Poincaré recurrences of DNA sequences

K. M. Frahm and D. L. Shepelyansky
Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France
(Received 2 September 2011; revised manuscript received 1 December 2011; published 27 January 2012)

We analyze the statistical properties of Poincaré recurrences of *Homo sapiens*, mammalian, and other DNA sequences taken from the Ensembl Genome data base with up to 15 billion base pairs. We show that the probability of Poincaré recurrences decays in an algebraic way with the Poincaré exponent $\beta \approx 4$ even if the oscillatory dependence is well pronounced. The correlations between recurrences decay with an exponent $\nu \approx 0.6$ that leads to an anomalous superdiffusive walk. However, for *Homo sapiens* sequences, with the largest available statistics, the diffusion coefficient converges to a finite value on distances larger than one million base pairs. We argue that the approach based on Poncaré recurrences determines new proximity features between different species and sheds a new light on their evolution history.

DOI: 10.1103/PhysRevE.85.016214 PACS number(s): 05.45.Tp, 87.14.gk, 05.40.Fb, 87.10.Vg

The Poincaré recurrence theorem of 1890 [1] states that, after a certain time, a dynamical Hamiltonian trajectory in a bounded phase space always returns to the close vicinity of an initial state. Even if recurrences definitely take place, the question about their properties, or more exactly, the question of what are the statistics of Poincaré recurrences and what are their correlation properties, still remain an unsolved problem for systems of dynamical chaos even after an impressive development of the theory of dynamical complexity [2–4]. The two limiting cases of periodic and fully chaotic motion are well understood: In the first case the recurrences are periodic while in the latter case the probability of recurrences $P(t)$ with time being larger than t drops exponentially at $t \to \infty$ [2–4]. Thus the latter case is similar to a coin flipping, where the probability to stay on the same side after more than t flips decays at 2^{-t}. However, in generic Hamiltonian systems the probability $P(t)$ decays algebraically with t, as $P(t) \sim 1/t^{\beta}$ due to long trappings in the vicinity of stability islands showing the Poincaré exponent $\beta \approx 1.5$ [5–10]. A detailed theoretical explication of this slow algebraic decay is still lacking. Usually the consecutive recurrences in dynamical systems are not correlated since a trajectory passes across the domains of a chaotic component.

The Poincaré recurrences represent a powerful tool for the analysis of statistical properties of symbolic trajectories of various types [2–4]. Surprisingly, this powerful tool of dynamical systems has not been applied for detailed statistical studies of the DNA sequence, which also can be viewed as a symbolic trajectory. There have been only a few earlier attempts going in this direction including researchers in dynamical systems [11] and bioinformatics [12–14]. However, in [11] only short recurrence times with $t \leqslant 4$ have been considered and it was concluded that the probability of recurrences decays exponentially. The studies in bioinformatics were not aware of the concept of Poincaré recurrences, but their approach had certain links with them aiming to use digital signal representations of genomic data [12]. The relative frequency analysis applied in [13,14] has certain similarities with the Poincaré recurrences approach, but the distance times still remain very short with $t \leqslant 20$ in [13] and $t \leqslant 100$ in [14]. No detailed comparative analysis with the exponential decay of Poincaré recurrences of random sequences or algebraic decay was presented there.

In this work, we apply the powerful approach of Poincaré recurrences to the available mammalian DNA sequences taken from the publicly available database [15]. The comparison with random data sequences and the known results for dynamical maps [5–10] allowed us to establish new, interesting features for the Poincaré recurrences of the DNA sequence. Our approach allowed to analyze the recurrences with time t being by five to six orders of magnitude larger than those reached in [11–14]. For the *Homo sapiens* (HS) database we performed statistical analysis for 1.5×10^{10} base pairs (bp). This amount of statistical data is four to five orders of magnitude larger compared to the previous studies of anomalous diffusion performed in [16–18] for DNA sequences. Using this large statistics we find that the DNA Poincaré recurrences are characterized by an algebraic decay with $\beta \approx 4$ for the HS database. For such a value of the Poincaré exponent β, the uncorrelated recurrences should lead to a usual diffusive random walk with a linear growth of the corresponding second moment $\sigma \sim Dt$ [6,7], with an effective time t given by the sequence length L measured in number of bp. At the same time the early studies for random walk in DNA sequences [16–18], with the total length of $t < 10^6$ bp, established that such a walk belongs to the Levy-type walks, with an anomalous superdiffusive growth of the second moment $\sigma \sim t^{1+\mu}$ and a growing diffusion coefficient $D(t) = \sigma/t \sim t^{\mu}$ with $\mu > 0$. Our studies show that this apparent contradiction is resolved by the presence of long-range correlations $C_P(t)$ between the Poincaré recurrences in DNA that make them different compared to dynamical chaos systems where such correlations are usually absent [5–10]. We show that $C_P(t)$ is characterized by a global algebraic decay with an exponent $\nu \approx 0.6$. Such a slow decay leads to an anomalous superdiffusion on scales of $t < 10^6$ bp with the exponent μ being in agreement with the previous studies [16–18]. However, for $t > 10^6$ bp the diffusion coefficient $D(t)$ for HS becomes finite due to cancellations of odd and even correlation terms which show a global algebraic decay with an exponent $\nu \approx 0.6$. We argue that the obtained results for the statistics of Poincaré recurrences of the DNA sequence open new possibilities for the genome evolution analysis.

To study the statistics of the Poincaré recurrence of mammalian DNA sequences we use the enormous database [15] considering a DNA sequence as a very long trajectory in

the space of four nucleobases A, G, C, T. Similar to [16], a walk along the DNA sequence length, marked as an effective time t, is described by a discrete variable $u(t)$ which takes values "+" for A, G of purine domain and "−" for C, T of pyrimidine domain (AG-CT). The differential distribution of Poincaré recurrences $p_1(t)$ is given by a relative number of segments of a fixed sign of length t while the integrated distribution $P(t)$ gives the relative number of recurrences with times larger than t. The probabilities of domains AG and CT are close to 0.5 for the HS and mammalian sequences. Thus the recurrences for both domains are very close to each other so that we show one average distribution $P(t)$ for AG-CT corresponding to the recurrences or crossings of the line $u = 0$. A similar situation takes place for AC and GT domains so that we show for them one average distribution $P(t)$ for AC-GT. For domains AT and CG the probabilities are approximately 0.6 and 0.4 and here we show separately the recurrence probability $P(t)$ for AT and CG domains. For Poincaré recurrences $P(t)$ of HS sequences these four cases are shown in Fig. 1 (left panel). On average we find an algebraic decay $P(t) \sim 1/t^{\beta}$ with $\beta \approx 4$. A formal fit for AG-CT data at $t > 10$ gives $\beta = 3.68 \pm 0.02$, but there are visible large-scale oscillations with a certain similarity to those seeing in dynamical maps [5,7,9]. The dependence $P(t) = 2^{-t}$ for a random sequence describes AG-CT and AC-GT data only on short times $t < 5$ while for larger times algebraic behavior becomes dominant.

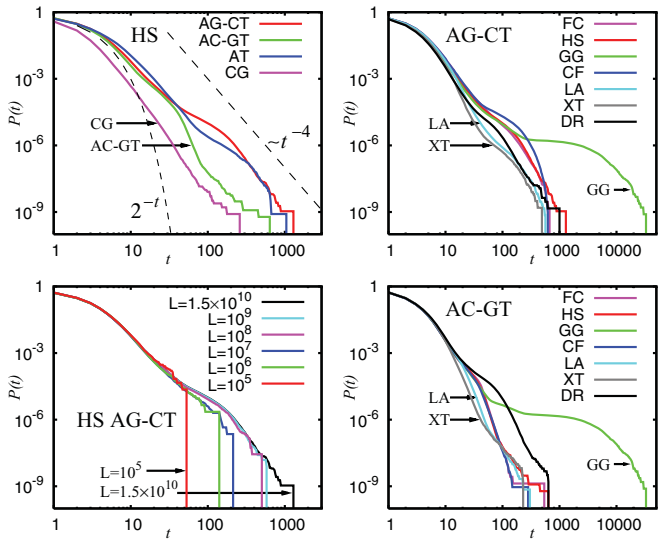

FIG. 1. (Color online) Statistics of Poincaré recurrences $P(t)$ for DNA sequences. Top left panel: DNA data of HS for Poincaré recurrences of domains AG-CT, AC-GT, AT, and CG (see text). The lower dashed curve shows the exponential behavior $P(t) = 2^{-t}$ valid for *random* sequences, the upper dashed line shows the average power law $P(t) \sim t^{-4}$ for comparison. Top right panel: AG-CT data for DNA sequences of the species: *Felis catus* (FC, Cat), *Homo sapiens* (HS, Human), Gorilla gorilla (GG, Gorilla), *Canis familiaris* (CF, Dog), *Loxodonta africana* (LA, Elephant), *Xenopus tropicalis* (XT, African Clawed Frogs) and *Danio rerio* (DR, Zebrafish). Bottom left panel: Convergence of the statistics of AG-CT Poincaré recurrences $P(t)$ for *Homo sapiens* as the length L of the considered DNA sequence increases from $L = 10^5$ to $L = 1.5 \times 10^{10}$. Bottom right panel: AC-GT data sets for the same species as in the top right panel.

We note that $P(t)$ is a positively defined quantity and thus it is statistically very stable: the sequences of size L well reproduce the initial part of $P(t)$ almost up to values $\sim 1/L$ as it is shown in Fig. 1 (bottom left panel), where L varies in a large interval of $10^5 \leqslant L \leqslant 1.5 \times 10^{10}$ bp. Thus our maximal size $L \approx 1.5 \times 10^{10}$ bp corresponds to five times the whole human genome size 3×10^9 bp that allows us to obtain more reliable statistical results for Poincaré recurrences. We note that we consider the recurrences' distances t not larger than 10^7 bp so that such distances remain significantly smaller than the size of one human genome and the size of the largest chromosome 2×10^8 bp.

It should be noted that there are other statistical studies of the DNA sequence which analyze the nearest-neighbor spacing distribution of a basis [19,20], being linked to level statistics of words [21], and the further analysis of $1/f$ like noise exponents as in [22]. We hope that the statistics of Poincaré recurrences will give important complementary tools for a deeper understanding of DNA sequence properties.

The comparison of statistics of Poincaré recurrences for HS, mammalian, and two other species are shown in Fig. 1 for the AG-CT case (a similar average behavior is found for AC-GT data). The total sequence lengths L for other species are by a factor of 3 shorter compared to the HS case. Up to $t \approx 20$ of all considered species show the same decay of $P(t)$, but at a larger value of t there is a separation of curves so that each species is characterized by its own statistics $P(t)$. On average, all species show an algebraic decay with $\beta \approx 4$ even if there is a strong oscillation with a flat region of $P(t)$ for GG sequence (the AC-GT data from Fig. 1 show a very similar behavior in this case). It is interesting to note that the curves of Poincaré recurrences are very close for HS and GG sequences up to $t \approx 200$ and for HS and FC sequences up to maximal $t \approx 10^3$. However, for the AC-GT data set the curves for these sequences become different for $t > 20$ (Fig. 1).

It is important to understand how the statistics of Poincaré recurrences is related to the anomalous superdiffusive walk discussed in [16–18]. The walk is described by a displacement variable $y(t) = \sum_{\tau=1}^{t} u(\tau)$ whose growth can be characterized by a diffusion coefficient defined as $D(t) = \sigma/t$ with the second moment $\sigma = \langle \Delta y(t)^2 \rangle$, $\Delta y(t) = y(t + t_0) - y(t_0) - \langle y(t + t_0) - y(t_0) \rangle$ and the average $\langle \cdots \rangle$ is done with respect to the initial position (or "time") t_0. In the case of a standard diffusive process the diffusion coefficient D converges to a finite value at large times. However, the results of [16] gave an algebraic superdiffusive growth $D(t) \sim t^{\mu}$ with the exponent $\mu \approx 0.34$ for the HS sequence of length $L \sim 10^5$ and $t \leqslant 10^3$. Our results are obtained on a significantly larger scale of t being four orders of magnitude larger compared to those reached in [16–18]. Our results for diffusion $D(t)$ are shown in Fig. 2. For the HS sequence we have large statistics and large exact segments without nondetermined bp marked as N in the database [15]. We find $\mu \approx 0.4$ for the range $10 < t < 10^6$ (fit gives $\mu = 0.349 \pm 0.001$) in a satisfactory agreement with previous studies [16–18]. Other species also show an algebraic growth of $D(t)$ with similar values of μ (Fig. 2). For the AC-GT data we also find a similar behavior with $\mu \approx 0.6$ for the HS sequence (Fig. 2). However, for the HS sequence with the most exact and long data set we find a saturation of $D(t)$ for large times $10^6 \leqslant t \leqslant 10^7$.

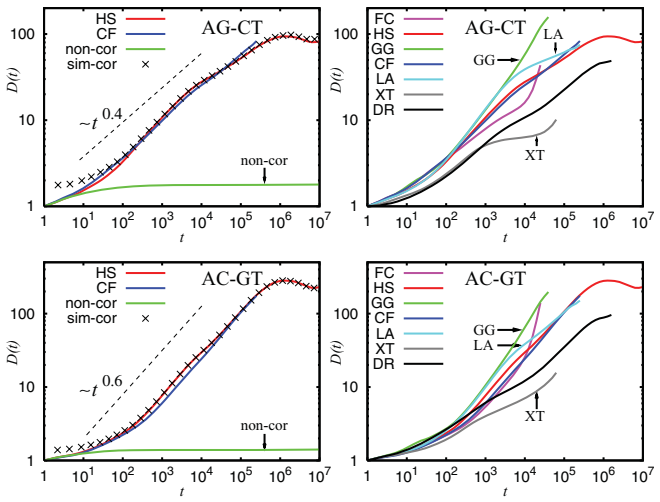

FIG. 2. (Color online) Top left panel: Diffusion coefficient $D(t) = \langle \Delta y^2(t) \rangle / t$ for AG-CT data sets of DNA sequences of HS and CF. The lower green curve (non-cor) is the diffusion coefficient obtained for a model with individual recurrences being distributed as the Poincaré recurrences of HS in Fig. 1, but *assuming* that subsequent Poincaré recurrences are *not correlated*. The black crosses (sim-cor) represent the diffusion coefficient obtained from Eq. (3) using the Poincaré recurrence correlation function $C_P(n)$ for HS (see text and Fig. 4 below). The dashed line shows a power law $D \sim t^{0.4}$. Top right panel: Diffusion coefficient $D(t)$ for AG-CT data sets of the same species as in the right panel of Fig. 1. Bottom panels: Diffusion coefficient $D(t)$ for AC-GT data sets for the same cases as in top panels; the dashed line in the left panel represents a power-law dependence $D(t) \sim t^{0.6}$.

The diffusion coefficient is related to the correlation function $c(t) = \langle u(t+t_0)u(t_0) \rangle$ as $D(t) = (1/t) \sum_{l=1}^{t} \sum_{j=-l+1}^{l-1} c(j)$ and hence a divergence of D implies a slow correlation decay $c(t) \sim t^{\mu-1}$ if $c(t)$ is monotonic. On the other hand, this correlation function can also be expressed as

$$c(t) = \sum_{n=1}^{\infty} (-1)^{n-1} \sum_{t_1+\ldots+t_n>t}^{\infty} (t_1 + \cdots + t_n - t)\, p_n(t_1, \ldots, t_n), \tag{1}$$

where $p_n(t_1, \ldots, t_n)$ is the joint distribution of n subsequent Poincaré recurrence times $t_1, \ldots, t_n$. In this sum each term represents the case where n subsequent recurrences are needed to cover the interval $0, 1, \ldots, t$ and the prefactor $t_1 + \cdots + t_n - t$ accounts for the number of different initial positions of the first recurrence to allow this. If we *assume* that subsequent Poincaré recurrences are *not correlated* [i.e., $p_n(t_1, \ldots, t_n) = p_1(t_1) \cdot \ldots \cdot p_1(t_n)$], and that $P(t_1)$ obeys the power law $P(t_1) \sim t_1^{-\beta}$ [i.e., $p_1(t_1) = P(t_1) - P(t_1+1) \sim t_1^{-\beta-1}$ we find that in the above expression the first term for $n = 1$ dominates the limit $t \to \infty$ and we find that $c(t) \approx \sum_{t_1=t+1}^{\infty} P(t_1) \sim t P(t) \sim t^{1-\beta}$]. We mention that this result was previously also obtained for chaotic Hamiltonian dynamics [6,7]. Therefore we should have a good convergence of D with $\beta \approx 4$. However, this relation is obtained for the case of *uncorrelated* Poincaré recurrences that may not be the case for DNA sequences. Indeed, if we generate uncorrelated recurrences with the distribution $P(t)$ being the same as in Fig. 1 for the AG-CT sequence of HS and compute with them the diffusion coefficient then we find a clear saturation of $D(t)$ at a finite value $D = 1.77$ (the green curve in Fig. 2, left panel), being significantly smaller then the actual data of $D(t) \sim 100$.

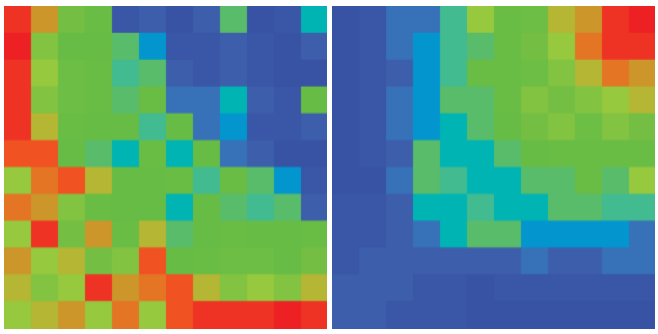

FIG. 3. (Color online) Left panel: Density plot of the normalized two point correlator $\tilde{p}_2(t_1,t_2)$ of two subsequent Poincaré recurrences t_1 and t_2 for AG-CT data sets of HS. The shown range $1 \leqslant t_1, t_2 \leqslant 12$ represents 99.4% of probability. Red (dark gray), green (light gray), and blue (black) colors represent maximal, zero, and minimal correlator values, respectively; horizontal and vertical axes show t_1 and t_2; maximal correlator values are located in the right half of the bottom line and the top half of the left column. Right panel: Normalized two-point correlator $\tilde{p}_2(t_1,t_3)$ of t_1 and t_3 for three subsequent Poincaré recurrences t_1, t_2, t_3 with t_1 and t_3 on the axes; maximal correlator values are located in the top right corner.

To visualize the correlations between Poincaré recurrences we also compute the joint probability $p_2(t_1,t_2)$ of two subsequent Poincaré recurrences t_1 and t_2 for the HS sequence of Fig. 1. The normalized two-point correlator is $\tilde{p}_2(t_1,t_2) = p_2(t_1,t_2)/[p_1(t_1)\, p_1(t_2)] - 1$, where $p_1(t_1) = P(t_1) - P(t_1+1)$ is the probability of one individual recurrence of length t_1. Its dependence on t_1, t_2 is shown in Fig. 3. The correlator is maximal for $t_1 = 1$ (i.e., below the average recurrence time $\langle t_1 \rangle = 2.27$) and $t_2 \geqslant 8$ (i.e., above average) or vice versa thus indicating anticorrelations between t_1 and t_2. In the right panel of Fig. 3 we show the normalized two-point correlator $\tilde{p}_2(t_1,t_3)$ for t_1 and t_3 taken from three subsequent Poincaré recurrence times t_1, t_2, t_3. In this case t_1 and t_3 are correlated (i.e., if t_1 is above average it is more likely that t_3 is also above average).

Thus in a sequence of Poincaré recurrences $t_1, t_2, t_3, \ldots$, the odd elements represent steps of length $t_1, t_3, \ldots$, of one sign of $u(t)$ and the even elements represent steps of length $t_2, t_4, \ldots$, of the other sign. The anticorrelations between t_1 and t_2 or t_2 and t_3 as well as the correlations between t_1 and t_3 indicate that once a preferential direction is chosen it is more likely for it to be enhanced thus explaining the diffusion enhancement compared to the uncorrelated Poincaré recurrences which give a finite coefficient $D \approx 1.7$. To work out this point on a more quantitative level we consider the displacement after n Poincaré recurrences at time $t = t_1 + \cdots + t_n \approx n \langle t_1 \rangle$. We can write for it

$$y(t_1 + \cdots + t_n) = (-1)^s \sum_{l=1}^{n} (-1)^{l-1} t_l, \tag{2}$$

where $(-1)^s$ is the sign of the first segment associated to t_1. For $n \gg 1$ this leads to

$$D(n\langle t_1\rangle) = \frac{1}{n\langle t_1\rangle}\sum_{l=1}^{n}\left(C_P(0) + 2\sum_{j=1}^{l-1}(-1)^j C_P(j)\right), \qquad (3)$$

where $C_P(j) = \langle t_1 t_{1+j}\rangle - \langle t_1\rangle^2$ is the Poincaré recurrence correlation function and the average is done over all recurrences [23]. We note that the above model of uncorrelated Poincaré recurrences corresponds to $C_P(j) = 0$ for $j > 0$. In this case Eq. (3) gives $D = C_P(0)/\langle t_1\rangle = 4.01/2.27 = 1.77$ in a perfect agreement with the data of Fig. 2.

The Poincaré recurrence correlation function $C_P(n)$ is computed from DNA sequence data and its dependence on the recurrence index or number $n \approx t/\langle t_1\rangle$ is shown in Fig. 4 for AG-CT data sets of HS and CF. For HS data this correlation function has alternate signs for odd and even n up to $n \approx 3 \times 10^3$. For larger n values these terms have the same sign and moreover these terms become approximately equal for $n > 10^5$. This leads to the cancellation of the odd and even terms in Eq. (3) and the saturation of the growth of diffusion coefficient at $t > 10^6$ as it is clearly seen in Fig. 2. Such a saturation of $D(t)$ takes place in spite of a rather slow algebraic decay of correlation $C_P(n) \sim n^{-\nu}$ with $\nu \approx 0.6$ (for even terms an error-weighted fit gives $\nu = 0.575 \pm 0.003$ at $10 \leqslant n \leqslant 3 \times 10^6$ and for odd terms $\nu = 0.479 \pm 0.005$ at $10 \leqslant n \leqslant 10^3$). From the found correlation function $C_P(n)$ we can determine the dependence $D(t)$ using Eq. (3) that gives a good agreement with the data obtained by a direct computation of $D(t)$ as it is shown in Fig. 2 (deviations at $t < 10$ are due to an approximate validity of the relation $t = t_1 + \cdots + t_n \approx n\langle t_1\rangle$ at small t). We note that the relation between exponents $\mu = 1 - \nu$, corresponding to a simple estimate $D \sim t\,|C_P(t)|$, remains valid in absence of odd or even terms cancellation at $t < 10^6$. For the CF data set we find approximately the same algebraic decay with $\nu \approx 0.6$ (Fig. 4, right panel). In this case the total number of recurrences N_r is statistically smaller compared to the HS case and in addition, undetermined letters N of bp are broadly scattered over the sequence. Due to that, here we do not find large number $N_r(n)$ of recurrence times at large n that force us to stop at $n < 2.5 \times 10^5$ where a saturation of $D(t)$ growth is not visible [for the HS case we have $N_r(n) \approx 5 \times 10^9$ recurrences at $n = 10^6$ but many of them are correlated and the statistical error of $C_P(n)$ is about 5% here while for smaller n it becomes smaller than the symbol size in Fig. 4]. For the AC-GT data sets, shown in Fig. 4, we find an algebraic decay with exponent $\nu \approx 0.4$ corresponding to the value $\mu \approx 0.6$ from corresponding Fig. 2. The convergence of the odd or even terms of $C_F(n)$ for the HS case takes place at $n > 10^5$ leading to saturation of the diffusion rate at $t > 10^6$ also visible for AC-GT data (Fig. 2). For CF data we have lower statistics for large n and t and the saturation of $D(t)$ remains invisible.

The analysis of the statistical accuracy of the computation of correlation function $C_P(n)$ is presented in Fig. 5. Here we show the variation of relative statistical error $\Delta C_P(n)/|C_P(n)|$ in the value of $C_P(n)$ as a function of n. This error increases from a level of 10^{-3} at $n < 100$ up to 0.1 at $n \approx 2 \times 10^6$

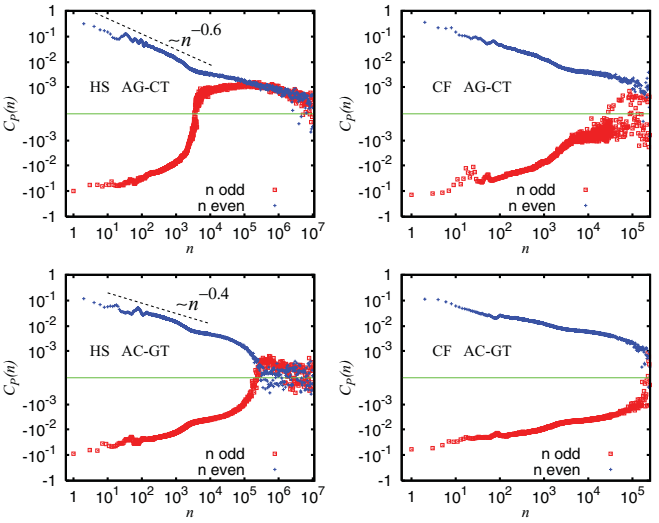

FIG. 4. (Color online) Top panels: Poincaré recurrence correlation function $C_P(n) = \langle t_1 t_{n+1}\rangle - \langle t_1\rangle^2$ of t_1 and t_{n+1} in a sequence of subsequent Poincaré recurrences t_j for HS (left panel) and CF (right panel) sequences for AG-CT data sets. Blue (black) crosses correspond to even n and red (gray) squares to odd n. The dashed line shows a power law $C_P(n) \sim n^{-0.6}$. For clarity, positive and negative values of $C_P(n)$ are shown on two separate logarithmic scales which are put together at $C_P = \pm 10^{-4}$ shown by the straight horizontal line. Bottom panels: Same as in top panels, but for AC-GT data sets for HS (left panel) and CF (right panel); the dashed line shows the dependence $C_P(n) \sim n^{-0.4}$.

for HS and at $n \approx 10^4$ for CF [a strong increase of error at $n \approx 3 \times 10^3$ for HS is related to a sign change of $C(n)$]. The relative error increases with n since at large n we have a smaller number of recurrences N_r contributing in the computation of $C(n)$. For the HS case the number of nondetermined N letters allows to have a significantly larger number of recurrences N_r compared to the CF case, and due to this we obtain statistically good values of $C(n)$ at significantly larger values of n.

Let us give now the formal fit parameter values for the dependencies discussed above. The fit of Poincaré recurrences for the data of Fig. 1 at $t > 10$ gives the Poincaré exponent $\beta = 3.68 \pm 0.02$ (AG-CT), 3.65 ± 0.04 (AC-GT), 3.75 ± 0.03 (AT), and 4.04 ± 0.05 (CG). The fit of $D(t) \sim t^{\mu}$ for the AG-CT data of HS in Fig. 1 gives $\mu = 0.3486 \pm 0.0008$ for the

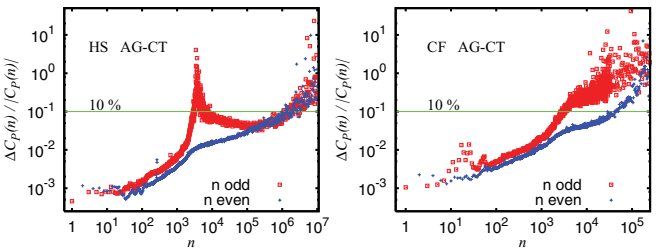

FIG. 5. (Color online) Relative statistical error $\Delta C_P(n)/|C_P(n)|$ of the Poincaré recurrence correlation function for HS (left panel) and CF (right panel) sequences for AG-CT data sets shown in top panels of Fig. 4. Blue (black) crosses correspond to even n and red (gray) squares to odd n. The straight horizontal line indicates the value of 10%.

range $10 \leqslant t \leqslant 10^6$, but there are two intervals with distinct values $\mu = 0.5010 \pm 0.0003$ for $10 \leqslant t \leqslant 3 \times 10^3$ and $\mu = 0.2859 \pm 0.0003$ for $3 \times 10^3 \leqslant t \leqslant 10^6$ so that we give in the text the average $\mu \approx 0.4$. For the AC-GT data of HS the whole range of $D(t)$ is well characterized by a fit exponent $\mu = 0.5553 \pm 0.0004$ for $100 \leqslant t \leqslant 10^6$ (see Fig. 2). Furthermore, for the AC-GT data the correlation function behaves also as $C_P(n) \sim n^{-\nu}$ where for HS the exponent obtained from an error-weighted fit is $\nu = 0.367 \pm 0.004$ for even terms and $\nu = 0.320 \pm 0.004$ for odd terms, both at $10 \leqslant n \leqslant 10^4$. Even if formal statistical errors are quite small we should note that there are rather pronounced oscillations and for that reason we give in the above discussions only approximate values of the exponents.

The presented results determine the statistics of Poincaré recurrences of DNA sequences and link their properties to the statistics of sequence walks studied previously [16–18]. The anomalous diffusion of walks is related to enormously long correlations between far away recurrences. For most detailed HS sequences the diffusion coefficient of these walks becomes finite due to cancellations of slow decaying correlations. For other species larger statistical samples are required to see if the diffusion coefficient saturation is present. The Poincaré recurrences $P(t)$ are statistically very stable and show a clear difference between various species. The statistical analysis of human and mammalian DNA sequences is now an active research field with links to genome evolution (see, e.g., [24–26]) and the approach based on Poincaré recurrences should bring here new useful insights.

The obtained properties of Poincaré recurrences can be used for the verification of various theories of genome evolution (see, e.g., [24–28]). Such theories should reproduce well the main statistical features of Poincaré recurrences described here. Indeed, the data of Fig. 1 show that for $t < 5$ the recurrences for all analyzed species behave like a random sequence of coin flipping. Thus the genome evolution generates random uncorrelated short-range recurrences. However, for the range $5 \leqslant t \leqslant 20$ we have a beginning algebraic decay of $P(t)$, but still all the species follow practically the same curve. This indicates the existence of a common period of initial evolution history. For $t > 20$ we observe a strong divergence of Poincaré curves of different species. Surprisingly, the curves of HS and FC (as well as LA and XT) remain very close to each other up to the largest recurrences with $t \approx 400$ for AG-CT data sets. At the same time, for AC-GT data sets a close proximity of recurrences is observed for HS, LA, and XT (as well as for FC and CF) up to the largest values $t \approx 300$. This shows the various aspects of proximity between species which should be investigated in further studies. We hope that the new tool of Poincaré recurrences will allow to analyze the proximity between species under a new angle, lightening new sides of life evolution.

[1] H. Poincaré, Acta Math. **13**, 1 (1890).

[2] V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics* (Benjamin, Paris, 1968).

[3] I. P. Cornfeld, S. V. Fomin, and Y. G. Sinai, *Erodic Theory* (Springer, New York, 1982).

[4] A. J. Lichtenberg and M. A. Lieberman, *Regular and Chaotic Dynamics* (Springer, Berlin, 1992).

[5] B. V. Chirikov and D. L. Shepelyansky, Physica D **13**, 395 (1984).

[6] J. D. Meiss and E. Ott, Phys. Rev. Lett. **55**, 2741 (1985).

[7] B. V. Chirikov and D. L. Shepelyansky, Phys. Rev. Lett. **82**, 528 (1999); **89**, 239402 (2002).

[8] E. G. Altman and H. Kantz, Europhys. Lett. **78**, 10008 (2007).

[9] G. Cristadoro and R. Ketzmerick, Phys. Rev. Lett. **100**, 184101 (2008).

[10] D. L. Shepelyansky, Phys. Rev. E **82**, 055202(R) (2010).

[11] L. Rossi and G. Turchetti, Physica A **338**, 267 (2004).

[12] A. S. S. Nair and T. Mahalakshmi, in Proceedings of IEEE Genomic Signal Processing. Bucharest, Romania (2005).

[13] V. Afreixo, C. A. C. Bastos, A. J. Pinho, S. P. Garcia, and P. J. S. G. Ferreira, Bioinformatics **25**, 3064 (2009)

[14] C. A. C. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. M. O. S. Rodrigues, and P. J. S. G. Ferreira, *Distances Between Dinucleotides in the Human Genome*, Adv. Intel. Soft Comp., edited by M. P. Rocha *et al.* (Springer, Berlin, 2011), Vol. 93, p. 205.

[15] Ensembl Genome data base [http://www.ensembl.org/] and [ftp://ftp.ensembl.org/pub/release-62/genbank/].

[16] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[17] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[18] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[19] M. F. Higareda, H. Hernandez-Saldana, and R. A. Mendez-Sanchez, Physica A **372**, 368 (2006).

[20] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martinez-Aroza, and J. L. Oliver, BMC Bioinformatics **7**, 446 (2006).

[21] P. Carpena, P. Bernaola-Galvan, M. Hackenberg, A. V. Coronado, and J. L. Oliver, Phys. Rev. E **79**, 035102(R) (2009).

[22] W. Li and D. Holste, Phys. Rev. E **71**, 041910 (2005).

[23] We note that is important to evaluate very carefully the Poincaré recurrence correlation function as $C_P(j) = \langle t_1 t_{1+j} \rangle - \langle t_1 \rangle \langle t_{1+j} \rangle$ where the averages $\langle t_1 \rangle$ and $\langle t_{1+j} \rangle$ are computed as *different* quantities using exactly the *same* data used to compute the average $\langle t_1 t_{1+j} \rangle$ (i.e., all sequences of at least $j + 1$ Poincaré recurrences $t_1, \ldots, t_{1+j}$ for which the covered DNA sequence is not interrupted by any nondetermined N letter entry). In principle, the average $\langle t_1 \rangle$ can be also computed for a larger data set (of simply all Poincaré recurrences available), but the resulting value would be slightly different and not appropriate for use in the correlation function.

[24] D. A. Wheller *et al.*, Nature (London) **452**, 872 (2008).

[25] J. Romiguier, V. Ranwez, E. J. P. Douzery, and N. Gatltier, Genome Res. **20**, 1001 (2010).

[26] Z. M. Frenkel, T. Bettecken, and E. N. Trifonov, BMC Genomics **12**, 203 (2011).

[27] M. Nei, *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).

[28] D. N. Cooper, *Human Gene Evolution* (Elsevier, Amsterdam, 1999).