Gene 447 (2009) 18–28

Contents lists available at ScienceDirect

Gene

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

Power-laws in the genomic distribution of coding segments in several organisms: An evolutionary trace of segmental duplications, possible paleopolyploidy and gene loss

Diamantis Sellis [1], Yannis Almirantis *

National Center for Scientific Research "Demokritos", Institute of Biology, 153 10 Athens, Greece

ARTICLE INFO

Article history:
Received 23 December 2008
Received in revised form 18 March 2009
Accepted 8 April 2009
Available online 8 July 2009

Received by J. Jurka

Keywords:
Power-laws
Genome structure
Long-range correlations
Genome shuffling
Genome evolution

ABSTRACT

Large-scale features of the spatial arrangement of protein-coding segments (PCS) are investigated by means of the inter-PCS spacers' size distributions, which have been found to follow power-laws. Linearity in double-logarithmic scale extends to several orders of magnitude in the genomes of organisms as disparate as mammals, insects and plants. This feature is also present in the most compact eukaryotic genomes and in half of the examined bacteria, despite their very limited non-coding space. We have tried to determine the sequence of events in the course of genomes' evolution which may account for the formation of the observed size distributions. The proposed mechanism essentially includes two types of events: (i) segmental duplications (and possibly paleopolyploidy), and (ii) the subsequent loss of most of the duplicated genes. It is shown by computer simulations that the formulated scenario generates power-law-like inter-PCS spacers' size distributions, which remain robust for a variety of parameter choices, even if insertion of external sequences, such as viruses or proliferating retroelements is included. Moreover, power-laws are preserved after most of the non-coding DNA has been removed, thus explaining the finding of this pattern in genomes as compact as that of *Takifugu rubripes*.

1. Introduction

The distributions of genes, regulatory regions and several types of repetitive DNA in the genomic landscape are highly non-random and inter-related in complex ways. One decade ago, based on the relatively small number of large and reliably annotated genomic sequences available at that time, we studied the size distribution of the distances separating the protein-coding segments (PCS) in some model organisms. Despite the limited size of the considered sequences (less than 1% of the length of a typical eukaryotic chromosome), we often observed the formation of power-laws in the size distribution of the inter-PCS spacers, i.e. linearity in double-logarithmic scale (Almirantis and Provata, 1999).

Recently, we have reported the existence of power-laws for several orders of magnitude in the size distribution of inter-repeat spacers in the human genome (Sellis et al., 2007) while preliminary results indicate that this feature is shared by different organisms and for several repeat classes.

In the present article, a systematic investigation of the size distribution of spacers between coding segments is undertaken in a whole-genome scale for several sequenced genomes. Its content is organized as follows: the statistical concepts which will be used are briefly reviewed in the beginning of the "Results" section. Then, evidence about power-laws in the inter-PCS size distributions in several organisms is presented. In the "Discussion" section, the proposed segmental duplication – gene loss model is described and combined with other components of genome dynamics: repeat insertions, indel dynamics and sequence loss. Computer simulations of the model are presented and compared with evidence from genomic architecture of several organisms. A short discussion on the results of this work and on its limitations is included in the "Concluding remarks" section.

2. Results

2.1. Preliminaries

We perform a systematic investigation of the size distribution of the distances separating consecutive Protein-Coding Segments (inter-PCS spacers) in the complete set of chromosomes of several genomes. In the Supplementary file 3 (size distributions) we show examples of the inter-PCS spacers' size distribution in both, semi-logarithmic and double-logarithmic scale. A linear region is observed in the double-log scale in most cases, but not in the semi-logarithmic scale. This

Abbreviations: E, extent of the linear region in a log-log plot; LINE, long interspersed repeats; Myr, million year(s); N(S), the number of spacers longer or equal to S; PCS, protein coding segments; r^2, regression coefficient; S, spacers' length; μ, the negative slope of the power-law in a cumulative log-log plot.

* Corresponding author. Tel.: +30 210 6503619; fax: +30 210 6511767.
E-mail addresses: dsellis@stanford.edu (D. Sellis), yalmir@bio.demokritos.gr (Y. Almirantis).
[1] Present address: Biology Department, Stanford University. Gilbert Building, Room 109 371 Serra Mall Stanford, CA 94305-5020, USA.

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.04.028

indicates that the genomic inter-PCS distances follow a power-law-like distribution:

$$N^*(S) \propto S^{-\zeta} = S^{-1-\mu} \quad \mu > 0$$

where, with $N^*(S)$ we denote the number of spacers with length around S, and with μ the power-law parameter which indicates how densely clustered are the PCSs.

Power-laws are often characterizing systems with long-range correlations, i.e. systems whose remote parts exhibit correlated structure or behavior. This is usually measured using some form of a correlation function and early in the nineties it has been shown that long-range correlations exist in the nucleotide sequences of non-coding genomic regions (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992). Power-laws in size distributions are in general associated with

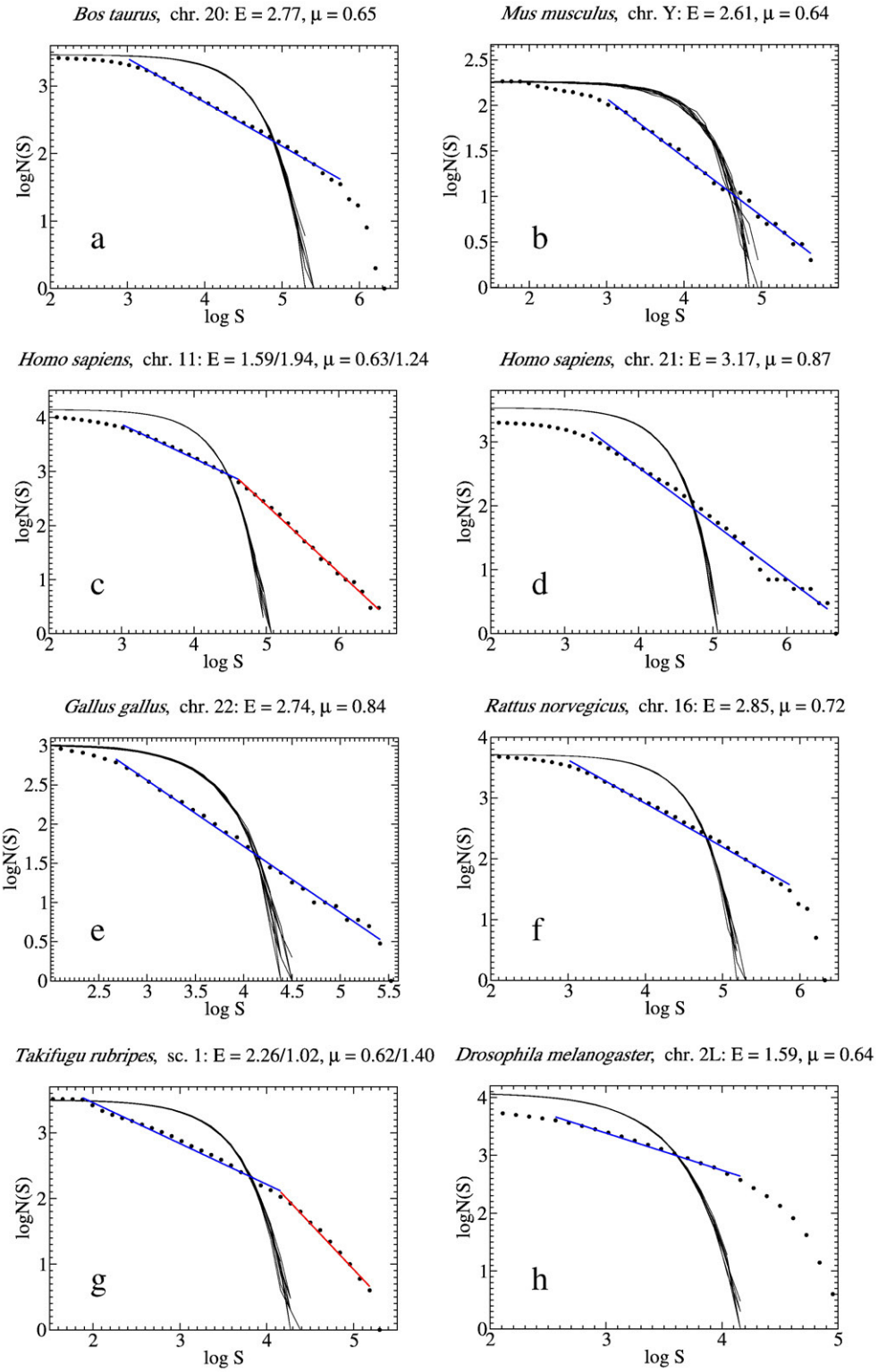

Fig. 1. Inter-PCS spacers' size distributions. Examples of eukaryotic inter-PCS spacers' cumulative size distributions accompanied by ten size distributions (continuous lines), corresponding to randomly distributed PCSs (for details see the Subsection 2.1 "Preliminaries").