ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Physica A ▮ (▮▮▮▮) ▮▮▮–▮▮▮

ELSEVIER

PHYSICA A

www.elsevier.com/locate/physa

Similarity and dissimilarity in correlations of genomic DNA

Boris Podobnik[a,b], Jia Shao[c,*], Nikolay V. Dokholyan[d], Vinko Zlatic[e], H. Eugene Stanley[c], Ivo Grosse[f]

[a]*Faculty of Civil Engineering, University of Rijeka, Rijeka, Croatia*
[b]*Zagreb School of Economics and Management, Zagreb, Croatia*
[c]*Center for Polymer Studies and Department of Physics, Boston University, Boston, USA*
[d]*School of Medicine, University of North Carolina, Chapel Hill, USA*
[e]*Rudjer Boskovic Institute, Zagreb, Croatia*
[f]*Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany*

Abstract

We analyze auto-correlations of human chromosomes 1–22 and rice chromosomes 1–12 for seven binary mapping rules and find that the correlation patterns are different for different rules but almost identical for all of the chromosomes, despite their varying lengths and GC contents. We propose a simple stochastic process for modeling these correlations, and we find that the proposed process can reproduce, quantitatively and qualitatively, the correlation patterns found in the genomes of human and rice.

© 2006 Elsevier B.V. All rights reserved.

PACS: 87.10; 05.40.+j

DNA is the carrier of genetic information of many living organisms. It is comprised of the four nucleotides A (adenine), T (thymine), C (cytosine), and G (guanine). In order to study correlations of DNA sequences one commonly maps each nucleotide to a binary number and then studies correlations of the resulting numerical sequence. There are three different partitions of four nucleotides into two subsets of two nucleotides each, and so there are three different binary mapping rules that map a nucleotide s onto a binary number $x \in \{-1, +1\}$:

- SW rule: $x = 1$ for C or G, and $x = -1$ otherwise.
- KM rule: $x = 1$ for A or C, and $x = -1$ otherwise.
- RY rule: $x = 1$ for A or G, and $x = -1$ otherwise.

There are many studies on long-range correlations in various DNA sequences using different binary representations of the four nucleotides [1–11] with partially contradicting results. As it is possible that a symbolic sequence shows long-range correlations for one rule, and no correlations for another rule, we conjecture that some of these apparent contradictions might be due to the fact that different research groups

[*]Corresponding author.
E-mail address: jiashao@buphy.bu.edu (J. Shao).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.05.041

have chosen different binary mapping rules. In the following, we analyze long-range correlations of DNA sequences using each of the three binary mapping rules.

First, we map a DNA sequence $(s_0, s_1, \ldots, s_{N-1})$ of length N to a binary sequence $(x_0, x_1, \ldots, x_{N-1})$ by one of the three binary mapping rules. Then, we compute correlations by employing the *detrended fluctuation analysis* (DFA) [12], a variant of the root-mean-square analysis of a random walk. In the DFA method, one measures the standard deviation $F(\ell)$ of the detrended fluctuations within a window of length ℓ as a function of ℓ [13]. If the auto-correlation function $C(\ell) \equiv \langle x_n x_{n+\ell} \rangle - \langle x_n \rangle \langle x_{n+\ell} \rangle$ can be approximated by a power-law with exponent γ, i.e., if $C(\ell) \propto \ell^{-\gamma}$, then $F(\ell)$ can be approximated by a power-law with exponent α, i.e., $F(\ell) \propto \ell^{\alpha}$, with $\alpha \approx 1 - \gamma/2$ [14]. If $\alpha > 0.5$, the time series is power-law correlated; if $\alpha = 0.5$, the time series is uncorrelated or short-range correlated; and if $\alpha < 0.5$, the time series is power-law anti-correlated.

By using the DFA method, we analyze auto-correlations from two completely sequenced genomes, one from the plant kingdom and one from the animal kingdom. Fig. 1(a) shows $F(\ell)$ versus ℓ for all 22 autosomes of homo sapiens using the SW rule, the KM rule, and the RY rule [15]. First, we find that for each binary mapping rule the $F(\ell)$ curves corresponding to different chromosomes are almost identical. Second, we find in agreement with Ref. [16,17] that for the SW rule the $F(\ell)$ curves can be approximated by a power-law with scaling exponent $\alpha \to 1$ for scales ℓ exceeding 10^6 bp. This states that SW rule asymptotically follows $1/f$ noise. Third, we find that for the KM rule and the RY rule the $F(\ell)$ curves can also be approximated by power-law, but with scaling exponents α_{KM} and α_{RY} that are substantially smaller than α_{SW}, i.e., $\alpha_{\mathrm{SW}} > \alpha_{\mathrm{KM}} \approx \alpha_{\mathrm{RY}}$.

In view of the many known differences among the 22 autosomes of homo sapiens it is surprising that their auto-correlations behavior is almost indistinguishable. In order to study if the phenomenon that auto-correlations of different chromosomes are almost indistinguishable whereas auto-correlations of different binary mapping rules are different from each other is ubiquitous, we analyze all 12 autosomes of *oryza sativa* using the SW rule, the KM rule, and the RY rule. From Fig. 1(b) we find that for each binary mapping rule the $F(\ell)$ curves corresponding to different chromosomes are almost indistinguishable. We find that the $F(\ell)$ curves for the KM rule and the RY rule can be approximated by a power-law with almost the same scaling exponent, i.e., $\alpha_{\mathrm{KM}} \approx \alpha_{\mathrm{RY}}$, and that for the SW rule there is a significant crossover in the $F(\ell)$ curves at approximately $\ell \approx 10^4$ bp. In agreement to Fig. 1(a) we find that also for rice the scaling exponent α_{SW} is greater than $\alpha_{\mathrm{KM}} \approx \alpha_{\mathrm{RY}}$ in the asymptotic regime.

In summary, we find that the $F(\ell)$ curves for all chromosomes are (i) identical and (ii) power-laws for each rule, (iii) power-law for SW rule is different from KM and RY rule, and (iv) power-laws of KM rule and RY rule are almost identical.

Several stochastic models have been proposed to generate long-range correlations, but most of them are focused on reproducing correlations for only one binary mapping rule [6,18]. In order to model the scaling

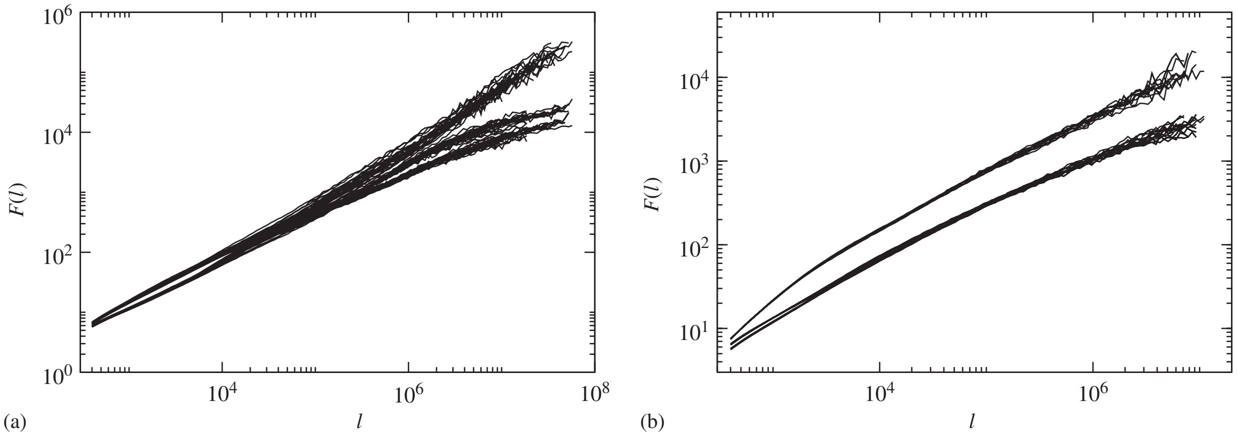

Fig. 1. Detrended fluctuation functions $F(\ell)$ versus ℓ. For the SW rule, the KM rule, and the RY rule we show the $F(\ell)$ curves (a) for all 22 human chromosomes (autosomes) and (b) rice. For both human and rice, for each rule $F(\ell)$ curves practically overlap. (a) For human, we find for the SW rule that $F(\ell)$ curves tends to $1/f$ law for very large scales. For very large scales we find $\alpha_{\mathrm{KM}} \approx \alpha_{\mathrm{RY}} < \alpha_{\mathrm{SW}}$. (b) For rice we find that the $F(\ell)$ curves for KM and RY rule collapse onto each other, where for large scales again we find $\alpha_{\mathrm{KM}} \approx \alpha_{\mathrm{RY}} < \alpha_{\mathrm{SW}}$.

properties observed in the genomes of human and rice, we proceed in two steps. First, we introduce two mutually independent ARFIMA processes [19] $z_n^{(\mathrm{SW})}$ and $z_n^{(\mathrm{KM})}$ defined by [20]

$$z_n^{(j)} = \sum_{\ell=1}^{\infty} a_\ell(\rho_j) z_{n-\ell}^{(j)} + e_n^{(j)}, \tag{1a}$$

$$a_\ell(\rho_j) = \frac{\Gamma(\ell-\rho_j)}{\Gamma(-\rho_j)\Gamma(1+\rho_j)}, \tag{1b}$$

where $j \in \{\mathrm{SW}, \mathrm{KM}\}$, $e_n^{(j)}$ are independent and identically distributed Gaussian variables, Γ denotes the Gamma function, and each stochastic process $z_n^{(j)}$ is parameterized by one parameter $\rho_j \in (-0.5, 0.5)$.

Second, we define the nucleotide

$$s_n = \begin{cases} \mathrm{A} & \text{if } z_n^{(\mathrm{SW})} < \delta \wedge z_n^{(\mathrm{KM})} < 0, \\ \mathrm{T} & \text{if } z_n^{(\mathrm{SW})} < \delta \wedge z_n^{(\mathrm{KM})} > 0, \\ \mathrm{G} & \text{if } z_n^{(\mathrm{SW})} > \delta \wedge z_n^{(\mathrm{KM})} < 0, \\ \mathrm{C} & \text{if } z_n^{(\mathrm{SW})} > \delta \wedge z_n^{(\mathrm{KM})} > 0, \end{cases} \tag{2}$$

where δ is the cutoff value, defined as a measure of excess of GC over AT nucleotides. This assignment can be also interpreted by two spin models, where each nucleotide is represented by two spins, one spin corresponding to the SW state of the nucleotide, and the other spin corresponding to the KM state of the nucleotide.

To simulate DNA sequences we choose two ARFIMA processes due to the fact that the ARFIMA process generates power-law correlated time series $z_n^{(j)}$ with scaling exponent $\alpha_j \approx 1 + \rho_j$ [19,21]. We find by numerical simulations that the scaling relation $\alpha_j \approx 1 + \rho_j$ also holds for the time series $\mathrm{sgn}(z_n^{(j)})$. Hence, two ARFIMA processes running simultaneously generate sequences with $\alpha_{\mathrm{SW}} \approx 1 + \rho_{\mathrm{SW}}$ and $\alpha_{\mathrm{KM}} \approx 1 + \rho_{\mathrm{KM}}$.

If δ in the above equation is chosen to be zero, the probability of occurrence of all nucleotides is identical, 0.25 due to the choice of mutually independent variables $\varepsilon_n^{(j)}$ in Eq. (1) taken to be out of symmetrical Gaussian distribution. Since for the Gaussian distribution probability of positive and negative values is equal, 0.5, probability of occurrence of any type of nucleotide is $0.5 \cdot 0.5 = 0.25$.

In order to model nonhomogenuous in occurrence of nucleotides in DNA sequences, probabilities of occurring spins, that we relate to nucleotides, must be different. Since the relative frequency of nucleotides C and G is equal (≈ 0.21 for chromosome 1), and the same holds for A and T (≈ 0.28 for chromosome 1), to obtain the probabilities $p_{\mathrm{S}} \equiv P(z_n^{(\mathrm{SW})} > \delta)$ and $p_{\mathrm{A}} \equiv P(z_n^{(\mathrm{A})} > 0)$, we equate relative frequency of nucleotide $\mathrm{A} \equiv (-1, +1)$ with probability $p_{\mathrm{W}} \cdot p_{\mathrm{A}}$, $\mathrm{T} \equiv (-1, -1)$ with $p_{\mathrm{W}} \cdot p_{\mathrm{K}}$, $\mathrm{C} \equiv (+1, +1)$ with $p_{\mathrm{S}} \cdot p_{\mathrm{A}}$, and $\mathrm{G} \equiv (+1, -1)$ with $p_{\mathrm{S}} \cdot p_{\mathrm{K}}$. From these equations, for chromosome 1, by using obvious relations $p_{\mathrm{A}} = 1 - p_{\mathrm{K}}$ and $p_{\mathrm{S}} = 1 - p_{\mathrm{W}}$, we easily obtain $p_{\mathrm{A}} = 0.5$ and $p_{\mathrm{W}} = 0.582$. Thus, asymmetry is needed only for SW rule, in accordance to Chergaff rule.

For this simple model, we can easily derive the auto-correlation function for the RY rule

$$C_{\mathrm{RY}}(\ell) = 4C_{\mathrm{SW}}(\ell)C_{\mathrm{KM}}(\ell) + C_{\mathrm{SW}}(\ell)(p_{\mathrm{K}} - p_{\mathrm{M}})^2 + C_{\mathrm{KM}}(\ell)(p_{\mathrm{W}} - p_{\mathrm{S}})^2, \tag{3}$$

where for $p_{\mathrm{K}} \approx p_{\mathrm{M}}$ and $p_{\mathrm{W}} \neq p_{\mathrm{S}}$, as is the case for DNA sequences, $C_{\mathrm{RY}}(\ell)$ reduces to $C_{\mathrm{KM}}(\ell)(p_{\mathrm{W}} - p_{\mathrm{S}})^2$. This implies

$$\alpha_{\mathrm{RY}} \approx \alpha_{\mathrm{KM}}. \tag{4}$$

Hence, the model predicts that $C_{\mathrm{RY}}(\ell)$ shows the same scaling behavior as $C_{\mathrm{KM}}(\ell)$. Interestingly, this behavior predicted for the model for asymptotically large ℓ is observed in real DNA (Fig. 1).

We perform numerical simulations with $\rho_{\mathrm{SW}} = 0.41$ and $\rho_{\mathrm{KM}} = 0.25$ in order to reproduce the observed power-law correlations of the human SW and KM rule, respectively, shown in Fig. 2(a). We generate a sequence of length $\mathrm{N} = 2 \times 10^7$ and compute $F(\ell)$, for ℓ ranging from $\ell = 10^3$ to 10^7 bp (asymptotic regime). In Fig. 2(b) we show $F(\ell)$ for SW, KM, and RY binary mapping rules. The model gives that $F(\ell)$ calculated for SW and KM rule are power-laws as expected, and predicts RY is specified also by power-law with an exponent smaller than expected. Next we perform numerical simulations with $\rho_{\mathrm{SW}} = 0.1$ and $\rho_{\mathrm{KM}} = 0.0$ in order to reproduce

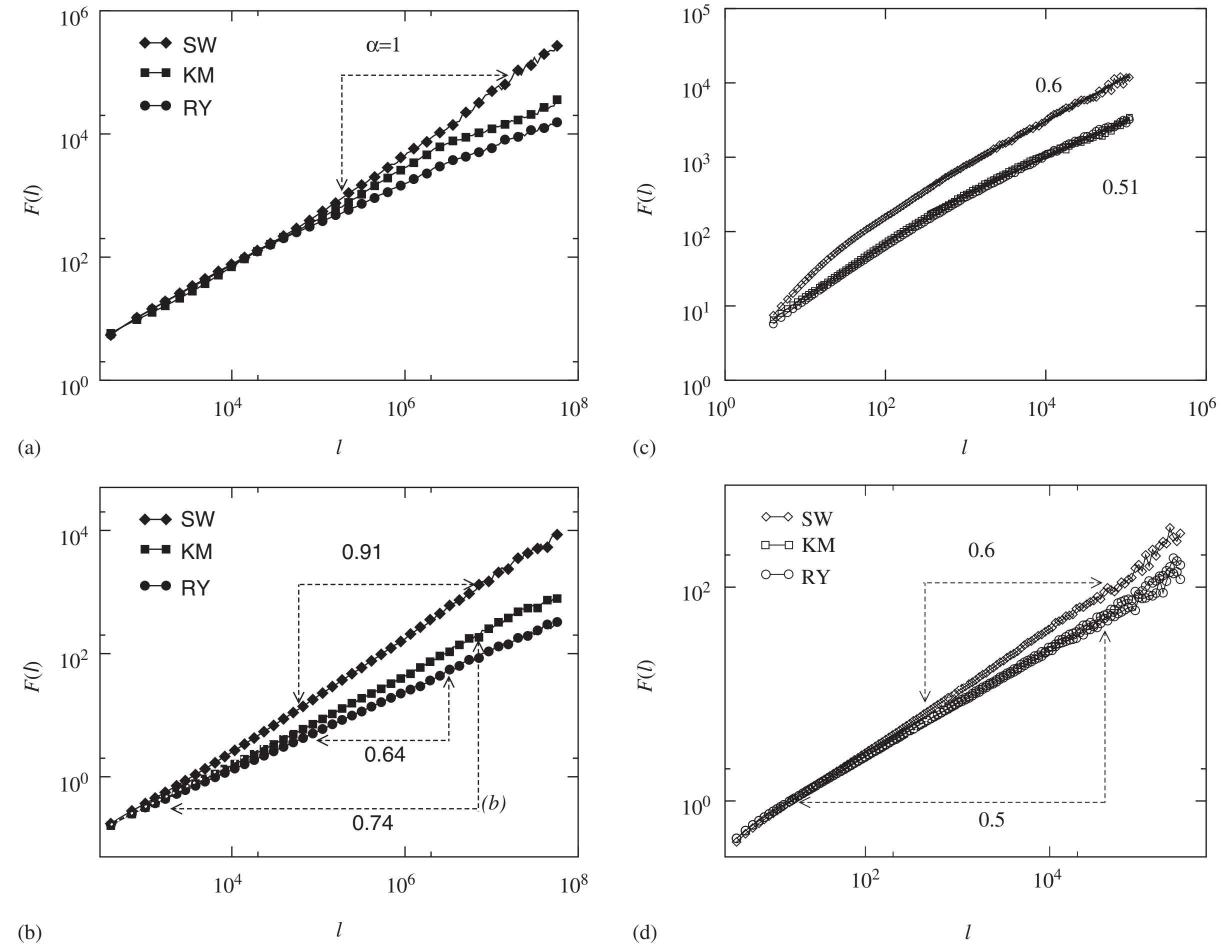

Fig. 2. $F(\ell)$ versus ℓ. (a) For chromosome 1, we find that RY and RY rule are approximately the same for very large scales, whereas for the SW rule $F(\ell)$ scales as $1/f$ [16]. We find that $F(\ell)$ curves for $A\bar{A}$, $C\bar{C}$, $G\bar{G}$, and $T\bar{T}$ rule practically overlap and scale asymptotically as SW rule. (b) Model simulations. For parameters $\alpha_{SW} = 0.41$ and $\alpha_{KM} = 0.25$ set to fit the scaling for the whole range for SW and KM rule, and the cutoff length $\delta = 0.2$ set to enable "AT rich" DNA, we perform numerical simulations and generate the sequences for different rules. We find the model reproduces the correlations scaling found in the empirical data. (c) For the rice, we find that again RY and KM rule are practically identical what we predict for the data where content of S and W are not the same. We find that $\alpha_{SW} > \alpha_{KM} \approx \alpha_{RY}$.

power-law correlations of the rice SW and KM rule, shown in Fig. 2(c). In Fig. 2(d) for the process we show that the $F(\ell)$ for SW rule is a power-law, while for KM and RY rule there are no correlations as we found in Fig. 2(c).

In order to test the validity of the model, we compute the auto-correlation function for the following four binary mapping rules:

- $A\bar{A}$ rule: $x = 1$ for A, and $x = -1$ otherwise.
- $G\bar{G}$ rule: $x = 1$ for G, and $x = -1$ otherwise.
- $C\bar{C}$ rule: $x = 1$ for C, and $x = -1$ otherwise.
- $T\bar{T}$ rule: $x = 1$ for T, and $x = -1$ otherwise.

For the model, for example, we can easily compute

$$C_{A\bar{A}}(\ell) = C_{SW}(\ell)C_{KM}(\ell) + C_{SW}(\ell)p_M^2 + C_{KM}(\ell)p_W^2, \tag{5a}$$

$$C_{T\bar{T}}(\ell) = C_{SW}(\ell)C_{KM}(\ell) + C_{SW}(\ell)p_K^2 + C_{KM}(\ell)p_W^2, \tag{5b}$$

$$C_{\mathrm{C\bar{C}}}(\ell) = C_{\mathrm{SW}}(\ell)C_{\mathrm{KM}}(\ell) + C_{\mathrm{SW}}(\ell)p_{\mathrm{M}}^2 + C_{\mathrm{KM}}(\ell)p_{\mathrm{S}}^2, \tag{5c}$$

$$C_{\mathrm{G\bar{G}}}(\ell) = C_{\mathrm{SW}}(\ell)C_{\mathrm{KM}}(\ell) + C_{\mathrm{SW}}(\ell)p_{\mathrm{K}}^2 + C_{\mathrm{KM}}(\ell)p_{\mathrm{S}}^2. \tag{5d}$$

We also derive that the following relation fold between the scaling exponents for all $\mathrm{N} \in \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$:

$$\alpha_{\mathrm{N\bar{N}}} = \max(\alpha_{\mathrm{SW}}, \alpha_{\mathrm{KM}}). \tag{6}$$

If $\alpha_{\mathrm{SW}} > \alpha_{\mathrm{KM}}$ one easily show that $C_{\mathrm{A\bar{A}}}(\ell)$ scales as $C_{\mathrm{SW}}(\ell)$, i.e., $\alpha_{\mathrm{A\bar{A}}} \approx \alpha_{\mathrm{SW}}$. This implies that for asymptotically large ℓ, $C_{\mathrm{N\bar{N}}}(\ell)$ has the same scaling behavior as $C_{\mathrm{SW}}(\ell)$. Note that all the analytical derivations just assume that there are two binary mapping rules characterized by power-law correlations and that there are four different constituents in the sequences.

In order to test if this predictions can be possibly be observed also for real DNA, we show in Fig. 3(a) $F(\ell)$ for all four $\mathrm{N\bar{N}}$ rules for all human autosomes. Interestingly, all autosomes are identical for each rule, for each rule $F(\ell)$ is a power-law, and for all rules exponents are identical and equal to α_{SW}. The same scaling behavior we find in Fig. 3(b) for rice chromosomes. $F(\ell)$ for all four $\mathrm{N\bar{N}}$ rules are identical and approximately they are power-law. Next we perform numerical simulations and generate sequences for all four $\mathrm{N\bar{N}}$ rules based on the parameters in Fig. 2. For human autosomes in Fig. 3(c) and for rice chromosomes in Fig. 3(d), we show that the $F(\ell)$ curves for the $\mathrm{N\bar{N}}$ rules are the same and scale similar as the real data. Thus, the model based on two ARFIMA processes is capable of reproducing qualitatively and to a large extent quantitatively with only three free parameters, α_{SW} α_{KM}, and δ set to fit KM and SW rule, the power-law correlations in human DNA also for RY rule and four $\mathrm{N\bar{N}}$ rules.

What is the importance of long-range correlations in natural phenomena and where this behavior is reflected in nature is still an open question. Generally, finding models for empirical observations has two main purposes. One is to understand the nature of observed phenomena, and the other is to enable predictions.

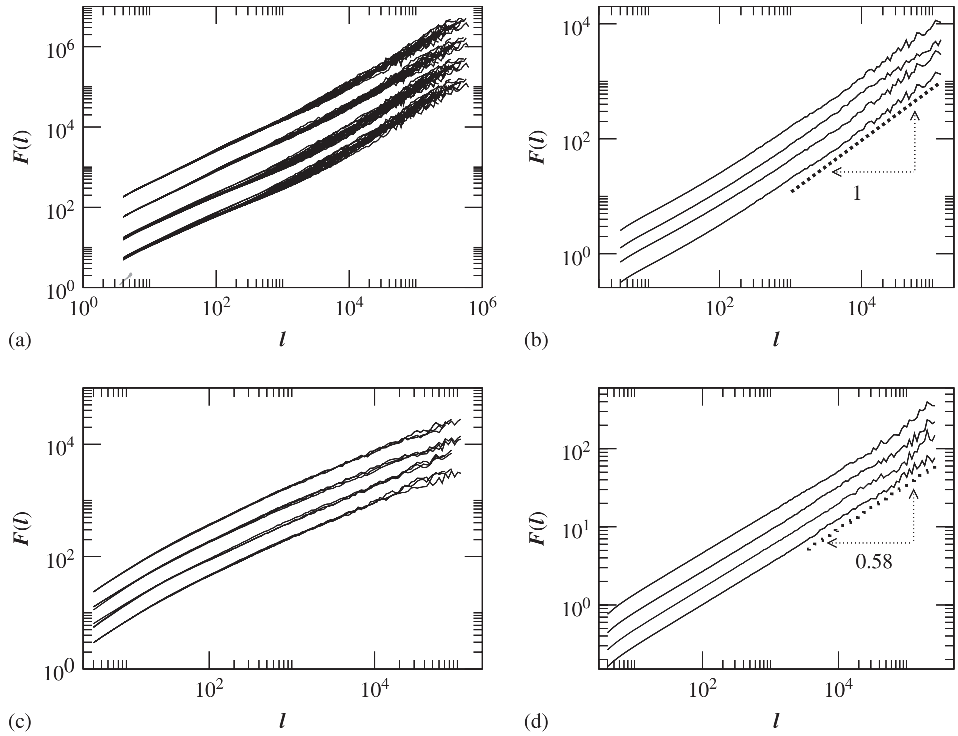

Fig. 3. $F(\ell)$ versus ℓ. For both (a) human and (b) rice chromosomes we find that $F(\ell)$ curves for $\mathrm{A\bar{A}}$, $\mathrm{C\bar{C}}$, $\mathrm{G\bar{G}}$, and $\mathrm{T\bar{T}}$ rules practically overlap. Curves for different rules are shifted. (c–d) Model simulations. For parameters set in Fig. 2, we find that $F(\ell)$ curves for $\mathrm{A\bar{A}}$, $\mathrm{C\bar{C}}$, $\mathrm{G\bar{G}}$, and $\mathrm{T\bar{T}}$ rule are identical and scale similar to the ones in the data.

To comply the second purpose, short-range correlated kth order Markov model [22,23] was proposed with the goal to predict the occurrence of repetitive $k$mers in genome sequences. But in opposite to our process specified by only three free parameters, kth order Markov model requires 4^{k+1} parameters to be estimated from occurrences of repetitive $k$mers in the empirical data. In future work, it would be worthwhile to put some effort on implementation of long-range correlations on Markov model to see how this might help increase predictive power of a new model. As a first step, allowing different cutoff lengths per each Gaussian distributions, our extended process may create not just nucleotides but also dimers or larger structures.

Acknowledgments

We thank S.V. Buldyrev for valuable discussions and the Ministry of Science of Croatia (Grant 0119623) and the German Federal Ministry of Education and Research (BMBF) for financial support.

References

[1] M. Ya Azbel, Phys. Rev. Lett. 31 (1973) 589.

[2] E.N. Trifonov, Bull. Math. Biol. 51 (1989) 417.

[3] C.-K. Peng, et al., Nature 356 (1992) 168.

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

[5] R. Voss, Phys. Rev. Lett. 68 (1992) 3805.

[6] S.V. Buldyrev, et al., Phys. Rev. 47 (1993) 4514.

[7] S.M. Osadnik, et al., Biophys. J. 67 (1994) 64.

[8] S.V. Buldyrev, et al., Phys. Rev. 51 (1995) 5084.

[9] P. Bernaola-Galvan, R. Roman-Roldan, J.L. Oliver, Phys. Rev. E 53 (1996) 5181.

[10] H. Herzel, I. Grosse, Physica A 216 (1995) 518.

[11] D. Holste, I. Grossse, H. Herzel, Phys. Rev. E 64 (2001) 041917;
D. Holste, et al., Phys. Rev. E 67 (2003) 061913.

[12] C.-K. Peng, et al., Phys. Rev. E 49 (1994) 1685.

[13] The auto-correlation function $C(\ell) \equiv \langle x_n x_{n+\ell}\rangle - \langle x_n\rangle\langle x_{n+\ell}\rangle$ and the variance function $F^2(\ell) \equiv (z_\ell - \langle z_\ell\rangle)^2$ are related by the Kubo formula $F^2(\ell) = \ell C(0) + 2\sum_{k=1}^{\ell-1}(\ell - k)C(k)$.

[14] Based on Ref. [13] one easily shows that, for asymptotically large sequence lengths N, if one of the functions $F(\ell)$ or $C(\ell)$ is of power-law form, then another one is also of power-law form where the exponents α, and γ of the scaling relations $F(\ell) \propto \ell^\alpha$ and $C(\ell) \propto \ell^{-\gamma}$ are related by $\alpha \propto 1 - \gamma/2$.

[15] The scaling exponents α are obtained by a least-square linear regression of $\ln F(\ell)$ versus $\ln \ell$. In order to make the fits $F(\ell) \propto \ell^\alpha$ comparable for all chromosomes, we always use the same fitting region ranging from $\ell = 10^3$ to 10^7 bp.

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

[17] S.V. Buldyrev, Power Laws, Scale-Free Networks and Genome Biology, in: E.V. Koonin, Y.I. Wolf, G.P. Karev (Eds.), Springer Science + Business Media.

[18] P. Allegrini, et al., Phys. Rev. 57 (1998) 4558.

[19] C.W.J. Granger, R. Joyeux, J. Time Series Anal. 1 (1980) 15;
J. Hosking, Biometrika 68 (1981) 165.

[20] For large values of ℓ, the weights $a_\ell(\rho_j)$ scales as $\ell^{-1-\rho_j}$.

[21] B. Podobnik, et al., Phys. Rev. E 71 (2) (2005) 025104 (R);
B. Podobnik, et al., Phys. Rev. E 72 (2) (2005) 026121.

[22] C.E. Lawrence, et al., Science 262 (1993) 208;
B. Lenhard, W.W. Wasserman, Bioinformatics 18 (2002) 1135.

[23] M. Borodovsky, D. Mcininch, Comput. Chem. 17 (1993) 123;
M. Borodovsky, et al., Nucleic Acids Res. 23 (1995) 2554;
S. Salzberg, et al., Nucleic Acids Res. 26 (1998) 544.