Ising model description of Long Range correlations in DNA sequences.

A. Colliva\textsuperscript{a}, R. Pellegrini\textsuperscript{b}, A. Testori\textsuperscript{a}, M. Caselle\textsuperscript{a}

\textsuperscript{a} Dipartimento di Fisica dell’Università di Torino and I.N.F.N. sez. di Torino, Via Pietro Giuria 1, I-10125 Torino, Italy

\textsuperscript{b} Physics Department, Swansea University, Singleton Park, Swansea SA2 8PP, UK

e–mail: colliva(ropelleg)(testori)(caselle)@to.infn.it

Abstract

We model long range correlations of nucleotides in the human DNA sequence using the long range one dimensional Ising model. We show that for distances between $10^3$ and $10^6$ bp the correlations show an universal behaviour 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.
1 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, 3, 4, 5, 6] to explain this behaviour and compared with the growing collection of genomic data [6, 7, 8].

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.

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

- They show a power law behaviour, with exponents characterized by a remarkable degree of universality, with very small variations across the human chromosomes [7] and between the human and mouse genomes [8].

- In the human case these exponents take values in the range 0.05-0.30 and show two rather non trivial features
  - they are correlated with the GC content of the chromosomes (see fig.2).
  - i.e. they are higher for the chromosomes which contain a larger amount of C and G nucleotides.
  - they are much smaller than any other power law exponent of correlators of genomic components (we shall explain this observation in sect.4 with a few examples).

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 modelling 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 fulfills these conditions is the so called "expansion-randomization" (E-R) model proposed by Li in 1989 [2] and solved exactly in [3]. The stationary state of the model is characterized by the expected long range correlations and is largely independent from microscopic details. [3]. The only problem of this model is that in order to match the observed exponents it requires a duplication rate much higher than the one 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 retrotransposons insertions. Thus it is likely that the E-R model should be complemented with some
other evolutionary process able to enforce long range correlation without requiring an expanding genome. In particular the unusual range of these correlations suggests the introduction of non-local interactions in the evolutionary process.

Following this line of reasoning in this paper we propose an evolutionary model based on non-local 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 non-local 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 Ising model", for which several exact and approximate results exist. In particular, differently from the ordinary (short range) Ising model, this model admits a critical point also in one dimension and in the neighbourhood 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 one 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 shall discuss the statistical analysis of the DNA sequences and recover the large distance correlations mentioned above. In the third section we shall 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.

2 Sequence Analysis

2.1 DNA correlators

We computed the base-base correlation function along the lines discussed for instance in [9, 10]. We defined a map from the 4 letter alphabet to a binary set as follows

\[ M = \begin{cases} 
\{A, T\} \rightarrow - \\
\{C, G\} \rightarrow + 
\end{cases} \]  

(1)

We shall 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 [9][10].

\[ \hat{\Gamma}_{\alpha\beta}(d) = \frac{N_{\alpha\beta}(d)}{N} - \frac{N_{\alpha}}{N} \frac{N_{\beta}}{N} \quad \alpha, \beta = \{+, -\} \]  

(2)

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

![Figure 1: Correlation functions for the human chromosomes in a log-log scale. Notice the impressive range of validity of the power law behaviour.](image)

Looking at the figure 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 \( 10^3 \) and \( 10^6 \) base pairs (bp) where a rather clear power law behaviour of the type

\[
\Gamma_{\alpha\alpha}(d) \sim d^{-\gamma}
\]

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

2.2 Power law fitting

Looking at figure [1] we notice a remarkable degree of similarity in the slope of the power law curves. In order to quantify this point we fitted the scaling exponent in the $10^3 - 10^6$ interval for all chromosomes using the the package related to the work by Newman et.al. [11].

The results of the power law fitting analysis, performed on human chromosomes (1-22), are reported in fig. [3]. Their distribution as a function of the CG content of the chromosomes is plotted in fig.[2]

![Figure 2: Distribution of the scaling exponents $\gamma$ for the human chromosomes as a function of their CG content.](image)
2.3 Stability of the power law behaviour under sequence coarsening.

A very interesting and non trivial feature of the DNA correlations that we are studying is that their power law behaviour is very stable against renormalization. We implement the renormalization transformation 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 figures 4 for various sizes of the renormalization window. All the other chromosomes show essentially the same behaviour. 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 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.

3 The model

The model that we propose is very simple. At each time step
Figure 4: Comparison of base-base correlations (lower part of the figure) for the chromosome 1, with their renormalized version (in the upper part of the figure, see text for definitions).

1. Randomly choose a spin $\sigma_i$ (i.e. a nucleotide) from the lattice.

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

$$p_{\pm} = \frac{e^{\beta|i-j|^\alpha \pm h}}{e^{\beta|i-j|^\alpha + h} + e^{-(\beta|i-j|^\alpha \pm h)}}$$  \hspace{1cm} (4)

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

$\beta, \alpha$ and $h$ are parameters which we shall discuss later but they will be always such that we may safely approximate the probability as

$$p_{\pm} \sim \frac{1}{2} + \frac{\beta}{|i-j|^\alpha} \pm h$$  \hspace{1cm} (5)

i.e 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:

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

More precisely, the above steps define one of the possible choices for a Montecarlo algorithm which simulates this particular model.

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 $\alpha$ which must be greater than one to have a well defined finite expression for the interaction energy. As $\alpha$ increases the model is characterized by three different behaviours.

- 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 $\alpha$. For $\beta > \beta_c$ the $Z_2$ symmetry of the model is spontaneously broken and the system is characterized by a non-zero magnetization. Using standard renormalization group analysis it can be shown that the critical point belongs to the mean field universality class.

- 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 $\alpha$.

- 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 $\beta$.

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 correlation are controlled by the scaling dimension $d_{\phi}$ of the spin operator

$$<\sigma_i \sigma_j> \sim \frac{1}{|i - j|^{2d_{\phi}}}. \quad (7)$$

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

$$\gamma \equiv 2d_{\phi} = 2 - \alpha \quad (8)$$

and leads to values of $\gamma$ in the range $2 > \gamma > 1$, i.e. much 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 $\epsilon$ expansion that $d_{\phi}$ is not renormalized up to the third order. Indeed it is commonly believed

---

\footnote{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 non-local model like this one and that cluster based models like the one discussed in [12] are much more efficient.}
that it should keep its mean field value to all orders in the $\epsilon$ expansion\(^2\). This is exactly what we need to fix the value of $\alpha$ in our model. In order to match the observed genomic correlation $\alpha$ 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 of fine tuning the value of the two relevant coupling constants ($\beta$ 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 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 $\beta \neq \beta_c$ using the very efficient cluster algorithm of \[12\]. We report as an example in fig. the result of a simulation performed at $h = 0$ and $\alpha = 1.75$ in the broken symmetric phase of the model, choosing the value of $\beta$ 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 $2 \times 10^5$ bp and that in the range $10 - 10^5$ a power law behaviour 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 $\gamma$ 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 $\gamma$ slightly smaller than the critical ones, but of the same order of magnitude.

4 Biological models

We see two possible (maybe related) biological realizations of the Markov process discussed above.

4.1 Chromatin contacts

Recent experiments with 3C (Chromosome Conformation Capture) based technologies \[15, 16\] allowed 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

\(^2\)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 $\alpha$ for which $d_\phi$ reaches the values it has in the corresponding short range Ising model. See \[14\] for an updated review of these results.
Figure 5: Comparison of the base-base correlations in the chromosome 17 (lower part of the figure) with the result of a simulation of the long range Ising model with $\alpha = 1.75$, $h = 0$ and $\beta > \beta_c$. The mean magnetization is $\sim 5\%$ in both cases of spatial scales. The most impressive results of these studies has been the discovery of the so called Topological Associated Domains (TAD)\cite{TAD1, TAD2} 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, $P_c(s) \sim s^{-\alpha_{TAD}}$, in the 0.5 - 7 Mb range. The values of $\alpha_{TAD}$ depend rather strongly on the cell line and condition, ranging for instance from $\alpha_{TAD} \sim 1.6$ for embryonic stem cells to $\alpha_{TAD} \sim 1.1$ for lymphoblastoid cells in the interphase (see \cite{TAD3} for a review). Several models have been proposed to describe this behaviour. Among the others an interesting proposal is the Strings and Binder Switch (SBS) model \cite{TAD4} which describes the chromosomes as self-avoiding walks 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 led 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 is for the moment only an indirect evidence. In order to substantiate it one should first identify these binding sequences and then verify that they are actually distributed following a power law and this goal seems for the moment out of reach.

Another mechanism which is suggested by the results of 3C experiments and 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 [20]). 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 behaviour.

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 $\alpha_{TAD}$ which is much larger than the exponent $\gamma$ of the DNA correlators. Our model, and in particular the relation $\gamma = 2 - \alpha$ of eq. (8) offers a nice explanation of this gap.

4.2 Retrotransposon insertion

Another possible realization of the evolutionary process discussed above is represented by transposon insertion. Transposons represent 45% of the human genome (for a review see for instance [21]). They are genetic elements which are able to duplicate themselves and insert in an (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 the 10% and 20% respectively of the human genome. Lines are autonomous retrotransposons, they are about 3kb long and are AT rich. They are very successful transposons which are present in almost 1.5 $\times$ 10$^6$ copies in the human genome. Alus are non autonomous elements which are retrotransposed by the Line machinery, they are about 300 bp long, CG rich and are probably the most successful non autonomous transposon in the primate lineage with more than 10$^6$ copies in the human genome. A very interesting feature of transposons distribution in the human genome is that the probability distribution $P(s)$ of the inter-transposon distance $s$ seems to behave as a power law [22, 23] $P(s) \sim s^{-\mu}$. This is true not only for Alu and Lines [22] but seems to be a generic feature of all transposon families [24]. The critical indices of this power law show a large degree of variability depending on the transposon family (and also on the subfamilies in the case of Alus) but most of these indices are in the range $1 < \mu < 2$ [22, 23]. Transposon insertion seems again a nice realization of the model that we propose. If we choose the first "spin" as one particular transposon, a new insertion of a transposon (which is identical to the first one) would represent a "successful" step of our Markov process in which the second spin is aligned with the first one.

---

Transposons distribution along the genome is a very interesting problem in itself. We plan to address it in more detail in a future publication [24]. As for the scope of the present paper let us only notice that, as it is easy to see from fig. 6, the behaviour of $P(s)$ is better described by the combination of an effective power law at short and intermediate distances and an exponential cut-off at larger distances.
with an excess probability with respect to a random choice which decreases with
the distance between the two spins exactly as in eq. (5). Depending on the AT rich
or CG rich nature of the transposon we shall have a positive or negative sign of $h$
in eq. (5). In order to test this conjecture we reproduced the analysis of [22] combining
together all the Alus and all the Lines. We report our results together with the fits
in fig.6. It is interesting to notice that the two families of transposons (probably
due to the different size) follow rather different slopes with an index $\mu_A \sim 1.5$ for
Alus and $\mu_L \sim 1.6$ for Lines. Both indices are slightly below, but not too far from
what would be needed to match the values of $\gamma$ that we observed in the human
genome. Moreover, as we have seen above, moving outside the critical point in
the scaling region slightly modifies the power law behaviour leading to an effective
(measured) value of $\gamma$ slightly below the critical one and thus perfectly compatible
with what we observed in the human chromosomes. The small difference between
the two values allows us to perform a rather non trivial test of this conjecture. It
is well known that different chromosomes have different content of Alus and Lines.
In particular CG rich chromosomes are Alu rich and AT rich chromosomes are Line
rich. If we assume that long range correlations are induced by retrotransposon in-
sertion, then we would expect to find different values of $\gamma$ in different chromosomes
depending on their Alu/Line content. In particular, for CG rich chromosomes,
correlations are driven by Alu repeats and due to eq.8 we expect higher values of
$\gamma$ since $\mu_A < \mu_L$ and viceversa for AT rich chromosomes where correlations are
driven by Line repeats. As it can be seen in fig.2 this prediction turns out to be in
nice agreement with the distribution of $\gamma$ values in the human case.

5 Concluding remarks and open issues.
There are a few open issues which we think should require further studies:

- As we 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 rate
  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.

- 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 and this could allow to test
  our hypothesis in a more extensive way.

- It would be interesting to test if the binding sequences postulated by the SBS
  model could be identified with the transposons which we discussed above.
  Indeed it is well known that transposons can bind several transcription factors
  [25] and are one of the tools the cell uses to convey genome-wide combinatorial
  regulation [26]. Moreover, as we have seen, inter-transposon distances follow
  power laws [22, 23] with indices which are similar to the measured values of
  $\alpha_{TAD}$. These indices depend in a significative way on the transposon type
  [23] and thus in principle $\alpha_{TAD}$ could be tuned to the different values in
Figure 6: Distribution of inter-transposon distances for the Alu and Line transposons in a log-log scale. The distribution is the combination of an effective power law behaviour at short distances and an exponential cut-off at larger distances \cite{24}. The best fit values for the critical indices are $\mu_A = 1.48$ for the Alus and $\mu_L = 1.59$ for the Lines.

presence of transcription factors with specific affinity for specific families of transposons.

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 constraint which shaped the genomes of higher eukaryotes.

References

[1] Peng C, Buldyrev S, Goldberger A, Havlin S, Sciortino F, et al. (1992) Long-range correlations in nucleotide-sequences. Nature 356: 168-170.

[2] Li W (1991) Expansion-modification systems: A model for spatial $1/f$ spectra. Physical Review A 43: 5240–5260.

[3] Messer PW, Arndt PF, Lassig M (2005) Solvable sequence evolution models and genomic correlations. Physical Review Letters 94: 138103.

[4] Grosberg A, Rabin Y, Havlin S, Neer A (1993) Crumpled globule model of the 3-dimensional structure of DNA. Europhysics Letters 23: 373-378.
[5] Buldyrev S, Goldberger A, Havlin S, Peng C, Stanley H, et al. (1993) Fractal landscapes and molecular evolution - modeling the myosin heavy-chain gene family. BioPhysical Journal 65: 2673-2679.

[6] Koroteev MV, Miller J (2011) Scale-free duplication dynamics: A model for ultraduplication. Physical Review E 84: 061919.

[7] Li W, Holste D (2005) Universal $1/f$ noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome. Physical Review E 71: 041910.

[8] Li W, Holste D (2004) Spectral analysis of guanine and cytosine fluctuations of mouse genomic DNA. arXiv:q-bio/0411017; arXiv:q-bio/0411017v1.

[9] Li W (1997) The study of correlation structures of DNA sequences: a critical review. Computers & Chemistry 21: 257–271.

[10] Messer PW, Arndt PF (2006) CorGen–measuring and generating long-range correlations for DNA sequence analysis. Nucleic Acids Research 34: W692–W695.

[11] Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Review 51: 661–703.

[12] Luijten E, Blote H (2002) Boundary between long-range and short-range critical behavior in systems with algebraic interactions. Physical Review Letters 89: 025703.

[13] Fisher ME, Ma Sk, Nickel B (1972) Critical Exponents for Long-Range Interactions. Physical Review Letters 29: 917-920.

[14] Chiara Angelini M, Parisi G, Ricci-Tersenghi F (2014) Relations between Short Range and Long Range Ising models. Physical Review E 89: 062120.

[15] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326: 289-293.

[16] Nagano T, Lubling Y, Stevens T, Schoenfelder S, Yaffe E, et al. (2013) Single-cell hi-c reveals cell-to-cell variability in chromosome structure. Nature 502: 59–64.

[17] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376-380.

[18] Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485: 381-385.

[19] Barbieri M, Chotalia M, Fraser J, Lavitas LM, Dostie J, et al. (2012) Complexity of chromatin folding is captured by the strings and binders switch model. Proceedings of the National Academy of Sciences 109: 16173-16178.
[20] Di Stefano M, Rosa A, Belcastro V, di Bernardo D, Micheletti C. (2013) Colocalization of Coregulated Genes: A Steered Molecular Dynamics Study of Human Chromosome 19. PLoS Comp. Biol. 9: e1003019.

[21] Cordaux R, Batzer MA (2009) The impact of retrotansposons on human genome evolution. Nature Reviews Genetics 10: 691-703.

[22] Sellis D, Provata A, Almirantis Y (2007) Alu and LINE1 distributions in the human chromosomes: Evidence of global genomic organization expressed in the form of power laws. Molecular Biology and Evolution 24: 2385-2399.

[23] Klimopoulou S, Sellis D, Almirantis Y (2012) Widespread occurrence of power-law distributions in inter-repeat distances shaped by genome dynamics. Gene 499: 88-98.

[24] Riba A, Testori A, Fumagalli M, Caselle M, Osella M In Preparation.

[25] Jordan I, Rogozin I, Glazko G, Koonin E (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics 19: 68-72.

[26] Testori A, Caizzi L, Cutrupi S, Friard O, De Bortoli M, et al. (2012) The role of Transposable Elements in shaping the combinatorial interaction of Transcription Factors. BMC Genomics 13: 400.