Low Frequency Rhythms in Human DNA Sequences: A Key to the Organization of Gene Location and Orientation?

S. Nicolay,[1] F. Argoul,[1] M. Touchon,[2] Y. d'Aubenton-Carafa,[2] C. Thermes,[2] and A. Arneodo[1]

[1]*Laboratoire de Physique, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France*
[2]*Centre de Génétique Moléculaire (CNRS), Allée de la Terrasse, 91198 Gif-sur-Yvette, France*

(Received 12 July 2003; published 31 August 2004)

We explore large-scale nucleotide compositional fluctuations of the human genome using multiresolution techniques. Analysis of the GC content and of the AT and GC skews reveals the existence of rhythms with two main periods of 110 ± 20 kb and 400 ± 50 kb that enlighten a remarkable cooperative gene organization. We show that the observed nonlinear oscillations are likely to display all the characteristic features of chaotic strange attractors which suggests a very attractive deterministic picture: gene orientation and location, in relation with the structure and dynamics of chromatin, might be governed by a low-dimensional nonlinear dynamical system.

DOI: 10.1103/PhysRevLett.93.108101 PACS numbers: 87.15.Cc, 05.45.–a, 87.10.+e

Understanding how chromatin is spatially and dynamically organized in the nucleus of eukaryotic cells and how this affects genome functions is one of the main challenges of cell biology. Recent technical developments in live cell imaging have confirmed that the structure and dynamics of chromatin play an essential role in regulating many biological processes, such as gene activity, DNA replication, recombination and DNA damage repair [1]. Actually, the structure and dynamics of chromatin are under the control of a number of mechanisms involving DNA-protein interactions, but the role of the DNA sequence itself in these processes remains controversial. On a local scale, specific sequence elements have been identified to interact with protein components of chromatin. For instance, some sequence motifs that favor the formation and positioning of nucleosomes, the basic unit of chromatin structure, were found to be regularly spaced [2]. Alternatively, similar motifs were shown to present long-range correlations (LRC) along the genome that are a signature of nucleosomes [3]. Other DNA regions, the scaffold or matrix attachment regions that constitute the anchor points of chromatin loop domains, are constituted by ~1 kbp AT-rich sequence patterns [4]. On larger scales, the folding of the nucleosomal strings into higher-order structures has been the issue of various models involving, e.g., random packing, coiling into hierarchical helical structures (solenoids), or loop-models, [5], but the DNA sequence itself was not taken into account. Recent results propose that loops are organized by the active transcription complexes [6]. Accordingly, gene positions and transcriptional activities would constitute major determinants of the microscopic structure of chromatin that would self-organize in a rather predictable way: the 3D structure would then result from the DNA primary sequence. We study here different compositional functions of the genome sequence that are known to be related to chromatin structure and to gene transcriptional activity, namely, the GC content and the intrastrand asymmetries between A and T (C and G).

Although the description of mammalian genomes in terms of isochores (domains of relatively constant GC) remains controversial [7], the large-scale heterogeneity of their GC content is fundamental for the understanding of chromosome organization, including gene density, replication timing, and chromatin packaging and positioning in the nucleus [1(c),8]. We report here the results for large GC-rich fragments in the human chromosomes (ch.) 11 (24 Mb, NT_033899.3), 14 (68 Mb, NT_02637.9), 21 (29 Mb, NT_011512.7), and 22 (23 Mb, NT_011520.8) that clearly reveal the existence of low-frequency rhythms. In Fig. 1(b) is shown a space-scale decomposition of the GC content of a portion of the ch. 22 [Fig. 1(a)]. This decomposition reveals that for distances larger than ~30 kb, the GC content can no longer be considered as fluctuating homogeneously; it instead displays rather regular nonlinear oscillatory behavior. The scale (or frequency^{-1}) content of this oscillating regime reveals the existence of two main broad peaks corresponding to the scales $\ell_1 = 100 \pm 20$ kb and $\ell_2 = 400 \pm 50$ kb respectively, that emerge from a continuous background. The former is the characteristic length of the basic oscillations obtained with the low-pass filtering scale $s_1^* = 40$ kb, although one may observe from time to time oscillations that have a larger length ($\sim 2\ell_1 = 200$ kb). If one uses a larger filtering scale $s_2^* = 160$ kb, in order to smooth out both the "small scales" (high frequencies) homogeneous long-range correlated GC fluctuations [3] and the basic oscillations of scale ℓ_1, one gets some oscillatory profile with a fundamental length $\ell_2 = 400$ kb [Fig. 2(a)]. Let us point out that similar periodicities are found for chs. 11 ($\ell_1 = 120 \pm 30$ kb, $\ell_2 = 410 \pm 60$ kb), 14 ($\ell_1 = 130 \pm 30$ kb, $\ell_2 = 420 \pm 60$ kb), and 21 ($\ell_1 = 110 \pm 20$ kb, $\ell_2 = 390 \pm 50$ kb), which indicates that these rhythms are likely to be robust characteristics of human DNA.

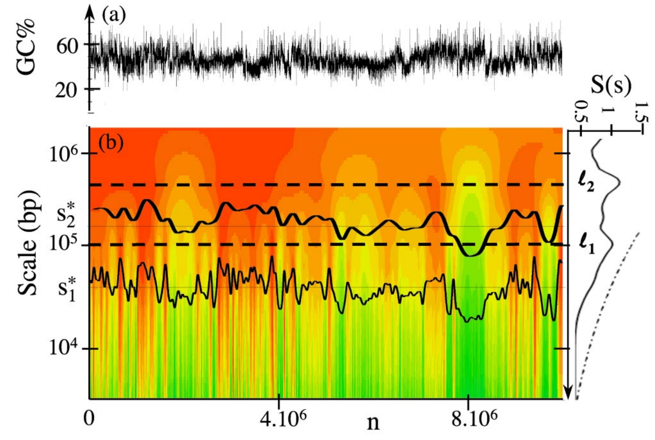

FIG. 1 (color online). Space-scale representation of the GC content of a 10 Mbp long fragment of the ch. 22 when using a Gaussian smoothing filter $g_s(x) = e^{-x^2/2s^2}/\sqrt{2\pi s^2}$. (a) GC content fluctuations computed in adjacent 1 kb boxes. (b) Color coding of the convolution product $T(n, s) = (GC\% * g_s)(n)$ using 256 colors from black (min) to red (max); superimposed are shown the smoothed GC profiles obtained at scales $s_1^* = 40$ kb and $s_2^* = 160$ kb. On the right is shown vertically the scale (frequency^{-1}) spectrum $S(s) = \sum |T(n, s)|$ computed with the complex Morlet wavelet over the entire ch. 22 (solid line); the dotted-dashed line corresponds to the extrapolation at large-scale of the power-law behavior (LRC) observed at scales $\lesssim 10$ kb. The horizontal dashed lines in the color picture correspond to the two main characteristic oscillation lengths $\ell_1 = 100$ kb and $\ell_2 = 400$ kb.

In parallel, we also examined the asymmetries between the two strands, i.e., deviations from intrastrand equimolarities between A and T and between G and C. These deviations have been extensively studied in prokaryotic, organelle, and viral genomes: the leading strand is relatively enriched in G over C and T over A in the weakly selected positions, and their properties have been used to detect the origins of replication [9]. In eukaryotes, recent studies have shown the existence of A/T and C/G skews associated to transcribed regions [10]. The space-scale analysis of the A/T and C/G skews of ch. 22 yields oscillatory profiles similar to those obtained for the GC content with still the two lengths $\ell_1 = 120 \pm 20$ kb and $\ell_2 = 375 \pm 50$ kb. In Fig. 2(b) is shown the oscillatory profile obtained for the smoothing scale s_2^* when adding the two skews. This profile displays rather regular oscillation trends of basic length ~ 375 kb. As compared to the smoothened GC profile [Fig. 2(a)], the oscillatory skew profile provides a remarkable guide for the organization of the spatial location and orientation of the (largest) genes: sense genes with the same orientation as the sequence are located around the negative minima of the oscillations (among transcribed sequences, this corresponds to $79.6 \pm 1.9\%$ (ch. 22), $84.0 \pm 2.6\%$ (ch. 11), $89.2 \pm 1.2\%$ (ch. 14) and $88.1 \pm 2.4\%$ (ch. 21) of 1 kb fragments that have the same orientation as the Watson strand), while antisense genes are quite symmetrically located around the maxima (mainly positive).

The two characteristic oscillation scales observed for both the GC content and the skews suggest two interpretations. The first one is of structural nature and is related to the hierarchical folding of chromatin into fibers and loops of different sizes: 100 kb corresponds to the size of DNA loops that have been observed by a number of experimental techniques [1,4,11]; 400 kb is likely to be the size of larger chromatin loops and/or may correspond to several basic loops or multi-loop subcompartments [5(c)]. The second interpretation is of functional nature and is based on the observation that these characteristic lengths correlate well to the replicon sizes observed in warm-blooded vertebrates [12]. Since these skew oscillations are also observed in large intergenic regions (but with smaller amplitude), they may arise from both transcription and replication mutation bias. Indeed, these oscillations are likely to reflect some correlation between gene organization into clusters with preferential gene orientation and replication.

As regards to the potential structural and dynamical significance of these large-scale GC and skew oscillations, one may raise the question of their stochastic or deterministic nature. Keeping in mind the possible sources of randomness in human DNA (repeated insertions, duplications, mutations, satellite DNA elongations, recombinations, translocations, ...), a random model seems quite natural. Here we use concepts and methods introduced in dynamical systems theory [13] to test whether deterministic chaos can be a realistic alternative. As shown in Fig. 3, this methodology consists in three

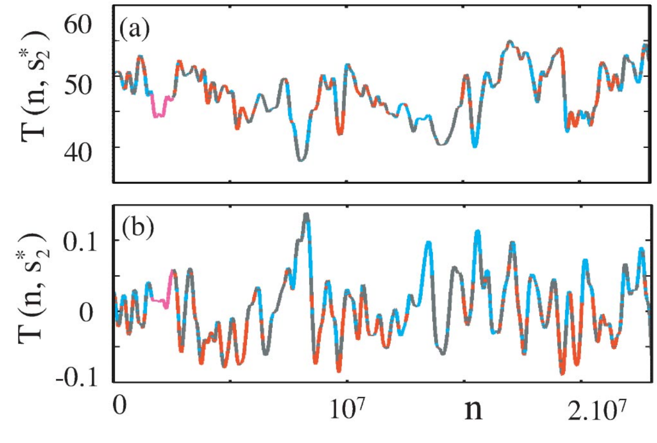

FIG. 2 (color online). Compositional oscillations observed in the ch. 22 fragment after low-pass filtering at scale $s_2^* = 160$ kb (see Fig. 1). (a) GC content; (b) deviation from intrastrand equimolarities $X = (A-T)/(A+T) + (C-G)/(C+G)$. The red (blue) portions of the profiles correspond to the location of sense (antisense) genes that have the same (opposite) orientation than the sequence. The location of the immoglobulin locus is shown in pink.

main steps: (i) the reconstruction of the trajectory, using the signal and its first $(d-1)$ derivatives as the coordinates in a d-dimensional phase-space; (ii) the definition of a Poincaré map from the successive crossings of the trajectory with a transversally intersecting hyperplane; (iii) for $d = 3$, the construction of a 1D map (iterative rule) by plotting W_{n+1} vs W_n, where W is some appropriate coordinate in the Poincaré section. In Fig. 3, we apply this methodology simultaneously to the ch. 22 skew oscillating profile [Fig. 2(b)] and to two numerically generated oscillatory profiles that will serve, respectively, as test deterministic and stochastic fluctuating signals. The former is the solution of the symmetric $(\theta \rightarrow -\theta)$ third-order nonlinear ordinary differential equation:

$$\dddot{\theta} + \mu_2\ddot{\theta} + \mu_1\dot{\theta} + \mu_0\theta + k\theta^3 = 0, \tag{1}$$

that has been emphasized to be the paradigm of nonlinear oscillators that display homoclinic chaos of Shil'nikov's type [14]. When adjusting the parameters to values ($\mu_0 = -5.5$, $\mu_1 = 3.5$, $\mu_2 = 1$, $k = 1$) close to homoclinic conditions, one expects to observe spiraling strange attractor behavior that is intermittently reinjected to the neighborhood of the saddle focus located at the origin after some oscillations around the two other saddle foci $\theta^*_{1,2} = \pm(-\mu_0)^{1/2}$ that ensure the nonlinear saturation of the dynamics. The second test numerical profile is obtained by a low-pass filtering (at scale s_2^*) of a Brownian motion that mimics a scale-invariant random walk profile.

In Fig. 3(a) is shown the 3D phase-portrait obtained from the ch. 22 skew oscillatory profile of Fig. 2(b). The topology of the corresponding trajectory is more regular and well organized than the structureless space-filling trajectory of the Brownian path shown in Fig. 3(a″). It strongly resembles the spiraling chaotic strange attractor generated by Eq. (1) in Fig. 3(a′). In particular, it displays a similar symmetric (skew → −skew) spiraling dynamics with episodic oscillations in the negative (positive) skew half-space mostly containing the large sense (antisense) genes. Using Shil'nikov's homoclinic chaos as a theoretical guide [14], we show in Fig. 3(b) the Poincaré map defined by the successive crossings of the ch. 22 skew trajectory [Fig. 3(a)] with a horizontal plane in the negative skew half-space where a negative skew saddle focus is likely to be located. One gets data points that are not at all spatially distributed in a random fashion, as observed for the Brownian motion trajectory in Fig. 3(b″). When distinguishing the crossings obtained from the successive trajectory loops around the lower negative skew saddle focus (green symbols), from those reinjected from the upper positive saddle focus (black symbols), one realizes that the sets of green and black crossings do no mix one with each other. In remarkable agreement with the two spiral arm strange attractors observed in Fig. 3(b′) for the Poincaré map of Shil'nikov homoclinic chaos, all the green crossings fall rather consistently on a spiraling pattern. Similarly, all the black crossings lie coherently on a geometrical curve that can again be approximated by a spiral having the same center as the previous one, but phase-shifted by π. Focusing on the green crossings only, we study their dynamics by plotting W_{n+1} vs W_n, where $W = \cos(\alpha)Y + \sin(\alpha)Z$. When tuning the parameter α to the value $\alpha = 302.5°$, then all the green crossing data points remarkably fall on a unique nonlinear curve [Fig. 3(c)], the hallmark of deterministic chaos [13]. The investigation of the black crossings leads to a similar diagnostic, contrasting with the featureless 1D map obtained in Fig. 3(c″) for the random walk trajectory. The fact that this nonlinear curve

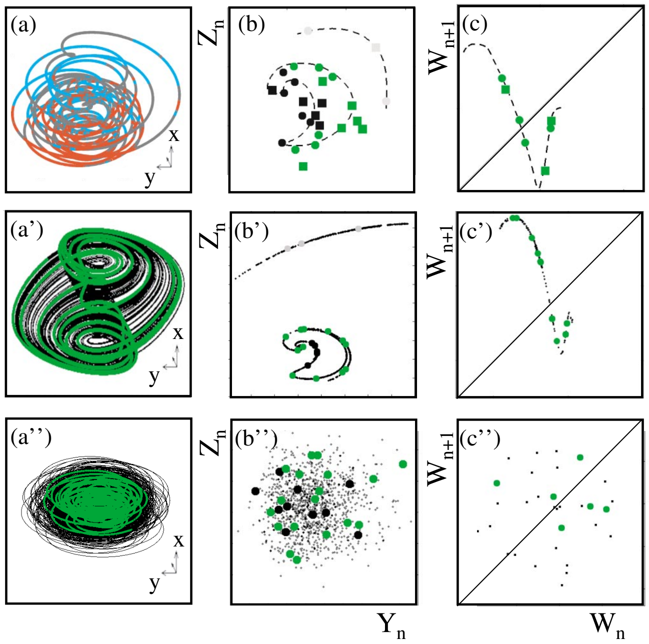

FIG. 3 (color online). Dynamical system analysis: (a)–(c) the 23 Mbp fragment of the ch. 22; in (b) and (c) disks and squares mean chs. 22 and 11, respectively; the dashed lines are drawn to guide the eyes. (a′)–(c′) the chaotic nonlinear oscillator [Eq. (1)]. (a″)–(c″) an uncorrelated random walk. Phase-portrait: trajectories reconstructed in a 3D (X, Y, Z) phase-space where $Y = \dot{X}$ and $Z = \ddot{X}$ from (a) the smoothed intrastrand assymmetry X defined in Fig. 2(b); (a′) the asymptotic solution $X = \theta$ of Eq. (1); (a″) the Brownian path X after the same low-pass filtering as in (a). Poincaré map: first return maps obtained from the crossings of the trajectories with some horizontal plane $X = X^* = \mathrm{Cst}$ along the direction $\dot{X} < 0$ where (b) $X^* = -0.16$; (b′) $X^* = -2$; (b″) $X^* = -0.01$. 1D map: 1D maps obtained from the green symbols in (b), (b′), and (b″), respectively, by plotting W_{n+1} vs W_n, where $W_n = \cos(\alpha)Y_n + \sin(\alpha)Z_n$, with (c) $\alpha = 302.5°$; (c′) $\alpha = 19.8°$; (c″) $\alpha = 115°$. In (b′) and (b″), the solid green and black symbols correspond to the green part of the trajectories in (a′) and (a″) which has a similar length as the ch. 22 trajectory in (a). In (b) and (b′) the green symbols correspond to successive loops of the trajectory around the lower $(X < 0)$ saddle focus without visiting the neighborhood of the upper focus.

looks like the multi-humped 1D map shown in Fig. 3(c′) strongly suggests that (i) when one knows some crossing of the ch. 22 trajectory with the Poincaré plane, one can predict the next one, and (ii) the succession of crossings is likely to obey the recursive dynamics of a homoclinic chaotic trajectory of Shil'nikov type [14]. In Figs. 3(b) and 3(c), we have superimposed data points coming from the analysis of the human ch. 11. They both display quite similar spatial arrangement [Fig. 3(b))] and dynamical characteristics [Fig. 3(c)]. In order to test whether deterministic chaos might be pertinent to describe the large-scale structure of human DNA sequences, we have computed the spectrum of Lyapunov exponents to confirm the existence of sensitivity to initial conditions. Actually, the maximal Lyapunov exponent is found positive, independently of the choice of the embedding dimension d ($4 \leq d \leq 7$) and this not only for the gene rich chs. 22 ($\lambda = 7.0 \pm 1.2 \times 10^{-3}$) and 11 ($\lambda = 7.5 \pm 1.5 \times 10^{-3}$) fragments, but also for the chs. 14 ($\lambda = 9.5 \pm 1.0 \times 10^{-3}$) and 21 ($\lambda = 8.8 \pm 1.8 \times 10^{-3}$) contigs. These results are quite consistent and in agreement with the estimate $\lambda = 6.5 \pm 1.0 \times 10^{-3}$ obtained for the numerical chaotic trajectory of Eq. (1) after rescaling θ and t in order to get amplitude and characteristic frequencies similar to those of the skew profile in Fig. 2(b). To what extent (i) this deterministic chaotic picture does apply to human chromosomes or only to some fragments of them and (ii) these large-scale chaotic oscillations in GC content and A/T (C/G) skews are common to other eukaryotic genomes, are fundamental issues that deserve further investigation.

Beyond the two fundamental periods ℓ_1 and ℓ_2, one can identify some other characteristic scales in the skew profile shown in Fig. 2(b). The 1 Mb episodes of 2, 3, or 4 successive positive or negative skew oscillations that correspond, respectively, to successive loops of the trajectory around the (hypothetical) positive and negative skew saddle foci in the reconstructed phase-portrait [Fig. 3(a)] might be associated to replication clusters or large replicons [6(b),12,15]. There is a well defined extra peak at $\ell_3 = 1100 \pm 200$ kb in the scale content of the A/T and C/G skews. Patterns of a few Mbp long might also be the signature of chromosome translocations. Understanding functional chromosome territory architecture therefore requires a definite answer to the question whether subchromosomal foci and/or replication foci are rather randomly arranged or highly organized structures that undergo more or less complicated dynamical changes during the cell cycle and cell differentiation. Most of the current computer models [5] of the high-order structure and dynamics of chromatin are inspired from polymer statistical physics and are definitely random. They all consist in modeling the structure of chromatin loops by a random walk and their dynamics by a diffusional motion. The results reported in this Letter put into light a very attractive alternative picture: the functional relationship between chromatin condensation and decondensation processes on the one hand and replication and gene regulation on the other hand, might be governed to some extent by a low-dimensional chaotic dynamical system.

This work was supported by the Action Bioinformatique Inter EPST2.

[1] (a) J. M. Bridger and W. A. Bickmore, Trends Genet. **14**, 403 (1998); (b) A. S. Belmont *et al.*, Current Opinion in Cell Biology **11**, 307 (1999); (c) T. Cremer and C. Cremer, Nature Reviews Genetics **2**, 292 (2001); (d) S. M. Gasser, Science **296**, 1412 (2002).

[2] (a) S. C. Satchwell, H. R. Drew, and A. A. Travers, J. Mol. Biol. **191**, 659 (1986); (b) H. Herzel, O. Weiss, and E. N. Trifonov, Bioinformatics **15**, 187 (1999).

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

[4] U. K. Laemmli *et al.*, Current Opinion in Genetics and Development **2**, 275 (1992).

[5] (a) R. K. Sachs *et al.*, Proc. Natl. Acad. Sci. U.S.A. **92**, 2710 (1995); (b) J. Y. Ostashevsky, Mol. Biol. Cell **9**, 3031 (1998); (c) C. Munkel *et al.*, J. Mol. Biol. **285**, 1053 (1999); (d) G. van den Engh, R. Sachs, and B. J. Trash, Science **257**, 1410 (1992); (e) G. Li, G. Sudlow, and A. S. Belmont, J. Cell. Biol. **140**, 975 (1998).

[6] (a) P. R. Cook, Nat. Genet. **32**, 347 (2002); (b) N. L. Mahy, P. E. Perry, and W. A. Bickmore, J. Cell. Biol. **159**, 753 (2002).

[7] (a) G. Bernardi, Gene **241**, 3 (2000); (b) *International Human Genome Sequencing Consortium*, [Nature (London) 409, 860 (2001)]; (c) W. Li *et al.*, Computational Biology and Chemistry **27**, 5 (2003).

[8] (a) J. M. Craig and W. A. Bickmore, BioEssays **15**, 349 (1993); (b) Y. Saitoh and U. K. Laemmli, Cell **76**, 609 (1994).

[9] (a) J. R. Lobry and N. Sueoka, Genome Biology **3**, 0058 (2002); (b) J. Mrazek and S. Karlin, Proc. Natl. Acad. Sci. U.S.A. **95**, 3720 (1998).

[10] (a) P. Green *et al.*, Nat. Genet. **33**, 514 (2003); (b) M. Touchon *et al.*, FEBS Lett. **555**, 579 (2003).

[11] M. G. Poirier *et al.*, Phys. Rev. Lett. **86**, 360 (2001).

[12] R. Berezney, D. D. Dubey, and J. A. Huberman, Chromosoma **108**, 471 (2000).

[13] (a) P. Bergé, Y. Pomeau, and C. Vidal, *Order Within Chaos* (John Wiley, New York, 1986); (b) *Homoclinic Chaos*, Physica D (Amsterdam) Vol. 62, edited by P. Gaspard, A. Arneodo, R. Kapral, and C. Sparrow (1993).

[14] (a) A. Arneodo, P. Coullet, and C. Tresser, Commun. Math. Phys. **79**, 573 (1981); (b) A. Arneodo, P. H. Coullet, and E. A. Spiegel, Geophys. Astrophys. Fluid Dyn. **31**, 1 (1985).

[15] D. A. Jackson and A. Pombo, J. Cell. Biol. **140**, 1285 (1998).