Repeats and correlations in human DNA sequences

Dirk Holste*
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Ivo Grosse
Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Stephan Beirer, Patrick Schieg, and Hanspeter Herzel[†]
Institute for Theoretical Biology, Humboldt University Berlin, Invalidenstrasse 43, D-10115 Berlin, Germany

(Received 22 January 2003; published 26 June 2003)

We study the nucleotide-nucleotide mutual information function $I(k)$ of the DNA sequences of the three completely sequenced human chromosomes 20, 21, and 22. We find in each human chromosome (i) the absence of the $k=3$ base pair (bp) sequence periodicity characteristic for protein coding regions, (ii) the absence of the $k=10-11$ bp sequence periodicity characteristic for both protein secondary structure and DNA bendability, and (iii) the presence of significant statistical dependencies at about $k=135$ bp and at about $k=165$ bp. We investigate to which degree the density and composition of interspersed repeats might explain these observed statistical patterns in all three human chromosomes. We use simple stochastic models to substitute known interspersed repeats and find by numerical studies that (iv) the presence of interspersed repeats dominates short-range correlations as measured by $I(k)$ on the scale of several hundred base pairs in human chromosomes 20, 21, and 22. On the other hand, we find that (v) interspersed repeats contribute only weakly to long-range correlations due to the clustering of highly abundant *Alu* repeats.

DOI: 10.1103/PhysRevE.67.061913 PACS number(s): 87.10.+e, 02.50.−r, 05.40.−a

I. INTRODUCTION

The analysis of statistical patterns in genomic DNA is of interest, since correlations may reflect biologically significant features of primary structures [1–4]. For instance, the sequence periodicity of 3 base pairs (bp) indicates the presence of protein coding sequences, such that this signal can be used to distinguish coding and noncoding DNA [5], and sequence periodicities of 10–11 bp reflect DNA bendability [6,7] as well as the secondary structure of proteins [8,9]. On the next length scale, correlations in the order of 10^2 bp have been found in random walk studies [10]. In subsequent studies it has been proposed that correlations on this length scale can be explained by the nucleosomal structure in eukaryotes [11,12]. Compositional heterogeneities on length scales exceeding hundreds of base pairs and ranging up to about 10^6 bp are a well-known biological phenomenon related to the presence of isochores [13,14] and these long-range correlations can be approximated by power laws [15–17].

Initial analyses of the first draft form of the human genome [18,19] show that protein coding regions constitute less than 3% of the total genome, which makes the complete annotation of protein coding and noncoding regions a difficult task [18–20]. In contrast to the small percentage of protein coding DNA, about 50% of the human genome consists of repetitive sequences [18]. Repeats are multiple approximate copies of patterns of nucleotides of various lengths, most of which are dispersed throughout the genome. Hence, more than the presence of protein coding regions, the presence of repetitive sequences can be expected to influence correlations in human DNA sequences to a large extent.

Here, we study the mutual information function $I(k)$ of the DNA sequences of chromosomes 20, 21, and 22 in order to investigate short- and long-range correlations in these three completed human chromosomes [21]. The paper is organized as follows: in Sec. II we introduce the notation and define $I(k)$, in Sec. III we study short-range $(10\ldots10^2$ bp) and long-range $(10^3\ldots10^6$ bp) nucleotide-nucleotide correlations by computing $I(k)$ of the DNA sequences of chromosomes 20, 21, and 22, and in Secs. IV–VI we discuss the presence of interspersed repeats and study to which degree these correlations might be related to repeats common in all the three human chromosomes.

II. SYMBOLS AND DEFINITIONS

The primary structure of DNA is polymeric and can be considered as a symbolic sequence of $\lambda=4$ symbols $\{A_1,A_2,A_3,A_4\}\equiv\{\mathrm{A,C,G,T}\}$, where 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 and by $p_{ij}(k)$ the relative frequency of the pair of symbols A_i and A_j in a distance k. Assuming that the DNA sequence under study can be considered as a realization of a stationary and ergodic process, one may associate with p_i the probability of finding, at any given sequence position, the symbol A_i, and one may associate with $p_{ij}(k)$ the joint probability of finding, at any given positions spaced by k symbols, the pair of symbols A_i and A_j. Two symbols in a distance k are statistically independent if $p_{ij}(k)$ factorizes to $p_{ij}=p_ip_j$ for all i and j. The nucleotide-nucleotide mutual information function [22,23]

*Electronic address: holste@mit.edu

[†]Electronic address: h.herzel@biologie.hu-berlin.de

TABLE I. Selected features for human chromosomes 20, 21, and 22 [26].

Feature	Chr20	Chr21	Chr22
Length (units of 10^6 bp)	59.1	33.8	33.8
G+C (%)	44	41	48
Genes	727	225	546
Repeats (%)	42	38	42
Alu repeats	27931	11874	22659
Alu repeats (%)	13	9	17

[a] Putative protein coding gene count.

[b] Based on annotation using REPEATMASKER [34].

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

quantifies in units of bits the amount of information that one can obtain about the identity of symbol A_j by learning the identity of symbol A_i located k symbols upstream of A_j. Clearly, $I(k)=0$ for random, uncorrelated sequences, and $I(k)>0$ if $p_{ij} \neq p_i p_j$ for some *i* and *j*, so $I(k)$ measures any deviation from statistical independence. Expanding $I(k)$ in terms of $p_{ij}-p_i p_j$, one finds that up to second order terms the mutual information function is proportional to the sum over all λ^2 squared correlation functions $C(k)$, and so a power-law decay of $C(k) \sim k^{-\gamma}$ is equivalently described by $I(k) \sim k^{-2\gamma}$ [24]. Finite sample effects bias estimates of $I(k)$ and the bias of the mutual information function increases with the distance *k*. Approximate analytic expressions for systematic and statistical errors are summarized elsewhere [24,25].

III. CORRELATIONS IN HUMAN CHROMOSOMES 20, 21, AND 22

In this section we study $I(k)$ of human chromosomes 20, 21, and 22, with the goal of quantifying short-range and long-range correlations in the DNA sequences of these first three finished human chromosomes [26]. Table I summarizes several statistical features of chromosomes 20, 21, and 22, including the proportion of various interspersed repeat categories. We note, in passing, that these three chromosomes contain about 40% interspersed repeats, and chromosome 22 has a higher concentration of G and C nucleotides than in chromosomes 20 and 21.

Figure 1 shows $I(k)$ in the DNA sequences of human chromosomes 20, 21, and 22 for $k=1,2,\ldots,10^6$ bp. We find significant statistical dependencies up to 10^6 bp in all the three chromosomes, and we find that the decay of $I(k)$ in these three human chromosomes can be approximated by power laws with exponents $2\gamma^{(\mathrm{Chr20})} \approx 0.5$, $2\gamma^{(\mathrm{Chr21})} \approx 0.5$, and $2\gamma^{(\mathrm{Chr22})} \approx 0.6$. The decay of $I(k)$ in chromosome 22 is steeper than $I(k)$ in chromosomes 20 and 21, and it decays up to $k=10^5$ bp, whereas in chromosomes 20 and 21 it ranges up to $k=10^4$ bp. This is in accord with previous analyses of the pronounced heterogeneity in human chromosome 22 [27–29].

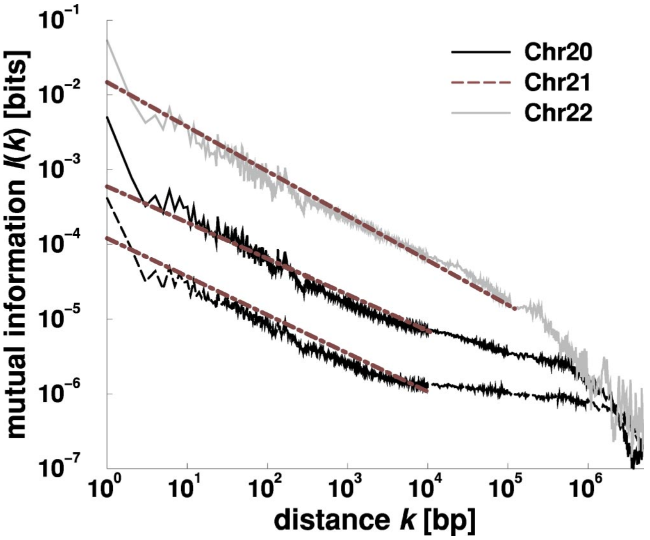

FIG. 1. Mutual information function $I(k)$ of the DNA sequences of human chromosomes (Chr) 20, 21, and 22. $I(k)$ has been multiplied by 10^{-1} (Chr20) and by 10^{-2} (Chr21), respectively, to allow for clear representation. The systematic error (standard deviation) due to finite sample size is in the order of 10^{-7} (10^{-10}) bits [24,25]. Lines represent the least-squares regressions for $k=1,2,\ldots,10^4$ bp with exponents $2\gamma^{(\mathrm{Chr20})} \approx 0.5$ and $2\gamma^{(\mathrm{Chr21})} \approx 0.5$, and for $k=1,2,\ldots,10^5$ bp with exponent $2\gamma^{(\mathrm{Chr22})} \approx 0.6$.

In a related study [12], the compositional heterogeneity in human chromosome 21 has been examined by spectral analysis [30]. Since the spectral exponent $S(f) \sim f^{-\beta}$ is related to $I(k) \sim k^{-2\gamma}$ via $\gamma = 1-\beta$, a value of $\beta^{(\mathrm{Chr21})} \approx 0.7$ corresponds to $2\gamma^{(\mathrm{Chr21})} \approx 0.6$, which is comparable to our findings from Fig. 1. It has been demonstrated that these long-range correlations are mainly due to the presence of isochores in human DNA [14,17,18,29].

In the remaining part of this paper we concentrate on the analysis of short-range correlations up to about $k=200$ bp. Figure 2 shows $I(k)$ of the DNA sequences of human chromosomes 20, 21, and 22 for $k=1,2,\ldots,200$ bp, and we find several statistically significant signals in this range. Figure 2 shows that in contrast to the correlation structure in yeast chromosomes [7], $I(k)$ in human chromosomes 20, 21, and 22 is neither dominated by sequence periodicities of $k=3$ bp nor by sequence periodicities of $k=10$–11 bp. Instead, in those three chromosomes $I(k)$ exhibits pronounced peaks at about $k=135$ bp and at about $k=165$ bp that cannot be attributed to statistical fluctuations. In the following sections we discuss the origin of these signals in connection with the presence of interspersed repeats.

IV. REPEAT CONTENT IN HUMAN CHROMOSOMES 20, 21, AND 22

Repetitive sequences form almost half of the human genome [18,19]. They can be classified into several categories [31,32], and about 45% of the human genome belongs to interspersed repeats [18,33]. Other repeat categories comprise, e.g., direct repetitions of oligonucleotides or tandemly repeated sequences. Interspersed repeats fall into several sub-

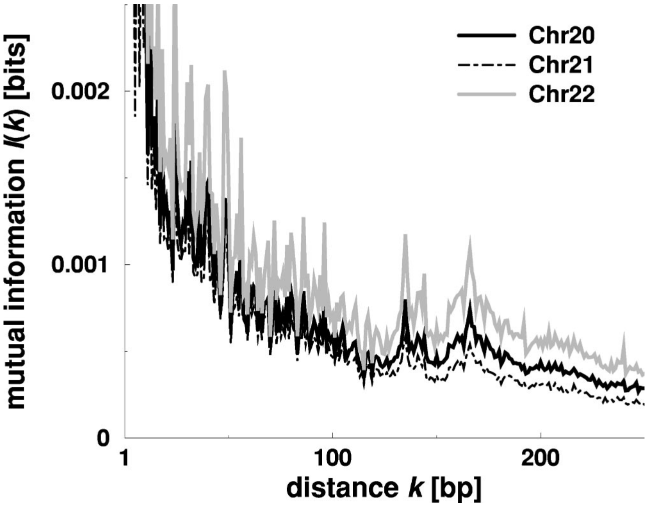

FIG. 2. Mutual information function $I(k)$ of the DNA sequences of human chromosomes 20, 21, and 22. $I(k)$ shows clear correlations over several hundred base pairs. On the length scale $100<k<200$ bp, two pronounced signals at about $k=135$ bp and at about $k=165$ bp are common to all three human chromosomes.

classes, such as long interspersed nucleic elements (LINEs) with a length of up to 6×10^3 bp and short interspersed nucleic elements (SINEs) with a length of several hundred base pairs.

One established procedure of finding repeats is based on a comprehensive organism-specific collection of presently known repetitive sequences, and the recognition of highly homologous stretches of DNA by similarity search has been implemented in REPEATMASKER [34], a program to locate and classify interspersed repeats in DNA sequences. Figure 3 shows histograms for the length distributions of interspersed repeats in human chromosomes 20, 21, and 22 as identified by REPEATMASKER.

A human-specific and very abundant family of SINEs are *Alu* repeats, which constitute about 10% of the human genome. *Alu*'s have been connected with sequence-specific integration [35,36], genome organization [37,38], regulation of enzyme activity [39], or human chromosome segregation [40].

An *Alu* is a dimer consisting of two approximately 130-bp-long monomeric units, with an insertion sequence (about 30 bp) in the second unit, and it is flanked by direct repeats [32,33]. Figure 4 sketches two *Alu* repeats consecutively. Most truncated repeats correspond to monomeric versions of *Alu* repeats. A sequence alignment of truncated *Alu*'s using the program CLUSTALX [41] reveals that many of them match the first or second units of complete *Alu* sequences. Correspondingly, the large peak in Fig. 3 at about $L=300$ bp is due to the high abundance of ‘‘complete’’ *Alu* repeats, while the broader peak at about $L=150$ bp is due to the high abundance of ‘‘partial’’ *Alu* repeats.

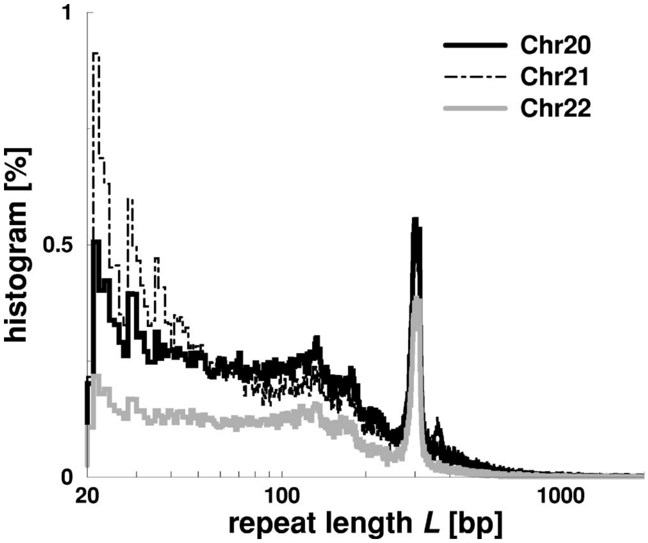

FIG. 3. Histograms of interspersed repeats of length L. We use equidistant bins of a single base pair for all annotated interspersed repeats in human chromosomes 20, 21, and 22. The histograms show a clear abundance of repeats at about $L=300$ bp corresponding to the presence of *Alu* repeats and other peaks at about $L=150$ corresponding to truncated *Alu*'s. In addition, all histograms exhibit long tails possibly due to the presence of LINEs as well as a number of repeats at smaller lengths.

V. EFFECTS OF REPEATS ON SHORT-RANGE CORRELATIONS

In this section, we study to which degree interspersed repeats contribute to the observed short-range correlations, by replacing repeats by simulated sequences while maintain-

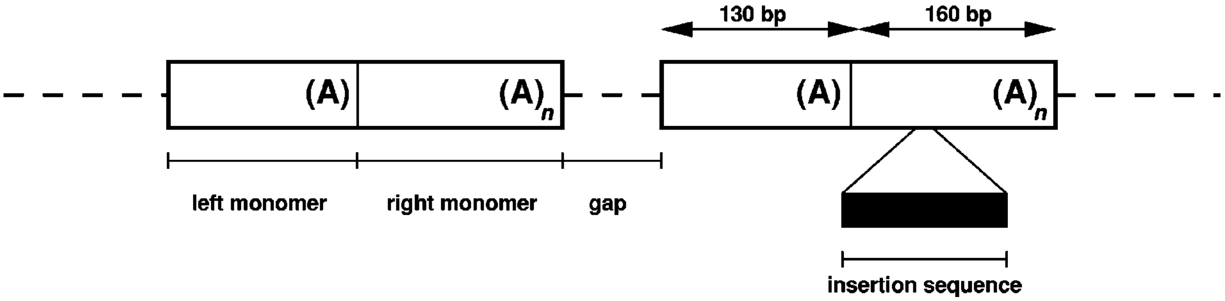

FIG. 4. Structure of two inserted, adjacent *Alu* repeats separated by a gap. Members of the *Alu* family are similar and exhibit about 87% homology to a consensus sequence, which is about 290 bp long and consists of a dimeric structure of two monomers, while one monomer maintains an insertion sequence. The monomers are linked by an adenine-rich tract (A) and contain a poly-(A) tract $(A)_n$ in the downstream region. In addition to internal correlations, the distribution of gaps between *Alu*'s can induce correlations in DNA sequences.

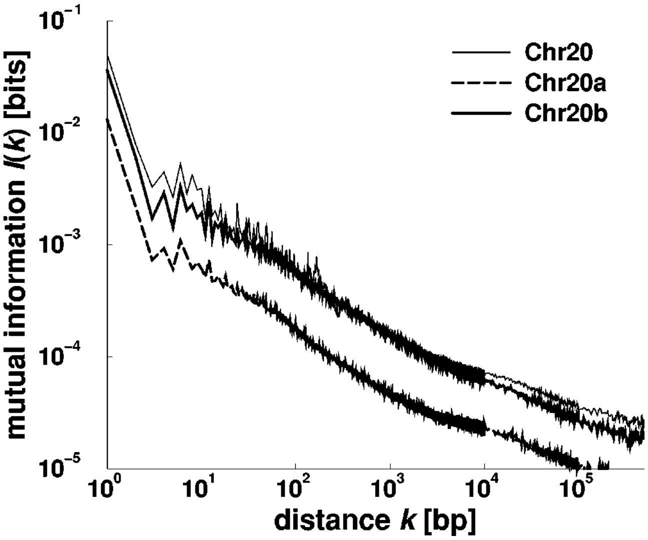

FIG. 5. Mutual information function $I(k)$ of the DNA sequences of human chromosomes 20, 20a (all interspersed repeats replaced), and 20b (*Alu* repeats replaced). $I(k)$ in chromosomes 20a and 20b is consistently smaller than $I(k)$ in chromosome 20 due to the substitution of 42% (20a), respectively, 13% (20b) sequence with random, uncorrelated nucleotides. The long-range correlations in the repeat-modified sequences persist in chromosomes 20a and 20b.

ing the actual chromosomal positions of repeats. In a first attempt, (i) we extract the relative frequencies $p_i^{(\mathrm{Chr20})}$, $p_i^{(\mathrm{Chr21})}$, and $p_i^{(\mathrm{Chr22})}$ from the nonrepetitive fractions of chromosomes 20, 21, and 22, (ii) generate random, uncorrelated sequences, and (iii) substitute repeats by simulated sequences in each chromosome. We distinguish two repeat-modified versions of human chromosomes: in chromosomes 20a, 21a, and 22a, we replace all interspersed repeats with random, uncorrelated nucleotides, and in chromosomes 20b, 21b, and 22b, we replace solely *Alu* repeats with random, uncorrelated nucleotides.

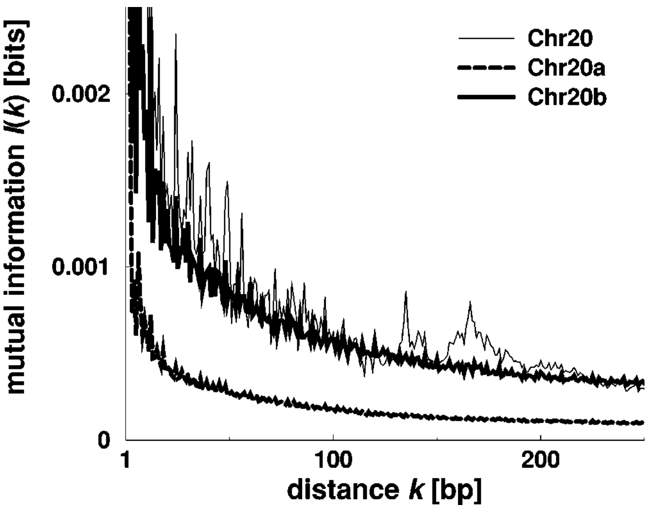

FIG. 6. Mutual information function $I(k)$ of the DNA sequences of human chromosome 20 and of the repeat-modified versions 20a and 20b. The suppression of $I(k)$ at about $k=135$ and at about $k=165$ bp in chromosomes 20a and 20b as compared to $I(k)$ in chromosome 20 relates these correlations to the presence of interspersed *Alu* repeats.

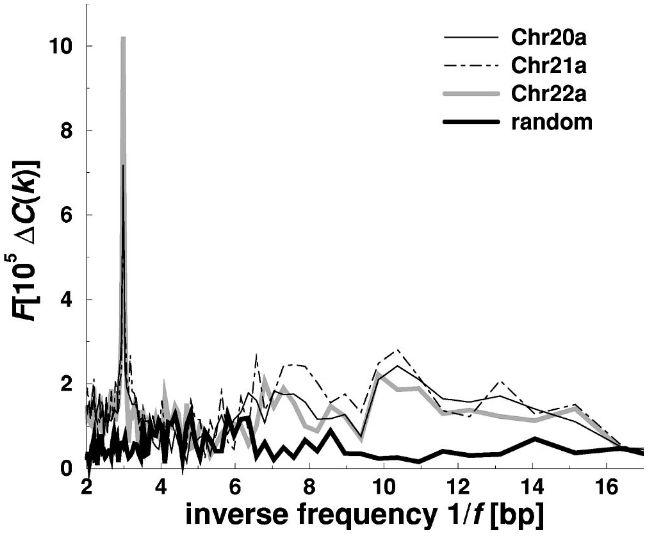

FIG. 7. Fourier transform F of the WW dinucleotide-dinucleotide correlation function $C(k)$ of the DNA sequence of human chromosomes 20a, 21a, and 22a [45]. We analyze first-order differences $\Delta C(k)$ of $C(k)$ in order to remove trends in the correlation function [46], and show the absolute values of $F[\Delta C(k)]$. The randomization of repeats suppresses signals in the range $k<200$ bp, and hence signals at 3 bp and at about 10–11 bp in nonrepetitive sequences become detectable. For the period $1/f=3$ bp, the height of the signal reflects the generally higher gene density in chromosomes 20 and 22 than in chromosome 21 (cf. Table I). For another sequence period at about $1/f=10$–11 bp, this signal is present in both coding and noncoding fractions of chromosomes 20a, 21a, and 22a, and may thus be attributed to DNA bendability.

Figure 5 shows a double-logarithmic plot of $I(k)$ in chromosomes 20, 20a, and 20b. We find that for $k=1,2,\ldots,10^5$ bp, $I(k)$ is larger in human chromosome 20 than in 20a and 20b, and that the slow (power-law) decay of $I(k)$ persists in both chromosomes 20a and 20b.

In Fig. 6, we show $I(k)$ for $k=1,2,\ldots,200$ bp, and contrast $I(k)$ in chromosome 20 with $I(k)$ in chromosomes 20a and 20b. We find that $I(k)$ in chromosomes 20a and 20b shows significantly less correlations (peaks) than in chromosome 20, and hence Fig. 6 demonstrates that short-range correlations are dominated by *Alu* repeats. It is also worth noting that repeat-induced peaks suppress other signals in this range, such as several pronounced signals at multiples of $k=6$ bp, and so we apply the Fourier transform to study periodic nucleotide variations in terms of frequencies.

Specifically, we test whether the Fourier transform F is able to reveal both the 3 bp and 10–11 bp sequence periodicities in repeat-modified sequences. Since it is known that DNA bendability is governed by dinucleotides [42,43] and that correlations of weakly binding nucleotides (W = A or T) give rise to pronounced signals [7], we calculate the dinucleotide-dinucleotide correlation function $C(k)$ between WW dinucleotide pairs [44]. Figure 7 shows $C(k)$ in chromosomes 20a, 21a, and 22a, and we find frequency components at $f^{-1}=3$ bp and at about $f^{-1}=10$–11 bp [45]. While a frequency of 3 bp is characteristic for protein coding regions, a frequency of about 10–11 bp is characteristic for DNA bendability.

On the length scale $100<k<200$, the mutual information function of original, un-modified DNA sequences exhibits two marked peaks at about $k=135$ bp and at about $k=165$ bp, which are common to all three chromosomes. Both peaks reflect the internal structure of *Alu* repeats (cf. Fig. 3). Even though *Alu* repeats constitute less than a half of the total fraction of interspersed repeats, they contribute strongly to the repeat-induced correlations. The first peak at about $k=135$ can be explained by the dimeric *Alu* repeat structure and correlated nucleotides in the duplicated monomers. The second peak at about $k=165$ can be explained by specific homopolymeric A-rich sequences within *Alu* repeats. Since each monomer has a poly-(A) downstream region, the excess of adenine nucleotides appears as a peak at about $k=165$ bp.

It is interesting that two other characteristic lengths in the human genome are on the same scale: (i) the average length of internal, fully protein coding regions (exons) is about 145 bp [18], and (ii) DNA stretches of about 150 bp are wrapped around nucleosomes [46]. Both features could potentially contribute to the observed short-range correlations. Earlier GC content oscillations with a period of 150–200 bp [47], as well as correlations up to 200 bp [11,12], have been attributed to DNA nucleosomal signals in human DNA sequences. Figure 6 shows that correlations at least up to $k\approx 200$ bp are strongly dominated by interspersed repeats. Hence, the interpretation of such correlations as characteristic for nucleosomes might have to be reconsidered.

VI. EFFECTS OF REPEATS ON LONG-RANGE CORRELATIONS

As shown in the preceding section, *Alu* repeats induce short-range correlations due to their internal dimeric structure, while Fig. 5 showed that the removal of any internal repeat structure has only minor effects on the power-law decay. However, a large fraction of repeats is dispersed throughout the human genome, and in this section we investigate to which degree their distribution may influence the slow decay of $I(k)$ in human chromosomes 20, 21, and 22 as demonstrated in Fig. 1.

In particular, the more than 10^6 *Alu* repeats within the human genome comprise a predominant category of repeats in human chromosomes and accumulate in the genome with a preference toward gene-rich chromosomal regions [18,33]. Histograms of distances between adjacent *Alu* repeats show significant deviations from an exponential decay expected for random chromosomal positions of repeats, and so we study the possible effects of the presence of clustering on long-range correlations.

In order to reveal effects of *Alu* repeat clustering on the decay properties of $I(k)$, we positionally randomize *Alu* repeats in chromosome 20 by reinserting *Alu*'s at randomly and uniformly chosen positions in chromosome 20r. Figure 8 shows a double-logarithmic plot of $I(k)$ in chromosome 20 and after the positionally randomization of about 28 000 *Alu* repeats in chromosome 20r. We find that short-range correlations up to about $k=300$ bp are hardly affected by the positional randomization *Alu* repeats as expected. Furthermore, we find that $I(k)$ shows only a weak, albeit significant, decrease of long-range correlations for increasing k beyond $k=300$ bp, and we obtain qualitatively similar results for chromosomes 21r and 22r.

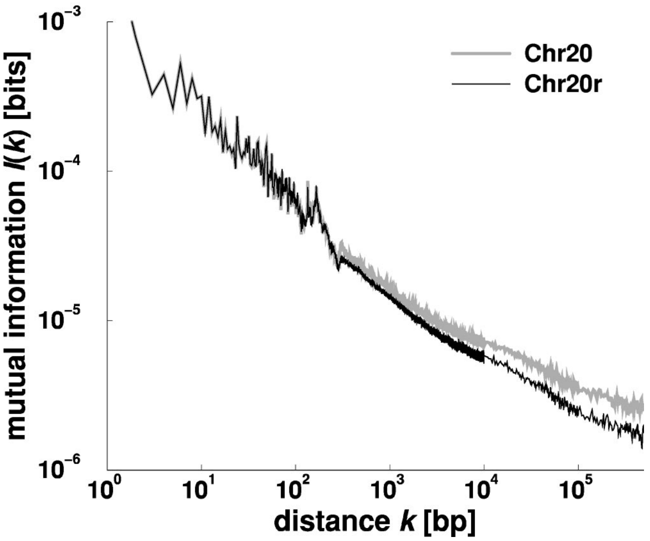

FIG. 8. Mutual information function $I(k)$ of the DNA sequence of human chromosomes 20 and of the positionally randomized version of chromosome 20r. We randomize the sequence by cutting *Alu* repeats from their original insert position and reinserting them at randomly and uniformly chosen positions without overlap. We find that for $k=1,2,\ldots,300$ bp $I(k)$ is virtually unaffected by positional randomization of interspersed *Alu* repeats, while $I(k)$ shows increasingly less pair correlations with increasing $k>300$ bp.

To clarify the effects of the presence of repeat clustering on the decay properties of $I(k)$ beyond $k>300$ bp, we examine $I(k)$ in simplified, binary translated DNA sequences $(\lambda=2)$. Given a DNA sequence of length N, we obtain a binary symbolic sequence by defining $A_i=1$ $(A_i=0)$ at position x_n, $n=1,2,\ldots,N$, if A_i is inside (outside) an *Alu* repeat. Figure 9 shows a double-logarithmic plot of $I(k)$ of the binary translated DNA sequence of chromosome 20 and of the binary translated DNA sequence of chromosome 20r. We find that $I(k)$ shows at about $k=300$ bp a pronounced drop in the binary translated sequence of chromosome 20r and is consistently smaller than in the binary translated sequence of chromosome 20.

Figure 9 shows for $k>300$ bp much more pronounced differences between $I(k)$ in the binary translated sequences of chromosomes 20 and 20r than between $I(k)$ in the DNA sequences of chromosomes 20 and 20r in Fig. 8. Note that the relationship between *Alu* repeat clustering and GC content is nontrivial [18,38,48], so the slow decay of $I(k)$ in Fig. 9 cannot be explained solely by considering long-range GC variations. Furthermore, *Alu* repeats are distributed over about a dozen subfamilies with varying degrees of similarity to their consensus, different ages, and different retrotransposition rates. One explanation for the less pronounced difference long-range correlations in Fig. 8 is that the amount of

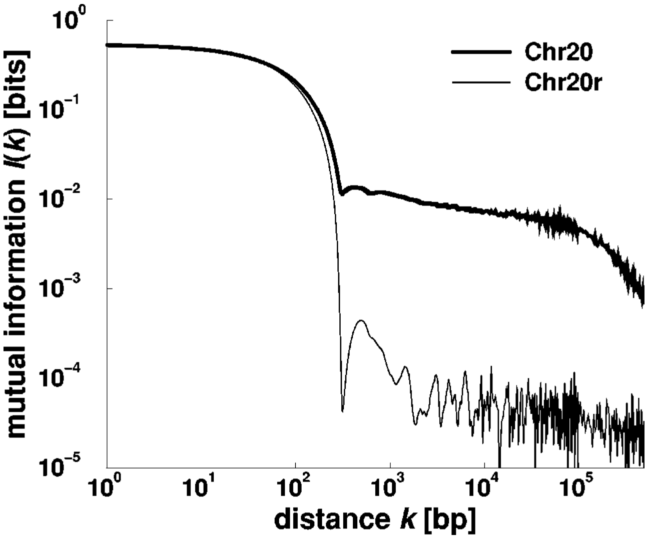

FIG. 9. Mutual information function $I(k)$ of the binary translated DNA sequences of human chromosomes 20 and 20r. We find that for $k=1,2,\ldots,300$ bp $I(k)$ shows similar decay properties in chromosomes 20 and 20r. It steeply drops at a distance of about $k=300$ bp, while $I(k)$ in chromosome 20 is consistently larger than $I(k)$ in chromosome 20r. Beyond $k>300$ bp (and up to about $k=10^5$ bp), the slow decay of $I(k)\sim k^{-2\gamma}$ in chromosome 20 can be approximated by a power law with exponent $2\gamma\approx 0.2$.

decrease of $I(k)$ depends on the overall GC content in human chromosomes. Most interspersed *Alu* repeats are GC rich [18,31,49], so the contrast between interspersed *Alu* repeats and nonrepetitive background is diminished in human chromosomes with a high GC content (such as chromosome 22), in which repeats and the nonrepetitive background have a similar DNA composition. On the other hand, a binary translation merges all *Alu* repeats families into one single type, such that the contrast between *Alu*'s and background sequence becomes enhanced.

VII. SUMMARY AND DISCUSSION

We study correlations in the DNA sequences of three completed human chromosomes, and we analyze the degree to which interspersed repeats influence short-range and long-range correlations by comparing $I(k)$ in the DNA sequences of chromosomes 20, 21, and 22 with both repeat-randomized and positionally randomized sequences.

We find that on the length scale of several base pairs, $I(k)$ shows no clear indications of sequence periodicities such as $k=3$ bp characteristic for protein coding sequences or $k=10-11$ bp characteristic for protein secondary structure and DNA bendability. These signals are hidden by internal correlations of interspersed repeats. After the randomization of interspersed repeats, signals at $k=3$ bp and at about $k=10-11$ bp become detectable in the Fourier spectrum of the dinucleotide correlation function $C(k)$. For the peak at about $k=10-11$ is present in the noncoding fraction of chromosomes 20, 21, and 22, this signal is indicative of DNA nucleosomal packaging. It is interesting that the $k=10-11$ bp sequence periodicity has previously been described in bacterial genomes, as well as in the yeast genome [7], and hence $C(k)$ gives first indications that such a signal might also be present in the human genome.

On the next length scale of several hundred base pairs, we observe that $I(k)$ is dominated by correlation within repetitive sequences. Specifically, we find that two statistically significant signals at about $k=135$ bp and at about $k=165$ bp are caused by internal correlations within the sequences of interspersed *Alu* repeats. While the peak at about $k=135$ bp can be explained by correlated repetitive nucleotides in the dimeric *Alu* structure, the second peak at about $k=165$ bp can be explained by an excess of homopolymeric poly-(A) sequences. It is known that the average length of internal, fully protein coding regions (exons) and DNA stretches wrapped around nucleosomes are in the same range of about 150 bp. Since the strength of repeat-induced correlations overrides sequence periodicities relevant to DNA bendability, the detection of such correlations as characteristic for nucleosomes might have to be reconsidered and alternatively be studied in repeat-randomized DNA sequences [11,12].

Human chromosomes 20, 21, and 22 exhibit long-range (power law) correlations exceeding several hundred base pairs and ranging over several orders of magnitude. These findings are in accord with previous studies showing GC content fluctuations over long distances [17,18,26,36]. On these length scales, we study repeat clustering [35,50] as one possible source of long-ranging instationarities and assess the impact of clustering by using the positional randomization of *Alu* repeats. We find that the clustering of *Alu*'s contributes only marginally to the observed long-range correlations in the DNA sequences of human chromosomes 20, 21, and 22, and so additional mechanisms need to be considered to explain the presence of long-range GC content variations of human chromosomes [51,52]. One possible mechanisms that could explain the origin of variations of GC nucleotides on a genomic scale is discussed elsewhere [53].

ACKNOWLEDGMENTS

We thank A. Smit and P. Green for providing REPEATMASKER and the National Science Foundation, National Institutes of Health, Cold Spring Harbor Association, and Deutsche Forschungsgemeinschaft for financial support.

[1] W.-H. Li, *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997).
[2] *Computational Models in Molecular Biology*, edited by S.L. Salzberg, D.B. Searls, and S. Kasif (Elsevier, Amsterdam, 1998).
[3] J.K. Percus, *Mathematics of Genome Analysis* (Columbia University Press, Cambridge, 2002).
[4] D.E. Reich *et al.*, Nat. Genet. **32**, 135 (2002).

[5] R. Staden and A.D. McLachlan, Nucleic Acids Res. **10**, 141 (1982); J.W. Fickett, *ibid.* **10**, 5303 (1982); J.W. Fickett and C.-S. Tung, *ibid.* **20**, 6441 (1992); M.S. Gelfand, J. Comput. Biol. **2**, 87 (1995); I. Grosse *et al.*, Phys. Rev. E **61**, 5624 (2000); D. Holste *et al.*, J. Theor. Biol. **206**, 525 (2000).

[6] E.N. Trifonov and J.L. Sussman, Proc. Natl. Acad. Sci. U.S.A. **77**, 3816 (1980).

[7] H. Herzel, O. Weiss, and E.N. Trifonov, Bioinformatics **15**, 187 (1999).

[8] V.B. Zhurkin, J. Biomol. Struct. Dyn. **4**, 785 (1981).

[9] O. Weiss and H. Herzel, J. Theor. Biol. **190**, 341 (1998).

[10] S.V. Buldyrev *et al.*, Biophys. J. **65**, 2673 (1993).

[11] B. Audit *et al.*, Phys. Rev. Lett. **86**, 2471 (2001); B. Audit *et al.*, J. Mol. Biol. **316**, 903 (2002).

[12] W. Li, Gene **300**, 129 (2002).

[13] J. Filipski, J.P. Thiery, and G. Bernardi, J. Mol. Biol. **80**, 177 (1973).

[14] G. Bernardi *et al.*, Science **288**, 953 (1985); G. Bernardi, Gene **241**, 3 (2000).

[15] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); C.-K. Peng *et al.*, Nature (London) **356**, 168 (1992); R.F. Voss, Phys. Rev. Lett. **68**, 3805 (1992); S.V. Buldyrev *et al.*, Phys. Rev. E **51**, 5084 (1995).

[16] W. Li, T.G. Marr, and K. Kaneko, Physica D **75**, 392 (1994); W. Li, Comput. Chem. (Oxford) **21**, 257 (1997).

[17] D. Holste, I. Grosse, and H. Herzel, Phys. Rev. E **64**, 041917 (2001).

[18] E. Lander *et al.*, Nature (London) **409**, 860 (2001).

[19] J.C. Venter *et al.*, Science **291**, 1904 (2001).

[20] G.K.-S. Wong *et al.*, Genome Res. **10**, 1672 (2000); M. Das *et al.*, Genomics **77**, 71 (2001); N. Katsanis, K.C. Worley, and J.R. Lupski, Nat. Genet. **29**, 88 (2001).

[21] We downloaded the DNA sequences of human chromosomes 20 21, and 22 from ftp://ftp.ensembl.org/pub/, which have been assembled in version v3.26.1. Human chromosomes 20 21, and 22 are almost complete and updated versions of these three chromosomes undergo only minor modifications. We excluded all presently unknown nucleotides due to gaps in the sequences.

[22] C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

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

[24] H. Herzel and I. Grosse, Phys. Rev. E **55**, 800 (1997).

[25] G.A. Miller, *Information Theory in Psychology*, edited by H. Quaster (Free Press, Glencoe, 1955); G.P. Basharin, Theor. Probab. Appl. **4**, 333 (1959); M.S. Roulston, Physica D **125**, 285 (1999).

[26] I. Dunham *et al.*, Nature (London) **402**, 489 (1999); M. Hattori *et al.*, *ibid.* **405**, 311 (2000); P. Deloukas *et al.*, *ibid.* **414**, 865 (2001); J.E. Collins *et al.*, Genome Res. **13**, 27 (2003).

[27] A. Nekrutenko and W.-H. Li, Genome Res. **10**, 1986 (2000).

[28] D. Häring and J. Kypr, Biochem. Biophys. Res. Commun. **280**, 567 (2001).

[29] O. Clay *et al.*, Gene **276**, 15 (2001).

[30] The power spectrum $S(f)$ can be calculated from the average of squared values of the Fourier transform. Define $f=n/N$ $(n=12,\dots,N/2)$ and enumerate nucleotide $x^{(l)}$ at sequence position l by $a_i^{(l)}=1$ if $x^{(l)}=A_i$, or 0 else, then it is defined as $S(f)\equiv\Sigma_{i=1}^{\lambda}N^{-1}|\Sigma_{l=1}^{N}a_i^{(l)}\mathrm{e}^{-i2\pi lf}|^2$.

[31] A.F.A. Smit, Curr. Opin. Genet. Dev. **6**, 743 (1996); A.F.A. Smit, *ibid.* **9**, 657 (1999).

[32] P. Sudbery, *Human Molecular Genetics* (Addison Wesley Longman, Singapore, 1998).

[33] M.A. Batzer and P.L. Deininger, Nat. Rev. Genet. **3**, 370 (2002).

[34] We use REPEATMASKER version v07072001. A.F.A. Smit and P. Green. Accessible at World-wide web http://repeatmasker.genome.washington.edu/

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

[36] A. Nekrutenko and W.-H. Li, Trends Genet. **17**, 619 (2001); R. Sorek *et al.*, Genome Res. **12**, 1060 (2002).

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

[38] P. Medstrand *et al.*, Genome Res. **12**, 1483 (2002).

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

[40] M.-A. Hakimi *et al.*, Nature (London) **418**, 994 (2002).

[41] F. Jeanmougin *et al.*, Trends Biochem. Sci. **23**, 403 (1998). Accessible at ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/

[42] C.R. Calladine and H.R. Drew, *Understanding DNA* (Academic Press, London, 1992).

[43] A. Bolshoy *et al.*, Proc. Natl. Acad. Sci. U.S.A. **88**, 2312 (1991).

[44] The dinucleotide-dinucleotide autocorrelation function is defined as $C(k)=\langle a_{ij}a_{nm}\rangle-\langle a_{ij}\rangle\langle a_{nm}\rangle$, with the average $\langle\cdot\rangle$ over all $\lambda^2\times\lambda^2$ pair probabilities $p_{ij,nm}(k)$ $(i,j=12,\dots,\lambda)$ at distance k, and $a_{ij}\in\{01\}$ enumerates all λ^2 dinucleotides. The combination of all possible weakly-binding (W = A or T) pairs ($a_{ij}=1$ if $A_i\cap A_j\in$ W, 0 else) defines the WW-WW autocorrelation function.

[45] Sequence periodicities at $k=3$ bp and at about $k=10$–11 bp become detectable in repeat-modified DNA sequences of chromosomes 20 21, and 22. We find evidence for sequence periodicities by (i) measuring the WW-WW autocorrelation function $C(k)$ for $k=2,3,\dots,100$ bp. Note that $C(k=1)$ is at least one order of magnitude larger than $C(k\geqslant2)$ and would hence obscure the analysis. After (ii) low-pass filtering $\Delta C(k)=C(k+1)-C(k)$, which suppresses frequency components proportional to $\sim f^{-2}$, and (iii) computing the discrete Fourier transform $F[\Delta C(k)]=(N-2)^{-1}\Sigma_{k=2}^{N}\Delta C(k)\mathrm{e}^{-i2\pi kf}$, the inspection of $F[\Delta C(k)]$ versus the frequency f reveals peaks at $f^{-1}=3$ bp and at about $f^{-1}=10$–11 bp.

[46] K.E. van Holde, *Chromatin* (Springer-Verlag, New York, 1988).

[47] R.L.P. Adams *et al.*, Eur. J. Biochem. **165**, 107 (1987).

[48] D. Takai and P.A. Jones, Proc. Natl. Acad. Sci. U.S.A. **99**, 3740 (2002).

[49] K. Jabbari and G. Bernardi, Gene **224**, 123 (1998).

[50] P. Green, Genome Res. **7**, 410 (1997).

[51] A. Eyre-Walker and L.D. Hurst, Nat. Rev. Genet. **2**, 549 (2001).

[52] K.J. Fryxell and E. Zuckerkandl, Mol. Biol. Evol. **17**, 1371 (2000).

[53] I. Grosse, M. Zhang, and H. Herzel (unpublished).