Ising-model description of long-range correlations in DNA sequences

A. Colliva,[1,*] R. Pellegrini,[2,†] A. Testori,[1,‡] and M. Caselle[1,§]
[1]*Dipartimento di Fisica dell'Università di Torino and I.N.F.N. sez. di Torino, Via Pietro Giuria 1, I-10125 Torino, Italy*
[2]*Physics Department, Swansea University, Singleton Park, Swansea SA2 8PP, UK*
(Received 29 September 2014; published 8 May 2015)

We model long-range correlations of nucleotides in the human DNA sequence using the long-range one-dimensional (1D) Ising model. We show that, for distances between 10^3 and 10^6 bp, the correlations show a universal behavior and may be described by the non-mean-field limit of the long-range 1D Ising model. This allows us to make some testable hypothesis on the nature of the interaction between distant portions of the DNA chain which led to the DNA structure that we observe today in higher eukaryotes.

DOI: 10.1103/PhysRevE.91.052703 PACS number(s): 87.10.Mn, 87.14.gk, 05.10.−a

I. INTRODUCTION

One of the most surprising features of higher eukaryotes genomes is the presence of long-range correlations in the composition of the DNA sequence. These correlations were discovered more than 20 years ago [1] when the first long continuous DNA sequences became available. Soon after this discovery several evolutionary models were proposed [2–13] to explain this behavior and compared with the growing collection of genomic data.

Thanks to next-generation sequencing projects an impressive amount of whole-genome sequences is now available, and the composition of genomic DNA can be studied systematically over a wide range of scales and organisms. This makes it now possible to assess the various models proposed for the description of these long-range correlations in a more extensive way. The statistical analysis is quite intricate since genomic DNA is a rather "patchy" statistical environment: it consists of genes, noncoding regions, repetitive elements, etc. Despite this complexity a few general results are by now well established.

(i) These correlations extend over a range much longer than previously expected and reach distances of the order of 10^7 bp (see Fig. 1).

(ii) They show a power-law behavior, with exponents characterized by a remarkable degree of universality, with very small variations across the human chromosomes [12] and between the human and mouse genomes [13].

(iii) In the human case these exponents take values in the range [0.05–0.30]. These values are much smaller than usual scaling exponents observed in genomic data. For instance they are much smaller than those of chromatine contacts within chromosomes as extracted from Hi-C experiments (we discuss in more detail this issue in Sec. IV).

These features pose severe constraints on the models proposed to explain the correlations. In particular, their universality and the unusual long-range scales are features typical of critical systems and suggest a modeling strategy characterized by scale invariance and universality, i.e., models which do not depend too much on the microscopic details and on genomic features with a fixed reference scale.

A very interesting model which fulfils these conditions is the so-called "expansion-randomization" (E-R) model proposed by Li [2] and solved exactly in Ref. [3]. The stationary state of the model is characterized by the expected long-range correlations and is largely independent of microscopic details [3]. The only problem of this model is that, in order to match the observed exponents, it requires duplication and insertion rates much larger than those derived from the actual expansion rate of our genome. Indeed, by comparing the human sequence with that of other mammals (and in particular with the mouse genome) one can see that in the last 100 Myr, i.e., since the mammalian radiation, the human genome was almost stable or at most expanded very slowly and that its expansion was mainly due to retrotransposon insertions. Thus, it is likely that the E-R model should be complemented with some other evolutionary process able to enforce long-range correlations without requiring an expanding genome. In particular, the unusual range of these correlations suggests the introduction of nonlocal interactions in the evolutionary process.

Following this line of reasoning in this paper we propose an evolutionary model based on nonlocal moves which—as the E-R one—is able to reproduce the large-distance correlations observed in the DNA sequences but does not require an expanding genome. Notwithstanding the intrinsic complexity of the nonlocal interactions, several features of the model, and in particular the scaling exponents, can be predicted very accurately because the stationary state of the model can be mapped into the equilibrium state of a (very peculiar) statistical model, the so-called "long-range one-dimensional (1D) Ising model," for which several exact and approximate results exist. In particular, different from the ordinary (short-range) Ising model, this model admits a critical point also in one dimension and in the neighborhood of this point displays long-range correlations exactly of the type observed in the human DNA sequence.

As we shall discuss below, we think that our model and the E-R model should be considered as complementary processes which were probably both active in the evolutionary path of higher eukaryotes and both contributed to shape the long-range features of the genome that we observe today.

This paper is organized as follows: In the next section we discuss the statistical analysis of the DNA sequences and

*colliva@to.infn.it
†ropelleg@to.infn.it
‡testori@to.infn.it
§caselle@to.infn.it

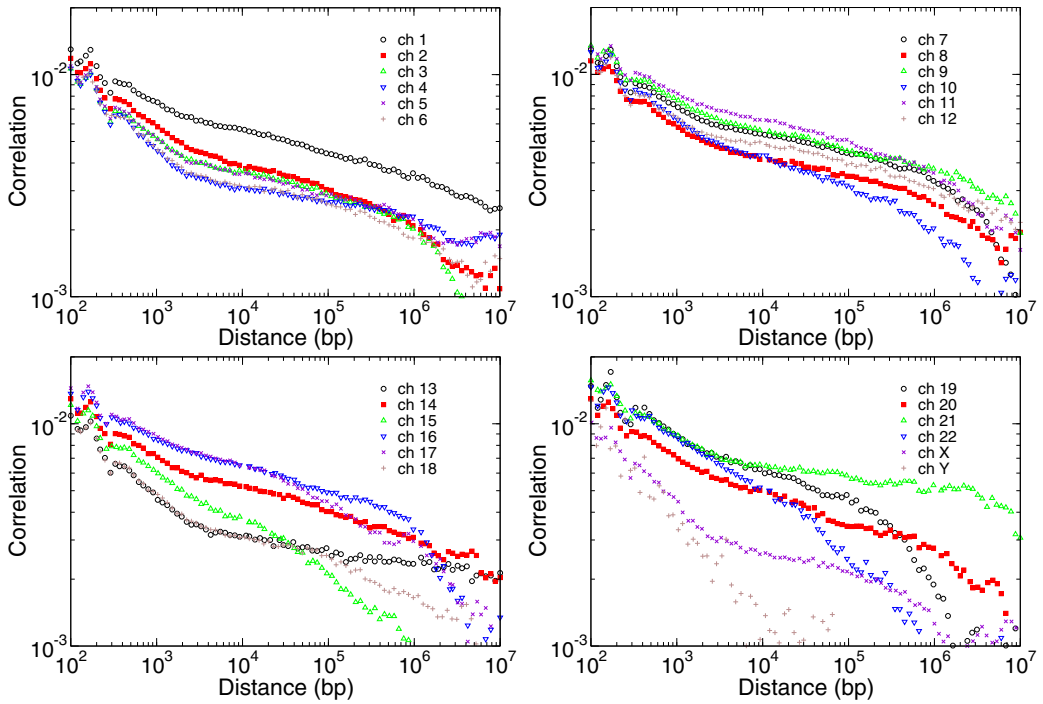

FIG. 1. (Color online) The combination $\Gamma_{++}(d) + \Gamma_{--}(d)$ for the human chromosomes on a log-log scale. Notice the similarity among the different curves and the wide range of validity of the power-law behavior.

recover the large-distance correlations mentioned above. In the third section we propose our model, map it into the 1D long-range Ising model and discuss its main properties. The fourth section is devoted to a tentative biological interpretation of our results while the last one is devoted to a few concluding remarks.

II. SEQUENCE ANALYSIS

A. DNA correlators

We computed the base-base correlation function along the lines discussed for instance in Refs. [14,15].

We defined a map from the four-letter alphabet to a binary set as follows:

$$\begin{cases} \{A,T\} \to -\\ \{C,G\} \to +. \end{cases} \tag{1}$$

We denote in the following these pairs of nucleotides as "spins" $\sigma = \pm 1$. This identification greatly simplifies the analysis while keeping the full complexity of long-range correlations of the genome and was adopted also in previous analyses of these correlations [3].

We computed the correlation function at a given distance d using a frequency-count estimator [14,15]:

$$\Gamma_{\alpha\beta}(d) = \frac{N_{\alpha\beta}(d)}{N} - \frac{N_\alpha}{N}\frac{N_\beta}{N}, \quad \alpha,\beta \in \{+,-\}, \tag{2}$$

where N is the total length of the sequence, $N_{\alpha\beta}(d)$ is the number of occurrences of α and β at distance d and N_α denotes the total number of spins of type α. Given the symmetries of the system it is enough for our purpose to compute only the positive correlators in which $\alpha = \beta$. The curves that we obtained for the combination $\Gamma_{++}(d) + \Gamma_{--}(d)$ are plotted on log-log scale in Fig. 1 for various human chromosomes.

Looking at the figures it is easy to identify three regimes. A short-range regime, below 1 kilobase (kb), which is dominated by the fine structure of the sequence (regulatory regions, correlations induced by nucleosomes, codon bias, ...), an intermediate regime between 2×10^3 and 10^5 (or 10^6 for the longest chromosomes) base pairs (bp) where the slopes of the correlators show a remarkable degree of similarity among different chromosomes and a rather clear power-law behavior of the type

$$\Gamma_{++}(d) \sim d^{-\gamma} \tag{3}$$

can be identified, and a large-distance region for $d > 10^6$ bases in which the correlation function drops drastically and no evidence of a universal behavior can be found. In the following we shall concentrate on the intermediate region. Our goal will be to construct an evolutionary model able to reproduce the observed power-law correlators.

It is natural to identify these three regimes with those which are typically observed in correlators of standard statistical mechanics models in the vicinity of a critical point [think, for instance, of the two-dimensional (2D) Ising model as an example]: a short-range regime which is dominated by "lattice artifacts" and depends on the precise microscopic definition of the model, a large-distance regime for distances larger than the correlation length whose behavior is dominated by the spectrum of the theory in which the correlation function decreases exponentially, and an intermediate "universal" regime in which the correlation function decreases with a power law and is dominated by the nearby critical point (and for this reason is universal, i.e., only depends on the universality class of the critical point).

The only nontrivial point of this identification is that, as is well known, no critical behavior (i.e., no long-range correlations) may exist in one-dimensional equilibrium statistical models with short-range interactions. This is the first indication that we shall have to consider in our analysis of one-dimensional models with long-range interactions. We shall come back to this point in the next section.

B. Power-law fitting

We fit $\Gamma_{++}(d)$ with Eq. (3) for all the chromosomes. In order to test the stability of the results with respect to the range of distances included in the fits we performed two different sets of fits, choosing first a conservative range $2\times10^3 < d < 10^5$ and then a much larger window $5\times10^2 < d < 10^6$.

The values of the scaling exponents extracted from the fits are reported in Table I and, for the more conservative choice of distances, in Fig. 2. Comparing the two columns of Table I, one may obtain a rough estimate of the systematic uncertainties induced by the fitting range.

It is interesting to notice that, with the exception of chromosomes 15, 22, and Y, the values of γ obtained with the more conservative choice of fitted distances are all contained in a rather narrow interval [0.06–0.13]. With the other choice of distances the range of values of γ increases, but it remains confined within the range [0.06–0.17]. The nontrivial behavior of chromosomes 15, 22, and Y had been already noticed and discussed in Ref. [12].

TABLE I. Values of the scaling exponents extracted from the fits. The first set of values, denoted by γ_1 correspond to the more conservative range of distances $2 \times 10^3 < d < 10^5$, while the second set of data, denoted by γ_2, corresponds to the larger range of fit distances $5 \times 10^2 < d < 10^6$. The statistical errors of the fits are reported in parentheses whereas the comparison between the two columns gives an idea of the systematic uncertainties induced by the fitting range.

Chromosome	γ_1	γ_2
1	0.0948(2)	0.1045(2)
2	0.1158(7)	0.1353(5)
3	0.0901(3)	0.1137(4)
4	0.0754(8)	0.0825(8)
5	0.0940(6)	0.1060(6)
6	0.0840(11)	0.1139(10)
7	0.0786(2)	0.0950(3)
8	0.0876(9)	0.1014(7)
9	0.0919(7)	0.0964(6)
10	0.1226(7)	0.1524(5)
11	0.0929(4)	0.1109(2)
12	0.0724(5)	0.0958(4)
13	0.0841(11)	0.0822(9)
14	0.0956(6)	0.1160(3)
15	0.1899(18)	0.2451(9)
16	0.1040(8)	0.1192(4)
17	0.1291(12)	0.1638(6)
18	0.0874(5)	0.1315(7)
19	0.1276(3)	0.1720(8)
20	0.1317(11)	0.1324(6)
21	0.0640(10)	0.0658(6)
22	0.2499(17)	0.2868(15)
X	0.0965(1)	0.1650(3)
Y	0.2970(16)	0.3650(8)

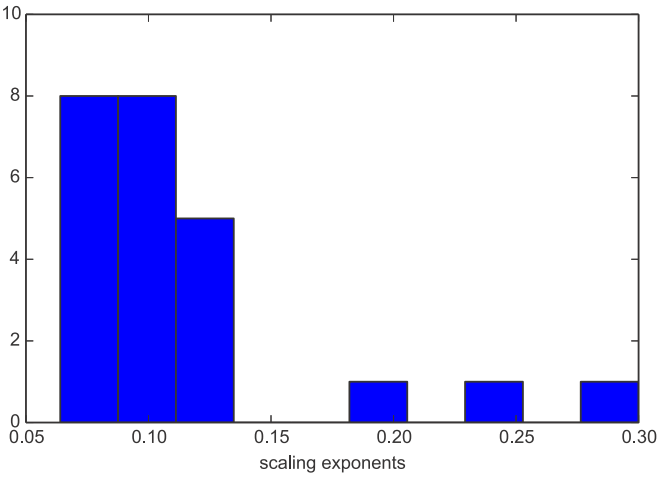

FIG. 2. (Color online) Distribution of the scaling exponents γ.

C. Stability of power-law behavior under sequence coarsening

A very interesting and nontrivial feature of the DNA correlations that we are studying is that their power-law behavior is very stable against renormalization. We implement the renormalization transformation by using a simple majority rule, coarse-graining the sequence and substituting each window with a sign chosen with a majority rule. The result of this process is represented in the case of chromosome 1 in the upper part of the graph in Fig. 3 for various sizes of the renormalization window. All the other chromosomes show essentially the same behavior. It is easy to see that, up to window sizes of 100 bp, nothing changes and that only for window sizes of 1000 bp one can observe some finite-size effect at short scale which disappears at larger distances where

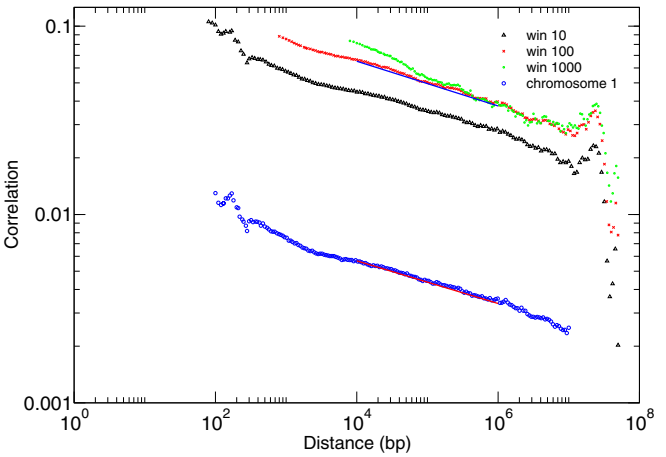

FIG. 3. (Color online) Comparison of base-base correlations in chromosome 1 (lower part of the figure, blue circles), with their renormalized version (in the upper part of the figure) obtained with a renormalization step of 10 bp (black triangles), 100 bp (red crosses), and 1000 bp (green dots).

the original exponent of the power-law decay is recovered also in this case. From a statistical mechanics point of view, this remarkable stability tells us that the original sequence is already very near to a critical point and that, irrelevant operators (i.e., subleading exponents), if present, should have an almost negligible coupling. From a biological point of view this is telling us that, more than the single nucleotide, the basic sequence element driving the observed correlations are sequences of intermediate length (from a few tens up to a few hundreds of bases) with a small AT or CG bias. We shall come back to this point in the last part of the paper.

III. THE MODEL

The model that we propose is very simple. It is defined on a one-dimensional lattice. At each time step, we

(1) randomly choose a spin σ_i (i.e., a nucleotide) from the lattice;

(2) randomly choose a second spin σ_j and fix its sign to be the same of σ_i with probability p_+ (p_-) if $\sigma_i = +1$ ($\sigma_i = -1$) defined as

$$p_\pm = \frac{e^{\beta|i-j|^{-\alpha}\pm h}}{e^{\beta|i-j|^{-\alpha}\pm h} + e^{-(\beta|i-j|^{-\alpha}\pm h)}}, \tag{4}$$

where $|i - j|$ denotes the distance between i and j.

β, α, and h are parameters which we shall discuss later but they will be always such that we may safely approximate the probability as

$$p_\pm \sim \frac{1}{2} + \frac{\beta}{|i-j|^\alpha} \pm h, \tag{5}$$

that is, the sum of a pure random choice with a drift plus an excess probability $p \sim \frac{\beta}{|i-j|^\alpha}$ to align the two spins between them.

These two steps define a Markov chain which has as stationary state the probability distribution of the 1D long-range Ising model defined by the following Hamiltonian:

$$\mathcal{H} = -\sum_{xy} J(x-y)^{-\alpha}\sigma_x\sigma_y - h\sum_x \sigma_x. \tag{6}$$

More precisely, the above steps define one of the possible choices for a Monte Carlo algorithm which simulates this particular model.[1]

This model is very interesting since it is the simplest example of a one-dimensional spin model with a critical point and has been the subject of considerable theoretical efforts in the last 50 years. Its phase diagram is rather complex and depends on the parameter α which must be greater than one to have a well-defined finite expression for the interaction energy. As α increases the model is characterized by three different behaviors:

(i) For $1 < \alpha < 1.5$ the model admits a second-order phase transition for $h = 0$ and for a critical value $\beta_c(\alpha)$ which depends on the precise value of α. For $\beta > \beta_c$ the Z_2 symmetry of the model is spontaneously broken and the system is characterized by a nonzero magnetization. Using standard renormalization group analysis [17], it can be shown that the critical point belongs to the mean-field universality class.

(ii) For $1.5 < \alpha < 2$ the model still has a second-order phase transition, but the universality class is not any more of the mean-field type. The critical exponents vary as functions of α.

(iii) For $\alpha > 2$ the system behaves as a short-range Ising model and since the lattice is one dimensional, there is no more a phase transition and the Z_2 symmetry is unbroken for any finite value of β.

The most important result for the scope of the present paper is that, due to the continuous nature of the phase transition, in the range $1 < \alpha < 2$ in the vicinity of the critical point we expect long-range correlations between spins. These correlations are controlled by the scaling dimension d_ϕ of the spin operator

$$\langle \sigma_i \sigma_j \rangle \sim \frac{1}{|i-j|^{2d_\phi}}. \tag{7}$$

The scaling dimension d_ϕ depends on α. In the mean-field regime, where it can be evaluated analytically, the relation is very simple: $d_\phi = 1 - \alpha/2$, which implies

$$\gamma \equiv 2d_\phi = 2 - \alpha, \tag{8}$$

and leads to values of γ in the range $1 > \gamma > 0.5$, i.e., larger than those which we have seen are typical of genomic correlators. Thus we are bound to study the model in the non-mean-field regime. Very few results are known exactly outside the mean-field regime but, remarkably enough, it can be shown using the ϵ expansion that d_ϕ is not renormalized up to the third order. Indeed, it is commonly believed that it should keep its mean-field value to all orders in the ϵ expansion.[2] This is exactly what we need to fix the value of α in our model. In order to match the observed genomic correlations α should range in the region $1.9 > \alpha > 1.7$ which gives for the anomalous dimensions values in the range $0.3 > \gamma > 0.1$ where most of the genomic correlators lie.

All the above considerations descend from well-known results on the 1D long-range Ising model. What is probably less known is that (contrary to the intuition we have from the short-range Ising models), this model, due to the peculiar long-range interaction term, can sustain long-range correlations in a very robust way without the need to fine tune the value of the two relevant coupling constants (β and h) to the critical value. This is due to the fact that, as we leave the critical point, the correlation length decreases much more slowly than in the short-range model, leaving a large window within which the spin-spin correlator decreases with the same power law of the critical point. For instance in the short-range 2D Ising model, for values of h such that the magnetization is of the

[1] In particular, the choice of Eq. (4) defines the so-called "heat-bath" algorithm. Notice, as a side remark, that if one is actually interested in simulating the model, the heat bath is not the best option for a nonlocal model like this one and that cluster-based models such as the one discussed in Ref. [16] are much more efficient.

[2] Notice that this result holds only in the one-dimensional case. In more than one dimension it can be shown that it holds only up to the value of α for which d_ϕ reaches the values it has in the corresponding short-range Ising model. See Ref. [18] for an updated review of these results.

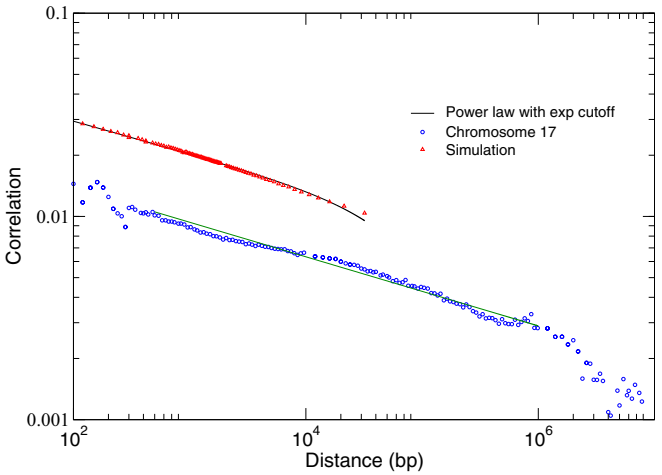

FIG. 4. (Color online) Comparison of the base-base correlations in the chromosome 17 (lower part of the figure, blue circles) with the result of a simulation (upper part of the figure, red triangles) of the long-range Ising model with $\alpha = 1.75$, $h = 0$, and $\beta > \beta_c$. The mean magnetization is ∼5% in both cases.

order of 5% the correlation length would be of few lattice spacings while in the long-range Ising model it reaches 10^5 lattice spacings. To support this observation we performed extensive simulations for $h \neq 0$ and $\beta \neq \beta_c$ by using the very efficient cluster algorithm of Ref. [16]. We report as an example in Fig. 4 the result of a simulation performed at $h = 0$ and $\alpha = 1.75$ in the broken-symmetry phase of the model, choosing the value of β so as to have a mean magnetization of 5% (red triangles in the figure). We see that the correlation length is of the order of 2×10^5 bp and that in the range 10 to 10^5 a power-law behavior with a value of the exponent (which we extracted from the simulations using exactly the same protocol which we used for the real DNA sequences) $\gamma \sim 0.16$ which is only slightly smaller than the one $\gamma = 2 - \alpha = 0.25$ predicted (and observed) at the critical point. We plot in the same figure for comparison the correlator of the chromosome 17 (which was chosen only because it has a value of γ similar to the one obtained in the simulation). The main lesson that we learn from these simulations is that, due to the long-range correlators in the Hamiltonian, the model is very robust, i.e., it is characterized by a large scaling region with correlation lengths which, even for values of the magnetization similar to the ones observed in the real sequences, reach hundreds of kilobases and with values of γ slightly smaller than the critical ones, but of the same order of magnitude.

IV. BIOLOGICAL MODELS

As we mentioned in the introduction, in the past 20 years there were several attempts to understand long-range correlations by linking them to other biological or physical characteristics of the DNA chain. These proposals range from the association to the bending [6] or elastic [7] properties of the DNA chain to the hierarchical organization of chromosomes [8,9], to the thermodynamic properties of DNA loops [10], to the role of folding and more generally of the three-dimensional (3D) structure of chromatin [4] or of transposable elements insertions in the E-R type of models [3,14]. Among these proposals these last two are particularly interesting for us due to the nonlocal nature of chromatin contacts and transposon insertions. Thanks to the recent advances in the study of the 3D structure of chromatin and of transposons classification [19] we are now in the position to test them in a more quantitative way.

A. Chromatin contacts

A possible role of chromatin contacts to explain DNA correlations was proposed very early [4] but was then somehow abandoned when it was realized that the scaling exponents of DNA correlations were much smaller than the typical scaling exponents of chromatin contacts that random polymer models would suggest. In the past few years thanks to the remarkable advances in chromosome conformation capture (3C) and Hi-C technologies [20,21] these chromatin contact exponents were observed experimentally and the gap with the DNA correlation scaling exponents was indeed confirmed. However, we have seen above that the 1D Ising model could nicely explain, thanks to Eq. (8), this gap. This prompted us to reanalyze in more detail this proposal.

Hi-C technologies allow us to obtain detailed genome-wide information on the physical contacts among distant genomic regions. The idea emerging from these experiments is that chromatin has a complex, "scale-free-like" organization across a range of spatial scales. The most impressive results of these studies has been the discovery of the so-called topological associated domains (TADs) [22,23] which are domains characterized by enriched levels of DNA-DNA contacts with an average contact probability $P_c(s)$, which decreases as a function of the genomic separation approximately as a power law, $P_c(s) \sim s^{-\alpha_{\rm TAD}}$, in the 0.5–7 Mb range. The values of $\alpha_{\rm TAD}$ depend rather strongly on the cell line and condition, ranging for instance from $\alpha_{\rm TAD} \sim 1.6$ for embryonic stem cells to $\alpha_{\rm TAD} \sim 1.1$ for lymphoblastoid cells in the interphase (see Ref. [24] for a review). Several models have been proposed to describe this behavior. Among the others an interesting proposal is the strings and binder switch (SBS) model [24] which describes the chromosomes as self-avoiding polymer chains with binding sites for diffusing molecules which mediate the DNA-DNA interactions. In order to create such an interaction the two portions of the chromosome should share the same binding sequence. The evolutionary process which leads to the formation of these pairs of similar binding sequences is very similar to the one we discussed above. In order to create a contact, the two binding sequences should evolve so as to become similar. The probability of this event to occur decreases as a function of the distance along the DNA chain following a power law exactly as in Eq. (5). Moreover depending on the CG content of the binding sequence we may have an overall drift modelled by the h term in Eq. (5). This identification is appealing, but it faces at least three problems:

(i) The TAD exponents show a cell line variability larger than the scaling exponents of the DNA correlations.

(ii) The TAD exponents are in general slightly smaller than what would be needed to explain DNA correlations.

(iii) The renormalization analysis discussed in Sec. II C suggests that the typical size of the sequences driving the

correlations should be of the order of 100 bp, much larger than the typical size of a transcription factor binding sequence.

There are a few possible directions which one could follow to address these issues. Let us briefly discuss them:

(i) TADs are likely to be the results of the action of different families of binding proteins. In each cell line only a subset of them is expected to be present, leading to a wide variety of TAD scaling exponents, while in the DNA sequence we see the signature of the union of all the corresponding binding sequences.

(ii) Our numerical experiment (see Sec. III) shows that, as we leave the critical point, we measure an "effective" scaling exponent slightly smaller than the one predicted by Eq. (8). This could explain the gap between the observed values of γ and those obtained from $\gamma = 2 - \alpha_{\mathrm{TAD}}$

(iii) It is well known that a few specific families of transposable elements (TEs) (we discuss in detail TEs in the next section) can bind in a specific way some transcription factors [25] and are one of the tools the cell uses to convey genome-wide combinatorial regulation [26]. If this holds true also for the proteins mediating the DNA-DNA interactions then we could identify their binding sequences with specific families of transposable elements whose typical size is exactly the one predicted by the renormalization analysis.

Overall we can say that a link between TAD organization and DNA correlations is plausible but that for the moment we are far from proving it. It is likely that future progress in understanding 3D chromatin contacts will also shed some light on this problem.

Before ending this section let us stress that direct chromatin contacts is not the only way one can use Hi-C results to explain DNA correlations. Another mechanism which could also lead to long-range sequence similarities is the colocalization of coregulated genes (see, for instance, the recent study in the case of the human chromosome 19 in Ref. [27]). This coregulation requires the presence of common regulatory sequences which, similarly to their target genes, should colocalize along the chromosome. Also in this case, however, in order to substantiate this hypothesis one should identify these regulatory sequences and test their power-law behavior.

The important observation for the scope of the present paper is that in both cases we expect that, if such a power decay exists, it should be driven by α_{TAD} which is much larger than the exponent γ of the DNA correlators. Our model, and in particular the relation $\gamma = 2 - \alpha$ of Eq. (8), offers a nice explanation of this gap.

B. Retrotransposon insertion as possible source of long-range correlations

Transposable elements, thanks to their repetitive sequence and their widespread presence in the genome are natural candidates as putative drivers of long-range DNA correlations. Moreover, it was reported in Refs. [28,29] that the probability distribution $P(s)$ of the intertransposon distance s follows a power-law behavior $P(s) \sim s^{-\mu}$. We devote this section to a critical examination of this issue.

Transposable elements (transposons) represent almost 45% of the human genome (for a review see, for instance, Ref. [30]). They are genetic elements which are able to duplicate themselves and insert in a (almost) random way in the hosting genome. They strongly contributed to shape the genome of all higher eukaryotes. Among the different transposon families a special role is played in the human case by the Alu and the LINE families which comprise 10% and 20%, respectively, of the human genome. LINEs are AT rich autonomous retrotransposons. Their full length is of about 3 kb, but most of their copies in the genome are truncated and thus the mean length is only of 430 bp. They are very successful transposons and are present in almost 1.5×10^6 copies in the human genome. Alus are nonautonomous elements which are retrotransposed by the LINE machinery, they are about 300 bp long, CG rich, and are probably the most successful nonautonomous transposons in the primate lineage with more than 10^6 copies in the human genome.

There are indeed a couple of features of transposable elements that support their possible role in driving DNA correlations:

(i) The renormalization analysis discussed in Sec. II C suggests that the typical size of the sequences driving the correlations should be of the order of 100–1000 bps, which is exactly the typical size of most of the existing transposable elements in the genome.

(ii) Transposon insertion can be described by using a nonlocal Markov process with some analogies with the one that we proposed in Sec. III. If we identify the first "spin" as one particular transposon, a new insertion of the same transposon would represent a "successful" step of the Markov process in which the second spin is aligned with the first one with an excess probability with respect to a random choice. There are, however, two main differences with respect to our model. First, while in the transposon case we have the insertion of a new spin and the total length of the sequences increases, in our case we substitute an existing spin with a new one and the total length is fixed. Second, while transposon insertion is certainly a nonlocal process, there is no obvious reason (except for the observation of Refs. [28,29] which, however, is limited to the short-scale behavior of intertransposon separation) to assume that the insertion probability should follow a power law as required in our model.

To test the role of transposons we performed the same correlation analysis discussed above in two artificial sequences obtained (using the REPEATMASKER software [19]) first by removing all the transposons (which were substituted with zero in the sequence thus giving no contribution to the correlators) and, second, by keeping only the transposons and substituting with zero the rest of the sequence. The results of the analysis are reported in Fig. 5.

Looking at the figure it is easy to see that the correlations survive almost unchanged when the transposons are removed and that the artificial chromosome made of only transposable elements shows no evidence of a long-range correlation beyond 10 kb, which is the typical scale of nearest-neighbor insertions. Together these two observations seem to exclude a role of transposon insertions in generating long-range correlations. However, it is important to stress that the above test involved the entire population of transposable elements as a whole. In view of the discussion of the previous section we cannot exclude that specific subfamilies of transposons could

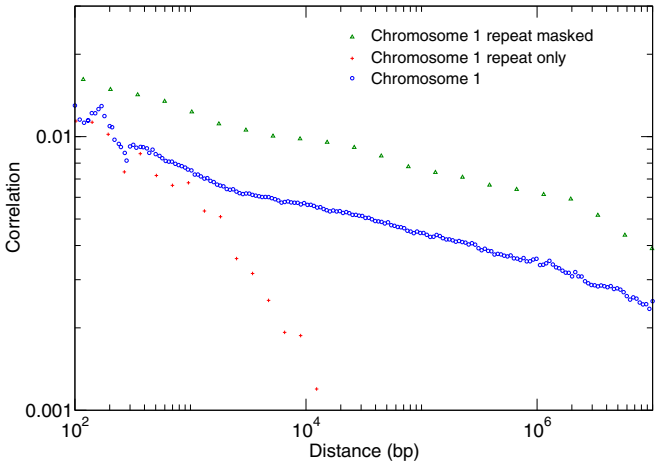

FIG. 5. (Color online) DNA correlations in chromosome 1 (blue circles) and in two artificial sequences obtained by eliminating transposons form the chromosome (green triangles), and by keeping only the transposons (red pluses). It is clear that DNA correlations are not affected by the elimination of transposons, while the artificial chromosome made of only transposable elements shows no evidence of a long-range correlation beyond 10 kb.

instead play a role both in the formation of TADs and of DNA correlations.

V. CONCLUDING REMARKS AND OPEN ISSUES

In this paper we show that long-range DNA correlations can be described rather well by the 1D long-range Ising model and that this description is rather robust, i.e., it does not need a specific fine tuning of the model parameters. We also show that a simple evolutionary model in which distant portions of DNA tend to have identical DNA sequences can be mapped to a Markov model which has the 1D long-range Ising model as stationary point.

There are a few open issues which we think should require further studies:

(i) As mentioned in the Introduction, the optimal description of the long-range nucleotide correlations could probably be achieved by a suitable combination of the expansion-randomization model with our 1D Ising proposal. To this end it would be important to evaluate the changes in the duplication and insertion rates with time and across the different species. Thanks to the increasing amount of sequencing data it is likely that precise measures of these rates will soon be available and will allow to tune the interplay between the two models.

(ii) It would be interesting to extend the present analysis to other organisms. Thanks to NGS studies we have now a rather precise knowledge of the retrotransposon repertoire of several organisms [31,32] and this could allow to test our results in a more extensive way.

(iii) It would also be important to further address the role of single families of transposable elements in this context. While the tests discussed in Sec. IV B exclude a role of repeated elements as a whole, one cannot exclude a possible role of selected subsets of transposons in shaping DNA correlations.

It is likely that, in the near future, thanks to the ongoing experimental efforts both on the sequencing side and on the reconstruction of the 3D structure of interphase chromosomes, much more data will be available to address these issues and to deepen our understanding of the physical and evolutionary constraints which shaped the genomes of higher eukaryotes.

ACKNOWLEDGMENTS

We thank E. Domany, C. Micheletti, M. Osella, and A. Rosa for useful discussions and suggestions. This work was partially funded by the FSP Grant GeneRNet.

[1] C.-K. Peng *et al.*, Long-range correlations in nucleotide sequences, Nature (London) **356**, 168 (1992).
[2] W. Li, Expansion-modification systems: A model for spatial $1/f$ spectra, Phys. Rev. A **43**, 5240 (1991).
[3] P. W. Messer, P. R. Arndt, and M. Lassig, Solvable sequence evolution models and genomic correlations, Phys. Rev. Lett. **94**, 138103 (2005).
[4] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, Crumpled globule model of the 3-dimensional structure of DNA, Europhys. Lett. **23**, 373 (1993).
[5] S. V. Buldyrev *et al.*, Fractal landscapes and molecular evolution: Modeling the myosin heavy chain gene family, Biophys. J. **65**, 2673 (1993).
[6] D. S. Goodsell and R. E. Dickerson, Bending and curvature calculations in B-DNA, Nucl. Acids Res. **22**, 5497 (1994).
[7] C. Vaillant, B. Audit, C. Thermes, and A. Arneodo, Influence of the sequence on elastic properties of long DNA chains, Phys. Rev. E **67**, 032901 (2003).
[8] B. Audit *et al.*, Long-range correlations in genomic DNA: A signature of the nucleosomal structure, Phys. Rev. Lett. **86**, 2471 (2001).
[9] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes, Long-range correlations between DNA bending sites: Relation to the structure and dynamics of nucleosomes, J. Mol. Biol. **316**, 903 (2002).
[10] C. Vaillant, B. Audit, and A. Arneodo, Thermodynamics of DNA loops with long-range correlated structural disorder, Phys. Rev. Lett. **95**, 068101 (2005).
[11] M. V. Koroteev, J. Miller, Scale-free duplication dynamics: A model for ultraduplication, Phys. Rev. E **84**, 061919 (2011).
[12] W. Li and D. Holste, Universal $1/f$ noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome, Phys. Rev. E **71**, 041910 (2005).
[13] W. Li and D. Holste, Spectral analysis of guanine and cytosine fluctuations of mouse genomic DNA, Fluct. Noise Lett. **04**, L453 (2004).
[14] W. Li, The study of correlation structures of DNA sequences: A critical review, Comput. Chem. **21**, 257 (1997).
[15] P. W. Messer and P. F. Arndt, CorGen–Measuring and generating long-range correlations for DNA sequence analysis, Nucl. Acids Res. **34**, W692 (2006).

[16] E. Luijten and H. Blote, Boundary between long-range and short-range critical behavior in systems with algebraic interactions, Phys. Rev. Lett. **89**, 025703 (2002).
[17] M. E. Fisher, S.-K. Ma, and B. G. Nickel, Critical exponents for long-range interactions, Phys. Rev. Lett. **29**, 917 (1972).
[18] M. Chiara Angelini, G. Parisi, and F. Ricci-Tersenghi, Relations between short-range and long-range ising models, Phys. Rev. E **89**, 062120 (2014).
[19] A. F. A. Smit, R. Hubley, and P. Green, REPEATMASKER Open-4.0. (2013-2015) *RepeatMasker Open-4.0.* http://www.repeatmasker.org
[20] E. Lieberman *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science **326**, 289 (2009).
[21] T. Nagano *et al.*, Single-cell Hi-C reveals cell-to-cell variability in chromosome structure, Nature (London) **502**, 59 (2013).
[22] J. R. Dixon *et al.*, Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature (London) **485**, 376 (2012).
[23] E. P. Nora *et al.*, Spatial partitioning of the regulatory landscape of the X-inactivation centre, Nature (London) **485**, 381 (2012).
[24] M. Barbieri *et al.*, Complexity of chromatin folding is captured by the strings and binders switch model, Proc. Natl. Acad. Sci. USA **109**, 16173 (2012).
[25] I. Jordan, I. Rogozin, G. Glazko, and E. Koonin, Origin of a substantial fraction of human regulatory sequences from transposable elements, Trends Genet. **19**, 68 (2003).
[26] A. Testori *et al.*, The role of transposable elements in shaping the combinatorial interaction of transcription factors, BMC Genom. **13**, 400 (2012).
[27] M. Di Stefano, A. Rosa, V. Belcastro, D. di Bernardo, and C. Micheletti, Colocalization of coregulated genes: A steered molecular dynamics study of human chromosome 19, PLoS Comput. Biol. **9**, e1003019 (2013).
[28] D. Sellis, A. Provata, and Y. Almirantis, Alu and LINE1 distributions in the human chromosomes: Evidence of global genomic organization expressed in the form of power laws, Mol. Biol. Evol. **24**, 2385 (2007).
[29] A. Klimopoulos, D. Sellis, and Y. Almirantis, Widespread occurrence of power-law distributions in inter-repeat distances shaped by genome dynamics, Gene **499**, 88 (2012).
[30] R. Cordaux and M. A. Batzer, The impact of retrotransposons on human genome evolution, Nat. Rev. Genet. **10**, 691 (2009).
[31] R. Kofler, A. J. Betancourt, and C. Schlötterer, Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in drosophila melanogaster, PLoS Genet. **8**, e1002487 (2012).
[32] A. S. Fiston-Lavier, M. G. Barrn, D. A. Petrov, and J. González, T-lex2, genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data, Nucl. Acids Res. **43**, e22 (2014).