Available online at www.sciencedirect.com

SCIENCE DIRECT®

Physica A 342 (2004) 270–280

PHYSICA A

www.elsevier.com/locate/physa

# From scale invariance to deterministic chaos in DNA sequences: towards a deterministic description of gene organization in the human genome

S. Nicolay[a], E.B. Brodie of Brodie[a], M. Touchon[b], Y. d'Aubenton-Carafa[b], C. Thermes[b], A. Arneodo[a,*]

[a] *Laboratoire de Physique, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France*

[b] *Centre de Génétique Moléculaire (CNRS), Allée de la Terrasse, 91198 Gif-sur-Yvette, France*

Received 24 November 2003

**Abstract**

We use the continuous wavelet transform to perform a space-scale analysis of the AT and GC skews (strand asymmetries) in human genomic sequences, which have been shown to correlate with gene transcription. This study reveals the existence of a characteristic scale $\ell_c \simeq 25 \pm 10$ kb that separates a monofractal long-range correlated noisy regime at small scales ($\ell < \ell_c$) from relaxational oscillatory behavior at large-scale ($\ell > \ell_c$). We show that these large scale nonlinear oscillations enlighten an organization of the human genome into adjacent domains ($\approx 400$ kb) with preferential gene orientation. When using classical techniques from dynamical systems theory, we demonstrate that these relaxational oscillations display all the characteristic properties of the chaotic strange attractor behavior observed nearby homoclinic orbits of Shil'nikov type. We discuss the possibility that replication and gene regulation processes are governed by a low-dimensional dynamical system that displays deterministic chaos.

*PACS:* 87.15.Cc; 05.45.−a; 87.10.+e

*Keywords:* Wavelet transform; Scale invariance; Long-range correlations; Colored noise; Deterministic chaos; DNA sequences; Human genome; Skewness; Genes

[*] Corresponding author.
*E-mail address:* alain.arneodo@ens-lyon.fr (A. Arneodo).

0378-4371/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2004.04.088

## 1. Introduction

Recently, the availability of fully sequenced genomes has enabled the statistical analysis of genomic sequences from the base pair scale up to the chromosome size. Different techniques including mutual information functions, entropies, auto-correlation functions, power spectra, "DNA walk" representation, Zipf analysis and wavelet transform (WT) have been used to investigate the structural complexity of DNA sequences and to test the possible relevance of scale invariance and fractal concepts (for reviews see Refs. [1,2]). One of the main messages that come out from these pioneering studies is the fundamental importance of the choice of the coding rule used to transform the DNA text into a signal that can be further processed. From a given sequence, different mono-, di- or tri-nucleotide coding rules can lead to different structural, dynamical or functional informations. In a previous work [3], the use of the WT to characterize the scale invariance properties of DNA bending profiles constructed from experimentally established tri-nucleotide structural tables, has revealed the existence of long-range correlations (LRC) over distances up to $\sim 20$ kb as related to the nucleosomal structure and dynamics of the chromatin fiber. Here we analyze the AT and GC skew profiles for large contigs in the human genome. These deviations from intrastrand equimolarities between A and T and between G and C have been extensively studied in prokaryotic genomes and particularly used to detect the origins of replication [4]. In the human genome, recent studies have shown that compositional strand asymmetries are coupled to transcription: on the coding strand, most genes present an excess of T over A and of G over C, exhibiting sharp transitions between transcribed and non-transcribed regions [5,6]. In addition, the values of these skews correlate with gene expression [7]. Taking advantage of these properties, we use here the WT microscope to perform a space-scale analysis of the AT and GC skew profiles in order to investigate the possible links between gene organization and gene expression.

## 2. Space-scale analysis based on the continuous wavelet transform

The WT is a space-scale analysis which consists in expanding signals in terms of wavelets that are constructed from a single function, the analyzing wavelet $\Psi$, by means of dilations and translations [2,8,9]. When using the successive derivatives of the Gaussian function, namely $g^{(N)}(x)=d^N g^{(0)}(x)/\mathrm{d}x^N$, with $g^{(0)}(x)=\mathrm{e}^{-x^2/2}/\sqrt{2\pi}$, then the WT takes the following simple expression:

$$T_{g^{(N)}}[s](x,a)=\frac{1}{a}\int_{-\infty}^{+\infty} s(y)g^{(N)}\left(\frac{y-x}{a}\right)\,\mathrm{d}y=\frac{d^N}{\mathrm{d}x^N}\,T_{g^{(0)}}[s](x,a)\,, \tag{1}$$

where $x$ and $a$ ($>0$) are the space and scale parameters respectively. Eq. (1) shows that the WT computed with $g^{(N)}$ is the $N$th derivative of the signal $s(x)$ smoothed by a dilated version $g^{(0)}(x/a)$ of the Gaussian function. This property is at the heart of various applications of the WT microscope as a very efficient multi-scale singularity tracking technique [2,8,9].

Actually, the skeleton of the WT (defined, at each scale $a$, by the set of all the points $x_i$ that correspond to a local maximum of $|T_\Psi[s](x,a)|$ and then by connecting these points across scales into the so-called maxima lines) provides a space-scale partitioning that is likely to contain all the information on the singularities of the signals [2,8,9]. Along the maxima line pointing to the point $x_o$ in the limit $a \to 0^+$, one can show that $|T_\Psi[s](x,a)| \sim a^{h(x_o)}$, provided that the order N of $\Psi$ be larger than the Hölder exponent $h(x_o)$ that characterizes the strength of the singularity located at $x_o$. According to the wavelet transform modulus maxima (WTMM) method [2,9], one can proceed to a statistical analysis of the singularities of $s$ via an investigation of the scaling behavior of some partition functions $\mathcal{Z}(q,a) = \sum_{x_i \in \mathcal{S}(a)} |T_\Psi(x_i,a)|^q \sim a^{\tau(q)}$ $(q \in \Re)$, where the sum is taken over the WT skeleton. The Legendre transform of the scaling exponents $\tau(q)$ is then the singularity spectrum $D(h) = \min_q(qh - \tau(q))$, defined as the Hausdorff dimension of the set of points $x$ where the Hölder exponent of $s$ is $h$ [2,9]. Homogeneous fractal distributions $s$ that involve singularities of unique Hölder exponent $h(x) = h$, are characterized by a linear $\tau(q)$ spectrum ($h = \partial\tau/\partial q$). On the contrary, a nonlinear $\tau(q)$ curve is the signature of nonhomogeneous distributions that display multifractal properties (i.e., $h(x)$ is a fluctuating quantity that depends upon $x$). Let us emphasize that one of the main advantages of the WT is its adaptative ability to perform time-frequency analysis [8] when using complex analyzing wavelets like the Morlet's wavelet $\Psi_M(t) = 1/\sqrt{2\pi}\mathrm{e}^{\mathrm{i}wt}\mathrm{e}^{-t^2/2}$.

## 3. Space-scale analysis of AT and GC strand asymmetries in the human genome

In this section, we report the results of a wavelet-based space-scale decomposition of the AT and GC skew profiles computed for large human contigs, namely a 23 Mb (resp. 24 Mb) long fragment of the chromosome 22 (resp. chromosome 11). $S_{AT}=(A-T)/(A+T)$ and $S_{GC}=(G-C)/(G+C)$ were calculated in adjacent 1 kb windows and plotted along the sequence. Since both profiles display similar noisy patterns (although generally $|S_{GC}| < |S_{AT}|$) with synchronous jumps that are likely to correspond to transitions between transcribed and non-transcribed regions, we will mainly consider $S=S_{AT}-S_{GC}$ (Fig. 1), in order to get larger fluctuation amplitudes than those displayed by each individual skew. The main outcome of our space-scale wavelet analysis is that the human skew profiles can be mainly decomposed into two components, a long-range correlated (colored) noisy background behavior on top of a large-scale oscillatory behavior with a pronounced relaxational character.

### 3.1. Colored noise behavior of human skew profiles ($a \lesssim 25$ kb)

When applying the WTMM method to the $S$ skew profiles of both the chromosomes 11 and 22 fragments, using $g^{(1)}(x)$ as analyzing wavelet (Fig. 1b), one clearly observes a scaling behavior of the partition functions $\mathcal{Z}(q,a)$ for scales $a \lesssim \ell_c$, where $\ell_c$=25±10 kb, above which some transition occurs that will be discussed in Section 3.2. The scaling exponent spectrum is clearly found linear, $\tau(q) = -0.18q - 1$, as the signature of monofractal scaling properties with Hölder exponent $h = -0.18 \pm 0.04$.

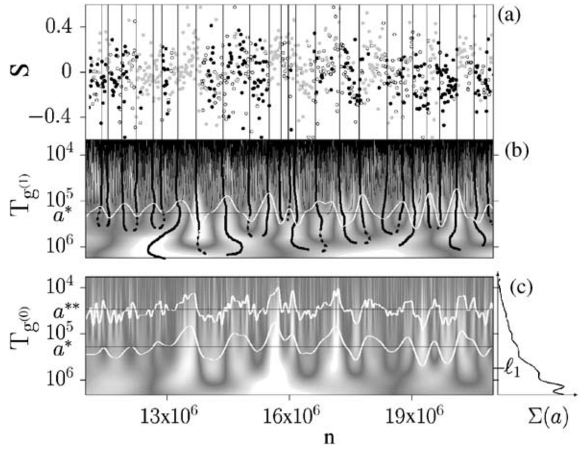

Fig. 1. Space-scale representation of the skew $S = S_{AT} - S_{GC}$ of a part of a 23 Mb long fragment (NT_011520.8) of the human chromosome 22 when using $g^{(N)}(x)$ as analyzing functions. (a) $S$ vs. $n$; the symbols correspond to (+) genes (black dots), (−) genes (grey dots) and intergene (circles); the vertical bars are defined from the extrema lines of $T_{g^{(1)}}[S]$ that exist at scale $a \geqslant a^* = 160$ kb (thick solid and dashed lines in (b)). (b) Wavelet representation of $S$ using $g^{(1)}(x)$; $T_{g^{(1)}}[S](n,a)$ is coded, independently at each scale $a$, using 256 grey levels from black (min) to white (max); the WT skeleton is shown in thin solid lines; the thick solid and dashed lines correspond respectively to WT local maxima and minima that exist at scales $a \geqslant a^* = 160$ kb (see text), as illustrated by the superimposed white oscillatory profile of $T_{g^{(1)}}[S](n,a^*)$. (c) Grey level coding of $T_{g^{(0)}}[S](n,a)$; superimposed to this picture is shown the skew profile smoothed at scale $a^*$ and $a^{**} = 40$ kb; at the right end of this figure is shown vertically the scale (frequency$^{-1}$) spectrum $\sum(a) = \sum |T_{\Psi_M}[S](n,a)|$ computed over the entire chromosome 22.

This result is corroborated by the study of the probability density functions (pdfs) of the WT modulus maxima that define the WT skeleton (Fig. 1b). As reported in Fig. 2a, these pdfs clearly evolve across scales but, as shown in Fig. 2b, they collapse onto a single curve, as predicted by the self-similarity relationship [2,9]:

$$a^h P_a(a^h|T|) = P(|T|)\ , \tag{2}$$

when adjusting $h{=}-0.18$. These observations converge to the diagnostic that, when seen at small scales ($a \lesssim 25$ kb), the skew profile behaves as a monofractal colored ($1/f^{\beta}$) noise that is singular everywhere with Hölder exponent $h = -0.18$ ($\beta = 2h + 1 = 0.64$). Note that consistent results are obtained when applying the WTMM method to the cumulative skew function; the corresponding “DNA walk” is found monofractal with a Hölder exponent $H(=h+1)=0.82$ that accounts for the existence of LRC as previously observed, in the range 0.5 kb $\lesssim a \lesssim$ 20 kb, with different coding rules [2,3].

### 3.2. Relaxational oscillatory behavior of human skew profiles ($a \gtrsim$ 25 kb)

When investigating the WTMM statistics on the WT skeleton at scales $a \gtrsim \ell_c$, one observes a drastic change in the sense that, as shown in Fig. 2c, the WTMM pdf no longer depends on the scale $a$. Without any rescaling, the data fall on the

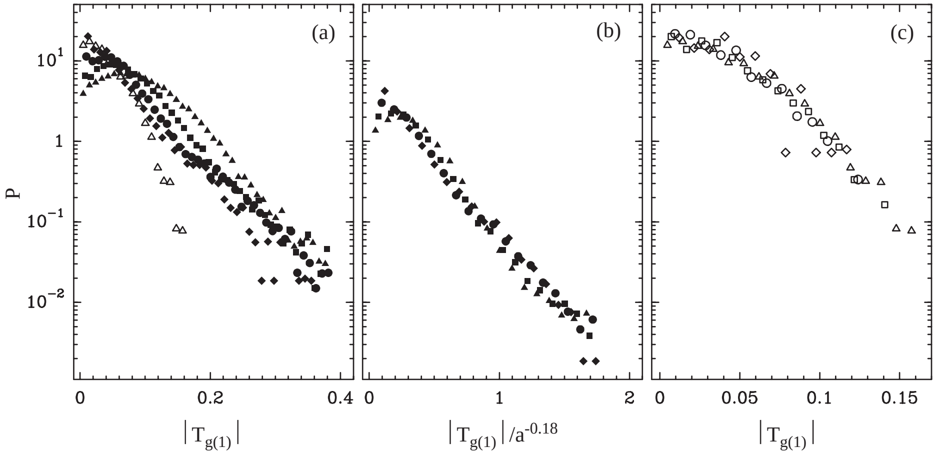

Fig. 2. WTMM pdfs of the skew profile $S = S_{AT} - S_{GC}$ of both the 23 Mb contig of the human chromosome 22 (Fig. 1) and a 24 Mb long fragment (NT_033899.3) of chromosome 11. The analyzing wavelet is $g^{(1)}(x)$. The symbols correspond to the following scales $a = 1$ (▲), 2 (■), 4 (●), 10 (◆), 40 (△), 80 (□), 160 (○), and 320 (◇) in kb units. (a) $P(|T_{g^{(1)}}|)$; (b) $P(|T_{g^{(1)}}|/a^{-0.18})$; (c) $P(|T_{g^{(1)}}|)$. The WT has been multiplied by a factor of 240 as in Fig. 3b.

same curve, which means that these pdfs satisfy Eq. (2) for $h = 0$. Actually, above $\ell_c$, the WTMM are approximately constant on every maxima lines that still exist at these large scales ($|T_{g^{(1)}}[S](n,a)| \sim a^{h=0}$). This result is confirmed when using the WTMM method; above 25 kb, one observes a cross-over in the scaling behavior of the partition function $\mathscr{Z}(q,a)$: the $\tau(q)$ spectrum no longer depends on $q$, as the signature of the presence of sharp jumps (discontinuities) in the skew profile that are detected by the WT microscope (at rather low magnification) as singularities of Hölder exponent $h = \partial\tau/\partial q = 0$. Indeed, as illustrated by the space-scale decomposition of the chromosome 22 skew fluctuations in Fig. 1c, when using the Gaussian function $g^{(0)}$ as smoothing filter, the $S$ profile obtained at the scale $a^{**} = 40$ kb (just above the critical cut-off scale $\ell_c$ required to smooth out the colored noisy background fluctuations) displays a rather regular nonlinear oscillatory behavior of relaxational nature. As shown vertically on the right of Fig. 1c, when using the complex Morlet's wavelet $\Psi_M$ to investigate the frequency content of this oscillatory regime, one gets a spectrum $\Sigma(a)$ that clearly reveals the existence of (at least) one main frequency corresponding to a characteristic length $\ell_1 = 400 \pm 50$ kb. This characteristic oscillatory period is put into light in Fig. 1c (see also Fig. 3a), where a large filtering scale $a^* = 160$ kb has been used to smooth out not only the high frequency noisy fluctuations but also some basic small amplitude oscillations of smaller characteristic length: a few (up to three or four) periods of positive skew values are followed by a few periods of negative skew values with, from time to time, some back and forth oscillations between positive and negative skew values. As illustrated in Fig. 3a, what is quite spectacular is the fact that this skew oscillatory profile provides a remarkable guide for the organization of the spatial location and orientation of the genes: genes with the same orientation as the sequence

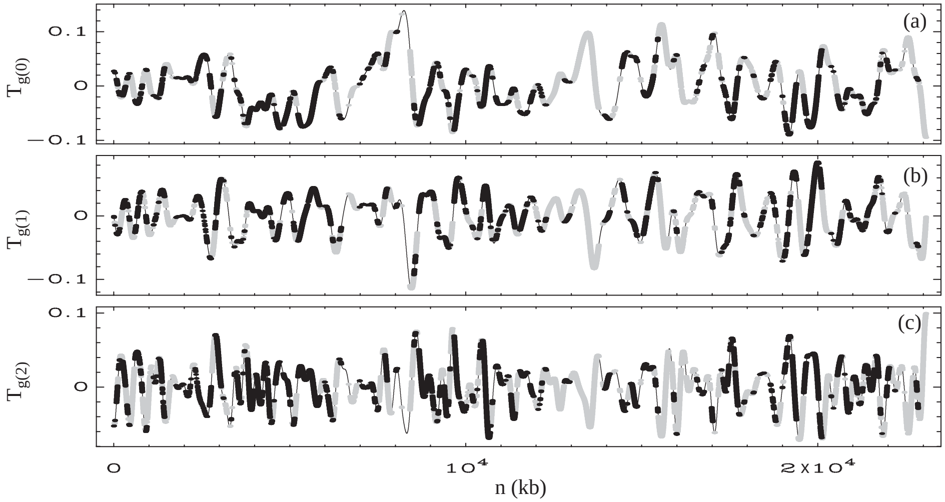

Fig. 3. Oscillatory deviations from intrastrand equimolarities: $S = S_{AT} - S_{GC}$, obtained for the 23 Mb long fragment of the human chromosome 22 after smoothing at scale $a^* = 160$ kb using (a) $g^{(0)}(x)$, (b) $g^{(1)}(x)$ and (c) $g^{(2)}(x)$. The thick black (resp. thick grey) portions of these profiles correspond to the locations of (+) (resp. (−)) genes. In (b) (resp. (c)), the WT has been multiplied by a factor of 240 (resp. 35000) for the sake of comparison.

(Watson orientation, noted as (+)) are located around the minima (mainly negative) of the oscillations while (−) genes are quite symmetrically located around the maxima (mainly positive) [10].

### 3.3. A multi-scale method to study gene clustering

As discussed above, the relaxational nature of the large-scale skew oscillatory behavior is likely to be the footprint of the sharp transitions ($h = 0$) between transcribed and nontranscribed regions as well as between adjacent groups of genes mostly transcribed in the same direction, successive domains corresponding generally to alternating directions. These sharp transitions can be detected by looking at the extrema (maxima and minima) of $T_{g^{(1)}}[S](n,a)$ (from Eq. (1), $T_{g^{(1)}}[S](n,a)$ is the derivative of $T_{g^{(0)}}[S](n,a)$) for a value of the scale $a \gtrsim a^*$ (Fig. 3b). Then by following the corresponding extrema lines across scales down to the resolution scale (1 kb), one determines the location of these jumps. As shown in Fig. 1a, for the chromosome 22 contig, this methodology allows us to determine the limits of adjacent domains whose mean size is 370 kb. Moreover, most of these domains correspond to clusters of genes mainly transcribed in the same direction. For example in domains of average orientation (+), 79.5% of 1 kb fragments have the (+) orientation with significant negative skew ($\bar{S} = -6.67 \pm 0.17\%$), while (−) fragments present small mean bias ($\bar{S} = 0.61 \pm 0.25\%$); domains with (−) orientation present symmetrical properties. In addition, the genes with the same orientation as the whole domains are actually longer that those in the reverse orientation. Several other gene-rich regions (i.e., regions presenting high mean GC content) of the human genome (situated in chromosomes 11, 14, 21 and 22) have been also examined and shown to display similar cooperative organization of gene location and orientation.

## 4. Deterministic chaos in the AT and GC skews of human genomic sequences

Regarding the potential functional significance of these large-scale skew oscillations, one may raise the question of their stochastic (random) or deterministic nature. In this section, we will use the concepts and methods introduced in dynamical systems theory [11–13] to suggest the possible existence of deterministic chaos in human skew profiles [10]. A classical way to reconstruct a phase portrait from a 1D time series consists in using the signal and its first $(d-1)$ derivatives as the coordinates in a $d$-dimensional phase space, where $d$ in commonly called the embedding dimension. According to the definition of the WT in Eq. (1), we show in Fig. 4a the 3D phase portrait obtained from the chromosome 22 smoothed skew profiles $T_{g^{(0)}}[S](n,a^*)$ (Fig. 3a), its first derivative $T_{g^{(1)}}[S](n,a^*)$ (Fig. 3b) and its second derivative $T_{g^{(2)}}[S](n,a^*)$ (Fig. 3c). The topology of the corresponding trajectory strikingly reminds of the spiraling chaotic strange attractors observed in low-dimensional nonlinear dynamical systems [11–13]. In Fig. 5a is shown for comparison the attractor generated by numerical integration of the symmetric $(\theta \to -\theta)$ third-order nonlinear ordinary differential equation:

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

which has been emphasized to be the paradigm of nonlinear oscillators that display homoclinic chaos of Shil'nikov's type [13,14]. When adjusting the model parameters to values ($\mu_0=-5.5$, $\mu_1=3.5$, $\mu_2=1$, $k=1$) close to homoclinic conditions (Fig. 5b),

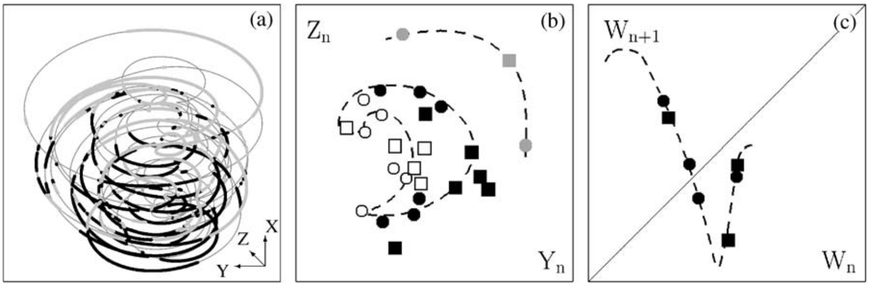

Fig. 4. Dynamical system analysis of the 23 Mb (resp 24 Mb) fragment of the human chromosome 22 (resp. 11). (a) Phase portrait reconstruction of the oscillatory profile of the chromosome 22 skew, where $X=T_{g^{(0)}}[S](n,a^*)$ (Fig. 3a), $Y=T_{g^{(1)}}[S](n,a^*)$ (Fig. 3b), $Z=T_{g^{(2)}}[S](n,a^*)$ (Fig. 3c) and $a^*=160$ kb is the smoothing scale. The thick black (resp. thick grey) portions of the trajectory correspond to (+) (resp. (−)) genes. (b) Poincaré map defined by the successive intersections of the trajectory with the plane $X=-0.16$ along the direction $\dot{X}<0$; the solid black symbols correspond to successive loops of the trajectory around the lower $(X<0)$ saddle focus without visiting the neighborhood of the upper focus. (c) 1D map obtained from the Poincaré map in (b) by plotting $W_{n+1}$ vs. $W_n$ where $W_n=\cos\alpha Y_n+\sin\alpha Z_n$, with $\alpha=302.5°$. In (b) and (c), disks and squares mean chromosome 22 and 11 respectively; the dashed lines are drawn to guide the eyes.

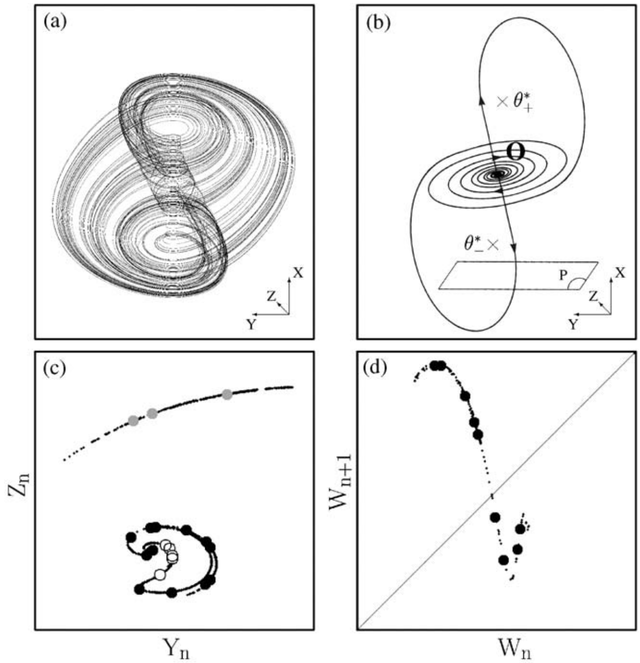

Fig. 5. Chaotic nonlinear oscillator of Shil'nikov type. (a) Trajectory of the nonlinear oscillator defined by Eq. (3) ($\mu_0 = -5.5$, $\mu_1 = 3.5$, $\mu_2 = 1$, $k = 1$), in the 3D phase-space $(X, Y, Z)$, where $X = \theta$, $Y = \dot{\theta}$ and $Z = \ddot{\theta}$. (b) Homoclinic conditions: two symmetric ($\theta \to -\theta$) homoclinic orbits biasymptotic to the saddle focus ($\gamma$, $\rho \pm i\omega$) located at the origin O and that satisfy the Shil'nikov's condition $-\gamma > \rho > 0$ for the existence of homoclinic chaos [14]. (c) Poincaré map obtained from the crossings of the trajectory with the plane P (see (b)) of equation $X = -2$ ($Y = \dot{X} < 0$); the black dots correspond to successive loops of the trajectory around the lower saddle focus $\theta^*_-$ ($X < 0$) without visiting the neighborhood of the upper focus $\theta^*_+$. (d) 1D map obtained from the black dots in the Poincaré map in (c) by plotting $W_{n+1}$ vs. $W_n$, where $W_n = \cos\alpha Y_n + \sin\alpha Z_n$ with $\alpha = 19.8°$.

one expects to observe spiraling strange attractor behavior that is intermittently reinjected to the neighborhood of the saddle focus located at the origin O after some oscillations around the two other saddle foci $\theta^*_\pm = \pm(-\mu_0)^{1/2}$ that ensure the nonlinear saturation of the dynamics. The chromosome 22 skew trajectory in Fig. 4a is likely to display similar symmetric ($S \to -S$) spiraling dynamics with episodic oscillations in the negative (resp. positive) skew half space corresponding to successive domains containing (+) (resp. (−)) gene clusters in Fig. 1a.

Using Shil'nikov homoclinic chaos as theoretical guide, we show in Fig. 5c the Poincaré map defined by the successive crossings of the numerical strange attractor trajectory in Fig. 5a with the plane $\theta = (-\mu_0)^{1/2} =$ Cst in the direction $\dot{\theta} < 0$. As predicted theoretically [14], all these crossings fall on a double multi-folded spiral whose two arms correspond respectively to crossings obtained from the successive trajectory loops around $\theta^*_-$ (●), and from the returns of the trajectory after visiting the neighborhood of $\theta^*_+$ (○). In Fig. 5d is shown the 1D map obtained by plotting $W_{n+1}$ vs.

$W_n$, where $W = \cos(\alpha)\dot{\theta} + \sin(\alpha)\ddot{\theta}$, when one retains only the former class of crossings of the Poincaré plane. With an appropriate choice of $\alpha = 19.8°$, all the data points fall on the graph of a nonlinear function that presents several extrema. This observation enlightens some deep mathematical results [14] showing that multi-humped 1D maps are the corner-stone of symbolic dynamics analysis of Shil'nikov homoclinic chaos. In Fig. 4b and c, we have reproduced this first-return map analysis for the chromosome 22 strange attractor like trajectory illustrated in Fig. 4a. When using a Poincaré plane susceptible to contain the hypothetical negative skew saddle focus (as in Fig. 5b) and further distinguishing (as in Fig. 5c) the crossings obtained from the trajectory loops remaining in the negative skew half-space (filled symbols) from those corresponding to the returns of the trajectory after some visit of the positive skew half-space (empty symbols), one gets data points that are not at all randomly distributed. Actually, the sets of "filled" and "empty" crossings do not mix with each other; all the "filled" crossings fall rather consistantly on a spiraling pattern. Similarly, all the "empty" crossings lie coherently on a geometrical curve that can again be approximated by a spiral having the same center, but phase-shifted by $\pi$. Focusing on the "filled" crossings only, one can study their dynamics on the corresponding spiral arm by plotting $W_{n+1}$ vs. $W_n$ as before (Fig. 5d). As shown in Fig. 4c, when tuning the parameter $\alpha$ to 302.5°, then all the data points fall onto a unique nonlinear curve, the hallmark of deterministic chaos [11–14]. Actually, the fact that this nonlinear curve behaves like the multi-humped 1D map shown in Fig. 5d strongly suggests that (i) when one knows some crossings of the chromosome 22 trajectory with the Poincaré plane, one can predict the next one, and (ii) the succession of crossings is very likely to obey the recursive dynamics of homoclinic chaotic trajectory of Shil'nikov type [10]. (Note that the investigation of the "empty" crossings leads to the same conclusion.)

Since chaos is well known to arise from exponential growth of infinitesimal perturbations together with some global nonlinear folding mechanism to guarantee the boundedness of the solution [11–13], we have further used the TISEAN package [15] to compute the spectrum of Lyapunov exponents. For the chromosome 22 skew trajectory (Fig. 4a), we have found that, independently of the choice of the embedding dimension ($4 \leqslant d \leqslant 7$), the maximal (and only the maximal) Lyapunov exponent $\lambda \simeq 7.0 \pm 1.0 \times 10^{-3}$ is definitely positive which confirms the existence of sensitivity to initial conditions. In Table 1 are reported the results of similar calculations for large contigs in human chromosomes 11 (Figs. 4b and c), 14, 21 and 22 (Fig. 4), which confirm that Shil'nikov chaotic strange attractor behavior of the type shown in Fig. 4, is likely to be a robust characteristic of human chromosomes since the only chromosomal regions where these skew large-scale oscillations are not visible (or of too small amplitude to be analyzed) are the several Mb long low GC regions [16] with low gene density (e.g. the 7 Mb low GC region of chromosome 21).

## 5. Discussion

In this work, we have shown that when (i) considering an original coding, namely the (AT–GC) skew, (ii) using a multi-scale methodology based on the space-scale

Table 1
Computation of the largest Lyapunov exponent ($\times 10^3$) using the TISEAN package [15] for a time delay $\tau=60$ kb and an embedding dimension $d$. The skew profiles were filtered at the smoothing scale $a^*=160$ kb. $\theta$ and $t$ in Eq. (3) were rescaled so that the chaotic trajectory displays amplitude and characteristic frequencies similar to those of the skew oscillatory profiles. The error on the estimate of the Lyapunov exponent ($\times 10^3$) is of the order of 1

| | $d$ | | | | |
|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 |
| Chromosome 11 (24 Mb) (NT_033899.3) | 12.6 | 8.9 | 6.1 | 6.9 | 7.9 |
| Chromosome 14 (68 Mb) (NT_026437.9) | 15.0 | 10.2 | 8.8 | 8.7 | 10.4 |
| Chromosome 21 (29 Mb) (NT_011512.7) | 12.2 | 8.7 | 7.4 | 8.6 | 11.3 |
| Chromosome 22 (23 Mb) (NT_011520.8) | 12.5 | 8.1 | 6.3 | 5.8 | 7.2 |
| Shil'nikov strange attractor (30 Mb) | 4.2 | 5.6 | 6.5 | 7.3 | 7.1 |

representation provided by the WT and (iii) applying the concepts and techniques issued from dynamical systems theory, we have been able to reveal the existence of chaotic strange attractor behavior in human DNA sequences. The observed skew nonlinear relaxational oscillations are likely to reflect a deterministic organization of the human genome into adjacent domains with preferential gene orientation. Since the mean size $\sim$ 350–400 kb of these domains is consistent with the debated size of replicons and replication foci [17], in a work under progress, we have examined the possibility that the observed domains might be related to replication. Surprisingly, the $\beta$-globin and lamin $B_2$ replication origins are found to coincide (within a few kb) with domain frontiers. This suggests some correlation between gene organization into clusters of coexpressed genes and replication in the human genome. A possible understanding of this functional relationship can be found in the recent experimental investigations [18] of the structure and dynamics of chromatin that play an essential role in regulating many biological processes, including gene activity and DNA replication. The LRC observed in DNA sequences up to distances $\sim$ 20 kb reflect the existence of LRC between the local bending properties of the double helix as a necessary ingredient for the structure and dynamics of the nucleosomal array [3]. The fact that, in parallel, we observe a strange attractor type behavior (with the same characteristic frequencie $\ell_1 \simeq 400$ kb) in the strand compositional asymmetry and other codings of structural nature [10] suggests that this nonlinear oscillatory behavior might be a characteristic of the nonhomogeneous mechanical properties of the chromatin fiber that naturally condition the high-order loop structure of chromatin and its dynamics. We hope that the present work will serve as a guide for future numerical and experimental studies with the specific goal of unravelling the existence of a low-dimensional chaotic dynamical system that would simultaneously exert a control on the structure and dynamics of chromatin loops as well as on the coordination of the replication and transcription timings.

## Acknowledgements

The authors are very grateful to F. Argoul and B. Audit for invaluable help and stimulating discussions. This work was supported by the Action Bioinformatique Inter EPST 2002.

## References

[1] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, S.M. Ossadnik, C.-K. Peng, M. Simons, Fractals 1 (1993) 283.

[2] A. Arneodo, B. Audit, N. Decoster, J.-F. Muzy, C. Vaillant, in: A. Bunde, J. Kropp, H.J. Schellnhuber (Eds.), The Science of Disasters: Climate Disruptions, Heart Attacks and Market Crashed, Springer, Berlin, 2002, p. 26.

[3] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J.-F. Muzy, A. Arneodo, Phys. Rev. Lett. 86 (2001) 2471;
B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, J. Mol. Biol. 316 (2002) 903.

[4] A.C. Frank, J.R. Lobry, Gene 238 (1999) 65.

[5] M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, FEBS Lett. 555 (2003) 579.

[6] P. Green, B. Ewing, W. Miller, P.J. Thomas, E.D. Green, Nat. Genet. 33 (2003) 514.

[7] J. Majewski, Am. J. Hum. Genet. 73 (2003) 688.

[8] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, New York, 1998.

[9] A. Arneodo, Y. d'Aubenton-Carafa, E. Bacry, P.V. Graves, J.-F. Muzy, C. Thermes, Physica D 96 (1996) 291.

[10] S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. Lett. (2003), submitted for publication.

[11] P. Bergé, Y. Pomeau, C. Vidal, Order Within Chaos, Wiley, New York, 1986.

[12] B.L. Hao (Ed.), Chaos II, World Scientific, Singapore, 1990.

[13] P. Gaspard, A. Arneodo, R. Kapral, C. Sparrow (Eds.), Homoclinic Chaos, Physica D 62 (1993).

[14] A. Arneodo, P. Coullet, C. Tresser, Commun. Math. Phys. 79 (1981) 573;
C. Tresser, Ann. Inst. H. Poincaré 40 (1984) 441;
A. Arneodo, P.H. Coullet, E.A. Spiegel, Geophys. Astrophys. Fluid Dyn. 31 (1985) 1.

[15] R. Hegger, H. Kantz, T. Schreiber, Practical implementation of nonlinear time series methods: the TISEAN package. Online documentation system available at http://www.mpiks-dresden.mpg.de/∼tisean.

[16] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome, Nature 409 (2001) 860.

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

[18] T. Cremer, C. Cremer, Nat. Rev. Genet. 2 (2001) 292.