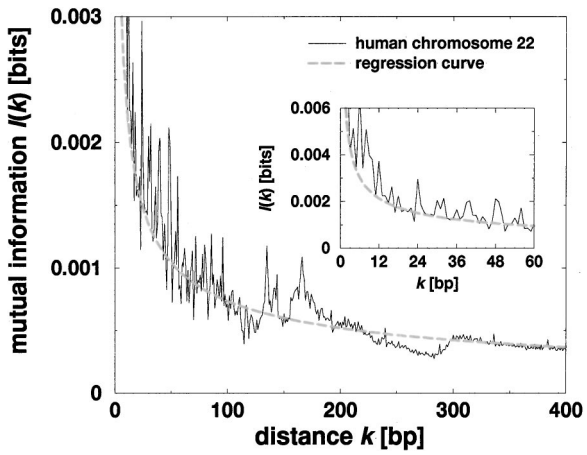

FIG. 1. Mutual information function $I(k)$ of the DNA sequence of human chromosome 22. We find that $I(k)$ shows clear correlations over several hundred bp, while it shows only weak periodicities of length 3 [27].

$$I(k) \equiv \sum_{i,j=1}^{\lambda} p_{ij}(k) \log_2 \frac{p_{ij}(k)}{p_i p_j}, \tag{1}$$

which quantifies the amount of information (in units of bits) that one can obtain about the identity of nucleotide A_j by learning the identity of nucleotide A_i located $k-1$ nucleotides upstream of A_j.

Figure 1 shows $I(k)$ of human chromosome 22 for $k=1,\ldots,400$ base pairs (bp). We find that $I(k)$ is significantly greater than zero and decays with increasing k. The inset of Fig. 1 shows that there are no pronounced period-3 oscillations, which confirms our expectation that the small fraction of only 3% of coding DNA sequences in human chromosome 22 [8] can only weakly influence the two-point correlations computed from the entire chromosome.

In order to study specific long-range correlations in human chromosome 22, we define for nucleotides A_i and A_j the deviation from statistical independence by the dependence matrix $D_{ij}(k) \equiv p_{ij}(k) - p_i p_j$, and we define the autocorrelation function [21],

$$C(k) \equiv \sum_{i,j=1}^{\lambda} a_i D_{ij}(k) a_j, \tag{2}$$

as the bilinear form of the matrix $D_{ij}(k)$ with the indicator variables $a_i \in \{0,1\}$. Specifically, the choice of $a_1 = a_3 = 1$ and $a_2 = a_4 = 0$ defines the purine-purine (R-R) autocorrelation function $C_{\mathrm{RR}}(k)$ [28], which has been intensively studied by random-walk analyses [29,30]. Another biologically relevant classification of nucleotides is the classification into weakly binding nucleotides (W=A or T) and strongly binding nucleotides (S=C or G) [31], and we obtain the corresponding autocorrelation function $C_{\mathrm{WW}}(k)$ by setting $a_1 = a_4 = 1$ and $a_2 = a_3 = 0$. $C_{\mathrm{MM}}(k)$ defines the third binary autocorrelation function for amino nucleotides (M=A or C), and it is obtained for $a_1 = a_2 = 1$ and $a_3 = a_4 = 0$. By Taylor-

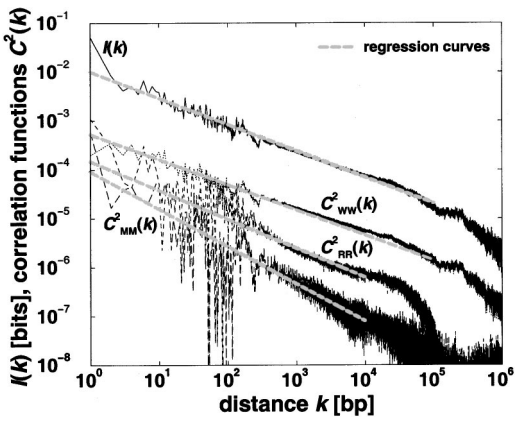

FIG. 2. Double-logarithmic plot of $I(k)$ and three binary autocorrelation functions $C^2_{\mathrm{WW}}(k)$, $C^2_{\mathrm{RR}}(k)$, and $C^2_{\mathrm{MM}}(k)$ of the DNA sequence of human chromosome 22. We find that $I(k)$ and $C^2_{\mathrm{WW}}(k)$ display power-law correlations over five orders of magnitude, while $C^2_{\mathrm{RR}}(k)$ and $C^2_{\mathrm{MM}}(k)$ decay to the noise level after about four orders of magnitude. The gray dashed curves represent the least-squares regressions with exponents $2\gamma = 0.53$ (I) and $2\gamma = 0.51$ (C^2_{WW}) using $k = 1,\ldots,10^5$ bp, and $2\gamma = 0.59$ (C^2_{RR}) and $2\gamma = 0.76$ (C^2_{MM}) using $k = 1,\ldots,10^4$ bp. The bias of $I(k)$ is of the order of 10^{-7} bits [32,33], and the bias of $C^2(k)$ is of the order of 10^{-9} [34].

expanding $I(k)$, we obtain (neglecting higher-order terms in D_{ij}) the following relation between $I(k)$ and the dependence matrix $D_{ij}(k)$:

$$I(k) \propto \frac{1}{2 \ln 2} \sum_{i,j=1}^{\lambda} \frac{D_{ij}^2(k)}{p_j p_j}, \tag{3}$$

which implies that a power-law decay of $C(k) \sim k^{-\gamma}$ leads to a power-law decay of $I(k) \sim k^{-2\gamma}$ [21].

Figure 2 shows a double-logarithmic plot of $I(k)$ and of all three binary autocorrelation functions $C^2_{\mathrm{RR}}(k)$, $C^2_{\mathrm{WW}}(k)$, and $C^2_{\mathrm{MM}}(k)$. We find that the decay of both $I(k)$ and $C^2_{\mathrm{WW}}(k)$ can be quantified by power laws up to $k = 100\,000$ bp with exponents $2\gamma \approx 0.5$, while the decays of $C^2_{\mathrm{RR}}(k)$ and $C^2_{\mathrm{MM}}(k)$ can be quantified by a power law up to $k = 10\,000$ bp. Figure 2 shows that R-R correlations are almost one order of magnitude weaker than W-W correlations and that M-M correlations are almost one order of magnitude weaker than R-R correlations, which is in agreement with previous observations of long-ranging G + C fluctuations [35–38] termed isochores [38]. In Sec. IV, we will study to which degree interspersed repeats can account for the decay of the mutual information function and the binary autocorrelation functions.

III. SHORT-RANGE OLIGONUCLEOTIDE DEPENDENCIES

In this section we analyze statistical dependencies within oligonucleotides of length $n = 1,2\ldots,10$, by computing the Shannon entropy H_n and the generalized Rényi entropies $H_n(q)$ of the distribution of oligonucleotides for $q \in [-20,20]$.

A. Entropy analysis

We denote an oligonucleotide (n-mer) of n consecutive nucleotides by $A_i^{(n)}$, and we denote the relative frequency of $A_i^{(n)}$ by $p_i^{(n)}$. Then the Shannon entropy of the frequency distribution $\{p_i^{(n)}\}$ is defined by [26]

$$H_n \equiv -\sum_{i=1}^{\lambda^n} p_i^{(n)} \log_2 p_i^{(n)}, \tag{4}$$

and measures the average uncertainty of an n-mer drawn randomly from the $p_i^{(n)}$ distribution. For a random uncorrelated sequence of λ equidistributed nucleotides, we obtain $H_n = n \log_2 \lambda$ bits. H_n is a frequently used statistical measure and has been applied to quantify the uncertainty in binding sites on DNA sequences [39], to the analysis of coding and noncoding regions [40], or to the decomposition of DNA sequences [41].

Let $p_{j|i}^{(1|n)} \equiv p(A_j|A_i^{(n)})$ denote the conditional probability of finding the $(n+1)$th nucleotide A_j following $A_i^{(n)}$. In analogy to Eq. (4), we define the entropy of the frequency distribution $\{p_{j|i}^{(1|n)}\}$ by

$$H_{1|n}(i) \equiv -\sum_{j=1}^{\lambda} p_{j|i}^{(1|n)} \log_2 p_{j|i}^{(1|n)}, \tag{5}$$

which measures the average uncertainty of A_j following $A_i^{(n)}$ in units of bits. For a random uncorrelated sequence of λ nucleotides, we obtain $H_{1|n}(i) = H_1$ for every oligonucleotide $A_i^{(n)}$.

The conditional entropy h_n is defined by [26]

$$h_n \equiv \sum_{i=1}^{\lambda^n} p_i^{(n)} H_{1|n}(i) = H_{n+1} - H_n, \tag{6}$$

which quantifies the average uncertainty of the $(n+1)$th nucleotide provided the preceding n nucleotides are known. The decay of the series h_n $(n=1,2,\ldots)$ reflects the statistical dependencies within oligonucleotides of length $n+1$. The quantity $h \equiv \lim_{n\to\infty} h_n$ is denoted as the entropy of the source, which plays an important role in coding theory [42], and is also related to the Kolmogorov entropy in dynamical systems theory [43]. A random uncorrelated sequence of λ nucleotides yields $h_1 = h_2 = \cdots = h = H_1$, a Markov chain of order m yields $h_m = h_{m+1} = \cdots = h$, and a periodic sequence with period p yields $h_p = h_{p+1} = \cdots = h = 0$ bits.

We study statistical dependencies within oligonucleotides of human chromosome 22 of length n by calculating the entropies H_n and the conditional entropies h_n. Due to the finite length $N = 33.4 \times 10^6$ bp, the estimates of H_n are biased and the bias of the entropy increases with the oligonucleotide length proportionally to $(\lambda^n - 1)/2N \ln 2$ [32,33]. Hence, weakly biased estimates of H_n can be obtained for $n \leqslant 10$.

Figure 3 shows H_n and h_n of the DNA sequences of human chromosome 22 for $n=1,2,\ldots,10$. We find that H_n is sublinearly increasing with increasing n, which indicates that there are weak, but nonvanishing, short-range statistical dependencies within n-mers for $n=1,2,\ldots,10$. The inset of

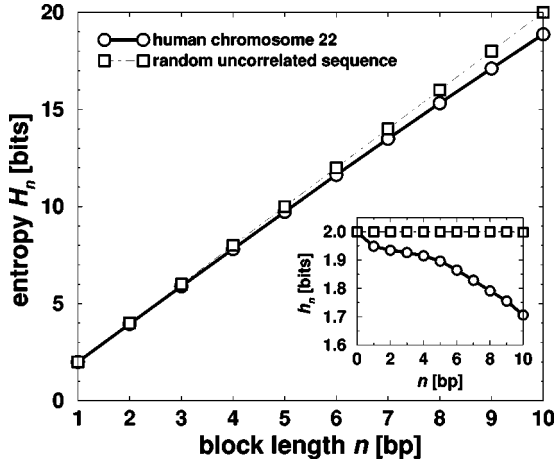

FIG. 3. Shannon entropy H_n of the DNA sequence of human chromosome 22 and a random uncorrelated sequence with $p_i = 1/4$. The inset shows the conditional entropy h_n $(h_0 \equiv H_1)$. H_n exhibits a sublinear growth with n, and $h_1 \approx 1.95$ bits is significantly smaller than 2 bits, which is related to pair correlations and to the nonuniformity of the $p_i^{(n)}$ distribution. We find that h_n decays below 1.75 bits for $n \geqslant 5$, which we will show might be explained by the redundancy due to repetitive sequences [44].

Fig. 3 shows the conditional entropies h_n versus n. We find that $h_n \geqslant 1.9$ bits for $n \leqslant 4$, which shows that there exist only weak statistical dependencies within 2-mers, 3-mers, and 4-mers. For increasing $n > 4$, we find a monotonic decay of h_n down to $h_{10} = 1.7$ bits. In Secs. IV, V, and VI, we will relate this decay of h_n with n to the presence of interspersed repeats, to the presence of monomeric tandem repeats, and to the suppression of CG dinucleotides.

B. Generalized entropy analysis

The Rényi entropies as a function of $q \in (-\infty, \infty)$ of the distribution of oligonucleotides of length n are defined by [45]

$$H_n(q) \equiv \frac{1}{1-q} \log_2 \left[\sum_{i=1}^{\lambda^n} (p_i^{(n)})^q \right]. \tag{7}$$

$H_n(q)$ is a widely used generalization of the Shannon entropy H_n, and has been applied to the analysis of nonlinear dynamical systems [46,47], time series [48,49], texts and DNA sequences [50,51], phylogenetic relationships [52], and in the context of the thermodynamic formalism [53]. The generalized entropies $H_n(q)$ are monotonically decreasing functions of q, and the decay of $H_n(q)$ with increasing q reflects the nonuniformity of the frequency distribution of oligonucleotides of length n. One can verify that $\lim_{q\to 1} H_n(q) = H_n$, and that $H_n(q)$ has the following noteworthy property: for large (small) values of q, the largest (smallest) probability $p_{\max}^{(n)}$ $(p_{\min}^{(n)})$ dominates $H_n(q)$, and in the limit $q \to \pm\infty$ we obtain $H_n(\infty) = -\log_2 p_{\max}^{(n)}$ and $H_n(-\infty) = -\log_2 p_{\min}^{(n)}$. For a random uncorrelated sequence of λ equidistributed nucleotides, the Rényi entropies are independent of q and equal to $H_n(q) = n \log_2 \lambda$ bits.

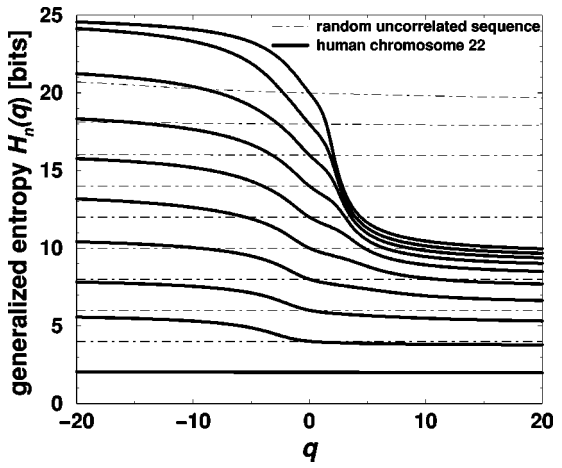

FIG. 4. Rényi entropies $H_n(q)$ of the DNA sequence of human chromosome 22 for $n=1,2,\ldots,10$ (from bottom to top) and a random uncorrelated sequence with $p_i=1/4$ [54]. We find that $H_1(q)$ is virtually independent of q, which indicates that the nucleotide distribution is approximately uniform. For $n\geqslant 2$ we find that $H_n(q)$ is dependent on q, which indicates that the $p_i^{(n)}$ distribution is nonuniform. For large q, we find that $H_n(q)$ assumes approximately 10 bits as n increases, while we find for small values of q that $H_n(q)$ is greater than expected for a random uncorrelated sequence with $p_i=1/4$.

We compute $H_n(q)$ as a function of q and n to characterize the nonuniformity of the $p_i^{(n)}$ distribution of oligonucleotides of lengths n. Figure 4 shows $H_n(q)$ of the DNA sequence of human chromosome 22 for $q\in[-20,20]$ and for $n=1,2,\ldots,10$. The bias of estimates of $H_n(1)$ transfers to estimates of $H_n(q)$ [55]. We find that $H_1(q)$ is almost independent of q, while $H_n(q)$ with $n\geqslant 2$ show a clear decay with increasing q, indicating that nucleotides are almost equidistributed, while the frequencies of di-, tri-, and longer oligonucleotides exhibit deviations from the equidistribution [56–58].

Figure 4 also shows that $H_n(20)\approx -\log_2 p_{\max}^{(n)}$ assumes about 10 bits for large n, which states that $p_{\max}^{(n)}\approx 10^{-3}$ becomes almost independent of n for $7\leqslant n\leqslant 10$. This finding implies that the most frequent oligonucleotide $A_{\max}^{(n)}$ can be extended by only one nucleotide, yielding the oligonucleotide $A_{\max}^{(n+1)}$ with relative frequency $p_{\max}^{(n+1)}\approx p_{\max}^{(n)}$ for $n\in[7,10]$. Indeed, we find for $n=1,2,\ldots,10$ that monomeric tandem repeats of poly-A and poly-T are the most frequent oligonucleotides in human chromosome 22, which is in agreement with previous observations [59,60].

In Secs. V and VI we will analyze systematically the effects of the presence of Alu repeats, the presence of monomeric tandem repeats, and the suppression of CG dinucleotides on short-range statistical dependencies within oligonucleotides.

IV. MODELING INTERSPERSED REPEATS

In this section we discuss some characteristics of repetitive sequences in human genomic DNA, introduce a simple stochastic model that incorporates the presence of randomly interspersed repeats, and illustrate some model predictions.

Up to 42% of human chromosome 22 is comprised of interspersed repeats. In order to test if and to which degree the presence of interspersed repeats may be responsible for the observed statistical patterns reported in Secs. II and III, we propose in Secs. IV–VI a repeat model and systematically compare the model predictions with experimental data.

Repetitive sequences are commonly classified according to their copy number into *intermediate repeats* and *highly repetitive DNA* [61]. Many repeats are associated with mobile elements that can copy themselves and insert their duplicate into a new location, by a process called retrotransposition. Retroposons are divided into short (SINEs) and long interspersed nucleic elements (LINEs) [62]. Typically, SINEs have no capability for retroposition on their own and are supposed to depend on LINEs for mobilization [62]. The evolutionary origin, role, and distribution of SINEs and LINEs in human DNA have received wide interest over recent years [61–63].

Alu repeats are an important family of SINEs found ubiquitously in mammalian genomes [64]. Human chromosome 22 contains about 2×10^4 Alu repeats, covering about 17% of the total chromosome. Members of the Alu family contain a common recognition sequence for the restriction enzyme AluI. They exhibit about 87% homology to a consensus sequence, which is about 300 bp long and consists of a dimeric structure flanked by direct repeats [2,3]. Most Alu subfamilies are old, inactive, and fixed in the human genome, while some younger subfamilies are thought to remain actively transposing. The biological role of Alu repeats is not yet fully understood. There are indications that Alu repeats may play a role in human genome organization [10,65], that they integrate at specific sites into their host genomes [66,64], or that they influence the regulation of enzyme activities [67].

Since the repeat density is about 42% and Fig. 1 shows almost no effect of the 3% protein-coding content in chromosome 22, we construct a simple stochastic model in the following two steps. (1) We characterize a repeat $\mathcal{R}$ by three parameters: (i) its relative frequency of occurrence ϱ, (ii) its length ℓ, and (iii) its fraction ε of random single-nucleotide substitutions. Consequently, $\varrho\ell$ gives the fraction of the total sequence covered by the repeat $\mathcal{R}$. One quarter of the substitutions replaces a nucleotide by itself, so effectively $\frac{3}{4}\varepsilon$ nucleotides mutate randomly to another one. For the sake of obtaining a simple description of each repeat $\mathcal{R}$, we neglect any internal correlations and construct a repeat $\mathcal{R}$ by a concatenation of independent nucleotides with $p_i=\frac{1}{4}$. (2) We intersperse a repeat ϱN times without overlap in a background sequence at randomly and uniformly chosen positions. In a first attempt, the background sequence we model by a Bernoulli sequence with equidistributed nucleotides.

While Alu repeats are well characterized and are known to have a length of about 300 bp, the length distribution of other SINEs and of LINEs is more complicated. Here we consider the presence of 17% Alu repeats in human chromosome 22 and discuss the superposition of different types of repeats afterward.

To elucidate the effect of interspersed repeats, let us consider a Bernoulli sequence of length $N(1-\varrho\ell)$ with a single

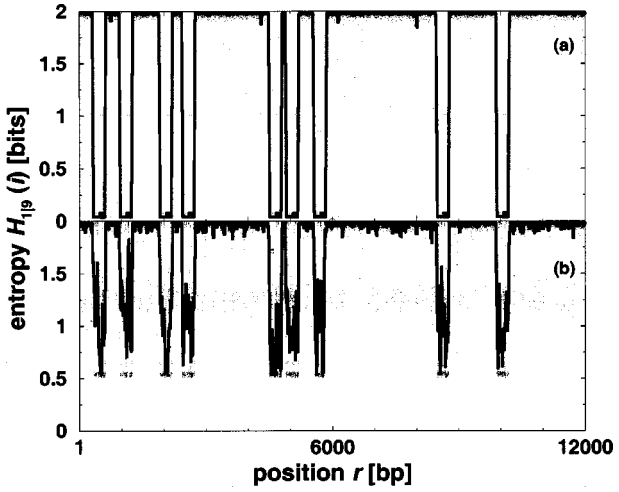

FIG. 5. Shannon entropy $H_{1|9}(i)$ for a realization of the repeat model with (a) $\varrho=0.0006$, $\ell=300$, and $\varepsilon=0$ and (b) $\varrho=0.0006$, $\ell=300$, and $\varepsilon=0.1$. We plot $H_{1|9}(i)$ from the preceding 9-mer i to the next nucleotide versus the sequence position r (dots) and display the 30 bp running average (solid line). (a) shows sharp drops of $H_{1|9}(i)$ from about 2 bits to nearly 0 bits, while (b) shows sharp drops of $H_{1|9}(i)$ by more than 1 bit when we mutate each repeat at randomly chosen positions.

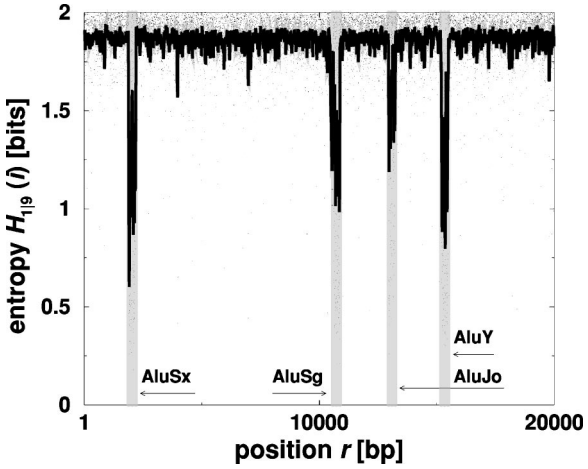

FIG. 6. Shannon entropy $H_{1|9}(i)$, from the preceding 9-mer i to the next nucleotide versus the sequence position r (dots) and the 30 bp running average (solid line) for $r=1,2,\ldots,20\,000$ bp of human chromosome 22. Any clear deviation of $H_{1|9}(i)$ from 2 bits is indicative of redundancy. The shaded regions give the locations of the first four Alu repeats (AluSx, AluSg, AluJo, and AluY) of human chromosome 22 [74], and they agree with the sharp drops of $H_{1|9}(i)$.

repeat $\mathcal{R}$ inserted at ϱN positions. We adopt the density $\varrho=0.0006$ [8] of Alu repeats in human chromosome 22, use $\ell=300$ bp as the length of Alu repeats, and let ε assume values 0 and 0.1, respectively. It is clear that n-mers can belong to one of the three following different classes: they are positioned either in the random part of the sequence ($\mathcal{A}$), or in the repetitive part of the sequence ($\mathcal{B}$), or they have an overlap within both parts of $\tilde{n}<n$ nucleotides ($\mathcal{C}$). The n-mer probability is not equally distributed among the classes $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. Due to the repeat density, n-mers in classes $\mathcal{B}$ and $\mathcal{C}$ occur with greater probability than expected by chance.

We calculate the entropy $H_{1|n}(i)$ and investigate the dependence of the $(n+1)$th nucleotide on the preceding n nucleotides. In Fig. 5, we show for $n=9$ the entropy $H_{1|n}(i)$ of a realization of the above repeat model for $\varepsilon=0$ and $\varepsilon=0.1$. We observe that $H_{1|9}(i)$ shows a distinct dependence on the sequence position. This positional dependence can be explained by the fact that 10-mers within a repeat occur with much higher frequency than in a random uncorrelated sequence of λ equidistributed nucleotides, and hence the corresponding $(n+1)$th nucleotides become more predictable. Figure 5(a) shows $H_{1|9}(i)$ for identical repeats ($\varepsilon=0$), and we find that $H_{1|9}(i)$ switches from about 2 bits to nearly 0 bits each time a repeat is detected.

We study the robustness of $H_{1|n}(i)$ by mutating a fraction ε of all copies of the repeat. Due to mutations the nucleotide following an n-mer within a repeat occurs with probability $1-\varepsilon$, and hence Eq. (5) gives $H_{1|n}(i)\approx-(1-\varepsilon)\log_2(1-\varepsilon)-\varepsilon\log_2(\varepsilon/3)$ for oligonucleotide $A_i^{(n)}$ within a repeat. Figure 5(b) shows for $\varepsilon=0.1$ and $n=9$ that $H_{1|9}(i)\approx0.6$ bits still indicates the approximate location of repeats that are close but not identical to the consensus sequence. This states that $H_{1|9}(i)$ is robust against moderate mutation rates in simulated repeats.

Common approaches of identifying repetitive sequences in DNA are based on the computation of dot matrices [68–70], screening a database of known repeats to which the sequence shows a significant sequence similarity [71–73], or searching for shared subsequences. In order to investigate how the property of the entropy $H_{1|n}(i)$ can be used to identify locations of interspersed repeats in large experimental data sets, we compute $H_{1|9}(i)$ for the first 20 000 positions of the DNA sequence of human chromosome 22. Figure 6 shows that the function of $H_{1|9}(i)$ versus i drops sharply at four sequence locations within the first 20 000 bp. Each drop has a length of about 300 bp. When we compare these locations of potential Alu repeats with the annotation of localized Alu repeats in the first 20 000 bp of human chromosome 22 [73,74], we find an agreement of the predicted and the actual locations. Hence, a plot of the entropy $H_{1|n}(i)$ versus the sequence position may provide one additional method for recognizing locations of interspersed repeats in long DNA sequences. Since the calculation of $H_{1|n}(i)$ requires no *a priori* knowledge of the repeat families and is robust against moderate levels of sequence deviation, it might be useful for complementing current methods of finding repetitive sequences of intermediate lengths in large genomes.

V. EFFECTS OF THE PRESENCE OF INTERSPERSED ALU REPEATS

In this section we examine to which degree the repeat model can explain the statistical properties of human chromosome 22 observed in Secs. II and III.

Figure 7 shows the mutual information function $I(k)$ of a simulated DNA sequence generated by the repeat model and

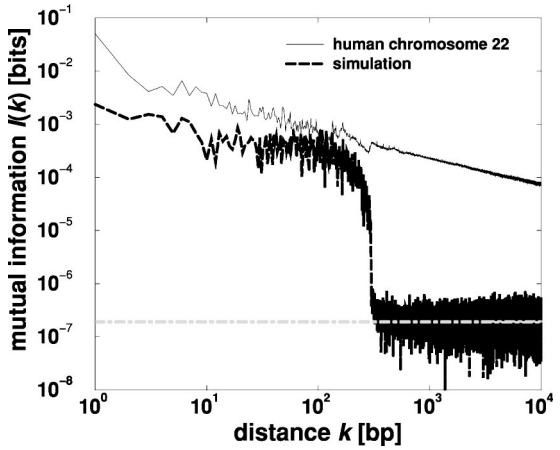

FIG. 7. Double-logarithmic plot of $I(k)$ for human chromosome 22 and a realization of the repate model with $\varrho = 0.0006$, $\ell = 300$, and $\varepsilon = 0.17$. We find for simulated DNA sequences that $I(k)$ shows some decay up to 300 bp, while it drops abruptly to the noise level for distances larger than $k \approx 300$. We obtain qualitatively the same result for all three binary autocorrelation functions. The gray dot-dashed line marks the bias of $I(k)$, and the increasing variance with increasing k is due to finite sequence length [23].

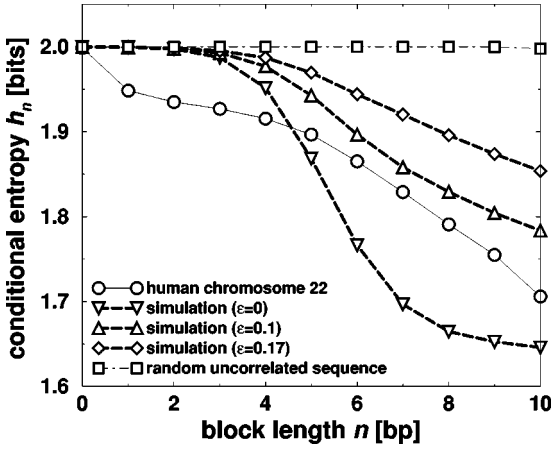

FIG. 8. Conditional entropy h_n ($h_0 \equiv H_1$) of human chromosome 22 and of three realizations of the repeat model with $\varrho = 0.0006$, $\ell = 300$, and $\varepsilon = 0$, 0.1, and 0.17. The decay of h_n of model sequences clearly demonstrates that the presence of interspersed Alu repeats decreases the conditional entropies h_n, which leads to a sublinear increase of H_n with n.

of the DNA sequence of human human chromosome 22. We find that there is some decay of $I(k)$ for distances $k < 300$ bp, and that $I(k)$ of the simulated DNA sequence drops abruptly to the noise level for $k > 300$ bp. This indicates that Alu repeats may explain about 10% of the statistical dependencies as measured by $I(k)$ in human chromosome 22 up to $k \approx 300$ bp but not the power-law decay beyond $k \approx 300$ bp. When we model the presence of Alu repeats together with 3% coding DNA content [23], we find that simulated coding DNA sequences have only minor effects on the decay of $I(k)$ of the model sequences, as anticipated. The full reconstruction of the decay as shown in Figs. 1 and 2, however, requires the incorporation of further biological features such as LINEs, clusters of CG dinucleotides, noncoding sequences, and chromosomal domains with a high or low G + C content (isochores).

Next, we study the effect of interspersed Alu repeats on short-range statistical dependencies of oligonucleotides. We compare the conditional entropies h_n of the DNA sequence of human chromosome 22 and of simulated DNA sequences according to the repeat model. Figure 8 shows h_n versus n for model sequences with different levels of mutation rates $\varepsilon = 0$, 0.1, and 0.17.

Taking $h_0 \equiv H_1$, we find that each curve starts at 2 bits due to a uniform nucleotide distribution $p_i = \frac{1}{4}$. For higher-order oligonucleotides we find for model sequences that each $h_n(\varepsilon)$ exhibits a pronounced decay for $n \geqslant 5$. These curves reflect the increased redundancy (or improvement of predictability) with increasing oligonucleotide length n due to the presence of interspersed repeats. For model sequences, the decay of the conditional entropies h_n can also be obtained from analytical expressions [44]. As intuitively expected the decay is reduced due to mutations. For $\varepsilon = 0.17$, which is a reasonable value for Alu repeats, we find at $n = 10$ a decay by about 0.15 bits. Comparing it to the decay of h_n of human chromosome 22 for $n \geqslant 5$, we find that a considerable fraction of the empirically observed redundancy can be attributed to the presence of Alu repeats.

Finally, we compare the Rényi entropies of the model and the empirical data. Figure 9 shows $H_n(q)$ of a simulated DNA sequence according to the repeat model with $\varrho = 0.0006$, $\ell = 300$, and $\varepsilon = 0.17$, and $H_n(q)$ of the DNA sequence of human chromosome 22. It turns out that there are clear discrepancies for large and small values of q. These observations point to over- and underrepresented oligonucleotides that are not explained by the Alu repeat simulations. In the next section we extend the repeat model by incorporating the presence of monomeric (poly-A and poly-T) tandem repeats and the suppression of CG dinucleotides.

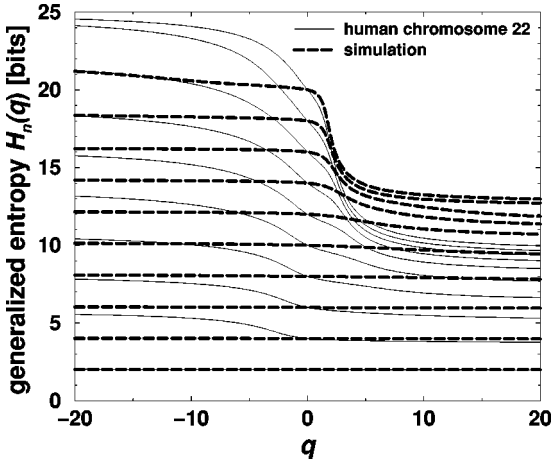

FIG. 9. Rényi entropies $H_n(q)$ of the DNA sequence of human chromosome 22 and a realization of the repeat model with $\varrho = 0.0006$, $\ell = 300$, and $\varepsilon = 0.17$ for $n = 1,2,\ldots,10$ (from bottom to top). We find clear discrepancies between the statistical properties of chromosome 22 (thin solid line) and the stochastic model with interspersed repeats (thick dashed line).

VI. EFFECTS OF THE PRESENCE OF MONOMERIC TANDEM REPEATS AND THE SUPPRESSION OF CG DINUCLEOTIDES

In this section we study the effects of monomeric tandem repeats and of over- and underrepresented dinucleotides by introducing two modifications to the repeat model studied above. (1) In analogy to interspersed repeats, we characterize a monomeric tandem repeat $\mathcal{M}$ by two parameters: (i) its relative frequency of occurrence ρ and (ii) its length w. In a first approximation we use a fixed length $w=w_0$ for all monomeric tandem repeats. (2) We intersperse a monomeric repeat ρN times nonoverlappingly in a background sequence at randomly and uniformly chosen positions. It is a general feature of the human genome that the CG dinucleotide [75] is underrepresented. To take this into account, we model the background sequence by a simple Markov process, defined by a 4×4 transition matrix

$$\Pi(\epsilon)=\begin{pmatrix} \frac{1}{4}-\frac{\epsilon}{9} & \frac{1}{4}-\frac{\epsilon}{9} & \frac{1}{4}+\frac{\epsilon}{3} & \frac{1}{4}-\frac{\epsilon}{9} \\ \frac{1}{4}+\frac{\epsilon}{3} & \frac{1}{4}+\frac{\epsilon}{3} & \frac{1}{4}-\epsilon & \frac{1}{4}+\frac{\epsilon}{3} \\ \frac{1}{4}-\frac{\epsilon}{9} & \frac{1}{4}-\frac{\epsilon}{9} & \frac{1}{4}+\frac{\epsilon}{3} & \frac{1}{4}-\frac{\epsilon}{9} \\ \frac{1}{4}-\frac{\epsilon}{9} & \frac{1}{4}-\frac{\epsilon}{9} & \frac{1}{4}+\frac{\epsilon}{3} & \frac{1}{4}-\frac{\epsilon}{9} \end{pmatrix}$$

whose elements Π_{ij} $(i,j=1,2,\ldots,\lambda)$ are the conditional probabilities of transition from nucleotide A_i to nucleotide A_j. The probability of transition from C to G (Π_{23}) has been reduced by the mutation rate ϵ, while the remaining elements have simply been chosen such that the nucleotide probabilities are $p_i=\frac{1}{4}$.

We use the reduction of $\epsilon=20\%$ of Π_{23} as compared to the expectation under a random uncorrelated model with λ equidistributed nucleotides [4,76]. Figure 10 shows that interspersed Alu repeats, monomeric tandem repeats (poly-A and poly-T), and the suppression of CG dinucleotides contribute to the overall statistical properties of human chromosome 22. In particular, Fig. 10(a) shows that the monomeric tandem repeats may partly account for the accumulation of $H_n(q)$ at large values of q, while Fig. 10(b) shows that the discrepancies of $H_n(q)$ for small values of q are diminished by including the suppression of CG dinucleotides.

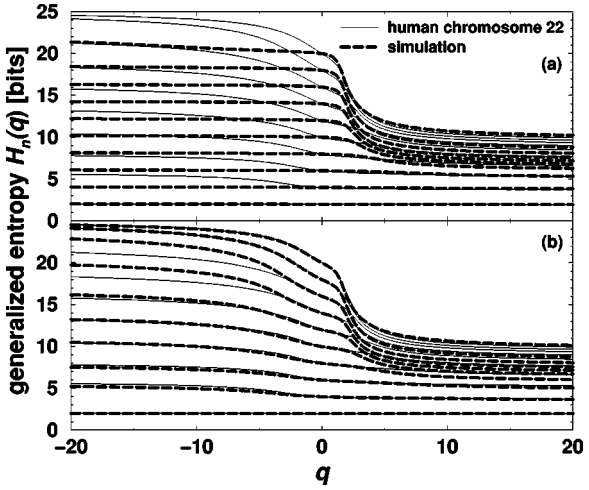

FIG. 10. Rényi entropies $H_n(q)$ of the DNA sequence of human chromosome 22 and of two realizations of the repeat model with interspersed repeats ($\varrho=0.0006$, $\ell=300$, $\varepsilon=0.17$), poly-A repeats ($\rho=0.001\,75$, $w=9$), and poly-T repeats ($\rho=0.001\,75$, $w=9$) for $n=1,2,\ldots,10$ (from bottom to top) and for $q\in[-20,20]$. The background sequences are modeled by (a) a Bernoulli process and (b) a first-order Markov process with $p_i=1/4$ and $p_{ij}=p_i\Pi_{ij}(\epsilon)$ with $\epsilon=0.2$. (a) shows that the presence of Alu, poly-A, and poly-T repeats influences the decay of $H_n(q)$ for $q\gg0$, while (b) shows that the incorporation of CG suppression results in a better approximation of $H_n(q)$ of the empirical data $q\ll0$. Hence, the superposition of Alu repeats, monomeric tandem repeats, and CG suppression causes a decay of $H_n(q)$ with q that is closer to that observed for human chromosome 22 than that for random uncorrelated sequences.

VII. SUMMARY AND DISCUSSION

We study short- and long-range correlations in the DNA sequence of human chromosome 22, using the mutual information function (I), three binary autocorrelation functions (C_{WW}, C_{RR}, and C_{MM}), and entropies (H_n). We find that $I(k)$ shows no pronounced period-3 oscillations, while $I(k)$ and $C^2_{\mathrm{WW}}(k)$ show a clear power-law decay over five orders of magnitude of k, and $C^2_{\mathrm{RR}}(k)$ and $C^2_{\mathrm{MM}}(k)$ show a power-law decay over four orders of magnitude of k. We further find that for $1\leqslant n\leqslant10$ the conditional entropies h_n decay with increasing oligonucleotide length n, that for $2\leqslant n\leqslant10$ the Rényi entropies $H_n(q)$ decay with increasing q, and that the decay of $H_n(q)$ becomes steeper with q as n increases. These findings show that there are short-range statistical dependencies within oligonucleotides for $2\leqslant n\leqslant10$ accompanied by an increasing nonuniformity of the frequency distribution of oligonucleotides.

We study a simple stochastic model that incorporates the presence of interspersed Alu repeats, and we find that this model can—to some degree—reproduce the decay of h_n with increasing n, but it cannot reproduce the decay of $H_n(q)$ with increasing q. However, when we extend the repeat model by incorporating the presence of monomeric tandem repeats and the suppression of CG dinucleotides, we find that the extended model can reproduce the observed decay of $H_n(q)$ with q better than the original model. In this study we use the assumptions that interspersed repeats have no internal structure and that monomeric tandem repeats have a fixed length. Yet certain types of repeats contain an internal structure (e.g., higher G + C content than the surrounding sequence or poly-A tails), and a precise knowledge of the length distribution of monomeric tandem repeats may yield a better quantitative agreement between the model and the observed data.

The origin of long-ranging correlations in the DNA sequence of human chromosome 22 remains to be uncovered. Human chromosome 22 is repeat rich, and it has a compara-

tively high gene density and G + C content. Repeat models with a fixed length of interspersed repeats induce statistical dependencies within the order of the repeat length. It can be anticipated that the incorporation of appropriate length distributions and the superposition of SINEs and LINEs, clusters of CG dinucleotides [76], noncoding regions, or chromosomal domains characterized by a distinct high or low G + C content [10,38] can extend the correlation length beyond several thousand bp.

A preliminary analysis of the draft version of the complete human genome yields an estimated fraction of protein-coding DNA sequences of about 1.5% and an estimated fraction of repeats of above 50%. Hence, we expect that the statistical patterns of the other 23 human chromosomes might be strongly influenced by the presence of repetitive sequences.

ACKNOWLEDGMENTS

We thank the Institute for Pure and Applied Mathematics (University of California, Los Angeles) for hospitality and support while part of this work was completed, and the Deutsche Forschungsgemeinschaft (DFG), the Graduate Program ''Dynamics and Evolution'' (GK 268), and the Cold Spring Harbor Laboratory Association for financial support.

[1] L. L. Gatlin, *Information Theory and the Living System* (Columbia University Press, New York, 1972).
[2] B. Lewin, *Genes VII* (Oxford University Press, Oxford, 2000).
[3] P. Sudbery, *Human Molecular Genetics* (Addison Wesley Longman, Singapore, 1998).
[4] B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland Publishing, New York, 1994).
[5] E. N. Trifonov and J. L. Sussman, Proc. Natl. Acad. Sci. U.S.A. **77**, 3816 (1980).
[6] J. S. Beckmann and E. N. Trifonov, Proc. Natl. Acad. Sci. U.S.A. **88**, 2380 (1991).
[7] V. V. Lobzin and V. R. Chechetkin, Phys. Usp. **43**, 55 (2000).
[8] I. Dunham *et al.*, Nature (London) **402**, 489 (1999).
[9] M. Hattori *et al.*, Nature (London) **405**, 311 (2000).
[10] International Human Genome Sequencing Consortium, Nature (London) **409,** 860 (2001).
[11] J. C. Venter *et al.*, Science **291**, 1904 (2001).
[12] J. Majewski and J. Ott, Genome Res. **10**, 1108 (2000).
[13] S. de Souza *et al.*, Proc. Natl. Acad. Sci. U.S.A. **97**, 12 690 (2000).
[14] E. Dawson *et al.*, Genome Res. **11**, 170 (2001).
[15] M. Scherf *et al.*, Genome Res. **11**, 333 (2001).
[16] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[17] C.-K. Peng *et al.*, Nature (London) **356**, 168 (1992).
[18] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[19] S. Buldyrev *et al.*, Phys. Rev. E **51**, 5084 (1995).
[20] J. W. C. Shepherd, Proc. Natl. Acad. Sci. U.S.A. **78**, 1590 (1981).
[21] H. Herzel and I. Grosse, Physica A **216**, 518 (1995).
[22] W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. **2**, 137 (1992).
[23] H. Herzel and I. Grosse, Phys. Rev. E **55**, 800 (1997).
[24] N. V. Dokholyan *et al.*, Phys. Rev. Lett. **79**, 5182 (1997).
[25] About 70% of the DNA sequence of yeast codes for proteins, and the mosaic alternation of coding and noncoding DNA sequences implies long-ranging correlations if the length distribution of coding DNA sequences is taken into account.
[26] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[27] We downloaded the DNA sequence of human chromosome 22 from the Sanger Center, http://www.sanger.ac.uk/HGP/Chr22/. The DNA sequence is not 100% complete but contains 11 gaps, accounting for about 3% of the sequence. We substituted the unknown nucleotides by independent identically distributed nucleotides drawn from a uniform distribution.
[28] The nucleotides adenine (A) and guanine (G) are purines (R), and the nucleotides cytosine (C) and thymine (T) are pyrimidines (Y). W stands for weakly binding nucleotides A and T, and S stands for strongly-binding nucleotides C and G. K stands for the keto nucleotides G and T, while M stands for amino nucleotides A and C.
[29] C.-K. Peng *et al.*, Phys. Rev. E **49**, 1685 (1994).
[30] S. V. Buldyrev *et al.*, Biophys. J. **65**, 2673 (1993).
[31] V. Brendel, J. S. Beckmann, and E. N. Trifonov, J. Biomol. Struct. Dyn. **4**, 11 (1986).
[32] G. A. Miller, in *Information Theory in Psychology*, edited by H. Quaster (Free Press, Glencoe, 1955).
[33] G. P. Basharin, Theor. Probab. Appl. **4**, 333 (1959).
[34] O. Weiss and H. Herzel, J. Theor. Biol. **190**, 341 (1998).
[35] J. W. Fickett, D. C. Torney, and D. R. Wolf, Genomics **13**, 1056 (1992).
[36] S. Karlin and V. Brendel, Science **259**, 677 (1993).
[37] R. A. Elton, J. Theor. Biol. **45**, 533 (1974).
[38] G. Bernardi *et al.*, Science **288**, 953 (1985).
[39] T. D. Schneider *et al.*, J. Mol. Biol. **188**, 415 (1986).
[40] R. N. Mantegna *et al.*, Phys. Rev. E **52**, 2939 (1995).
[41] P. Bernaola-Galván *et al.*, Phys. Rev. E **53**, 5181 (1996).
[42] A. Khinchin, *Mathematical Foundations of Information Theory* (Dover, New York, 1967).
[43] J. P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).
[44] H. Herzel, W. Ebeling, and A. O. Schmitt, Phys. Rev. E **50**, 5061 (1994).
[45] A. Rényi, *Probability Theory* (North-Holland, Amsterdam, 1970).
[46] H.-G. Schuster, *Deterministic Chaos. An Introduction* (Physik Verlag, Weinheim, 1984).
[47] T. C. Halsey *et al.*, Phys. Rev. A **33**, 1141 (1986).
[48] K. Pawelzik and H.-G. Schuster, Phys. Rev. A **35**, 481 (1987).
[49] B. Pompe, Chaos, Solitons Fractals **4**, 83 (1994).
[50] W. Ebeling, T. Pöschel, and K.-L. Albrecht, Int. J. Bifurcation Chaos Appl. Sci. Eng. **5**, 51 (1995).
[51] D. Holste *et al.*, J. Theor. Biol. **206**, 525 (2000).
[52] J. A. Glazier *et al.*, Phys. Rev. E **51**, 2665 (1995).
[53] C. Beck and F. Schlögl, *Thermodynamics of Chaotic Systems* (Cambridge University Press, Cambridge, 1993).
[54] We estimate $p_i^{(n)}$ by $\hat{p}_i^{(n)}=(N_i+\alpha)/(N+\lambda^n\alpha)$, where N_i is the

number of oligonucleotides $A_i^{(n)}$ $(i=1,2,\ldots,\lambda^n)$, $N=\Sigma_i^{\lambda_n} N_i$, and α is the number of pseudocounts. We set $\alpha=0$ for H_n and $\alpha=1$ for $H_n(q)$ to prevent for $q<0$ the divergence of $H_n(q)$ in case of unobserved oligonucleotides; R. Durbin *et al.*, *Biological Sequence Analysis* (Cambridge University Press, Cambridge, 1998).

[55] Approximate analytic expressions for the bias of $H_n(q)$ show for i.i.d. sequences that the systematic error of estimates of $H_n(q)$ is, in general, dependent on p_i, and weakly biased estimates for $q\in[-20,20]$ can be obtained for $n=1,2,\ldots,10$ given $N=33.4\times10^6$ bp; P. Grassberger, Phys. Lett. A **128**, 369 (1988); D. Holste, I. Grosse, and H. Herzel, J. Phys. A **31**, 2551 (1998).

[56] D. Tautz, M. Trick, and G. Dover, Nature (London) **322**, 652 (1986).

[57] P. Bucher and G. Yagil, DNA Seq. **1**, 27 (1991).

[58] G. I. Bell and D. C. Torney, Comput. Chem. (Oxford) **17**, 185 (1993).

[59] S. Bonhoeffer *et al.*, Phys. Rev. Lett. **76**, 1977 (1996).

[60] R. F. Voss, Phys. Rev. Lett. **76**, 1978 (1996).

[61] A. F. A. Smit, Curr. Opin. Genet. Dev. **6**, 743 (1996).

[62] A. M. Shedlock and N. Okada, BioEssays **22**, 148 (2000).

[63] A. M. Weiner, Nature Genet. **24**, 332 (2000).

[64] J. E. Stenger *et al.*, Genome Res. **11**, 12 (2001).

[65] J. R. Korenberg and M. C. Rykowski, Cell **53**, 391 (1988).

[66] Y. Toda, R. Saito, and M. Tomita, J. Mol. Evol. **50**, 232 (2000).

[67] W. M. Chu *et al.*, Mol. Cell. Biol. **18**, 58 (1998).

[68] W. M. Fitch, Biochem. Genet. **3**, 99 (1969).

[69] A. J. Gibbs and G. A. McIntyre, Eur. J. Biochem. **16**, GC1 (1970).

[70] E. L. L. Sonnhammer and R. Durban, Gene **167**, 1 (1996).

[71] J.-M. Claverie, Methods Enzymol. **266**, 212 (1996).

[72] J. Jurka *et al.*, Comput. Chem. (Oxford) **20**, 119 (1996).

[73] A. F. A. Smit and P. Green, computer code REPEATMASKER, http://ftp.genome.washington.edu/RM/.

[74] Sanger Center, http://www.sanger.ac.uk/cgi-bin/cwa/22cwa.pl.

[75] Since the nucleotides C and G are connected by a phosphodiester bond, these dinucleotides are traditionally termed CpG.

[76] M. Gardiner-Garden and M. Frommer, J. Mol. Biol. **196**, 261 (1987).