Gene 499 (2012) 88–98

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Widespread occurrence of power-law distributions in inter-repeat distances shaped by genome dynamics

Alexandros Klimopoulos [a], Diamantis Sellis [b], Yannis Almirantis [a,*]

[a] National Center for Scientific Research "Demokritos," Institute of Biology, 153 10 Athens, Greece
[b] Department of Biology, Stanford University, Stanford, CA 94305-5020, USA

ARTICLE INFO

Article history:
Accepted 6 February 2012
Available online 18 February 2012

Keywords:
transposable elements
power-law distribution
genome evolution
fractal globule

ABSTRACT

Repetitive DNA sequences derived from transposable elements (TE) are distributed in a non-random way, co-clustering with other classes of repeat elements, genes and other genomic components. In a previous work we reported power-law-like size distributions (linearity in log–log scale) in the spatial arrangement of Alu and LINE1 elements in the human genome. Here we investigate the large-scale features of the spatial arrangement of all principal classes of TEs in 14 genomes from phylogenetically distant organisms by studying the size distribution of inter-repeat distances. Power-law-like size distributions are found to be widespread, extending up to several orders of magnitude. In order to understand the emergence of this distributional pattern, we introduce an evolutionary scenario, which includes (i) *Insertions* of DNA segments (e.g., more recent repeats) into the considered sequence and (ii) *Eliminations* of members of the studied TE family. In the proposed model we also incorporate the potential for transposition events (characteristic of the DNA transposons' life-cycle) and segmental duplications. Simulations reproduce the main features of the observed size distributions. Furthermore, we investigate the effects of various genomic features on the presence and extent of power-law size distributions including TE class and age, mode of parental TE transmission, GC content, deletion and recombination rates in the studied genomic region, etc. Our observations corroborate the hypothesis that insertions of genomic material and eliminations of repeats are at the basis of power-laws in inter-repeat distances. The existence of these power-laws could facilitate the formation of the recently proposed "fractal globule" for the confined chromatin organization.

1. Introduction

Interspersed repeats are mostly inactive repetitive sequences derived from initially active transposable elements (TEs), which are found in almost all eukaryotic genomes (Jurka, 1998). In some species they account for a large percentage of the genome's size, as in human ~45% (Lander et al., 2001; Makalowski, 2003), opossum ~52% (Gentles et al., 2007) and mouse ~38% (Waterston et al., 2002).

In a previous work (Sellis et al., 2007) we investigated the large-scale features of Alu and L1 spatial arrangement in the human genome by studying the size distribution of inter-repeat distances. In most cases, we found power-law-like size distributions often spanning several orders of magnitude. We also proposed a model for the emergence of this type of distributions consisting of neutral processes of genomic evolution based on TE turnover.

Here, we extend this study into 14 genomes from various taxonomic groups, for TE families from every major TE class. After detecting power-law-like inter-repeat size distributions in several cases, we focus on the correlation between distribution features and various genomic properties. Thus, we attempt to understand the dynamics at the origin of this particular pattern and using computer simulations test the plausibility of the proposed model, as extended here in order to apply for both retroelements and DNA transposons. Two questions may be formulated: (a) whether this model can explain the observed distributions and (b) why in some cases power-laws are more extended, while in others linearity is rudimentary or completely absent. We will address again these questions in the final section.

1.1. Background information on genomic clustering of transposable elements

The proliferation of transposable elements interacts in a complex way with other aspects of whole-genome evolutionary dynamics, improving in several cases genome functionality (see Bowen and Jordan, 2002; Jurka, 2008, and references given therein). Repeats can be

Abbreviations: TE(s), Transposable element(s); SIM, Subsequently Inserted genomic Material. All repeated DNA incorporated in the genome after the proliferation of a TE subfamily; I-D skew, (D-I)/(D+I), where D and I are the populations of Direct and Inverted repeat pairs respectively; FrPL, All cases showing a power law over all examined cases for a given organism and a given TE class; ME, Mean Extent of the linear region of a distribution in log-log scale.

* Corresponding author. Tel.: +30 2106503619; fax: +30 2106511767.
E-mail addresses: aklimop@bio.demokritos.gr (A. Klimopoulos), dsellis@stanford.edu (D. Sellis), yalmir@bio.demokritos.gr (Y. Almirantis).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.02.005

partially incorporated into coding sequences (Deininger and Batzer, 1999; Volff and Brosius, 2007; Abrusan et al., 2008). As shown in both cases of retrotransposition, by either retrotransposons or rertotransposed genes (Okamura and Nakai, 2008) new promoters are distributed into the genome, thus modifying its regulatory pattern (Medstrand et al., 2005). Repeats may also be recruited for playing an often unknown functional role, as there are repeats known to form groups of conserved not expressed (CNE) sequences (see e.g., Xie et al., 2006; Lowe et al., 2007). It is very probable that the proliferation of several TE families has provided a variety of advantages to host genomes, without positive selection of newly transposed copies. In the well-studied case of the distribution of the old Alus, which in the human genome is clearly skewed towards high-GC content regions, it has been shown (Brookfield, 2001) that this GC preference cannot be due to positive selection (for comprehensive reviews for several TE types, see Kapitonov and Jurka, 2006; Jurka et al., 2007).

The distribution of most classes of transposable elements in the genome often deviates from randomness. In the human genome, LINE1s (L1) are found with a higher probability in the AT-rich genomic compartments, while the older Alu subfamilies have a clear preference for the GC-rich genomic regions (see e.g., Brookfield, 2001; Pavlicek et al., 2001; Deininger and Batzer, 2002; Jurka et al., 2004; Hackenberg et al., 2005). This tendency seems to be shared by LINEs and SINEs in several organisms. MIRs, which are ancient SINEs, follow a distribution similar to old Alus with respect to GC content, despite their composition: Alus are GC-rich, while MIRs are AT-rich (Matassi et al., 1998; Jurka and Kapitonov, 1999). DNA transposons do not show any marked preference for GC content or gene density, at least when studied all together (Ovcharenko et al., 2005). However, some DNA transposons, e.g., the non-autonomous DNA transposon MER53 show some target site preference (Jurka and Kapitonov, 1999). LTR retrotransposons also seem to show little dependence on GC content, at least in the rat, mouse and human genomes (Yang et al., 2004). Several other tendencies for co-localization of repeats with functional regions, like *cis*-regulatory modules and conserved regions have been found (see, e.g., Table 2 in Jurka et al., 2007) concerning mainly retroelements.

2. Results and discussion

2.1. Power-law-like distribution of transposable elements in several genomes

The purpose of the present article is to systematically investigate the distribution of representative TE populations at chromosomal scale for various species. To this end, we study the size distribution of the inter-repeat spacers for a given TE population (mostly inactive copies) using the *cumulative size distribution* (see supplementary file 6). We define N(S) as the number of spacers with length larger or equal to S (in nucleotides, nt) and plot the logarithm of N(S) versus the logarithm of S. Linearity in such log–log plots indicates a power-law distribution (Newman, 2005). Power-laws in nature usually have an upper and a lower cutoff, which determine the linear region in log–log scale, where self-similarity is observed (Sellis et al., 2007; Clauset et al., 2009). In what follows, each figure that represents inter-repeat size distributions also includes curves for 10 random surrogate data sets (for details, see "Methods"). Notice also that for each TE population an upper limit in the divergence from the consensus sequence or a lower limit in the length (excluding very truncated copies) is imposed.

The principal finding of the present study is that for all major TE classes, power-law-like distributions are of widespread occurrence. The extent (E) of the linear region in log–log plots provides a simple measure for judging the validity of a power-law-based description. An additional quantity characterizing a power-law is the negative exponent (slope) of the cumulative distribution μ. The slope for a typical power-law does not exceed the value of $\mu=2$, as $\mu<2$ is a condition leading to a non-convergent standard deviation (Newman, 2005). Fig. 1 presents some examples of power-law-like distributions formed by inter-repeat distances. "Supplementary Table I" in supplementary file 1 includes all cases of studied transposable element populations forming a power-law, sorted by organism and chromosome (330 cases). Representative plots are given in the supplementary file 2: "Plots." In supplementary file 3: "Supplementary Table II," the full list of examined transposable elements, including ones without a power-law, is given (1331 cases). Table 1 presents an abridged version of our results. Two quantities summarizing the characteristics of power-laws in whole TE classes per organism are used: (i) Mean extent of the linear region of the distribution in log–log scale (ME) per organism for the three TE classes (SINEs, LINEs and DNA elements), which are mostly studied in this work. (ii) The fraction (FrPL) of all cases showing a power-law over all examined cases for a given organism and a given TE class. Here as individual "case" we consider the size distribution of spacers for a given TE family or subfamily in a given chromosome.

2.2. The "insertion–elimination model"

Our working hypothesis is that the power-law-like distributions of TEs can be explained with simple neutral processes of genomic evolution based on TE turnover. The "insertion–elimination model," as introduced in Sellis et al. (2007), is based on models initially formulated for the explanation of fractality in aggregation patterns in physicochemical systems (Takayasu et al., 1991). Our suggestion takes into account the one-dimensional topology of DNA and includes molecular events well established to occur in genome dynamics in the course of evolutionary time.

We here summarize the main points of the proposed model: Let us consider a sequence where a population of "markers" (representing the members of a TE population) is distributed, initially at random. Then, we assume the following molecular events, each with an assigned probability of incidence: [1] Elimination of a marker (repeat) of the initial population, occuring either by recombinational excision or due to progressive decomposition by nucleotide substitutions and/or indel events. This leads to the aggregation (merging together) of the spacers initially separated by the eliminated repeat. [2] Incorporation into existing spacers of subsequently inserted genomic material (SIM) such as repeats of more recent TE families, viral or other exogenous DNA, etc. [3] A fraction of the total chromosome (chosen at random) is copied with certain probability and then the copy is inserted at a random position. [4] In some simulations, the random insertion of markers is continued during the action of the previously described types of events. In these cases, a number of random transposition events are also introduced. This combination of duplication and transposition events is introduced in order to simulate the life-cycle of DNA transposons.

Using the insertion–elimination model, we simulate genomes evolving under a variety of parameter values. Initially we include only the type [1] and [2] events (see Fig. 2). We find that both a large number of elimination events [1] and large values of SIM [2] result in a more extended linear region in the transient power-law-like distributions (Fig. 2a, b). On the contrary, inclusion of transpositions [4] reduces the extent of the linear region (Fig. 2c). This is intuitively expected, because the power-law extent depends on the formation of long spacers and the larger the spacer, the higher is the probability of a random re-insertion to split the spacer into shorter segments.

All of the above simulations were repeated with the inclusion of segmental duplication events, which increased the final sequence length by 20%. This modification (i.e., inclusion of events of type [3]) did not qualitatively affect the derived results (figures not shown), thus verifying that the proposed model remains robust if segmental duplications, which are frequent in the history of several genomes, are taken into account. Details of the simulations are provided in supplementary file 5.