Bifractality of human DNA strand-asymmetry profiles results from transcription

S. Nicolay,[1,*] E. B. Brodie of Brodie,[1] M. Touchon,[2] B. Audit,[1] Y. d'Aubenton-Carafa,[2] C. Thermes,[2] and A. Arneodo[1]

[1]*Laboratoire Joliot-Curie and Laboratoire de Physique, UMR 5672, CNRS, ENS-Lyon,
46 Allée d'Italie, 69364 Lyon Cedex 07, France*

[2]*Centre de Génétique Moléculaire, UPR 2167 CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France*

(Received 8 July 2006; published 12 March 2007)

We use the wavelet transform modulus maxima method to investigate the multifractal properties of strand-asymmetry DNA walk profiles in the human genome. This study reveals the bifractal nature of these profiles, which involve two competing scale-invariant (up to repeat-masked distances $\lesssim 40$ kbp) components characterized by Hölder exponents $h_1 = 0.78$ and $h_2 = 1$, respectively. The former corresponds to the long-range-correlated homogeneous fluctuations previously observed in DNA walks generated with structural codings. The latter is associated with the presence of jumps in the original strand-asymmetry noisy signal S. We show that a majority of upward (downward) jumps colocate with gene transcription start (end) sites. Here 7228 human gene transcription start sites from the refGene database are found within 2 kbp from an upward jump of amplitude $\Delta S \geqslant 0.1$ which suggests that about 36% of annotated human genes present significant transcription-induced strand asymmetry and very likely high expression rate.

DOI: 10.1103/PhysRevE.75.032902 PACS number(s): 87.15.Cc, 05.40.−a, 05.45.Df, 87.10.+e

The existence, the nature, and the origin of long-range correlations (LRC's) in DNA sequences has been the subject of considerable recent interest [1]. Most of the investigations of DNA sequences were originally performed with different techniques, which all consisted in measuring a unique scaling exponent related to the roughness (Hurst) exponent H of the corresponding DNA walk profile [1,2]. But the measurement of a unique exponent fails to resolve multifractality [2] and provides limited information upon the nature of the LRC. In a previous work [3], we introduced the wavelet transform modulus maxima (WTMM) method, which allows computing the so-called multifractal spectra—e.g., the $D(h)$ singularity spectrum of Hölder exponent (h) values. A comparative statistical analysis of DNA walks generated from eukaryote and eubacterial sequences using structural (curvature) coding tables has shown that the corresponding DNA chain bending profiles are monofractal (homogeneous) and that there exist two scaling regimes: (i) in the 10–200-bp range, LRC's are observed for eukaryotic ($H \simeq 0.6$) [and not for eubacterial ($H = 0.5$)] sequences as the signature of the nucleosomal structure and (ii) over larger distances (200–20 000 bp), stronger LRC's ($H \simeq 0.8$) seem to exist in any sequence [4]. Recently, investigation of the thermodynamical properties of DNA chains [5] has revealed that the presence of these two LRC structural disorder regimes is likely to predispose DNA to form small loops favoring chromatin condensation and decondensation processes.

Here we generalize the application of the WTMM method to genome-wide multifractal sequence analysis when using alternative codings related to biological properties. According to the second parity rule [6], under no-strand-bias conditions, each genomic DNA strand should present equimolarities of A and T and of G and C [7]. Deviations from intrastrand equimolarities have been extensively studied during the past decade, and the observed skews have been attributed to asymmetries intrinsic to replication and transcription. Actually, during these processes, mutational events and repair mechanisms can affect the two strands differently, leading to transcription- and replication-associated asymmetries originally observed in bacteria [8,9]. It is only recently that (i) eukaryotic transcription-associated strand asymmetries have been established [10,11] and (ii) replication-associated strand asymmetries in mammals have been revealed by chromosome-wide multiscale analysis [12].

We investigate strand asymmetries along human chromosomes via the computation of TA and GC skews in nonoverlapping 1-kbp windows: $S_{TA} = (T-A)/(T+A)$ and $S_{GC} = (G-C)/(G+C)$. Because of the observed correlation between TA and GC skews [11], we will consider the total skew $S = S_{TA} + S_{GC}$ [Fig. 1(a)] and its corresponding DNA walk obtained by cumulating S values along the sequences: $\Sigma(n) = \Sigma_{j=1}^{n} S(j)$ [Fig. 1(b)]. Our goal is to show that the skew DNA walks of the 22 human autosomes display an unexpected (with respect to previous monofractal diagnosis [3]) bifractal scaling behavior [13] in the range 10–40 kbp as the signature of the presence of transcription-induced jumps in the LRC noisy S profiles. Sequences and gene annotation data ("refGene") were retrieved from the UCSC Genome Browser (May 2004). We used RepeatMasker to exclude repetitive elements that might have been inserted recently and would not reflect long-term evolutionary patterns.

The WT is a space-scale analysis which consists in expanding signals in terms of wavelets that are constructed from the analyzing wavelet ψ by means of dilations and translations [2,3]. The WT of a function Σ is defined as

$$T_{\psi}(x_0, a) = \frac{1}{a} \int_{-\infty}^{+\infty} \Sigma(x) \psi\left(\frac{x - x_0}{a}\right) dx, \qquad (1)$$

where x_0 and a (>0) are the space and scale parameters, respectively. The main advantage of using the WT for analyzing the regularity of a function Σ is its ability to eliminate order-n polynomial behavior (low-frequency components induced by genome compositional heterogeneity) by simply

[*] Permanent address: Institut de Mathématique, Université de Liège, Grande Traverse 12, 4000 Liège, Belgium

choosing a wavelet ψ whose $n+1$ first moments are zero $[\int x^m \psi(x)dx=0,\ 0 \leqslant m \leqslant n]$ [2,3]. In this work, we will use the second derivative of the Gaussian function: $\psi^{(2)}(x) = d^2(e^{-x^2/2}/\sqrt{2\pi})/dx^2$ which has two vanishing moments. The WTMM method [2,3] consists in investigating the scaling behavior of some partition functions defined in terms of wavelet coefficients:

$$Z(q,a) = \sum_{l \in \mathcal{L}(a)} \Big[\sup_{\substack{(x,a') \in l \\ a' \leqslant a}} |T_\psi(x,a')| \Big]^q \sim a^{\tau(q)}, \tag{2}$$

where $q \in \mathbb{R}$. The sum is taken over the WT skeleton [Fig. 1(c)] defined at each fixed scale a by the local maxima of $|T_\psi(x,a)|$; these WTMM are disposed on curves connected across scales called maxima lines; the set $\mathcal{L}(a)$ of all maxima lines that exist at scale a indicates how to position the wavelets in order to obtain a partition of the set of singularities of Σ at this scale. Indeed, the Legendre transform of $\tau(q)$ is the singularity spectrum $D(h) = \min_q[qh - \tau(q)]$ defined as the Haussdorf dimension of the set of points x where the Hölder exponent value is h [2,3]. Homogeneous fractal functions (i.e., functions with a unique Hölder exponent H) are characterized by a linear $\tau(q)$ curve $(\partial\tau/\partial q = h = H)$. Since $Z(q,a)$ amounts to computing the q-order moment of the WTMM probability density function (pdf) $\rho_{|T_\psi(.,a)|}$ at scale a, monofractal scaling implies that the shape of this pdf does not depend on the scale a, formally expressed by the self-similarity relationship [2,3]

$$\rho_{|T_\psi(.,a)|/a^H}(t) = \rho_{|T_\psi(.,1)|}(t). \tag{3}$$

On the contrary, a nonlinear $\tau(q)$ is the signature of nonhomogeneous functions displaying multifractal properties [$h(x)$ is a fluctuating quantity that depends upon x].

When computing $Z(q,a)$ [Eq. (2)] from the WT skeletons of the skew DNA walks Σ of the 22 human autosomes, we get convincing power-law behavior for $-1.5 \leqslant q \leqslant 3$ (data not shown). In Fig. 2(a) are reported the $\tau(q)$ exponents obtained using a linear regression fit of $\ln Z(q,a)$ vs $\ln a$ over the range of scales 10 kbp $\leqslant a \leqslant$ 40 kbp. All the data points remarkably fall on two straight lines $\tau_1(q) = 0.78q - 1$ and $\tau_2(q) = q - 1$ which strongly suggests the presence of two types of singularities $h_1 = 0.78$ and $h_2 = 1$, respectively, on two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ with the same Haussdorf dimension $D = -\tau_1(0) = -\tau_2(0) = 1$, as confirmed when computing the $D(h)$ singularity spectrum in Fig. 2(b). This observation means that $Z(q,a)$ can be split into two parts [14]:

$$Z(q,a) = C_1(q)a^{qh_1-1} + C_2(q)a^{qh_2-1}, \tag{4}$$

where $C_1(q)$ and $C_2(q)$ are prefactors that depend on q. Since $h_1 < h_2$, in the limit $a \mapsto 0^+$, the partition function is expected to behave like $Z(q,a) \sim C_1(q)a^{qh_1-1}$ for $q > 0$ and like $Z(q,a) \sim C_2(q)a^{qh_2-1}$ for $q < 0$, with a so-called phase transition [14,15] at the critical value $q_c = 0$. Surprisingly, it is the contribution of the weakest singularities $h_2 = 1$ that controls the scaling behavior of $Z(q,a)$ for $q > 0$ while the strongest ones $h_1 = 0.78$ actually dominate for $q < 0$ [Fig. 2(a)]. This inverted behavior originates from finite (1-kbp) resolution which prevents the observation of the predicted scaling behavior in the limit $a \mapsto 0^+$. The prefactors $C_1(q)$ and $C_2(q)$ in Eq. (4) are sensitive to (i) the number of maxima lines in the WT skeleton along which the WTMM behave as a^{h_1} or a^{h_2} and (ii) the relative amplitude of these WTMM. Over the range of scales used to estimate $\tau(q)$, the WTMM along the maxima lines pointing (at small scale) to $h_2 = 1$ singularities are significantly larger than those along the maxima lines associated to $h_1 = 0.78$ [see Figs. 2(c) and 2(d)]. This implies that the larger $q > 0$, the stronger the inequality $C_2(q) \gg C_1(q)$ and the more pronounced the relative contribution of the second term on the right-hand side of Eq. (4). On the opposite for $q < 0$, $C_1(q) \gg C_2(q)$ which explains that the strongest singularities $h_1 = 0.78$ now control the scaling behavior of $Z(q,a)$.

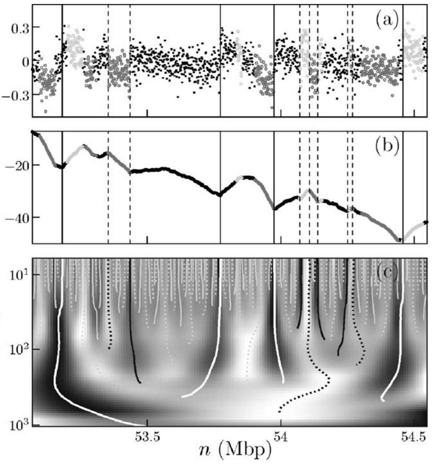

FIG. 1. (a) Skew profile $S(n)$ of a repeat-masked fragment of human chromosome 6; light gray (dark gray) 1-kbp window points correspond to sense (antisense) genes lying on the Watson (Crick) strand; black points to intergenic regions. (b) Cumulated skew profile $\Sigma(n)$. (c) WT of Σ; $T_{\psi^{(2)}}(n,a)$ is coded from white (min) to black (max); the WT skeleton defined by the maxima lines is shown in solid (dashed) lines corresponding to positive (negative) WT values. For illustration black solid (dashed) maxima lines are shown to point to the positions of 4 upward (3 downward) jumps in S [vertical dashed lines in (a) and (b)] that coincide with gene transcription starts (ends). Thick white lines correspond to maxima lines that persist above $a \geqslant 200$ kbp and that point to sharp upward jumps in S [vertical solid lines in (a) and (b)] that are likely to be the locations of putative replication origins [12]; note that three out of those four jumps are colocated with transcription start sites.

In Figs. 2(c) and 2(d) are shown the WTMM pdf's computed at scales $a = 10$, 20, and 40 kbp after rescaling by a^{h_1} and a^{h_2}, respectively. We note that there does not exist a value of H such that all the pdf's collapse on a single curve as expected from Eq. (3) for monofractal DNA walks. Consistently with the $\tau(q)$ data in Fig. 2(a) and with the inverted scaling behavior discussed above, when using the two exponents $h_1 = 0.78$ and $h_2 = 1$, one succeeds in superimposing respectively the central (bump) part [Fig. 2(c)] and the tail [Fig. 2(d)] of the rescaled WTMM pdf's. This corroborates the bifractal nature of the skew DNA walks that display two competing scale-invariant components of Hölder exponents: (i) $h_1 = 0.78$ corresponds to LRC homogeneous fluctuations

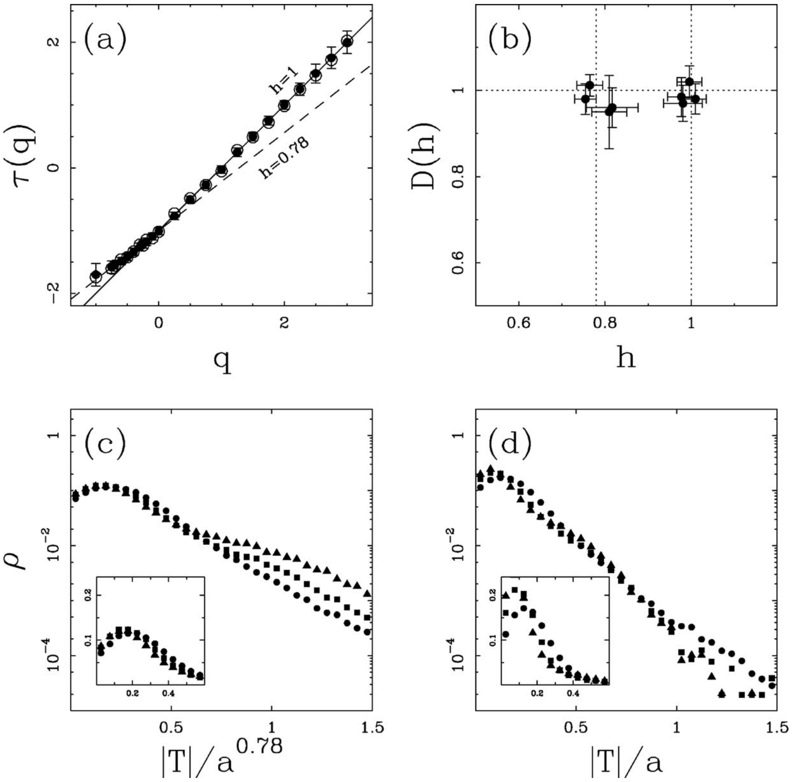

FIG. 2. Multifractal analysis of $\Sigma(n)$ of the 22 human (solid symbols) and 19 mouse (open circles) autosomes using the WTMM method with $\psi^{(2)}$ over the range 10 kbp$\leqslant a \leqslant$40 kbp. (a) $\tau(q)$ vs q. (b) $D(h)$ vs h. (c) WTMM pdf: ρ is plotted versus $|T|/a^H$ where $H=h_1=0.78$, in semilogarithmic representation; the inset is an enlargement of the pdf central part in linear representation. (d) Same as in (c) but with $H=h_2=1$. In (c) and (d), the symbols correspond to scales $a=10$ (●), 20 (■), and 40 (▲) kbp.

previously observed over the range 200 bp$\lesssim a \lesssim$20 kbp in DNA walks generated with structural codings [2,4] and (ii) $h_2=1$ is associated to convex ∨ and concave ∧ shapes in the DNA walks Σ indicating the presence of discontinuities in the derivative of Σ—i.e., of jumps in S [Figs. 1(a) and 1(b)]. At a given scale a, a large value of the WTMM in Fig. 1(c) corresponds to a strong derivative of the smoothed S profile and the maxima line to which it belongs is likely to point to a jump location in S.

In Fig. 3 are reported the results of a statistical analysis of the jump amplitudes in human S profiles. For maxima lines that extend above $a^*=10$ kbp, the histograms obtained for upward and downward variations are quite similar, especially their tails that are likely to correspond to jumps in the S profiles [Fig. 3(a)]. When computing the distance between upward or downward jumps ($|\Delta S|\geqslant 0.1$) to the closest transcription start (TSS) or end (TES) sites [Fig. 3(b)], we reveal that the number of upward jumps in close proximity ($|\Delta n| \lesssim 3$ kpb) to the TSS overexceeds the number of such jumps close to the TES. Similarly, downward jumps are preferentially located at the TES. These observations are consistent with the steplike shape of skew profiles induced by transcription: $S>0$ ($S<0$) is constant along a sense (antisense) gene and $S=0$ in the intergenic regions [11]. Since a steplike pattern is edged by one upward and one downward jump, the set of human genes that are significantly biased is expected to contribute to an even number of $\Delta S>0$ and $\Delta S<0$ jumps when exploring the range of scales $10\lesssim a \lesssim 40$ kbp, typical of human gene size. Note that in Fig. 3(a), the number of sharp upward jumps actually slightly exceeds the number of sharp downward jumps, consistently with the experimental observation that whereas the TSS's are well defined, the TES's may extend over 5 kbp, resulting in smoother downward skew transitions [11]. This TES particularity also explains the excess of upward jumps found close to TSS's as compared to the number of downward jumps close to TES's [Fig. 3(b)].

In Fig. 4(a), we report the analysis of the distance of the TSS to the closest upward jump. For a given upward jump amplitude, the number of TSS's with a jump within $|\Delta n|$ increases faster than expected (as compared to the number found for randomized jump positions) up to $|\Delta n|\simeq 2$ kbp. This indicates that the probabilty to find an upward jump

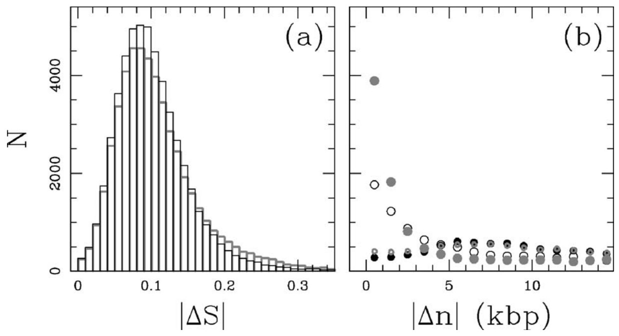

FIG. 3. Statistical analysis of skew variations at the singularity positions determined at scale 1 kbp from the maxima lines that exist at scales $a\geqslant 10$ kbp in the WT skeletons of the 22 human autosomes. For each singularity, we computed the variation amplitudes $\Delta S=\bar{S}(3')-\bar{S}(5')$ over two adjacent 5-kbp windows, respectively in the $3'$ and $5'$ directions and the distances Δn to the closest TSS (TES). (a) Histograms $N(|\Delta S|)$ for upward ($\Delta S>0$, gray) and downward ($\Delta S<0$, black) skew variations. (b) Histograms of the distances Δn of upward (gray) or downward (black) jumps with $|\Delta S|\geqslant 0.1$ to the closest TSS (filled black bullet, filled gray bullet) and TES (empty black circle, empty gray circle).

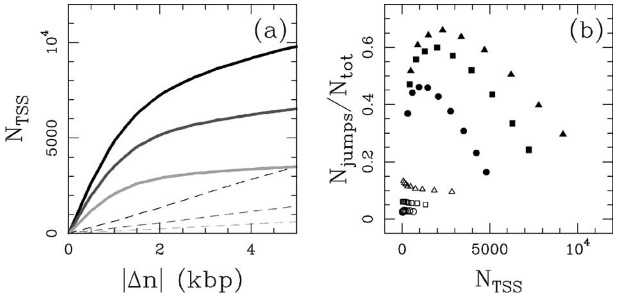

FIG. 4. (a) Number of TSS's with an upward jump within $|\Delta n|$ (abscissa) for jump amplitudes $\Delta S > 0.1$ (black), 0.15 (dark gray), and 0.2 (light gray). Solid lines correspond to true jump positions while dashed lines to the same analysis when jump positions were randomly drawn along each chromosome. (b) Among the $N_{tot}(\Delta S^*)$ upward jumps of amplitude larger than some threshold ΔS^*, we plot the proportion of those that are found within 1 kbp (●), 2 kbp (■), or 4 kbp (▲) of the closest TSS vs the number N_{TSS} of the so-delineated TSS's. Curves were obtained by varying ΔS^* from 0.1 to 0.3 (from right to left). Open symbols correspond to similar analyses performed on random upward jump and TSS positions.

within a gene promoter region is significantly larger than elsewhere. For example, out of 20 023 TSS's, 36% (7228) are delineated within 2 kbp by a jump with $\Delta S > 0.1$. This provides a very reasonable estimate for the number of genes expressed in germline cells as compared to the 31.9% recently experimentally found to be bound to Pol II in human embryonic stem cells [16].

Combining the previous results presented in Figs. 3(b) and 4(a), we report in Fig. 4(b) an estimate of the efficiency/coverage relationship by plotting the proportion of upward jumps ($\Delta S > \Delta S^*$) lying in TSS proximity as a function of the number of so-delineated TSS's. For a given proximity threshold $|\Delta n|$, increasing ΔS^* results in a decrease of the number of delineated TSS's, characteristic of the right tail of gene bias pdf. Concomitant to this decrease, we observe an increase of the efficiency up to a maximal value corresponding to some optimal value for ΔS^*. For $|\Delta n| < 2$ kbp, we reach a maximal efficiency of 60% for $\Delta S^* = 0.225$; 1403 out of 2342 upward jumps delineate a TSS. Given the fact that the actual number of human genes is estimated to be significantly larger (~30 000) than the number provided by refGene, a large part of the 40% (939) of upward jumps that have not been associated to a refGene could be explained by this limited coverage. In other words, jumps with sufficiently high amplitude are very good candidates for the location of highly biased gene promoters. Let us point that out of the above 1403 (2342) upward jumps, 496 (624) jumps are still observed at scale $a^* = 200$ kbp. According to Ref. [12], these jumps are likely to also correspond to replication origins underlying the fact that large upward jumps actually result from the cooperative contributions of both transcription- and replication-associated biases. The observation that 80% (496/624) of the predicted replication origins are colocated with TSS's enlightens the existence of a remarkable gene organisation at replication origins [12].

To summarize, we have demonstrated the bifractal character of skew DNA walks in the human genome. When using the WT microscope to explore (repeat-masked) scales ranging from 10 to 40 kbp, we have identified two competing homogeneous scale-invariant components characterized by Hölder exponents $h_1 = 0.78$ and $h_2 = 1$, which, respectively, correspond to LRC-colored noise and sharp jumps in the original DNA compositional asymmetry profiles. Remarkably, the so-identified upward (downward) jumps are mainly found at the TSS (TES) of human genes with high transcription bias and thus very likely higly expressed. This study also underlines that most replication origins are important organizing centers for transcription mechanisms. As illustrated in Fig. 2(a), similar bifractal properties are also observed when investigating the 19 mouse autosomes. This suggests that the results reported in this work are general features of mammalian genomes.

This work was supported by the ACI IMPBio 2004, the Agence Nationale de la Recherche (NT05-3_41825), the Conseil Régional Rhône-Alpes and the PAI Tournesol.

[1] H. E. Stanley *et al.*, Fractals **1**, 283 (1993); S. V. Buldyrev *et al.*, Phys. Rev. E **51**, 5084 (1995); W. Li, Comput. Chem. (Oxford) **21**, 257 (1997).

[2] A. Arneodo *et al.*, in *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes*, edited by A. Bunde, J. Kropp, and H. Schellnhuber (Springer-Verlag, Berlin, 2002), p. 26.

[3] A. Arneodo *et al.*, Phys. Rev. Lett. **74**, 3293 (1995); A. Arneodo *et al.*, Physica D **96**, 291 (1996).

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

[5] C. Vaillant *et al.*, Phys. Rev. Lett. **95**, 068101 (2005); C. Vaillant *et al.*, Eur. Phys. J. E **19**, 263 (2006).

[6] E. Chargaff, Fed. Proc. **10**, 654 (1951); R. Rudner *et al.*, Proc. Natl. Acad. Sci. U.S.A. **60**, 921 (1968).

[7] J. W. Fickett *et al.*, Genomics **13**, 1056 (1992); J. R. Lobry, J. Mol. Evol. **40**, 326 (1995).

[8] J. M. Freeman *et al.*, Science **279**, 1827 (1998); A. Beletskii *et al.*, J. Mol. Biol. **300**, 1057 (2000); M. P. Francino and H. Ochman, J. Mol. Evol. **18**, 1147 (2001).

[9] J. Mrázek and S. Karlin, Proc. Natl. Acad. Sci. U.S.A. **95**, 3720 (1998); A. C. Frank and J. R. Lobry, Gene **238**, 65 (1999); E. P. Rocha *et al.*, Mol. Microbiol. **32**, 11 (1999); E. R. M. Tillier and R. A. Collins, J. Mol. Evol. **50**, 249 (2000).

[10] P. Green *et al.*, Nat. Genet. **33**, 514 (2003).

[11] M. Touchon *et al.*, FEBS Lett. **555**, 579 (2003); M. Touchon *et al.*, Nucleic Acids Res. **32**, 4969 (2004).

[12] M. Touchon *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 9836 (2005); E.-B. Brodie of Brodie *et al.*, Phys. Rev. Lett. **94**, 248103 (2005).

[13] U. Frisch, *Turbulence* (Cambridge University Press, Cambridge, 1995); see Sec. 8.5.2.

[14] J.-F. Muzy *et al.*, Int. J. Bifurcation Chaos Appl. Sci. Eng. **4**, 245 (1994); A. Arneodo *et al.*, Physica A **213**, 232 (1995).

[15] T. Bohr and T. Tel, in *Direction in Chaos*, edited by B. L. Hao (World Scientific, Singapore, 1988), Vol. 2, p. 194.

[16] T. I. Lee *et al.*, Cell **125**, 301 (2006).