Large-scale oscillation of structure-related DNA sequence features in human chromosome 21

Wentian Li*
The Robert S. Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, New York 11030, USA

Pedro Miramontes†
Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior, Ciudad Universitaria, D.F. 04510, Mexico
and Departamento de Matemáticas, Universidad de Sonora, Encinas y Rosales, Hermosillo 83000 Sonora, Mexico
(Received 24 March 2006; published 9 August 2006)

Human chromosome 21 is the only chromosome in the human genome that exhibits oscillation of the (G+C) content of a cycle length of hundreds kilobases (kb) (500 kb near the right telomere). We aim at establishing the existence of a similar periodicity in structure-related sequence features in order to relate this (G+C)% oscillation to other biological phenomena. The following quantities are shown to oscillate with the same 500 kb periodicity in human chromosome 21: binding energy calculated by two sets of dinucleotide-based thermodynamic parameters, AA/TT and AAA/TTT bi- and tri-nucleotide density, 5′-TA-3′ dinucleotide density, and signal for 10- or 11-base periodicity of AA/TT or AAA/TTT. These intrinsic quantities are related to structural features of the double helix of DNA molecules, such as base-pair binding, untwisting or unwinding, stiffness, and a putative tendency for nucleosome formation.

DOI: 10.1103/PhysRevE.74.021912 PACS number(s): 87.14.Gg, 87.15.−v, 87.16.Sr, 02.50.−r

I. INTRODUCTION

DNA sequences are full of features at small, intermediate, and large scales [1]. At short distances, there is strong periodicity-of-three-nucleotide signal in protein-coding regions (but absent in noncoding regions)[2] and a weaker but ubiquitous 10-11 bases signal in many genomes [3]. At intermediate length scales, there are Alu sequences of about 300 bases long [4] and nucleosome-forming sequences of around 120–200 bases [5]. At large length scales, the most well-known features are the existence of alternating (G+C)%-high and (G+C)%-low "isochores"[6] and the distribution of sine waves that prefers long-wavelength signals (the so-called "1/*f*" spectra when viewed in the spectral space [7]).

A recent survey of (G+C)% fluctuation in all human (*Homo sapiens*) chromosomes revealed that chromosome 21 exhibits a unique 500 kilobases (kb) oscillation in (G+C)% [8]. This oscillation starts around the position of 43.5×10^6 bases (Mb) and lasts five cycles [with five (G+C)%-low six (G+C)%-high peaks]. No other human chromosomes exhibit similar periodicities with such a long cycle length.

Human chromosome 21 has other special properties as compared to the rest of the human chromosomes. First, it is the shortest human chromosome. Second, its (G+C)% increases stepwise from left (centromeric) to right (telomeric—i.e., close to the end of the chromosome), with three distinct "super" isochore regions [see, e.g., Fig. 3 of Ref. [6(b)]]. The 500 kb oscillation of (G+C)% described above appears in the third region with the highest (G+C)% and the highest gene content. Third, the failure rate in segregating homologous chromosomes during meiosis is the highest among surviving infants in human chromosome 21 than any other human chromosomes. When this happens, the surviving infants typically carry three copies of chromosome 21 ("trisomy 21") instead of one copy [9]. The resulting Down syndrome is the leading case of birth defects [10].

The uniqueness of the 500 kb oscillation in (G+C)% in human chromosome 21 and highest trisomy rate in chromosome 21 among surviving infants motivated us to speculate the possibility that this 500 kb oscillation might be somewhat related to the trisomy risk. An argument is that the periodicity in (G+C)% is a basis for certain structural periodicity, which in turn might interfere with the proper segregation of chromatids during meiosis. One intriguing observation is that for younger mothers with trisomy 21, the placement of meiosis exchange tends to be telomeric [11].

In this paper, we examine whether sequence-based structure features oscillate with the 500 kb cycle length in the telomeric region of human chromosome 21. The structural features we focus on include the helix binding energy, flexibility or stiffness in secondary structure of DNA helix, tendency for nucleosome formation based on a periodicity of 10-11 bases, and a tendency for anchoring DNA loops.

Note that only the intrinsic quantities are calculable here: chromatin structures that depend on extrinsic protein factors require experimental data, and these data are not yet conclusive. Also note that the sequence-to-structure connections in some models are based on simplified assumptions and our calculation may only give a partial picture of DNA helix structure properties. Our hope is for this work to contribute to the eventual establishment of a sequence-function connection.

II. DNA BINDING ENERGY AND STABILITY

It has been well known that base pairs with strong bases (G-C) are more stable than base pairs with weak bases (A-T),

*Electronic address: wli@nslij-genetics.org
†Electronic address: pmv@fciencias.unam.mx

TABLE I. Free energy (ΔG) of helix binding in nearest-neighbor models at 37 °C with Breslauer-SantaLucia parameters (kcal/mol).

5′/3′	G	A	T	C
G	2.75/1.84	1.41/1.30	1.13/1.44	2.82/2.24
A	(see CT)	1.66/1.00	1.19/0.88	(see GT)
T	(see CA)	0.76/0.58	(see AA)	(see GA)
C	3.28/2.17	1.80/1.45	1.35/1.28	(see GG)

due to the presence of three versus two hydrogen bonds. This single-base model of binding energy has been extended to dinucleotide models where a dinucleotide step (two neighboring basepairs) contributes an amount to the total binding energy [12]. There are two commonly used parameter value sets in the dinucleotide model: one by Breslauer and his colleagues [13] and another summarized by SantaLucia, also known as the unified parameters [14]. The nearest-neighbor free energy ΔG parameter values at 37 °C are listed in Table I for all 16 dinucleotide steps.

A 3.9 Mb sequence from the NCBI Build 35 (May 2004, hg17) of human chromosome 21 is downloaded from the UCSC genome browser [15], starting from the position 43 Mb and ending at the right telomere, of position 46.944 323 Mb.

Figure 1 shows the (G+C)% and averaged binding free energy ΔG calculated by the dinucleotide model with the parameters of Breslauer *et al.* and SantaLucia, using nonoverlapping windows of 2 kb. It is clear that the binding energy is higher in (G+C)%-high peak regions and thus also oscillates with the 500 kb periodicity. However, the magnitude of oscillation is larger in the free energy based on the parameters of Breslauer *et al.* than that using SantaLucia's parameters [range of (1.51–2.23) versus (1.18–1.69)].

Among the values of ΔG in Table I, the highest helix binding energies are usually associated with two strong bases (G or C), with the exception of 1.84 kcal/mol for GG/CC dinucleotide in SantaLucia's parameters. The lowest binding energies tend to be associated with two weak bases (A or T), but with the exceptions of AA/TT (1.66 kcal/mol) and AT (1.19 kcal/mol) dinucleotides in the parameters of Breslauer *et al.* The difference between the two sets of parameters is the largest for CG (1.11 kcal/mol, 40.7% of the average between the two parameters), GG/CC (0.91 kcal/mol, 39.7%), and AA/TT (0.66 kcal/mol, 49.6%) dinucleotides. With these exceptions, one may not automatically assume the binding energy to fluctuate the same way as (G+C)%. What Fig. 1 shows is that the difference between the single-base model (counting the number of weak and strong bases) and the dinucleotide models is not large enough to destroy the 500 kb oscillation in the binding energy.

The correlation coefficient between the windowed energy values and the (G+C)% values was calculated (the first two lines in Table II). These correlation values show that SantaLucia parameters are more correlated with the GC% than the parameters of Breslauer *et al.* (correlation coefficient of 0.998 versus 0.981 using the 2 kb window). By examining the two sets of free energy parameters in Table I closely, it is clear that the difference can be traced to the fact that the parameters of Breslauer *et al.* assign a higher energy value

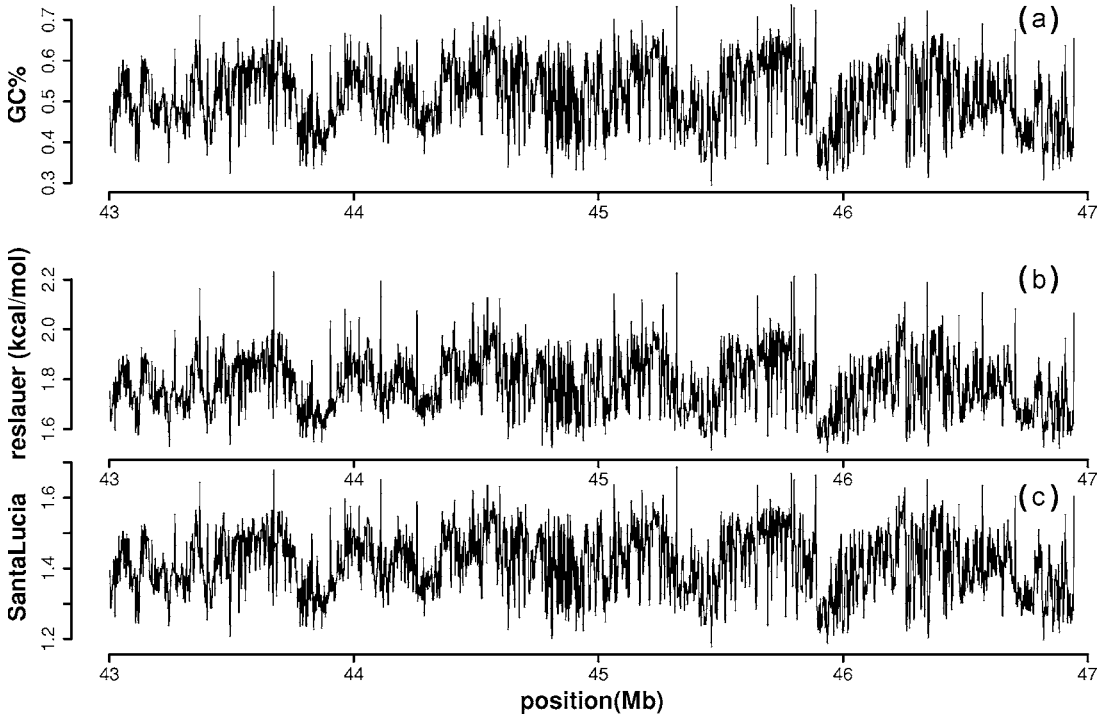

FIG. 1. (a) (G+C)% calculated in nonoverlapping windows of size 2 kb, (b) free energy ΔG in the nearest-neighbor model with Breslauer's parameter values, and (c) free energy ΔG in the nearest-neighbor model with SantaLucia's parameter values. The x axis is the chromosome position, in Mb.

TABLE II. Correlation coefficients of 11 quantities obtained from nonoverlapping 2 kb windows: GC%, bindinger energy by the models of Breslauer *et al.* and SantaLucia, densities of 5′-YR-3′, AA/TT, AAA/TTT, 5′-TA-3′, AA-10b-AA/TT-10b-TT, AAA-10b-AAA/TTT-10b-TTT, and VWG-10b-VWG, and density of top S/MAR hexamers. Testing of correlation coefficient equal to zero is significant at p value=0.01 level for all pairs except those marked by asterisks (YR-AA p=0.064, YR-AAA p=0.049, YR-AA10AA p=0.056, and YR-SMAR p=0.93).

	GC	Breslauer *et al.*	SantaLucia	5′YR3′	AA	AAA	5′TA3′	AA10AA	AAA10AAA	VWG10VWG
Breslauer *et al.*	0.981									
SantaLucia	0.998	0.985								
5′YR3′	−0.133	−0.195	−0.103							
AA	−0.960	−0.896	−0.950	−0.042*						
AAA	−0.917	−0.844	−0.903	−0.044*	0.974					
5′TA3′	−0.946	−0.915	−0.947	0.183	0.912	0.858				
AA10AA	−0.864	−0.791	−0.851	−0.043*	0.922	0.956	0.810			
AAA10AAA	−0.610	−0.545	−0.595	−0.064	0.683	0.789	0.557	0.866		
VWG10VWG	0.526	0.398	0.514	0.279	−0.657	−0.637	−0.574	−0.601	−0.458	
S/MAR	−0.881	−0.807	−0.868	−0.002*	0.929	0.967	0.854	0.947	0.810	−0.617

for two AT-rich dinucleotides than SantaLucia's parameters: 5′-AA-3′ and 5′-AT-3′. It is still debatable whether the parameters of Breslauer*et al.* or SantaLucia reflect the *in vivo* situation of helix local thermodynamics [16], and the issue may not be settled soon [17].

III. DNA FLEXIBILITY, STIFFNESS, AND UNTWISTING

Without an actual measurement of the DNA polymer mechanic properties, we rely on dinucleotides and trinucleotides that are known to be related to the DNA flexibility, stiffness, and untwisting to study the variation of these properties along the chromosome. For example, the A…A/T…T tract is known to have a stiff configuration because of an additional hydrogen bond between adjacent pairs along two diagonally located bases [18]. This hypothesis had been confirmed for AA/TT dinucleotide by their limited range of roll and slide values [19]. We use the AA/TT dinucleotide and AAA/TTT trinucleotide densities in a moving window as an indicator of the intrinsic stiffness of the double helix.

Unlike A/T-tracts, 5′-pyrimidine-purine-3′ (5′-YR-3′) steps can adopt two possible configurations and thus they are flexible [20]. In a simplified approach, we use the 5′-YR-3′ density as an indicator of the flexibility of the DNA double helix.

Among the four 5′-YR-3′ steps (CA, CG, TA, TG), 5′-TA-3′ has the weakest base pair binding. The biconfiguration nature and weak binding make 5′-TA-3′ one of the best candidates for untwisting initiation sites of the double helix [20]. We use the 5′-YR-3′ and 5′-TA-3′ densities in moving windows as an indicator of an untwisting potential.

Figure 2 shows the densities of the above-mentioned di- and tri-nucleotides: AA/TT%, AAA/TTT%, 5′-YR-3′%, and 5′-TA-3′%. The 500 kb oscillation in the first two densities is clearly seen. The 5′-YR-3′ density does not exhibit any regular oscillation of 500 kb, whereas the 5′-TA-3′ density does oscillate with the 500 kb wavelength.

Note that the signal we are measuring by the di- and tri-nucleotide densities is different from that of CpG islands [21]. In detecting CpG islands, the density of 5′-CG-3′ dinucleotide is normalized by the square of GC% (the observed over expected, or O/E), and the presence of a signal requires the 5′-CG-3′ density to be at least a quadratic function of GC%. In fact, it was known that the O/E signal increases with the GC%, indicating a cubic relationship between the 5′-CG-3′ density and GC% in CpG islands [22]. Here only the "linear" signal was measured.

IV. PERIODICITY-10-BASE SIGNAL AND NUCLEOSOME-FORMING POTENTIAL

It has been known that almost all genomes contain a AA-10b-AA/TT-10b-TT signal [3], where the "10b" can be 10 or 11 bases for individual cases, but after averaging becomes a real number between 10 and 11. This periodic signal is also present in the aligned nucleosome-forming sequences [23]. We count the number of occurrences of AA-10-AA, TT-10-TT, AA-11-AA, and TT-11-TT in a moving window, then convert to the density (a similar calculation for the AAA-10b-AAA/TTT-10b-TTT density is also carried out). As a crude approximation, this density is used to indicate the region's tendency for nucleosome formation.

Figure 3(a) and 3(b) show the AA-10b-AA/TT-10b-TT and AAA-10b-AAA/TTT-10b-TTT densities in a 2 kb nonoverlapping moving window. The 500 kb oscillation is clearly seen and may support the idea that the nucleosome-forming strength also oscillates with that wavelength in this region.

However, it was suggested that the regular spacing of 10 bases of another triplet motif, [not-T][A/T][G], can be considered as a nucleosome formation signal (called "VWG" signal) [24]. We count the occurrences of [not-T][A/T][G]-10/11-[not-T][A/T][G] and [C][A/T][not-A]-10/11-[C][A/T][not-A] in a moving window, whose density is plotted in Fig. 3(c). This VWG signal does not exhibit a 500 kb oscillation in this region.

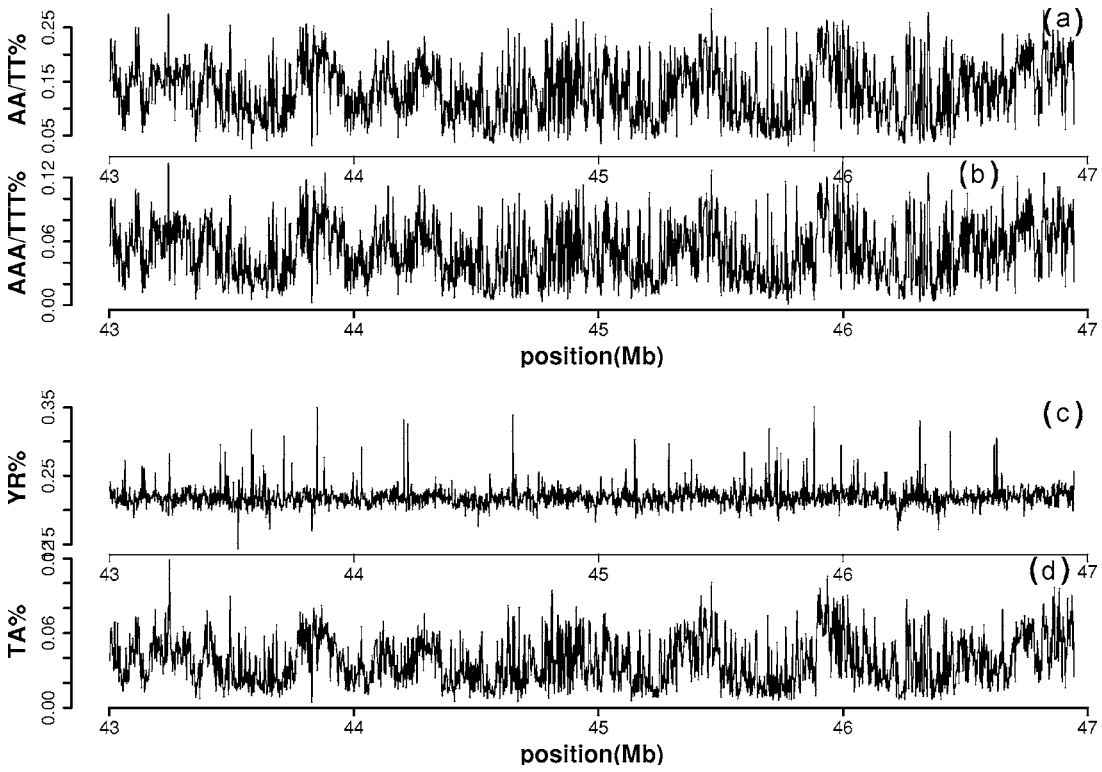

FIG. 2. (a) Density of AA/TT in nonoverlapping windows of size 2 kb, (b) AAA/TTT density, (d) 5′-YR-3′ density, and (d) 5′-TA-3′ density.

In a more sophisticated study based on discriminant analysis, a composite measure called "nucleosome formation potential" (NFP) was proposed [25]. As shown in Fig. 1 of Ref. [26], this NFP value decreases with GC%. Since the AA-10b-AA/TT-10b-TT and AAA-10b-AAA/TTT-10b-TTT densities also decrease with GC%, the two measures are con-

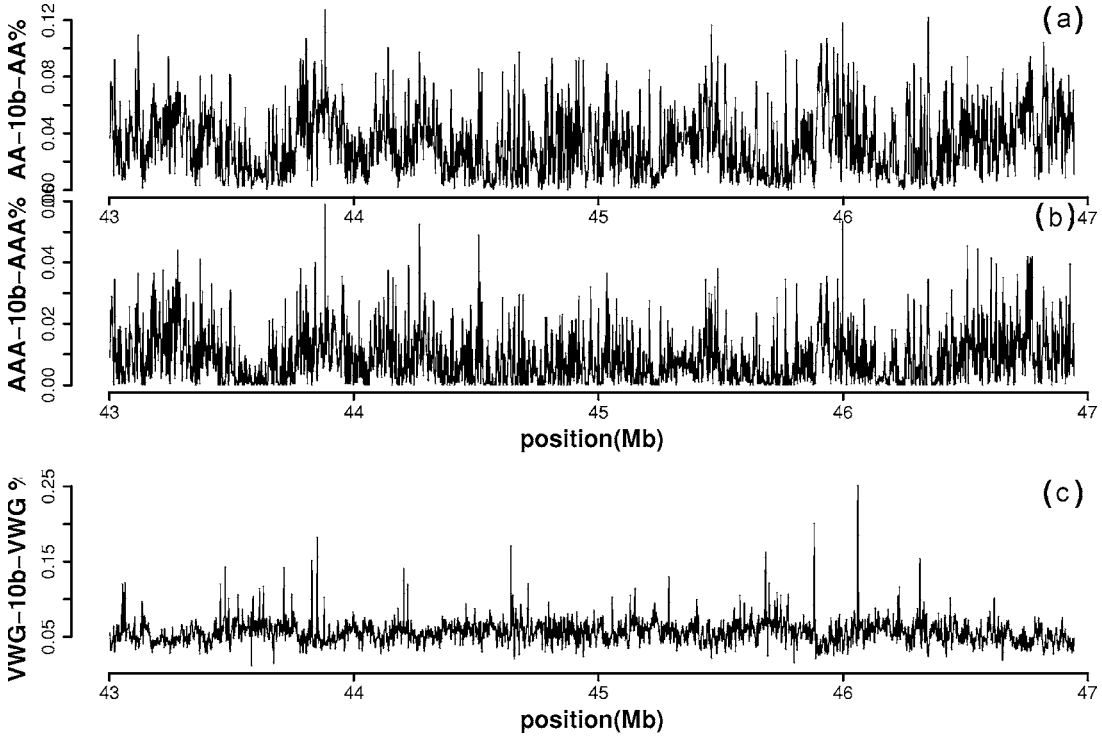

FIG. 3. (a) Density of AA-10b-AA/TT-10b-TT in nonoverlapping windows of size 2 kb, (b) AAA-10b-AAA/TTT-10b-TTT density, and (c) YWG-10b-VWG density, where VWG indicates [not-T][A/T][G] or its reverse complement triplet [C][A/T][not-A].

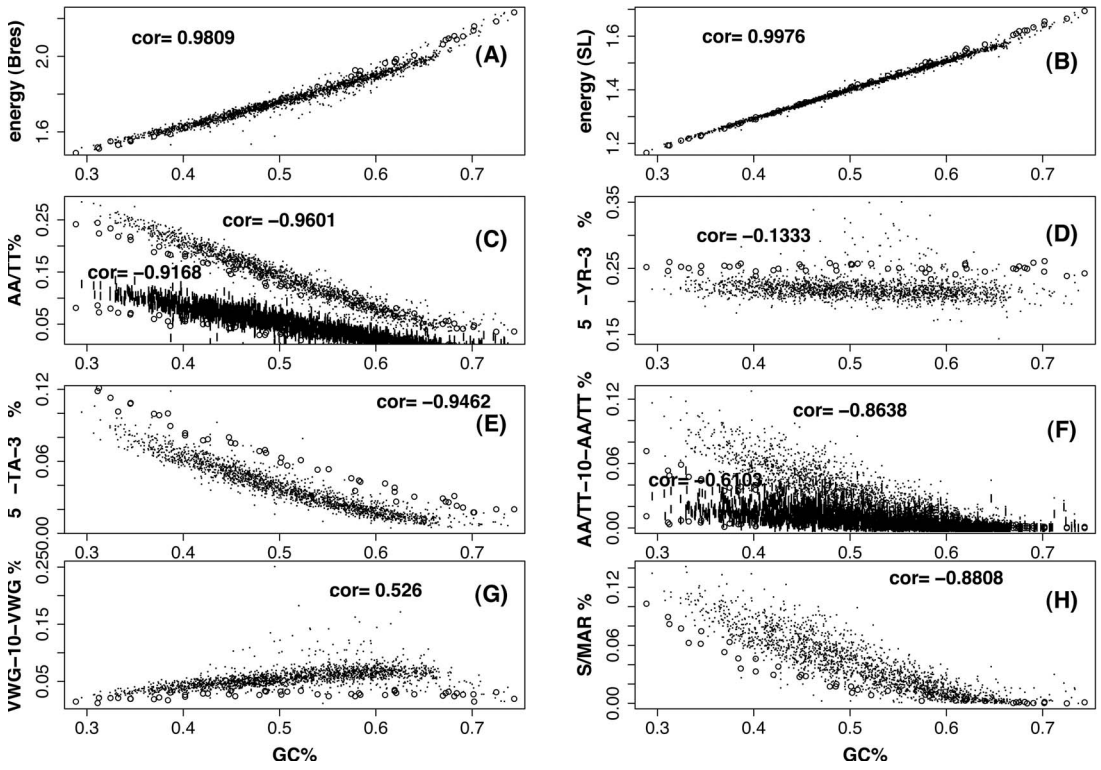

FIG. 4. Scatter plots of ten quantities versus GC%: (a) helix binding energy by the model of Breslauer *et al.*, (b) binding energy by SantaLucia's model, (c) AA/TT (upper) and AAA/TTT (lower, using the symbol |) densities, (d) 5′-YR-3′ density, (e) 5′-TA-3′ densities, (f) AA-10b-AA/TT-10b-TT (upper) and AAA-10b-AAA/TTT-10b-TTT (lower, using the symbol |) densities, (g) VWG-10-VWG densities, and (h) density of the top 34 hexamers in known S/MAR sequences and their reverse complements. The corresponding values for randomized sequences are also shown (grey circles). The correlation coefficient between these quantities and GC% is indicated on the plot.

sistent. The VWG signal, however, does not have a simple relationship with GC%, though mostly it increases with GC%. Whether one can predict the nucleosome-forming potential of a DNA sequence accurately and whether such an intrinsic potential really exists seem still to be open questions, and it is possible that either the AA/TT-10b-AA/TT or VWG-10b-VWG signal does not present the whole picture of nucleosome formation.

V. DISCUSSION AND CONCLUSION

Besides the helix-structure-related intrinsic features, the scaffold/matrix-attached regions (S/MAR's) are another pattern that can be determined from the DNA sequence. S/MAR's are the base or foundation of DNA loops [27], and S/MAR sequences can be obtained from S/MAR databases such as the one developed at the University of Göttingen [28].

By examining the top 34 most frequent hexamers in S/MAR sequences [Table 2 of [28(b)]], it is clear that S/MAR's are AT rich [29]. In fact, only 11 hexamers contain one G or C, ranked 10, 16–18, 21, 22, 25–27, 29, 30 in the top 34, and the rest consist exclusively of A and T [28]. It is not surprising that the S/MAR hexamer density (percentage of hexamers that match the top 34 most frequent S/MAR hexamer motifs and their reverse complement) also oscillates with a 500 kb wavelenegth in this region [30].

The existence of 500 kb oscillations in most of the quantities we have examined indicates that these structure-related sequence features are correlated with GC%. To assess this correlation directly, Fig. 4 shows the scatter plot of ten quantities used in Figs. 1–3 as versus GC%, and Table II lists the correlation coefficients of all pairs among these 11 quantities. Figure 4 and Table II have confirmed that these structure-based sequence features are highly correlated (the test results of these correlation coefficients are all significant with the exception of a few pairs involving 5′-YR-3′) and GC% can be used as a good surrogate for these features (with the exception of 5′YR-3′).

The density of 5′-YR-3′ is not correlated with other quantities studied (four correlation coefficients are not significant at the p value=0.01 level, and five other correlation coefficients, though significant, are rather weak). The next group of quantities that have weak correlation with others are the AAA-10b-AAA/TTT-10b-TTT and VWG-10b-VWG densities, with several correlation coefficients in the 0.4–0.5 range.

One may ask the question whether the correlation between these quantities and GC% is "trivial," because these patterns are either dominated by GC-rich or AT-rich di- and tri-nucleotides. This question can be addressed by examining the GC%-preserving random sequences. In Fig. 4 the ten structure-related quantities for the random sequences are shown as a function of GC% (circles). Several interesting observations can be made.

(i) The binding energies calculated on real DNA sequences are very close to those calculated on randomized sequences. However, the binding energy of real DNA sequences is slightly lower than that of random sequences at high GC% values. A similar observation was made in [31] [Fig. 1(c) of Ref. [31]] on the "relative" thermostability.

(ii) The A/T-tract density is higher in real DNA sequences than randomized sequences, mainly in the AT-rich ranges. It indicates that DNA sequences are more rigid than randomized sequences in general.

(iii) The biconfigurational 5′-YR-3′ dinucleotide density is lower in real DNA sequences than randomized sequences (with some exceptions for DNA segments with GC% around 50%–60%). It indicates that DNA sequences are less flexible than randomized sequences.

(iv) The 5′-TA-3′ density is lower in DNA sequences than random sequences, making them less susceptible to helix untwistings.

(v) The periodicity of 10/11 bp signal for both AA/TT, AAA/TTT, and VWG triplet has a stronger presence in real DNA sequences than random sequences, probably making them more likely to form nucleosomes.

(vi) The S/MAR potential is higher in DNA sequences than randomized sequences.

From these observations, one may expect that the binding energy faithfully follows the same variation and oscillation as GC%; A/T tract density, TA density, AAA-10b-AAA signal, and S/MAR signal more or less follow the same oscillation as GC%; YR density, AAA-10b-AAA signal, and YWG-10b-YWG signal may not follow the same oscillation as GC%.

It has been known that GC% conveys biological information [6(c)]. For example, the Giemsa-dark chromosome staining band, or G band, is AT rich, whereas the Giemsa-light band or R band is GC rich [32], or by a new hypothesis, AT rich and GC rich relative to its neighboring bands [33]. The gene density is another example, with GC-rich regions being relatively gene rich [34]. Fluorescence microscopy images show that chromosomes inside the nucleus are organized in a radial order, called "chromosome territories" [35]. The GC-rich, gene-rich regions tend to be located towards the center of the nucleus [36], and the corresponding chromatin compartments are more "open" [35].

Without experimental evidence, it is difficult to speculate what type of high-order chromatin structure this 500 kb oscillation might cause. According to the chromatin structure model summarized in [37], there could be multiple levels of foldings in the hierarchical structure of a chromatid: Watson and Crick's double helix (10 bp for one helix turn), nucleosomes (~200 bp per unit), solenoids (6 nucleosome units per helix turn or 1.2 kb) that twist to form a loop of ~50 kb, rosettes that consist of 6 loops (~300 kb), coils that consist of 30 rosettes (~9 Mb), and finally the chromatids consist of, for a medium-sized human chromosome, ~10 coils. Within the framework of this model, our 500 kb oscillation matches roughly the size of a rosette. However, we should caution that the exact figure for the size of these hierarchical units is illustrative and the model itself may be too much based on *in vitro* experiments and on inactive cells [38].

The unique large-scale oscillation of GC% in human chromosome 21 studied in this paper and in [8] can be further analyzed from several perspectives. One is about its evolutionary presevation in other species. Due to the high degree of similarity between humans and chimpanzees, it is natural to assume that the same 500 kb oscillation would also be present in the chimpanzee genome. Indeed, it was shown that a 500 kb oscillation exists in chimpanzee chromosome 22 [30]. On the other hand, no such 500 kb oscillation was observed in the mouse genome. It would be interesting to check its existence in species in between mouse and human.

It was suggested for the yeast genome [39] that the transcription direction of open reading frame (ORF) points from GC-rich to GC-poor regions. Combined with the general picture that the DNA loop anchored in AT-rich regions whereas the GC-rich part of the loop is exposed to the outside, transcription likely starts from the top of the DNA loop to the loop base. Although the length scale between two GC-rich regions analyzed in the yeast genome (~10 kb) is much shorter than the GC% oscillation length studied here, there is some evidence of gene density on two opposite strands alternating in this region [Fig. 5(c) of Ref. [8]]. A more careful analysis is needed to confirm the similarity between the human and yeast genomes, and the regular oscillation of GC% discussed here provides an ideal test ground.

In conclusion, the 500 kb oscillation in GC% as reported in [8] was shown to lead to a similar oscillation of some intrinsic structure-related patterns. And we hypothesize that a regular oscillation in chromatin structure with the same wavelength is also present in this region.

ACKNOWLEDGMENTS

W.L. acknowledges financial support from the The Robert S Boas Center for Genomics and Human Genetics. P.M. thanks the support of DGAPA Project No. IN111003.

[1] W. Li, Comput. Chem. **21**, 257 (1997).

[2] J. W. Fickett, Nucleic Acids Res. **10**, 5303 (1982); V. R. Chechetkin, L. A. Knizhnikova, and A. Y. Turygin, J. Biomol. Struct. Dyn. **12**, 271 (1994); G. Gutierrez and A. Marin, J. Theor. Biol. **167**, 413 (1994); S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, CABIOS, Comput. Appl. Biosci. **13**, 263 (1997); W. Lee and L. Luo, Phys. Rev. E **56**, 848 (1997).

[3] E. N. Trifonov and J. L. Sussman, Proc. Natl. Acad. Sci. U.S.A. **77**, 3816 (1980); J. Widom, J. Mol. Biol. **259**, 579 (1996); V. R. Chechetkin and V. V. Lobzin, J. Biomol. Struct. Dyn. **15**, 937 (1998); H. Herzel, O. Weiss, and E. N. Trifonov, Bioinformatics **15**, 187 (1999); E. Larsabal and A. Danchin, BMC Bioinf. **6**, 206 (2005).

[4] C. W. Schmid and W. R. Jelenik, Science **216**, 1065 (1982); C. Willard, H. T. Nguyen, and C. W. Schmid, J. Mol. Evol. **26**,

180 (1987); M. A. Batzer and P. L. Deininger, Nat. Rev. Genet. **3**, 370 (2002).
[5] D. Hewish and L. Burgoyne, Biochem. Biophys. Res. Commun. **52**, 504 (1973); H. R. Widlund *et al.*, J. Mol. Biol. **267**, 807 (1997); J. Widom, Q. Rev. Biophys. **34**, 269 (2001).
[6] (a) G. Macaya, J. P. Thiery, and G. Bernardi, J. Mol. Biol. **108**, 237 (1976); (b) G. Bernardi, Gene **276**, 3 (2001); (c) G. Bernardi, *Structural and Evolutionary Genomics* (Elsevier, Amsterdam, 2004).
[7] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992); X. Lu, Z. Sun, H. Chen, and Y. Li, Phys. Rev. E **58**, 3578 (1998); A. Fukushima *et al.*, Gene **300**, 203 (2002); W. Li and D. Holste, Fluct. Noise Lett. **4**, L453 (2004); Phys. Rev. E **71**, 041910 (2005).
[8] W. Li and D. Holste, Comput. Biol. Chem. **28**, 393 (2004).
[9] T. Hassold and P. Hunt, Nat. Rev. Genet. **2**, 280 (2001).
[10] S. E. Antonarakis *et al.*, Nat. Rev. Genet. **5**, 725 (2004); D. Patterson and A. C. S. Costa, *ibid.* **6**, 137 (2005).
[11] N. E. Lamb *et al.*, Am. J. Hum. Genet. **76**, 91 (2005).
[12] H. DeVoe and I. Tinoco, Jr., J. Mol. Biol. **4**, 500 (1962).
[13] K. J. Breslauer, R. Frank, H. Blöcker, and L. A. Marky, Proc. Natl. Acad. Sci. U.S.A. **83**, 3746 (1986).
[14] J. SantaLucia, Jr., Proc. Natl. Acad. Sci. U.S.A. **95**, 1460 (1998).
[15] Genome browser from the University of California at Santa Cruz (UCSC) Genome Bioinformatics Site. URL: http://genome.ucsc.edu/.
[16] P. Miramontes and G. Cocho, Physica A **321**, 577 (2003).
[17] A. Panjkovich and F. Melo, Bioinformatics **21**, 711 (2005).
[18] H. C. M. Nelson, J. T. Finch, B. F. Luisi, and A. Klug, Nature (London) **33**, 221 (1987).
[19] M. A. El Hassan and C. R. Calladine, Philos. Trans. R. Soc. London, Ser. A **355**, 43 (1997).
[20] C. R. Calladine, H. R. Drew, B. F. Luisi, and A. A. Travers, *Understanding DNA—The Molecule and How It Works*, 3rd ed. (Elsevier, Amsterdam, 2004).
[21] M. Gardiner-Garden and M. Frommer, J. Mol. Biol. **196**, 261 (1987); F. Larsen, G. Gundersen, R. Lopez, and H. Prydz, Genomics **13**, 1095 (1992).
[22] K. Matsuo *et al.*, Somatic Cell Mol. Genet. **19**, 535 (1993).
[23] S. C. Satchwell, H. R. Drew, and A. A. Travers, J. Mol. Biol. **191**, 659 (1986).
[24] P. Baldi, S. Brunak, Y. Chauvin, and A. Krogh, J. Mol. Biol. **263**, 503 (1996); A. Stein and M. Bina, Nucleic Acids Res. **27**, 848 (1999).
[25] V. G. Levitsky, O. A. Podkolodnaya, N. A. Kolchanov, and N. L. Podkolodny, Bioinformatics **17**, 998 (2001); **17**, 1062 (2001).
[26] A. E. Vinogradov, Nucleic Acids Res. **33**, 559 (2005).
[27] J. Mirkovitch, M. E. Mirault, and U. K. Laemmli, Cell **39**, 223 (1984).
[28] (a) I. Liebich, J. Bode, M. Frisch, and E. Wingender, Nucleic Acids Res. **30**, 307 (2002); (b) I. Liebich, J. Bode, I. Reuter, and E. Wingender, *ibid.* **30**, 3433 (2002).
[29] Y. Saitoh and U. K. Laemmli, Cell **76**, 609 (1994).
[30] W. Li (unpublished).
[31] A. E. Vinogradov, Nucleic Acids Res. **31**, 1838 (2003).
[32] D. E. Coming, Annu. Rev. Genet. **12**, 25 (1978); T. Ikemura and S. Aota, J. Mol. Biol. **203**, 1 (1988).
[33] Y. Niimura and T. Gojobori, Proc. Natl. Acad. Sci. U.S.A. **99**, 797 (2002).
[34] D. Mouchiroud *et al.*, Gene **100**, 181 (1991); S. Zoubak, O. Clay, and G. Bernardi, *ibid.* **174**, 95 (1996).
[35] N. Sadoni *et al.*, J. Cell Biol. **146**, 1211 (1999); T. Cremer *et al.*, Crit. Rev. Eukaryot Gene Expr **10**, 179 (2000); R. R. Williams, Trends Genet. **19**, 298 (2003).
[36] S. Saccone, C. Federico, and G. Bernardi, Gene **300**, 169 (2002).
[37] J. Filipski *et al.*, EMBO J. **19**, 1319 (1990).
[38] K. Van Holde and J. Zlatanova, J. Biol. Chem. **270**, 8373 (1995).
[39] J. Filipski and M. Mucha, Gene **300**, 63 (2002); A. Marin, M. Wang, and G. Gutierrez, *ibid.* **333**, 151 (2004).