Poincaré recurrences of DNA sequence

K.M.Frahm and D.L.Shepelyansky
Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France
(Dated: September 2, 2011)

We analyze the statistical properties of Poincaré recurrences of Homo sapiens, mammalian and other DNA sequences taken from Ensembl Genome data base with up to fifteen billions base pairs. We show that the probability of Poincaré recurrences decays in an algebraic way with the Poincaré exponent $\beta \approx 4$ even if oscillatory dependence is well pronounced. The correlations between recurrences decay with an exponent $\nu \approx 0.6$ that leads to an anomalous super-diffusive walk. However, for Homo sapiens sequences, with the largest available statistics, the diffusion rate converges to a finite value on distances larger than million base pairs.

PACS numbers: 87.14.gk,05.40.Fb,05.45.Tp,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 a close vicinity of an initial state. Probability of such Poincaré recurrences $P(t)$ drops exponentially at large return times t for fully chaotic systems [2, 3] being similar to a coin flipping with an exponential decay of one side stay probability. However, in generic Hamiltonian systems the probability $P(t)$ decays algebraically with t due to long trappings in a vicinity of stability islands showing the Poincaré exponent $\beta \approx 1.5$ [4–9]. We apply such an approach to available mammalian DNA sequences [10] and show that their Poincaré recurrences are characterized by an algebraic decay with $\beta \approx 4$ for Homo sapiens (HS) database of $1.5 \cdot 10^{10}$ base pairs (bp). In contrast to Hamiltonian systems [4–9], the Poincaré recurrences in DNA exhibit long range correlations leading to an anomalous super-diffusion on scales of $t < 10^6$ bp in agreement with previous studies [11–13]. 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 $\nu \approx 0.6$.

To study the statistics of Poincaré recurrence of mammalian DNA sequences we use the enormous database [10] considering a DNA sequence as a very long trajectory in the space of four nucleobases A, G, C, T. Similar to [11], 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 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 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 recurrences or crossings of 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 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). In 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 [4, 6, 8]. 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. 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 $10^5 \leq L \leq 1.5 \cdot 10^{10}$ bp.

The comparison of statistics of Poincaré recurrences for HS, mammalian and two other species are shown in Fig. 1 for AG-CT case (similar average behavior is found for AC-GT data). The total sequence lengths L for other species are by a factor 3 shorter compared to HS case. Up to $t \approx 20$ all considered species show the same decay of $P(t)$ but at larger value of t there is a separation of curves so that each species is characterized by its own statistics $P(t)$. In 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 (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 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 super-diffusive walk discussed in [11–13]. The walk is described by a displacement variable $y(t) = \sum_{\tau=1}^{t} u(\tau)$ whose growth can be characterized by a diffusion rate defined as $D(t) = \langle \Delta y(t)^2 \rangle / t$ with $\Delta y(t) = y(t + t_0) - y(t_0) - \langle y(t+t_0) - y(t_0) \rangle$ and the average $\langle \cdots \rangle$ is done with

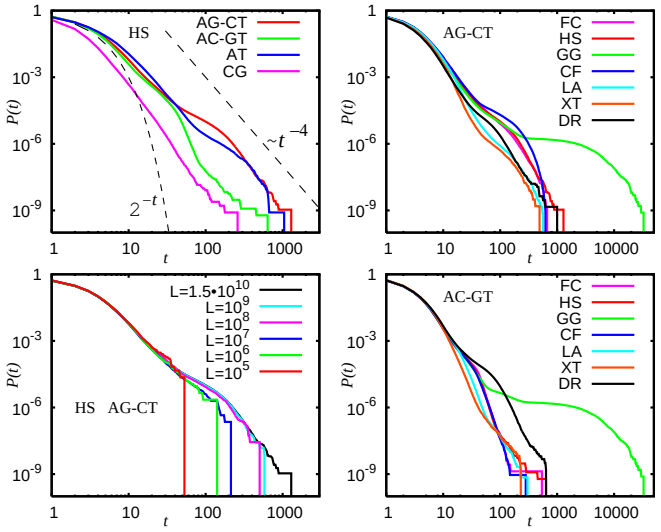

FIG. 1: Statistics of Poincaré recurrences $P(t)$ for DNA sequences. *Top left panel*: DNA data of Homo sapiens (HS) for Poincaré recurrences of domains AG-CT, AC-GT, AT and CG (see text). The lower dashed curve shows the exponential behavior $P(t) = 2^{-t}$ valid for *random* sequences, the upper dashed 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 $L = 1.5 \cdot 10^{10}$. *Bottom right panel*: AC-GT data sets for the same species as in the top right panel.

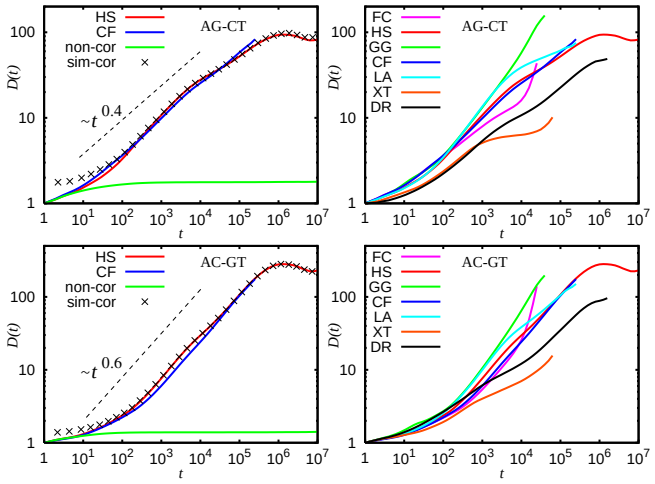

FIG. 2: *Top left panel*: Diffusion rate $D(t) = \langle \Delta y^2(t) \rangle / t$ for AG-CT data sets of DNA sequences of HS and CF. The lower green curve (non-cor) is the diffusion rate obtained for a model with individual recurrences being distributed as the Poincaré recurrences of HS in Fig. 1 but *assuming* that subsequent Poincaré recurrences are *not correlated*. The black crosses (sim-cor) represent the diffusion rate obtained from (2) 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 rate $D(t)$ 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 $D(t) \sim t^{0.6}$.

respect to the initial position (or "time") t_0. In case of a standard diffusive process the diffusion rate D converges to a finite value at large times. However, the results of [11] give an algebraic super-diffusive growth $D(t) \sim t^{\mu}$ with the exponent $\mu \approx 0.34$ for HS sequence of length $L \sim 10^5$ and $t \leq 10^3$. Our results obtained on a significantly larger scale are shown in Fig. 2. For HS sequence we have large statistics and large exact segments without non-determined bp marked as N in database [10]. 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 [11–13]. Other species also show an algebraic growth of $D(t)$ with similar values of μ (Fig. 2). For AC-GT data we also find a similar behavior with $\mu \approx 0.6$ for HS sequence (Fig. 2). However, for HS sequence with most exact and long data set we find a saturation of $D(t)$ for large times $10^6 \leq t \leq 10^7$.

The diffusion rate 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 the results obtained for chaotic Hamiltonian dynamics show that $c(t) \sim tP(t) \sim t^{1-\beta}$ [5, 6] so that 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 AG-CT sequence of HS and compute with them the diffusion rate then we find a clear saturation of $D(t)$ at a finite value $D = 1.77$ (green curve in Fig. 2, left panel), being significantly smaller then the actual data of $D(t) \sim 100$.

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 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 $t_2 \geq 8$ (i.e. above average) or vice-versa thus indicating anti-correlations 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

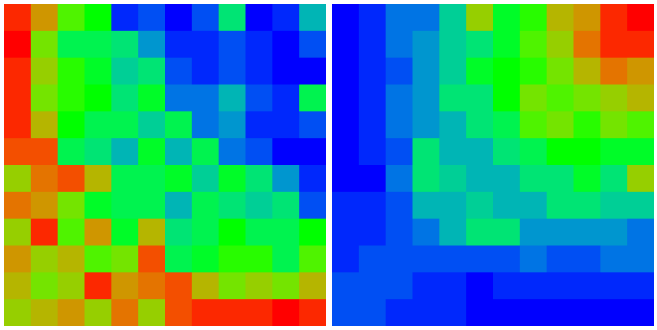

FIG. 3: *Left panel*: Density plot of the normalized two point correlator $\tilde{p}_2(t_1, t_2)$ of two subsequent Poincaré recurrences t_1 and t_2 for AG-CT data sets of HS. The shown range $1 \leq t_1, t_2 \leq 12$ represents 99.4% of probability. Red (green, blue) color represents maximal (zero, minimal) values, horizontal/vertical axes show t_1 and t_2. *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.

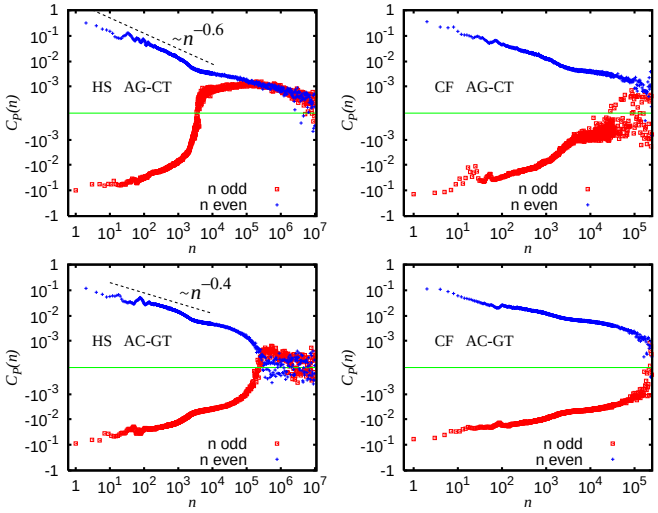

FIG. 4: *Top panels:* Poincaré recurrence correlation function $C_P(n) = \langle t_1\, t_{n+1}\rangle - \langle t_1\rangle^2$ of t_1 and t_{n+1} in a sequence of subsequent Poincaré recurrences t_j for HS (left panel) and CF (right panel) sequences for AG-CT data sets. Blue crosses correspond to even n and red 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 green 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 $C_P(n) \sim n^{-0.4}$.

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, ... the odd elements represent steps of length t_1, t_3, ... of one sign of $u(t)$ and the even elements represent steps of length t_2, t_4, ... of the other sign. The anti-correlations 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 $t = t_1 + \ldots + t_n \approx n\langle t_1\rangle$. We can write for it

$$y(t_1 + \ldots + t_n) = (-1)^s \sum_{l=1}^{n} (-1)^{l-1} t_l \quad , \qquad (1)$$

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

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

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. We note that the above model of uncorrelated Poincaré recurrences corresponds to $C_P(j) = 0$ for $j > 0$. In this case Eq.(2) 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 recurrence index/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 $n \approx 3 \cdot 10^3$. For larger n values these terms have the same sign and moreover these terms become approximately equal for $n > 10^5$. This leads to cancellation of odd and even terms in (2) and 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 $10 \leq n \leq 3 \cdot 10^6$ and for odd terms $\nu = 0.479 \pm 0.005$ at $10 \leq n \leq 10^3$). From the found correlation function $C_P(n)$ we can determine the dependence $D(t)$ using (2) 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 $t = t_1 + .. + t_n \approx n\langle t_1\rangle$ 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/even terms cancellation at $t < 10^6$. For 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 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 $n < 2.5 \cdot 10^5$ where a saturation of $D(t)$ growth is not visible (for HS case we have

$N_r(n) \approx 5 \cdot 10^9$ 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 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 odd/even terms of $C_F(n)$ for HS case takes place at $n > 10^5$ leading to saturation of 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 saturation of $D(t)$ remains invisible.

Let us give here the formal fit parameter values for the dependences 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), 4.04 ± 0.05 (CG). The fit of $D(t) \sim t^{\mu}$ for AG-CT data of HS in Fig. 1 gives $\mu = 0.3486 \pm 0.0008$ for the range $10 \leq t \leq 10^6$ but there are two intervals with distinct values $\mu = 0.5010 \pm 0.0003$ for $10 \leq t \leq 3 \cdot 10^3$ and $\mu = 0.2859 \pm 0.0003$ for $3 \cdot 10^3 \leq t \leq 10^6$ so that we give in the text the average $\mu \approx 0.4$. For 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 $100 \leq t \leq 10^6$ (see Fig. 2). Furthermore for 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 $10 \leq n \leq 10^4$. Even if formal statistical errors are quite small we should note that there are rather pronounced oscillations and due to that reason we give in the above discussions only approximate values of the exponents.

The presented results establish the properties of statistics of Poincaré recurrences of DNA sequences and link their properties to the statistics of sequence walks studied previously [11–13]. 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 rate saturation is present. The Poincaré recurrences $P(t)$ are statistically very stable and show 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. [14–16]) and the approach based on Poincaré recurrences should bring here new useful insights.

[1] H. Poincaré, *Sur le problème des trois corps et les equations de la dynamique*, Acta Math. **13**, 1 (1890).

[2] I.P. Cornfeld, S.V. Fomin and Y.G. Sinai, *Erodic theory*, Springer, N.Y. (1982).

[3] A.J. Lichtenberg and M.A. Lieberman, *Regular and chaotic dynamics*, Springer, Berlin (1992).

[4] B.V. Chirikov and D.L. Shepelyansky, *Correlation properties of dynamical chaos in Hamiltonian systems*, Physica D **13**, 395 (1984).

[5] J.D. Meiss and E. Ott, *Markov-tree model of intrinsic transport in Hamiltonian systems*, Phys. Rev. Lett. **55**, 2741 (1985).

[6] B.V. Chirikov and D.L. Shepelyansky, *Asymptotic statistics of Poincaré recurrences in Hamiltonian systems with divided phase space*, Phys. Rev. Lett. **82**, 528 (1999); **ibid. 89**, 239402 (2002).

[7] E.G. Altman and H. Kantz, *Hypothesis of strong chaos and anomalous diffusion in coupled symplectic maps*, Europhys. Lett. **78**, 10008 (2007).

[8] G. Cristadoro and R. Ketzmerick, *Universality of algebraic decays in Hamiltonian systems*, Phys. Rev. Lett. **100**, 184101 (2008).

[9] D.L. Shepelyansky, *Poincaré recurrences in Hamiltonian systems with a few degrees of freedom*, Phys. Rev. E **82**, 055202(R) (2010).

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

[11] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, *Long-range correlations in nucleotide sequences*, Nature **356**, 168 (1992).

[12] W. Li and K. Kaneko, *Long-range correlation and partial* $1/f^{\alpha}$ *spectrum in a noncoding DNA sequence*, Europhys. Lett. **17**, 655 (1992).

[13] R.F. Voss, *Evolution of long-range fractal correlations and* $1/f$ *noise in DNA base sequences*, Phys. Rev. Lett. **68**, 3805 (1992).

[14] D.A. Wheller *et al.*, *The complete genome of an individual by massively parallel DNA sequencing*, Nature **452**, 872 (2008).

[15] J. Romiguier, V. Ranwez, E.J.P. Douzery and N. Gatltier, *Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes*, Genome Res. **20**, 1001 (2010).

[16] Z.M. Frenkel, T. Bettecken and E.N. Trifonov, *Nucleosome DNA sequence structure of isochores*, BMC Genomics **12**, 203 (2011).