Correlations in DNA sequences: The role of protein coding segments

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

Ivo Große
Center for Polymer Studies and the Department of Physics, Boston University, Boston, Massachusetts 02215

(Received 2 May 1996)

Protein coding segments (exons) exhibit persistent correlations between their nucleotides with a pronounced period three. It is shown in this paper that this periodicity induced by the nonuniform codon usage implies long-range correlation over hundreds of base pairs if the length distribution of exons is taken into account. We derive expressions which relate the length distribution of exons to the correlation decay and find agreement with numerical simulations. Finally, we analyze the decay of the mutual information function in yeast chromosomes, in an *E. coli* chromosome region, and in myosin heavy chain genes as representative examples. It turns out that in these cases we can explain most of the long-range statistical dependences even quantitatively. [S1063-651X(97)00101-3]

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

I. INTRODUCTION

The statistical analysis of DNA sequences is of importance for understanding the structure and function of genomes [1–8]. Statistical dependences between nucleotides have been analyzed for decades in various contexts [9–15]. Among physicists the detection of long-range correlations has attracted much attention during the past years [16–23]. Using mutual information functions [16,20,23], autocorrelation functions [19,22], spectra [18,24], and random walk analyses [17,25,26], correlations ranging from a few base pairs (bp) up to 10^4 bp have been analyzed. However, the biological interpretation of most of these findings remains still speculative.

From a molecular biological point of view, long-range correlations are not surprising since the complex organization of genomes involves many different scales. In fact, large variations in base composition on scales of thousands of base pairs have been discussed extensively in the literature (see, e.g., [27–33]). For example, Elton [27] reviews experimental data showing that DNA fragments up to 10^4 bp have rather large variances of the guanine (G)+cytosine (C) content. He points out that these variations cannot be explained by short-correlated fluctuations. In this way, long-range correlations were already indicated decades ago. Explicit examples of pronounced fluctuations of the G+C content together with the gene distribution with an approximate period of 10^5 bp were provided by the recent sequencing of yeast chromosomes [34,35].

It has been pointed out by several authors that the mosaic structure of genomes is presumably responsible for long-range correlations [20,28,33]. Indeed, the organization of the genome is very complex: eukaryotic genes usually consist of several protein coding segments (*exons*) interrupted by intervening sequences (*introns*). Moreover, there are regulatory elements such as promoters, splice sites, enhancers, and silencers, which are sometimes up to thousands of base pairs away from exons. Genomes of higher eukaryotes also comprise long stretches of DNA without any obvious biological function containing, e.g., *pseudogenes* and various types of *repeats* [8,23,36].

There are several models of DNA where a segmentational structure is postulated [27,32,37–39]. Elton discusses, for example, the variance of the G+C content for a model with constant and exponentially distributed fragments [27], and Buldyrev *et al.* study a Lévy-walk model [38]. However, hypothetical length distributions of fragments have to be postulated in these papers.

Contrarily, we will show in this paper that already the well-known length distribution of exons generates long-ranging correlations. As a first step we demonstrate in Sec. III that a nonuniform *codon usage* in protein coding segments induces persistent period-three oscillations. In that section we introduce a model by which we generate artificial DNA sequences called *pseudoexons*—a concatenation of statistically independent codons chosen randomly from a given codon usage probability table. In Secs. IV and V, we emphasize the central role of the length distribution of exons. We derive analytic expressions which relate the exon length distribution to the correlation decay and show that these analytic results are in perfect agreement with numerical simulations. In Sec. VI, we apply these theoretical considerations to several DNA sequences (yeast chromosomes, *E. coli* DNA, and a myosin heavy chain gene).

We show that correlations on scales of hundreds of base pairs can be simulated even quantitatively by taking into account solely the nonuniformity of the codon usage and the length distribution of exons. In this way we relate well-known biological facts to observed long-range correlations between nucleotides.

II. CORRELATION MEASURES

DNA sequences can be viewed as symbolic strings composed of the four ‘‘letters’’ $(A_1,A_2,A_3,A_4)\equiv(\text{A,C,G,T})$ re-

*Electronic address: herzel@itp1.physik.tu-berlin.de ferring to the nucleotides adenine, cytosine, guanine, and thymine. The probability of finding the nucleotide A_i is denoted by p_i $(i=1,2,3,4)$. Pair correlations within sequences can be measured by the joint probabilities $p_{ij}(k)$ of finding the symbol A_i and k letters downstream the symbol A_j. Then, statistical independence of symbols in a distance k is defined by $p_{ij}(k)=p_ip_j$, which leads to the mutual information function $I(k)$ [6,12,22,40–42] as a measure of statistical dependence

$$I(k)=\sum_{i,j=1}^{4} p_{ij}(k)\log_2\frac{p_{ij}(k)}{p_ip_j}. \tag{1}$$

By choosing the logarithm to base 2, $I(k)$ is measured in bit and gives the information on the letter A_j knowing the letter A_i. The mutual information $I(k)$ vanishes if, and only if, statistical independence holds, i.e., if all 16 joint probabilities $p_{ij}(k)$ factorize. Consequently, the mutual information allows us to detect any pair correlation.

More specific indicators of dependences are correlation functions. Their definition requires an assignment of numbers a_i to the corresponding symbols A_i. Assuming ergodicity and stationarity, the usual estimation of autocorrelation functions via averages over the sequence

$$C(k)=\langle a(n)a(n+k)\rangle-\langle a(n)\rangle\langle a(n+k)\rangle \tag{2}$$

can be written in terms of the probabilities defined above,

$$\begin{aligned}C(k)&=\left(\sum_{i,j=1}^{4} p_{ij}(k)a_ia_j\right)-\left(\sum_{i=1}^{4} p_ia_i\right)\left(\sum_{j=1}^{4} p_ja_j\right)\\&=\sum_{i,j=1}^{4}[p_{ij}(k)-p_ip_j]a_ia_j.\end{aligned} \tag{3}$$

By definition, correlation functions measure only linear dependences. However, for quaternary sequences such as DNA, six properly chosen autocorrelation functions and three cross-correlation functions can guarantee the statistical independence between all nucleotide pairs [22].

Long-range correlations are often characterized by power laws

$$C(k)\propto k^{-\gamma}. \tag{4}$$

Such a scaling behavior can also be analyzed by using power spectra or the random walk approach with related scaling exponents [26]. A power law (4) implies also a power-law decay of the mutual information function

$$I(k)\propto k^{-2\gamma}. \tag{5}$$

This can be easily derived using a Taylor expansion in terms of

$$D_{ij}(k)=p_{ij}(k)-p_ip_j, \tag{6}$$

which measure deviations from statistical independence. Since $I(k)$ has a minimum at $D_{ij}\equiv 0$, the sum over all linear terms vanishes, and we obtain

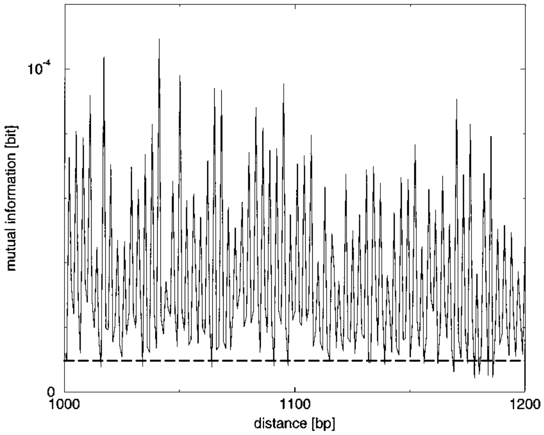

FIG. 1. Mutual information function of the yeast chromosome XI (666 448 BP). The periodicity due to the triplet code is visible even for distances above 1000 bp. The dashed line marks the bias according to Eq. (8).

$$I(k)=\frac{1}{2\ \ln 2}\sum_{i,j=1}^{4}\frac{D_{ij}^2(k)}{p_ip_j}+O(D_{ij}^3). \tag{7}$$

In the Appendix we use this relation to discuss finite sample effects. Equation (7) illustrates that the mutual information $I(k)$ accumulates all pair correlations in a distance k. For DNA sequences, the above second-order approximation is extremely close to the actual mutual information because of the weakness of correlations (see, e.g., Fig. 1). Since correlation functions can be written as quadratic forms of the *dependence matrix* D_{ij} [cf. Eqs. (3) and (6)], a scaling exponent γ of correlation functions leads to an exponent 2γ for the mutual information.

In this paper we study mainly the decay of the mutual information function as an overall measure of statistical dependences. In contrast to entropies of long ‘‘words’’ [20,36] the statistical and systematic errors of the mutual information are relatively small since only 16 probabilities have to be estimated from samples of thousands of nucleotides. For example, the bias of the mutual information for a sample of size N has been calculated [12,22] to be

$$\Delta I=\frac{9}{2\ \ln 2N}, \tag{8}$$

which is marked in some figures by a dashed line. Though this bias is small, it becomes relevant for very weak correlations. Therefore, we discuss finite sample effects in some detail in the Appendix.

III. EFFECTS OF A NONUNIFORM CODON USAGE

Analyses of DNA sequences revealed that their correlation functions often exhibit strong period-three components, which are induced by the genetic code [9–11,24,43]. Figures 1–3 exemplify these periodicities for yeast chromosome XI, an *E. coli* chromosome region, and a myosin heavy chain gene.

In protein coding segments, 61 of the possible 64 codons (three-symbol words) encode 20 different amino acids whereas the remaining 3 are used as stop codons. It has been

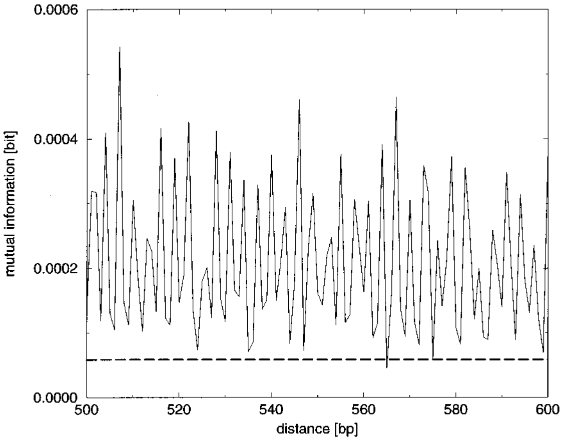

FIG. 2. Period-three oscillations of the mutual information for a chromosome region of *Escherichia coli* (strain K-12, 111 401 bp).

discussed [28,44–46] that these codons are used with quite different frequencies for several reasons. (i) There are specific amino acid compositions for proteins. (ii) The number of triplets encoding an amino acid is different. (iii) For any amino acid, a preference of certain codons over others exists. (iv) The G+C content of the third codon position is correlated to the G+C content of the surrounding DNA region [35].

In general, a nonuniform codon usage causes the concentration of each nucleotide to be different in all three positions of the reading frame. As we will show in the following, it is exactly this *position asymmetry* of all four nucleotides that introduces the pronounced period-three pattern of correlation functions as well as the mutual information function.

In order to quantify the effect of a nonuniform codon usage on correlation measures, we introduce a stochastic model that randomly concatenates subsequent codons. In the following, we term the model sequences of independent codons chosen from a given codon usage table *pseudoexon*. As we will see, these pseudoexons, which consist of statistically independent codons, display periodic long-range correlations between their nucleotides. Our next task is to analytically calculate the strength of these correlations, which was shown to be a prominent long-range correlation pattern of real DNA (cf. Figs. 1–3).

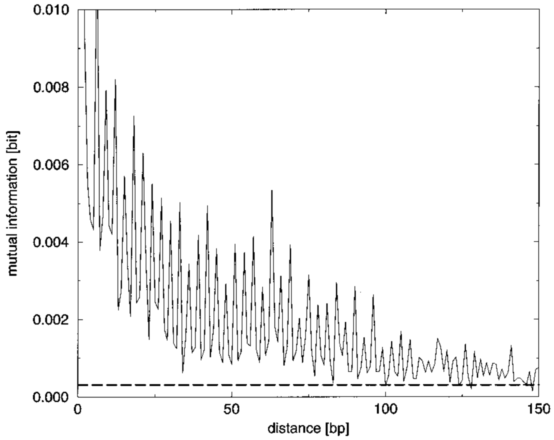

FIG. 3. Mutual information function of the HUMBMYH7 gene (20 855 bp from the first to the last exon). The mean exon length is about 150 bp which is the characteristic length of the decay of the pronounced period-three oscillations.

Let us start with the calculation of the mutual information of an infinitely long pseudoexon generated by such a Bernoulli-like process on the level of codons. We denote the frequency of the ith nucleotide at the mth position by $p_i^{(m)}$ $(m=1,2,3)$. The overall probability of symbol i follows directly by averaging over the three positions

$$p_i=\frac{p_i^{(1)}+p_i^{(2)}+p_i^{(3)}}{3} \quad (i=1,\ldots,4). \tag{9}$$

The following table displays the 12 frequencies $p_i^{(m)}$, which are obtained from the 5805 bp of the protein coding segments from the intensively studied [15,47] human β-myosin heavy chain (HUMBMYH7) gene.

	Position 1	Position 2	Position 3
A	0.296	0.437	0.079
C	0.248	0.184	0.343
G	0.351	0.123	0.471
T	0.105	0.256	0.107

The joint probabilities $p_{ij}(k)$ can be obtained directly from tables as shown above. For $k\geqslant 3$, the corresponding probabilities factorize due to our assumption of independence. First we consider $k=3,6,9,\ldots$; i.e., the two symbols of the pair are in the same position within the frame

$$p_{ij}(k)=\frac{p_i^{(1)}p_j^{(1)}+p_i^{(2)}p_j^{(2)}+p_i^{(3)}p_j^{(3)}}{3}. \tag{10}$$

For $k=4,7,10,\ldots,$ we obtain

$$p_{ij}(k)=\frac{p_i^{(1)}p_j^{(2)}+p_i^{(2)}p_j^{(3)}+p_i^{(3)}p_j^{(1)}}{3}, \tag{11}$$

and distances $k=5,8,11,\ldots,$ lead to

$$p_{ij}(k)=\frac{p_i^{(1)}p_j^{(3)}+p_i^{(2)}p_j^{(1)}+p_i^{(3)}p_j^{(2)}}{3}. \tag{12}$$

Inspection of the last two expressions reveals that $p_{ij}(k=4,7,\ldots,)=p_{ji}(k=5,8,\ldots,)$. Consequently, the values of the mutual information at these positions are identical. The above expressions allow us to calculate the *in-frame mutual information*

$$I_{in}\equiv I(k=3,6,9,\ldots,) \tag{13}$$

and the *out-of-frame mutual information*

$$I_{out}\equiv I(k=4,5,7,8,10,11,\ldots,). \tag{14}$$

For example, the table given above yields

$$I_{in}=0.0247,$$

$$I_{out}=0.0083.$$

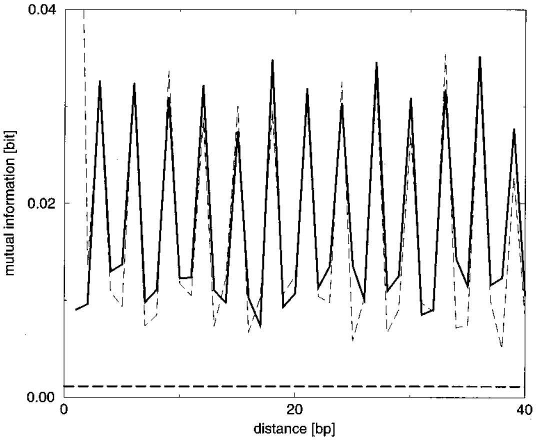

FIG. 4. Dashed line: Mutual information of a concatenation of all 40 exons (5 805 BP) of the HUMBMYH7 gene (compare Fig. 3). Full line: Corresponding pseudoexon (5 805 bp) generated from the codon usage table of the HUMBMYH7 gene.

The corresponding high-low-low pattern is indeed obvious in the examples graphed in Figs. 1–4. Figure 4 displays the period-three oscillations of a pseudoexon that is indeed quite similar to the mutual information of the corresponding exons.

In summary, for a single protein coding segment, a given codon usage table allows us to analytically calculate the resulting period-three oscillations. However, genomes contain many exons, introns, and intergenic sequences. Moreover, protein coding segments are found in all three reading frames and on both DNA strands. Therefore, we are not surprised by the fact that the mutual information function plotted in Figs. 1–3 are decaying and thus deviate from a purely repeated high-low-low pattern.

Section IV is devoted to the role of the length distribution of exons, which indeed strongly affects the decay properties of correlation measures. Taking into account these length distributions, we can generalize the pseudoexon model to stochastic models of genes and even of whole chromosomes termed *pseudochromosomes*.

IV. LENGTH DISTRIBUTION OF EXONS

We have discussed in the preceding section that the joint probabilities $p_{ij}^{(k)}$ calculated within an exon reflect the nonuniform codon usage. Since long stretches of DNA include many different exons, only a fraction of pairs A_i and A_j are located on the same exon. More precisely, an exon of length l contains $l-k$ pairs contributing to the codon usage induced periodicity. Consequently, the length distribution $\rho(l)$ of exons in a given DNA will be considered in this section.

We define $\rho(l)$ as the probability distribution that an exon has a length l. In Sec. V we discuss, for instance, a fixed length $l=L$, exponential, and power-law distributions $\rho(l)$. Figure 5 shows a histogram of the lengths of exons for yeast chromosomes. It can be seen that there are rather long protein coding segments. Regression reveals that the empirical distribution can be approximated by an exponential decay (full line) and by a power law (dashed line) as well. Hence, we discuss both cases in some detail.

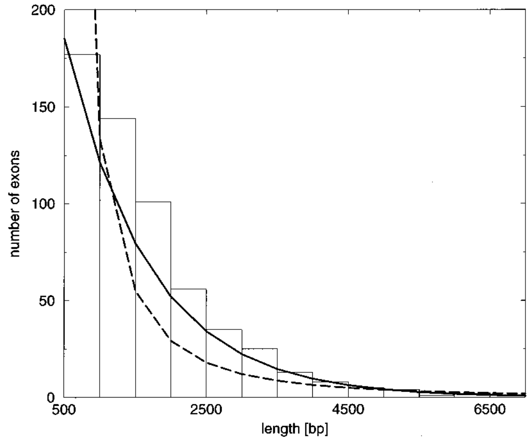

FIG. 5. Histogram of open reading frames (ORF's) longer than 500 bp from the yeast chromosomes III, IX, and XI. Regression by an exponential function and a power-law decay are indicated by full and dashed lines, respectively.

For the sake of simplicity, we assume below that all exons are characterized by a single codon usage table. This is, of course, a strong assumption since it is known that the codon usage depends, e.g., on the degree of gene expression [44,46]. However, Sharp and Li claim that ''within species the differences are largely in the degree rather than the direction of codon usage bias'' [46]. If whole chromosomes are analyzed, one has to take into account that genes are located on both strands. Therefore we use in our simulations of *pseudochromosomes* (see Sec. VI) also complementary codon usage tables.

As discussed in the preceding section the nonuniform codon usage leads to specific statistical dependences within exons. These are quantified below by the dependence matrix

$$D_{ij}^{exon}(k)=p_{ij}^{exon}(k)-p_i^{exon}p_j^{exon}. \tag{15}$$

In the following we denote the total fraction of protein coding sequences in a given DNA sequence by F. For the yeast chromosomes we have, for example, $F\approx 0.7$ [34]. The task is now to estimate the correlation decay for a given sequence length N, fraction of coding segments F, and probability distribution $\rho(l)$.

The mean exon length is given by

$$\overline{l}=\sum_l l\rho(l). \tag{16}$$

For yeast DNA, where genes exhibit only a few introns, the mean exon length is about 1400 bp. The typical length scale of human exons is a few hundred base pairs. However, there are also exons with a length of several thousand base pairs (e.g., exon 11 of the BCRA1 gene comprises 3426 bp).

The expectation value $\overline{n}$ of the number of exons in a sequence of length N and an exon fraction F is

$$\overline{n}=\frac{FN}{\overline{l}}. \tag{17}$$

Consequently, the average number of exons with a length l is given by

$$n(l)=\rho(l)\overline{n}=\frac{FN\rho(l)}{\Sigma_l l\rho(l)}. \tag{18}$$

Since we focus in this paper on correlations due to the nonuniform codon usage, we neglect statistical dependences of pairs A_i and A_j which are not within the same exon. This implies, for example, that the base composition in exons and introns is considered to be the same. Generalizations of this simplified approach are discussed in the final section.

Since every exon contributes $l-k$ pairs, we obtain the number $Z(k)$ of pairs which are located in the same exon

$$Z(k)=\sum_{l=k+1}^{l_{max}}(l-k)n(l). \tag{19}$$

Here, overlaps of protein coding segments have been neglected. The total number of pairs in a distance k is $N-k$, and hence, the overall deviations $D_{ij}(k)$ from statistical independence are given by

$$D_{ij}(k)=\frac{Z(k)}{N-k}D_{ij}^{exon}(k). \tag{20}$$

This result can now explain the decay of correlation functions and the mutual information function, since both measures can be obtained from the decay of the $D_{ij}(k)$ [cf. Eqs. (3)–(7)]. It can be seen that beside the internal period-three oscillations described by $D_{ij}^{exon}(k)$ we obtain a k-dependent prefactor related via $Z(k)$ to the exon length distribution. Since we are primarily interested in the long-ranging correlations, we focus in the following on the envelope:

$$E(k)\propto\sum_l\frac{l-k}{N-k}n(l). \tag{21}$$

This formula is a central result of this paper. It elucidates the immediate effect of the length distribution of exons on the decay properties of correlation measures, which we will exemplify in Sec. V.

V. MODELS OF LENGTH DISTRIBUTIONS

Now we illustrate the considerations of the preceding section for three representative probability distributions $\rho(l)$, namely, a uniform, an exponential, and a power-law distribution.

We analytically derive the corresponding decay laws and test the predictions using *pseudogenes*. (We use this terminus in analogy to *pseudoexons* and *pseudochromosomes* for corresponding stochastic model sequences. It should not be confused with knocked out genes which are termed pseudo genes as well.) These consist of interspersed pseudoexons within a *random sea*, i.e., statistically independent letters with the same base composition as the pseudoexons. The length of each exon is chosen randomly from the distribution $\rho(l)$ under consideration. In all simulations in this section we have chosen the codon usage table of the HUMBMYH7 gene studied in Sec. III. Details of the simulations are described in the figure captions.

As a first model we discuss a fixed length of all protein coding segments $l=L$,

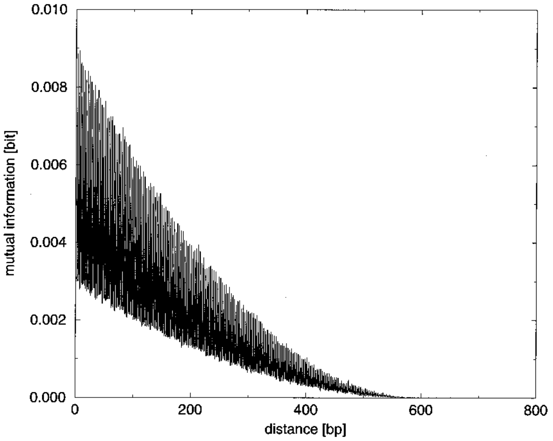

FIG. 6. Mutual information of a 10^6 bp long random sequence. Within a ‘‘random sea’’ of independent letters A, C, G, and T, 1000 pseudoexons of a length 600 bp have been interspersed. For small k, we observe the expected period-three oscillations between F^2I_{in} and F^2I_{out} [see Eqs. (7) and (20)]. Please note that Eq. (24) predicts exactly the parabolic decay between $k=0$ and $k=600$.

$$\rho(l)=\delta_{lL}. \tag{22}$$

This yields

$$\overline{n}=n(L)=\frac{FN}{L}. \tag{23}$$

For $k<L$ we obtain essentially a linear decay of the envelope $E(k)$

$$E(k)\propto\frac{FN}{N-k}\frac{L-k}{L}\approx F\frac{L-k}{L}. \tag{24}$$

The k dependence of the denominator can be neglected for $N\gg L$. The resulting linear decay of $D_{ij}(k)$ implies a quadratic decay of the mutual information function [cf. Eq. (7)]. Such a parabola is seen in Fig. 6 for a pseudogene with constant exon length.

Of course, it is more realistic to assume an exponentially decaying length distribution (compare Fig. 5). As above in Eq. (24), we neglect the k dependence of the denominator. For the sake of simplicity, we further replace the summation in Eq. (21) by an integration from k to infinity. Then an exponential length distribution

$$\rho(l)=\lambda\ \exp(-\lambda l) \tag{25}$$

gives an exponential decay of the envelope

$$E(k)\propto F\lambda\int_k^\infty(l-k)\rho(l)dl=F\ \exp(-\lambda k). \tag{26}$$

Figure 7 displays the results for a corresponding simulation of a pseudogene.

As a last example we consider a power-law decay from a lower cutoff length L_{min} with an exponent $\beta>2$

$$\rho(l)=(\beta-1)L_{min}^{\beta-1}l^{-\beta}\quad\text{for } l\geq L_{min} \tag{27}$$

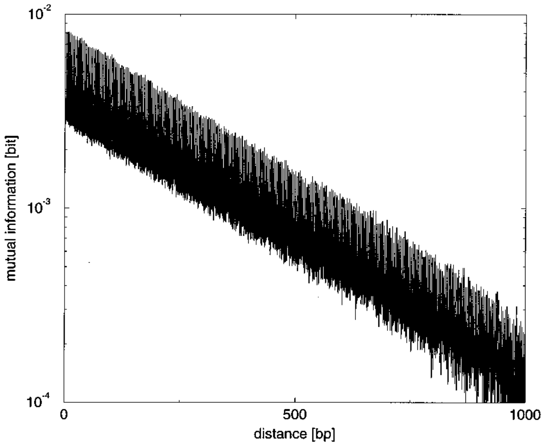

FIG. 7. Mutual information of a 10^6 bp long sequence containing 1000 pseudoexons with exponentially distributed lengths (mean value 600 bp). The logarithmic vertical scale reveals the predicted exponential decay.

and zero otherwise. The mean value of the length is then given by

$$\overline{l}=\frac{\beta-1}{\beta-2}L_{min}. \tag{28}$$

After integration we obtain a power-law decay of the envelope for $k>L_{min}$

$$E(k)\propto\frac{FL^{\beta-2}}{\beta-1}k^{2-\beta}. \tag{29}$$

The log-log presentation of the mutual information in Fig. 8 indicates indeed a power law for a simulation of a corresponding pseudogene.

These examples show how the length distribution of exons affects the decay of correlations, which are due the nonuniform codon usage. In Sec. VI we show that our considerations apply to DNA sequences and that a considerable amount of observed correlations can be predicted just by knowing codon usage tables and the length distribution of exons.

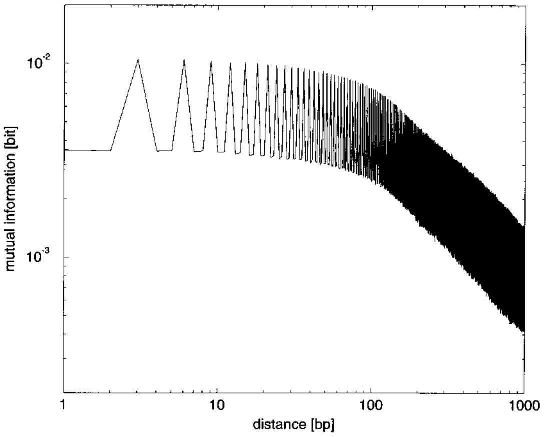

FIG. 8. Mutual information of a 7×10^6 bp long sequence with 7000 pseudoexons. The parameters of the exon length distribution are $L_{min}=150$ and $\beta=\frac{9}{4}$.

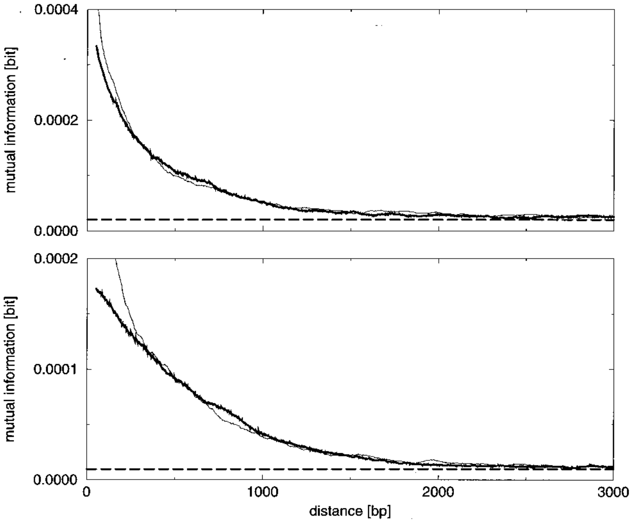

FIG. 9. Decay of the mutual information function for yeast chromosomes (thin lines) and the corresponding pseudochromosomes (thick lines). In order to reduce the strong fluctuations (compare Fig. 1) and to focus on the decay we have applied a 99 bp running average. Upper graph: Chromosome III. The codon usage table was taken from the temperature-sensitive lethal TSM1 protein (4 221 bp). Lower graph: Chromosome XI, table from the ORF which encodes dynein (12 276 bp).

VI. APPLICATIONS TO DNA SEQUENCES

In this section we apply our concept to representative DNA sequences. It was already demonstrated in Fig. 1 that the periodicity due to the nonuniform codon usage plays a significant role in yeast DNA. This is due to large fraction of coding sequences ($F\approx0.7$) and rather long exons (compare Fig. 5). In order to quantify the effect of exons on correlations we generate *pseudochromosomes* as follows: codon usage tables are taken from long yeast genes as a basis for the simulation of pseudoexons (see Sec. III). In order to simulate strand symmetry, 50% of the pseudoexons are generated with the complementary codon usage table. The empirical histogram from the corresponding chromosome is taken as length distribution for the interspersed pseudoexons. In between the pseudoexons Bernoulli sequences with the same base composition are inserted. In this way a stochastic model of a chromosome is defined which incorporates only well-known features—the nonuniform codon usage and the alternation of coding segments and intervening sequences.

Figure 9 reveals that the decay for the pseudochromosomes is quite similar to the actual decay for the yeast chromosomes. Only for small k additional correlations can be seen which are discussed in Sec. VII.

Similar agreement was also found for codon usage tables from other protein coding segments and for some strand asymmetry.

Significant long-range correlations in the yeast chromosome III up to several kilo base pairs have been reported by Munson, Taylor, and Michaels [48]. The existence of such

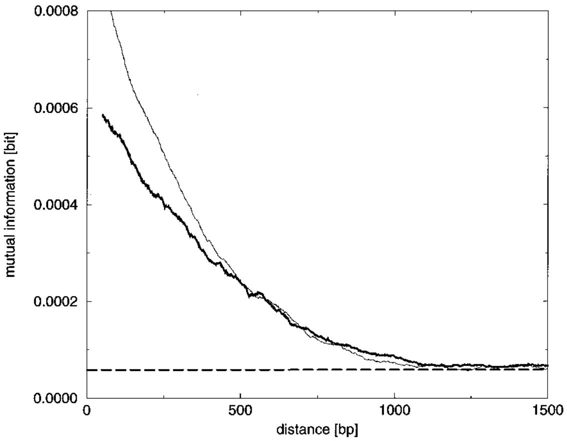

FIG. 10. Mutual information decay for the *E. coli* chromosome region (thin line) and a corresponding pseudoregion with the same length distribution of (pseudo-) exons. The codon usage table was taken from the isoleucil-tRNA ligase (2 811 bp). As in Fig. 9 a 99 bp running average was applied.

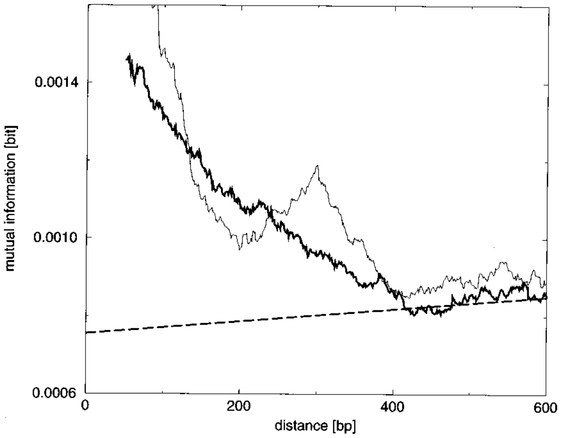

FIG. 11. Comparison of the smoothed mutual information (99 bp running average) of Brugia malayi myosin heavy chain gene (8 600 bp from the first to the last exon) and a corresponding random sequence with the same exon length distribution and codon usage. Since the sample size decreases with the distance there is a clear increase of the bias (see also the Appendix).

correlations is indeed corroborated by our mutual information analysis. However, they exist also in a pseudochromosome (see Fig. 9), and hence, the length distribution of exons is sufficient to explain these correlations.

In the same way as for the yeast chromosome, we generated a stochastic model of a DNA region of *E. coli* (see Fig. 2). Figure 10 shows a comparison of the mutual information functions.

Finally, we discuss the correlation decay in the myosin heavy chain gene M74000 of *Brugia malayi*. We have chosen this gene since the 15 exons constitute about 68% of the total gene. Consequently, the correlations due to the exons and their length distribution are more pronounced then in genes with only a few percent of exons. (In fact, the decay for the human myosin heavy chain depicted in Fig. 3 is also strongly influenced by correlations within its introns.) The codon usage table and empirical length distribution of the analyzed gene are taken to generate a pseudogene as described in Sec. V. Since there are fairly long exons in this gene, Fig. 11 displays the expected long tail of the envelope. Quite similar correlations are found in the corresponding pseudogene (thick line) pointing to the fact that most correlations are solely due to the length distribution of exons. It turns out that for such relatively short DNA sequences a careful calculation of the bias (dashed line in Fig. 11) is necessary for a correct interpretation of the decay.

VII. SUMMARY AND DISCUSSION

Our paper was devoted to relating a significant part of observed long-range correlations to the pattern of protein coding segments. We have shown that the triplet code induces via a nonuniform codon usage persistent oscillations of correlation measures. By taking into account the length distributions of exons, a long-ranging decay of the mutual information function and correlation functions could be predicted. For example, a power-law distribution of the exon length implies a power-law decay of correlation measures.

Pseudochromosomes based on the empirical length distribution in yeast chromosomes exhibit a quite similar decay of correlations and, therefore, most of the correlations in yeast DNA could be traced back to a simple origin. Our considerations apply to all parts of genomes where coding segments constitute a significant portion of the DNA such as bacteria or retroviruses. This was exemplified for a DNA region of *E. coli* and for a myosin heavy chain gene with a large fraction of exons.

Typically, in higher eukaryotes only a few percent of the DNA are protein coding regions. Consequently, observed long-range correlation in DNA as the human β-globin region [17] or in genes with very long introns [16] cannot be explained simply by the nonuniform codon usage within exons. Moreover, the well-known compositional variations along chromosomes on scales above 10^5 bp [28,34,35] are beyond the scope of our analysis.

Our concept is, however, more generally applicable. It can be formulated as follows: (i) look for fragments of differing statistical properties, (ii) analyze its length distribution, (iii) define appropriate (stochastic) pseudosequences, and (iv) analyze their correlation decay, and (v) compare it with the empirical mutual information. Related stochastic models of the DNA heterogeneity have a long tradition [27,32,36,38], but these models are based on hypothetical length distributions of fragments. Contrarily, our approach simply exploits the well-known length distribution of exons.

As a first step of a more general approach, Schmitt, Ebeling, and Herzel [49] recently studied length distributions of over-represented ''words'' termed *modules*. We suggest analyzing also length distributions of—for example—isochores [30], gene clusters, dispersed repeats, simple-sequence DNA, or CpG islands. If one takes into account different compositions of exons and introns, the length distribution of introns comes into play as well. We expect that stochastic models which include the actual length distributions of all these segments can relate most observed long-range correlations to known biological structures.

Though we have quantitatively explained the origin of long-range correlations in mostly protein coding sequences,

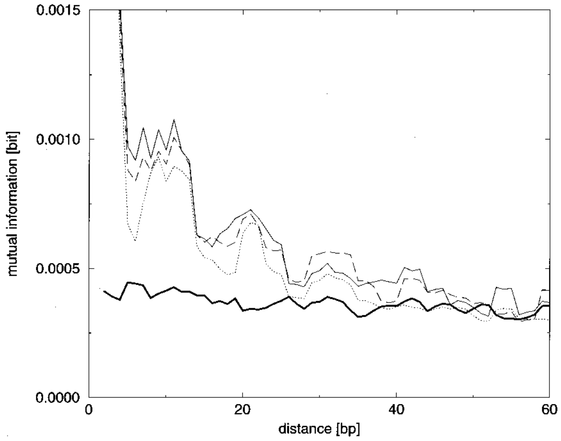

FIG. 12. Mutual information of yeast chromosomes III (full line), IX (dashed line), XI (dotted line) for short distances. In order to eliminate the dominating period-three oscillations, we apply a running average over 3 bp. The comparison with a pseudochromosome (thick line) reveals additional correlations (in particular, a 10–11 bp period).

many questions remain open. For example, correlations within introns and intergenic sequences were not the subject of this paper. Moreover, we have seen in Fig. 9 additional correlations in yeast DNA for small distances, which cannot be explained by our pseudoexon concept. Figure 12 displays an example of such a peak structure with a periodicity of about 10 BP. These peaks may reflect the *pitch* of DNA, i.e., a 10.5 BP periodicity that has been found in curved DNA [50,51] and DNA folded into nucleosomes [52]. Additionally, the well-known three-four amino acid periodicities in α-helical proteins [12,53–55] are a possible source of the observed peak structure.

In summary, we have shown in this paper that the length distribution of exons in real DNA induces long-range correlations which can be described by appropriate stochastic models. We stress, finally, that beyond these correlations other DNA base pair fluctuations exist on various scales [13,16–18,30,35]. Their role for the chromosome organization and gene expression has still to be explored.

ACKNOWLEDGMENTS

We thank S. Buldyrev, W. Ebeling, J. Fickett, R. Guigo, W. Li, E. Stanley, E. Trifonov, and O. Weiss for stimulating discussions and acknowledge support by the Deutsche Forschungsgemeinschaft, the Studienstiftung des Deutschen Volkes, and the National Institutes of Health.

APPENDIX: ESTIMATION OF THE MUTUAL INFORMATION FROM FINITE SAMPLES

In this Appendix, we derive analytic expressions for statistical and systematic errors that occur by estimating the mutual information function $I(k)$ from finite sequences. The estimator we use throughout our paper is the so called natural estimator $\hat{I}(k)$ of the mutual information function, which is defined as

$$\hat{I}(k)=\sum_{i,j=1}^{4}\hat{p}_{ij}(k)\ln\frac{\hat{p}_{ij}(k)}{\hat{p}_i\hat{p}_j}, \tag{A1}$$

where $\hat{p}_{ij}(k)$ and $\hat{p}_i$ denote the relative frequencies rather than the (unknown) probabilities $p_{ij}(k)$ and p_i defined in Sec. II. Note that we have measured, so far, the mutual information in bits which corresponds to the logarithm of base 2 in Eq. (1). In this Appendix we use the natural logarithm for convenience.

As the estimates $\hat{p}_{ij}(k)$ and $\hat{p}_i$ vary from sequence to sequence, the values of $\hat{I}(k)$ also fluctuate. Our task will be to find approximate closed form expressions for the mean and the variance of the distribution of $\hat{I}(k)$ as well as to derive its asymptotic form in the limit of large sequence lengths.

1. The mutual information bias

Let us start with expressing the natural mutual information function estimator $\hat{I}(k)$ in terms of the natural estimators of the one-gram and two-gram Shannon entropies, which we denote by $\hat{H}_1$ and $\hat{H}_2(k)$, respectively.

$$\hat{I}(k)=\sum_{i,j=1}^{4}\hat{p}_{ij}(k)\ln\hat{p}_{ij}(k)-2\sum_{i=1}^{4}\hat{p}_i\ln\hat{p}_i \tag{A2}$$

$$=2\hat{H}_1-\hat{H}_2(k). \tag{A3}$$

By expanding $\ln\hat{p}_{ij}(k)$ and $\ln\hat{p}_i$ about $p_{ij}(k)$ and p_i, we obtain a power series expansion of the expectation value $E(\hat{I}(k))$ in terms of moments of the multinomial distribution, all of which can be derived by elementary methods. Using the biases of $\hat{H}_1$ and $\hat{H}_2(k)$ derived in [12,56,57], we obtain

$$E(\hat{I}(k))=2E(\hat{H}_1)-E(\hat{H}_2(k)) \tag{A4}$$

$$=2\left(H_1-\frac{3}{2N}\right)-\left(H_2(k)-\frac{15}{2N}\right)+O(1/N^2) \tag{A5}$$

$$=I(k)+\frac{9}{2N}+O(1/N^2). \tag{A6}$$

This states, that (on average) we overestimate the mutual information by an amount of $9/2N$ nits due to the finite length of the studied sequence.

2. The mutual information variance

For the sake of simplicity, we omit the distance k as the argument of the mutual information function $I(k)$, the Shannon entropy $H_2(k)$, the probabilities $p_{ij}(k)$, and their estimators. According to Eq. (A2) and denoting the covariance between two random variables a and b by $\mathrm{cov}(a,b)$, we obtain

$$\sigma^2(\hat{I})=4\sigma^2(\hat{H}_1)+\sigma^2(\hat{H}_2)-4\mathrm{cov}(\hat{H}_1,\hat{H}_2) \tag{A7}$$

for the variance of the mutual information estimate $\hat{I}$. The first two terms in this equation are already given in [56,57], who derive that

$$\sigma^2(\hat{H}_1)=\frac{1}{N}\left(\sum_{i=1}^{4} p_i \ln^2 p_i - H_1^2\right)+O(1/N^2) \qquad \text{(A8)}$$

as well as

$$\sigma^2(\hat{H}_2)=\frac{1}{N}\left(\sum_{i,j=1}^{4} p_{ij} \ln^2 p_{ij} - H_2^2\right)+O(1/N^2). \qquad \text{(A9)}$$

Therefore, we dedicate the following paragraph to deriving the covariance between the Shannon entropy estimates $\hat{H}_1$ and $\hat{H}_2$, which appear to be not at all independent but highly correlated as observed in [22].

Following the lines in [58], we obtain

$$\begin{aligned}\mathrm{cov}(\hat{H}_1,\hat{H}_2)&=E((\hat{H}_1-H_1)(\hat{H}_2-H_2))\\&\quad -E(\hat{H}_1-H_1)E(\hat{H}_2-H_2) \qquad \text{(A10)}\\&\propto E\left(\sum_{i,j=1}^{4}(\ln p_{ij})(\hat{p}_{ij}-p_{ij})\right.\\&\quad\left.\times\sum_{k=1}^{4}(\ln p_k)(\hat{p}_k-p_k)\right). \qquad \text{(A11)}\end{aligned}$$

The symbol $\propto$ indicates that $O(1/N^2)$ terms are neglected. Further calculations yield

$$\begin{aligned}\mathrm{cov}(\hat{H}_1,\hat{H}_2)&=\sum_{i,j,k,l=1}^{4}\ln p_{ij}\ln p_k E((\hat{p}_{ij}-p_{ij})(\hat{p}_{kl}-p_{kl}))\\&=\sum_{i,j=1}^{4}\ln p_{ij}\ln p_i\frac{p_{ij}(1-p_{ij})}{N} \qquad \text{(A12)}\\&\quad-\sum_{i,j,k,l=1}^{4}(1-\delta_{ik})(1-\delta_{jl})\\&\quad\times\ln p_{ij}\ln p_k\frac{p_{ij}p_{kl}}{N} \qquad \text{(A13)}\\&=\frac{1}{N}\sum_{i,j=1}^{4}\ln p_{ij}\ln p_i p_{ij}\\&\quad-\frac{1}{N}\sum_{i,j,k,l=1}^{4}\ln p_{ij}\ln p_k p_{ij}p_{kl} \qquad \text{(A14)}\\&=\frac{1}{N}\mathrm{cov}(\ln p_i,\ln p_{ij}), \qquad \text{(A15)}\end{aligned}$$

which relates the covariance of the natural Shannon entropy estimates to the covariance of the logarithms of the underlying probabilities: the covariance between the observed two-gram Shannon entropy and its marginal one-gram Shannon entropy observed from the same sample of size N is, in a first-order approximation, equal to the covariance between the logarithms of the joint probabilities p_{ij} and the logarithms of their marginal probabilities p_i divided by N.

Let us eventually derive an approximation for the correlation coefficient r between $\hat{H}_1$ and $\hat{H}_2$, which is defined as the normalized covariance

$$\begin{aligned}r(\hat{H}_1,\hat{H}_2)&\equiv\frac{\mathrm{cov}(\hat{H}_1,\hat{H}_2)}{\sqrt{\sigma^2(\hat{H}_1)\sigma^2(\hat{H}_2)}} \qquad \text{(A16)}\\&\propto\frac{\mathrm{cov}(\ln p_i,\ln p_{ij})}{\sqrt{\sigma^2(\ln p_i)\sigma^2(\ln p_{ij})}} \qquad \text{(A17)}\\&=r(\ln p_i,\ln p_{ij}). \qquad \text{(A18)}\end{aligned}$$

This is a really noticeable result, since the right hand side of this equality does not depend on the sample size N. It states that the correlation coefficient between the natural estimates of the statistics $\hat{H}_1$ and $\hat{H}_2$ is independent of the sequence length and given by the correlation coefficient between the logarithm of the joint probabilities p_{ij} and the logarithm of their marginal probabilities p_i. Since $\hat{H}_1$ and $\hat{H}_2$ of DNA sequences are strongly correlated, we understand why the mutual information fluctuations are small compared to the fluctuations of both $\hat{H}_1$ and $\hat{H}_2$ [22].

By combining Eqs. (A7)–(A9) with Eq. (A15), we obtain

$$\begin{aligned}\sigma^2(\hat{I})&\propto\frac{4}{N}\sigma^2(\ln p_i)+\frac{1}{N}\sigma^2(\ln p_{ij})-\frac{4}{N}\mathrm{cov}(\ln p_i,\ln p_{ij}) \qquad \text{(A19)}\\&=\frac{1}{N}\sigma^2(\ln p_{ij}-2\ln p_i) \qquad \text{(A20)}\\&=\frac{1}{N}\sigma^2\left[\ln\left(\frac{p_{ij}}{p_ip_j}\right)\right]. \qquad \text{(A21)}\end{aligned}$$

Note again that this equality relates the *sample variance* of the mutual information estimates to the variance of the 16 numbers $\ln[(p_{ij}/p_ip_j)]$ as worked out in more detail in [58].

3. The asymptotic mutual information distribution

In the following, we denote the statistical dependences by $\hat{D}_{ij}=\hat{p}_{ij}-\hat{p}_i\hat{p}_j$ and expand the mutual information $\hat{I}$ in a Taylor series about $\hat{D}_{ij}$:

$$\begin{aligned}\hat{I}&=\sum_{i,j=1}^{4}\hat{p}_{ij}\ln\frac{\hat{p}_{ij}}{\hat{p}_i\hat{p}_j} \qquad \text{(A22)}\\&=\sum_{i,j=1}^{4}(\hat{D}_{ij}+\hat{p}_i\hat{p}_j)\ln\left(1+\frac{\hat{D}_{ij}}{\hat{p}_i\hat{p}_j}\right) \qquad \text{(A23)}\\&=\sum_{i,j=1}^{4}(\hat{D}_{ij}+\hat{p}_i\hat{p}_j)\left(\frac{\hat{D}_{ij}}{\hat{p}_i\hat{p}_j}-\frac{\hat{D}_{ij}^2}{2\hat{p}_i^2\hat{p}_j^2}+\cdots\right) \qquad \text{(A24)}\\&=\sum_{i,j=1}^{4}\frac{\hat{D}_{ij}^2}{2\hat{p}_i\hat{p}_j}+O(\hat{D}_{ij}^3). \qquad \text{(A25)}\end{aligned}$$

The quantity

$$\chi^2 \equiv N \sum_{i,j=1}^{4} \frac{\hat{D}_{ij}^2}{\hat{p}_i \hat{p}_j} \tag{A26}$$

is known as χ^2 *statistics*, which asymptotically approaches a χ^2-probability distribution with nine degrees of freedom [40]. For Bernoulli sequences with vanishing D_{ij}, the resulting asymptotic probability density of the natural mutual information estimates $\hat{I}$ reads

$$P(\hat{I}) = \frac{N^{9/2} \hat{I}^{7/2}}{\Gamma(9/2)} e^{-N\hat{I}}. \tag{A27}$$

Otherwise, $2N\hat{I}$ asymptotically approaches a noncentral χ^2-probability distribution for nonvanishing D_{ij} [40,59].

These reviewed expressions for the bias, variance, and asymptotic distribution provide a firm statistical basis for applications of the mutual information function in sequence analysis.

[1] L. L. Gatlin, *Information Theory and the Living System* (Columbia University Press, New York, 1972).

[2] E. N. Trifonov and V. Brendel, *Gnomic–A Dictionary of Genetic Codes* (Balaban, Rehovot, 1986).

[3] G. von Heijne, *Sequence Analysis in Molecular Biology–Treasure Trove or Trivial Pursuit* (Academic, San Diego, 1987).

[4] *Computers and DNA*, edited by G. Bell and T. Marr (Addison-Wesley, Reading, 1990).

[5] J. D. Watson, M. Gilman, J. Witkowski, and H. Zoller, *Recombinant DNA* (Freeman, New York, 1992).

[6] H. P. Yockey, *Information Theory and Molecular Biology* (Cambridge University Press, Cambridge, England, 1992).

[7] N. A. Kolchanov and H. A. Lim, *Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution* (World Scientific, Singapore, 1994).

[8] B. Lewin, *Genes V* (Oxford University Press, London, 1994).

[9] J. W. C. Shepherd, J. Mol. Evol. **17**, 94 (1981).

[10] J. W. Fickett, Nucl. Acid Res. **10**, 5303 (1982).

[11] W. Ebeling, R. Feistel, and H. Herzel, Phys. Scr. **35**, 761 (1987).

[12] H. Herzel, Syst. Anal. Mod. Sim. **5**, 435 (1988).

[13] E. N. Trifonov, Bull. Math. Biol. **51**, 417 (1989).

[14] J. W. Fickett and C.-S. Tung, Nucl. Acid Res. **20**, 6441 (1992).

[15] I. Große, H. Herzel, S. V. Buldyrev, and H. E. Stanley (unpublished).

[16] W. Li, Int. J. Bif. Chaos **2**, 137 (1992).

[17] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortina, M. Simons, and H. E. Stanley, Nature **356**, 186 (1992).

[18] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[19] B. Borštnik, D. Pumpernik, and D. Lukman, Europhys. Lett. **23**, 389 (1993).

[20] H. Herzel, A. O. Schmitt, and W. Ebeling, Chaos, Solitons Fractals **4**, 97 (1994).

[21] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[22] H. Herzel and I. Große, Physica A **216**, 518 (1995).

[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, 1996).

[24] V. R. Chechetkin and A. Yu. Turygin, J. Phys. A **27**, 4875 (1994).

[25] C. A. C. Dreismann and D. Larhammer, Nature **361**, 212 (1993).

[26] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng, and M. Simons, Physica A **205**, 214 (1994).

[27] R. A. Elton, J. Theor. Biol. **45**, 533 (1974).

[28] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier, Science **228**, 953 (1985).

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

[30] G. Bernardi, Ann. Rev. Genet. **23**, 637 (1989).

[31] T. Ikemura, K.-N. Wada, and S.-I. Aota, Genomics **8**, 207 (1990).

[32] J. W. Fickett, D. C. Torney, and D. R. Wolf, Genomics **13**, 1056 (1992).

[33] S. Karlin and V. Brendel, Science **259**, 677 (1993).

[34] H. Feldmann *et al.*, EMBO J. **13**, 5795 (1994).

[35] B. Dujon *et al.*, Nature **369**, 371 (1994).

[36] H. Herzel, W. Ebeling, and A. O. Schmitt, Phys. Rev. E **50**, 5061 (1994).

[37] G. A. Churchill, Bull. Math. Biol. **51**, 79 (1989).

[38] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).

[39] W. Li, T. G. Marr, and K. Kaneko, Physica D **75**, 392 (1994).

[40] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).

[41] H. Herzel and W. Ebeling, Phys. Lett. A **111**, 1 (1985).

[42] W. Li, J. Stat. Phys. **60**, 823 (1990).

[43] L. Luo and H. Li, Bull. Math. Biol. **53**, 345 (1991).

[44] T. Ikemura, J. Mol. Biol. **146**, 1 (1981).

[45] R. Staden, Nucl. Acid Res. **12**, 551 (1984).

[46] P. M. Sharp and W.-H. Li, Nucl. Acid Res. **15**, 1281 (1987).

[47] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and M. Simons, Biophys. J. **65**, 2673 (1993).

[48] P. J. Munson, R. C. Taylor, and G. S. Michaels, Nature **360**, 636 (1992).

[49] A. O. Schmitt, W. Ebeling, and H. Herzel, BioSystems **37**, 199 (1996).

[50] E. M. Trifonov and J. L. Sussman, Proc. Natl. Acad. Sci. USA **77**, 3816 (1980).

[51] A. K. Konopka and G. M. Smythers, CABIOS **3**, 193 (1987).

[52] I. Ioshikhes, A. Bolshoy, and E. N. Trifonov, J. Biomol. Struct. Dyn. **9**, 1111 (1992).

[53] M. I. Kanehisa and T. Y. Tsong, Biopolymers **19**, 1617 (1980).

[54] S. H. White, Annu. Rev. Biophys. Biomol. Struct. **23**, 407 (1994).

[55] A. O. Schmitt, E. Kolker, and E. N. Trifonov (unpublished).

[56] G. P. Basharin, Theory Prob. Appl. **4**, 333 (1959).

[57] B. Harris, Topics Inf. Theory (Keszhtely) **16**, 323 (1975).

[58] I. Große, *Statistical Analysis of Biosequences*, Diplomthesis, Humboldt University, Berlin, 1995.

[59] N. L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Univariate Distributions* (Houghton-Mifflin, Boston, 1970).