Gene 584 (2016) 148–160

Contents lists available at ScienceDirect

Gene

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

Research Paper

Fractality and entropic scaling in the chromosomal distribution of conserved noncoding elements in the human genome

Dimitris Polychronopoulos [a,1], Labrini Athanasopoulou [b], Yannis Almirantis [a,*]

[a] Institute of Biosciences and Applications, National Center for Scientific Research “Demokritos”, 15310 Athens, Greece
[b] Department of Theoretical Physics, Jožef Stefan Institute, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 7 December 2015
Received in revised form 22 January 2016
Accepted 14 February 2016
Available online 17 February 2016

Keywords:
Fractality
Block-entropy
Shannon entropy
Entropic scaling
Box-counting
Conserved noncoding elements
Ultraconserved elements
CNE

ABSTRACT

Conserved non-coding elements (CNEs) are defined using various degrees of sequence identity and thresholds of minimal length. Their conservation frequently exceeds the one observed for protein-coding sequences. We explored the chromosomal distribution of different classes of CNEs in the human genome. We employed two methodologies: the scaling of block entropy and box-counting, with the aim to assess fractal characteristics of different CNE datasets. Both approaches converged to the conclusion that well-developed fractality is characteristic of elements that are either extremely conserved between species or are of ancient origin, i.e. conserved between distant organisms across evolution. Given that CNEs are often clustered around genes, we verified by appropriate gene masking that fractal-like patterns emerge even when elements found in proximity or inside genes are excluded. An evolutionary scenario is proposed, involving genomic events that might account for fractal distribution of CNEs in the human genome as indicated through numerical simulations.

1. Introduction

Long-range correlations were shown to be present in the nucleotide sequence of the non-protein-coding part of eukaryotic genomes, soon after such genomes were sequenced (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992). In previous studies, we investigated the distributional features that extend at a large-scale, of genomic elements such as protein coding sequences (Sellis and Almirantis, 2009), transposable elements (Sellis et al., 2007; Klimopoulos et al., 2012) and conserved noncoding elements (Polychronopoulos et al., 2014a), by exploring the size distribution of inter-exon, inter-repeat and inter-CNE distances respectively. In most cases we observed power-law-like size distributions that often span several orders of magnitude.

In information theory, entropy was conceived by Claude Shannon (Shannon, 1948) to be an estimator of the amount of information that is carried in a transmitted message. During the last decades, scale invariance and fractality have been found in time series from signal transmission in electronic engineering, earthquakes, economy, social sciences and many other fields. Very often, such studies have been carried out using the standard box-counting technique and, in several cases of systems characterized by long range correlations, Shannon entropy has also been used in applications including biological sequence analysis (Vinga, 2014; Polychronopoulos et al., 2014b; Carbone, 2013).

In a previous work (Athanasopoulou et al., 2010) we studied the scaling properties of the block entropy of the distribution of genes in whole chromosomes of eukaryotic genomes through a coarse-graining reduction of the DNA sequence into a symbol sequence. The convention we followed was that zeros “0” in the symbol sequence stood for non-protein-coding nucleotides and ones “1” for nucleotides belonging to Protein Coding Segments (coding exons, denoted as PCSs). Several studies have shown that a linear scaling of the Shannon-like (or block) entropy $H(n)$ with the length n of the word (called hereafter n-word or block of length n) in semi-logarithmic plots is a clear indication of long-range order and fractality, as we are going to discuss in the next section (Grassberger, 1986; Ebeling and Nicolis, 1991; Ebeling and Nicolis, 1992; Ebeling et al., 1996). We verified this conjecture numerically in the case of finite Cantor-like symbol sequences (Athanasopoulou et al., 2010). Then, we showed that the genomic distribution of Protein Coding Segments often exhibits this particular scaling. In a more recent work (Athanasopoulou et al., 2014), we studied the scaling properties of the block entropy in the chromosomal distribution of transposable elements (TEs) and again we found the occurrence of the aforementioned scaling. The observed linearity in semi-

Abbreviations: CNEs, conserved noncoding elements; E, extent of linearity in semi-logarithmic scale; FrL, Fractality-related Linearity; GRBs, Genomic Regulatory Blocks; PCSs, Protein Coding Segments; SD, segmental duplication; TEs, Transposable Elements; UCEs, UCNEs, ultraconserved elements.

[*] Corresponding author.
E-mail address: yalmir@bio.demokritos.gr (Y. Almirantis).

[1] Present address: Department of Molecular Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London and MRC Clinical Sciences Centre, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK.

http://dx.doi.org/10.1016/j.gene.2016.02.022
0378-1119/© 2016 Elsevier B.V. All rights reserved.

logarithmic plots in the two types of genomic components follows different modalities. We have been able to attribute the observed distributional patterns, as expressed by entropic scaling and their differences, to the different evolutionary history of PCSs and TEs by means of a simple model. The model is shown, through computer simulations, to reproduce the observed pattern and includes key evolutionary events characterizing both genomic elements highly conserved in the course of evolution (e.g. Protein Coding Segments and CNEs) and genomic elements mostly non-conserved (the studied populations of TEs). The proposed model is composed, in both of its variations (the one for conserved and the other for non-conserved elements), of biologically plausible molecular events and is based on a previous model formulated in the framework of aggregative dynamics (Takayasu et al., 1991).

In Athanasopoulou et al. the entropic scaling analysis of the considered TE chromosomal distributions is accompanied by a box-counting study throughout (Athanasopoulou et al., 2014). Box-counting verified the appearance of fractality and self-similarity extending to several orders of magnitude in most cases where the aforementioned linear entropic scaling in semi-logarithmic scale was observed. In an older work of our group studying only chromosomal region (parts of chromosomes) where annotation about protein coding was available at the time, box-counting revealed indications of fractality in the distributions of genes (Provata and Almirantis, 2000).

In references (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992) and in numerous other later works, several research groups investigated aspects of genomic/nuclear structure at several length scales and levels of organization. Their converging results indicate that long-range order, correlations extending at several length scales and fractality are ubiquitous in the nucleotide juxtaposition and in the distribution of functional elements or compositional inhomogeneities in the genome. In a study of Lieberman-Aiden et al. (2009) the organization of the eukaryotic nucleus according to the 'fractal globule' model has been proposed, through the combination of novel experimental and computational techniques. In this case, the fractal pattern has been revealed because the contact probability as a function of genomic distance across the genome shows a power law scaling at a wide range of lengths. It is beyond the scope of the present article to exhaustively review this rapidly growing domain of research. In numerous works, Shannon entropy, fractality and related concepts (lacunarity, succolarity) are employed in order to describe genomic structure and to derive information which can be used from understanding functional and evolutionary aspects of genomic organization up to medical and diagnostic purposes. To mention a few, Cattani and Pierro (see Cattani and Pierro (2013) and references given therein) present a fractality and Shannon entropy analysis of whole chromosomes at the level of nucleotide distributions, deriving results which demonstrate the complementarity of the used methodologies. This study also suggests a framework applicable to the classification of DNA sequences. At a different level, evidence of fractality could be searched and quantified by means of an analysis of digitalized microscopic images of chromatin. Metze, in a comprehensive review (Metze, 2013), describes how changes in the fractal dimension derived from image analysis can be applied in cancer prognosis. It has to be noted that such techniques are also applicable outside the chromatin research, as fractality appears in a multiplicity of microscopic and middle scale biological patterns with crucial roles in the understanding of the micro-anatomy of relevant tissues; see e.g. the work of Pantic et al. (2014)) where fractality analysis of digital micrographs succeeds to systematically distinguish between histologically similar brain regions.

In the present work we focus on the study of several collections of conserved noncoding elements (CNEs) using entropic scaling analysis and box-counting. The genesis and evolutionary dynamics of conserved noncoding elements remain an enigma. It has been calculated that approximately 5.5% of the human genome is under selective constraint; of that, 1.5% is believed to encode for polypeptide chains, and 3.5% is assigned to known regulatory functions, while there is little evidence suggesting the possible roles for the remaining part (Lindblad-Toh et al., 2011). The discovery of 481 ultraconserved elements (UCEs) of more than 200 bp in length that are identical among human, mouse and rat genomes paved the way for a series of efforts with the aim to identify long sequences showing extreme levels of conservation (Bejerano et al., 2004). Roughly 25% of those UCEs fall within known protein-coding sequences. Since the identification of UCEs, researchers have focused on identifying conserved elements based on (i) lower thresholds of sequence similarity over whole genome alignments of two or more organisms, (ii) several thresholds of minimal length of conserved sequence, and (iii) the filtering of elements located inside exonic sequences (Elgar and Vavouri, 2008; Harmston et al., 2013). Throughout this article, we use the term CNE(s) for conserved noncoding elements to refer to all such elements despite their specific characterization as UCNEs, CNEs, etc. in the related literature. We use a particular name only whenever we want to refer to a specific class of elements.

It is believed that CNEs are enriched in gene deserts (Kim and Pritchard, 2007; Stephen et al., 2008) while, in mammalian genomes, a large number of those elements is often located at such distances from the closest genes that exceed in some cases 2 Mb, which is the limit for any known cis regulatory element (Woolfe and Elgar, 2008; Lettice et al., 2003). Little could be conjectured concerning what those distant CNEs actually perform in the cell; there is evidence, however, showing that they form an essential part of Genomic Regulatory Blocks (GRBs) and that they could synergistically function alongside with their target genes (Kikuta et al., 2007; Dimitrieva and Bucher, 2012). Furthermore, the literature suggests that CNEs are selectively constrained and not mutational cold spots (Drake et al., 2006).

Since it is generally believed that sequence conservation across genomes is a key indication of functional relevance, the study of the sequence-specific characteristics of different classes of CNEs, as well as the mechanisms that might have led to their genesis, would be of paramount importance in an effort to crack the so far enigmatic regulatory code of our genome.

In a recent study of our group (Polychronopoulos et al., 2014a), the size distribution of the inter-CNE distances of a variety of CNE collections has been investigated and power-law-like distributions have been found in most cases. This means that the abundance of inter-CNE spacers depends linearly on the spacers' length in double-log scale, and this is found to occur for a range of spacer sizes often higher that two decimal orders of magnitude and in some cases exceeding the three orders. In the study presented herein we include entropic scaling and box-counting of four additional data sets not present in Polychronopoulos et al. (2014a). For completeness, we include plots for the complementary cumulative size distribution of the inter-CNE distances for these CNE collections not studied earlier (see Supplementary data file). We also refer the interested reader to this recent work for further details about aspects of CNE biology and several conjectures about their role and organization in the vertebrate genome.

2. Methods

2.1. Box-counting method for the determination of fractality

Box-counting is widely used for assessing the fractality of symbol sequences and of other types of discrete datasets (Mandelbrot, 1982; Feder, 1998). Here we describe a one-dimension implementation of this method. We cover the chromosome with one dimensional "boxes" of length δ. The number of boxes overlapping CNEs is assumed to be the chromosomal length $L(\delta)$ occupied by CNEs. In a fractal structure the length measured in this way does not reach a fixed value as δ decreases (Feder, 1998). This length scales as: $L(\delta) \sim \delta^D$. The exponent D is the negative fractal dimension D_f of the fractal pattern we consider. The plots depicting how $L(\delta)$ scales as a function of δ are shown in log–log scale. The slope of the linear part of the curve and the extent of the linearity are both informative for the characterization of the fractal