Orphan and gene related CpG Islands follow power-law-like distributions in several genomes: Evidence of function-related and taxonomy-related modes of distribution

Giannis Tsiagkas, Christoforos Nikolau, Yannis Almirantis

A R T I C L E   I N   F O

Article history:
Available online xxx

Keywords:
CpG-Islands
CGIs
Orphan CpG-Islands
CpG dinucleotide
Power-law-like distribution
Genome evolution

A B S T R A C T

CpG Islands (CGIs) are compositionally defined short genomic stretches, which have been studied in the human, mouse, chicken and later in several other genomes. Initially, they were assigned the role of transcriptional regulation of protein-coding genes, especially the house-keeping ones, while more recently, there is found evidence that they are involved in several other functions as well, which might include regulation of the expression of RNA genes, DNA replication etc. Here, an investigation of their distributional characteristics in a variety of genomes is undertaken for both whole CGI populations as well as for CGI subsets that lie away from known genes (gene-unrelated or “orphan” CGIs). In both cases power-law-like linearity in double logarithmic scale is found. An evolutionary model, initially put forward for the explanation of a pattern similar found in gene populations is implemented. It includes segmental duplication events and eliminations of most of the duplicated CGIs, while a moderate rate of non-duplicated CGI eliminations is also applied in some cases. Simulations reproduce all the main features of the observed inter-CGI chromosomal size distributions. Our results on power-law-like linearity found in orphan CGI populations suggest that the observed distributional pattern is independent of the analogous pattern that protein coding segments were reported to follow. The power-law-like patterns in the genomic distributions of CGIs described herein are found to be compatible with several other features of the composition, abundance or functional role of CGIs reported in the current literature across several genomes, on the basis of the proposed evolutionary model.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Genomic CpG islands, in which CpG dinucleotides are abundant and non-methylated, have been initially detected experimentally and defined as short (hundreds of nucleotides-long) stretches in vertebrate genomes, thus offered as cleavable sites for mCpG-sensitive restriction enzymes (HTF islands, see e.g. Bird, 1986). As lengthy DNA sequences and whole genomes became progressively available, sequence-based definitions of CpG Islands (CGIs) and computational algorithms using a sliding window combined with threshold values for key quantities were put forward. Thresholds for sequence-based search of CGIs are considered for: (i) the minimal island length \( L_{\text{min}} \); (ii) the percentage of cytosine and guanine content \( C+G \), abbreviated in the following as \( CG \); and (iii) the observed over expected frequency of occurrence of the CpG dinucleotide \( \text{CpG}(e) \). Gardiner-Garden and Frommer (1987) introduced the first widely used set of thresholds \( L_{\text{min}} = 200 \text{bp}, \ CG > 50\% \), \( \text{CpG}(e) > 0.6 \); to which we will hereafter refer as “relaxed criteria”; \( \text{bp} \) stands for ‘base pairs’, while later, Takai and Jones (2002) used more conservative threshold values, named hereafter “stringent criteria” \( L_{\text{min}} = 500 \text{bp}, \ CG > 55\% \), \( \text{CpG}(e) > 0.65 \). In the following, these threshold choices will be abridged as G–C\&F and T\&J respectively. Alternatively, methods based on the degree of CpG dinucleotide clustering (Hackenberg et al., 2006; Glass et al., 2007) and on entropic edge detection (Luque-Escamilla et al., 2005) have also been introduced. In order to avoid false positives (e.g. due to high C+G Alu sequences in the human genome), especially when relaxed criteria are used, repeat-masking is usually applied before searching for CGIs.

CGIs are widely accepted as markers for the existence of protein-coding genes, the promoter regions of which are usually proximal to or overlapping with the islands. Consequently, the search for CGIs was motivated by the need of detection of yet unannotated protein coding genes. Therefore, most of the
introduced CGI finding algorithms were judged upon their ability to find CGIs in the proximity of known genes (see e.g. Bird, 1987; Han and Zhao, 2008, 2009).

The functional character of a CGI is related to the condition that its CpG dinucleotides are predominantly unmethylated. Genome-wide methylation data were not available until recently. The CpG depletion of the vertebrate genome is explained on the basis of heavy methylation and the consequent mutation of 5-methylcytosine to thymine, verified by the observed increase in TpG and CpA abundances and is extensively discussed in relation to CGIs evolution and function. See Bird (1986), Antequera (2003) and Jabbari and Bernardi (2004), where however the opposite position (i.e. the independence of CGI deficiency and TpG, CpA excess on the level of DNA methylation) is held. The propensity of 5-methylcytosine to quickly mutate to thymine has driven to the conjecture that the methylation is hindered in genomic sequences where CpG frequency is close to the expected one, at least in the germ line. This lack of methylation on specific locations may be seen as the result either of protection from point-mutations (mutational “cold spots”) or of purifying selection due to functional roles which are fulfilled only if specific compositional traits are preserved in the underlying sequences. CGIs cannot be generally seen as mutational cold spots, as there are converging data that many CGIs in vertebrates have been lost when they ceased to be under purifying selection. These are retained for long evolutionary time in other species where they remain functional, e.g. in human and mouse genome comparisons; see Antequera and Bird (1993), Matsuo et al. (1993).

After the systematic search for CGIs in the human and other genomes using the simple sequence criteria described above, researchers attempted to introduce several forms of epigenomic information. Bock et al. (2007) used a method combining a standard sliding window algorithm with available epigenomic data for human chromosomes 21 and 22. Tanay et al. (2007), based on human and chimpanzee genomes’ comparison, found regions with reduced CpG mutability, while the same domains largely coincide with Polycomb-binding sites. Later on, Illingworth et al. (2010) provided a comprehensive list of CGIs in the human and mouse genomes based principally on methylation data. The most important of this work’s observations was that there exist numerous “orphan” CpG Islands (i.e. not connected to a known protein-coding gene), independently of the method used for their detection. Illingworth and co-workers formulated a hypothesis according to which most of orphan CGIs are related to functional promoters of unknown genes, many of which could be functional RNA genes. More recently, CGIs were shown to often coincide with Origins of Replication (ORIs) thus suggesting their functional roles may extend beyond transcriptional regulation (Cayrou et al., 2011). To what extent the colocalization of CGI and origins of replication is due to a direct causal link or an indirect effect mediated by positional preferences of gene transcription start sites remains, nonetheless, unclear.

In previous works (Sellis et al., 2007; Sellis and Almirantis, 2009; Klimopoulos et al., 2012) we have observed that distances between transposable elements (TE) belonging to the same family, as well as distances between protein-coding segments (PCS) often follow “power-law-like” size distributions in entire chromosomes. Such distributions present extended linear regions in log–log scale (see Methods). Our principal result herein is that CGIs also, either studied as entire populations or when only orphan islands are considered, follow power-law-like size distributions. We have put forward an evolutionary model, including biologically plausible steps, for the temporal evolution of genomic components that are subjected to purifying selection. This model, when tested numerically, is shown to systematically generate power-law-like size distributions, like the ones found in the CGI chromosomal distributions.

Here we investigate the large-scale genomic distribution of CpG Islands in several genomes, after masking for known repeated sequences (for all organisms apart from S. cerevisiae). Power-law-like distributions are extensively observed and through the proposed model we attempt to extend our understanding of the evolution and function of CGIs. In the cases of human and mouse genomes, the data of Illingworth et al. (2010) are also used alongside with CGI coordinates derived from the implementation of compositional thresholds. For all studied organisms, besides the “standard” G–C=G and T=C threshold sets, several additional choices of sequence composition are also tested and the resulting inter-CGI distances’ distributions are critically presented. A separate study of populations of orphan CGIs is performed, principally in order to exclude that the observed power-law-like distributions in whole CGI populations are a mere consequence of similar distributions already known to exist for the inter-genic distances (Sellis and Almirantis, 2009). The persistence of orphan CGIs’ chromosomal distributions to form power-laws when, taken alone, apart from an indication of their functional role and conservation through purifying selection, can provide valuable insight into the overall organization of genome architecture and the mechanisms under which sequences are attributed with functionality.

2. Methods

2.1. Origin of the genomic and epigenomic data

The genomes of the following organisms are used in this study: Apis melifera, Bos taurus, Caenorhabditis elegans, Canis familiaris, Danio rerio, Drosophila melanogaster, Gallus gallus, Homo sapiens, Monodelphis domestica, Mus musculus, Saccharomyces cerevisiae. Several or all chromosomes of each genome are studied. Only entire chromosomes are considered. Information about the origin of the genomic sequences used may be found in the supplementary file “genome data”. These data are used for the sequence-based determination of CGIs, by means of the algorithm described in the following sub-Section 2.2.

There is not an unambiguous consensus about what an orphan CpG-Island really is. The term has been widely used to denote all CGIs which are not directly involved in the regulation of nearby genes, but the exact proximity relationships between CGI and gene bodies vary according to the study. In the following, depending on genome sizes, we use different distance thresholds from the nearest gene in order to characterize a CGI as orphan. More specifically, in large genomes, we have considered, in some cases, full-length gene masking, with flanks of 5kbp and 2kbp (thousands of base pairs) at 5’ and 3’ ends respectively; for more information see later on (sub-Section 2.5) and the supplementary file “Supplementary Table”.

Data for coordinates of CGIs derived from the methylation state of the sequence, as determined by Illingworth et al. (2010) were downloaded from the additional material of this article. Information about the origin of the genomic coordinates used for masking of transcription start sites (TSS) and of genes may also be found in “genome data”. Additional data for gene or CDS (i.e. protein-coding exon) coordinates were downloaded from the FTP site of NIH [ftp.ncbi.nih.gov/genomes/MapView] for each of the organisms we have studied. We used the annotation of these files in order to define the subset of CGIs which may be considered as orphan.

2.2. Algorithm for the extraction of CGIs coordinates in genomic sequences

The algorithm we have used for the localization of CGIs is an implementation of the method described by Takai and Jones.
2.3. Power-laws in size distributions

2.3.1. The equations in terms of probabilities

Suppose there is a large collection of \( n \) objects (in our case spacers between CGIs), each characterized by its length \( S \). In such random collections of objects we can approximate the distribution of their sizes with an exponential distribution (like the runs of heads in a coin-tossing experiment). Let \( p(S) \) be the probability of a spacer having length between \( S - s/2 \) and \( S + s/2 \), (where \( s \) is the size of the bin width) and \( N(S) \) the number of spacers:

\[
N'(S) = np(S) \propto e^{-as}, a > 0
\]

When scale-free clustering appears, long-range correlations extend to several length scales (ideally, in our case for the whole examined genomic region) and the spacers’ size distributions follow a power-law, which corresponds to a linear graph in a double logarithmic scale:

\[
N'(S) = np(S) \propto S^{-\zeta} = S^{1-\mu}, \mu > 0
\]

2.3.2. Original vs. cumulative size distribution

In the present work we use the “complementary cumulative distribution function” (Clauset et al., 2007) for the sizes of spacers (distances) between CpG Islands, defined as follows:

\[
P(S) = \int_S^\infty p(r)dr,
\]

where, \( p(r) \) is the original spacers’ size distribution. The cumulative distribution forms smoother “tails”, less affected by statistical fluctuations. Although the cumulative form of the distribution may be seen as presenting more “inertia” thus overshadowing local features, the inclusion of ten surrogate data distribution curves, along with every genomic one (see below for details), allows for an objective estimation of the linearity trend in each case. As it is pointed out by Sims et al. (2007) the use (in double-logarithmic plots) of the slope of the cumulative form of the distribution for the estimation of the exponent of the power-law gives much more precise results than the use of the slope of the original one. It is also, by definition, independent of any binning choice: in a cumulative curve the value of \( P(S) \) for length \( S \) is not associated with the subset of spacers whose length falls in the same bin, as in the original distribution, but it corresponds to the number of all spacers longer than \( S \).

The cumulative form of a power-law size distribution is again a power-law characterized by an exponent (slope) equal to that of the original distribution plus 1 (i.e. minus one in absolute value): if \( p(r) \propto r^{-1-\mu} \), then

\[
N(S) = np(S) \propto \int_S^\infty (r^{-1-\mu})dr \propto S^{-\mu},
\]

where \( N(S) \) is the number of spacers longer or equal to \( S \).

In finite size collections taken from physical systems, we cannot speak about “power-laws” but rather about “power-law-like” distributions. This is particularly important for our study (see Section 4 for further details, where the proposed model is explained) as there appears to be no particular tendency for a unique (universal) exponent \( \mu \). The intensity of the power-law-like pattern is quantified through the value of the extent (E) of the linear region in double logarithmic scale.

Additionally to CGIs genomic distributions, all figures also include a bundle of ten simulated size distributions (continuous lines) where an equal number of markers, corresponding to the CGIs of the original natural sequence, are randomly positioned in a sequence of equal length. The inclusion of these random (surrogate) data sets in the figures allows for a direct visualization of the difference between observed and random distribution patterns.

2.4. Methodology for the search of threshold values for CGIs localization in several genomes

As discussed in the introduction, the threshold values for CGIs detection in the cited literature are mostly determined for the purpose of the localization of new protein-coding genes, which are known to have promoters often intersecting with CGIs. Most of this work concerns the human and a few other, warm-blooded, animal genomes. Here, in order to further elaborate on alternative modes of CGI detection, we form a two-dimensional “parameter grid” of threshold values for CG% and CpGo/e (around the G&G and T&J threshold choices) for either \( L > 200 \) bp or \( L > 500 \) bp, for some chromosomes of the most widely studied genomes. These grids are given in the supplementary file “grid”, tabulating the used thresholds alongside with the extent (E) of the complementary cumulative inter-CGI distances size distributions. This is done in an attempt to determine the threshold values corresponding to functional islands, which are thus preserved in evolutionary time, (see in the Discussion). The criterion adopted for choosing chromosomes for the creation of their parameter grids is to include the chromosome with the most extended linearity for the standard threshold choices G&G&F or T&J, and one or more chromosomes (taken at random) with a medium E value. In the case of *C. elegans* we derived a grid only for one chromosome, because power-laws for the others are found to be rudimentary, i.e. with a linear extent below unity. The grid-driven search is done only for a limited number of chromosomes, for 8 organisms out of the 11 overall studied.

Also provided in the Supplementary Table (see also in the Results and Table 1) are values of \( E \) given for the standard G&F and T&J threshold sets and for alternative sets for the same values of CG% and CpGo/e, interchanging the minimal length thresholds, denoted by altG-G&F (for \( L > 500 \) bp) and altT&J (for \( L > 200 \) bp).

2.5. Details of the masking procedure

In order to define the “orphan” CGI subsets, for large genomes (actually, all the vertebrates included in our list): (i) either we masked the total length of genes by additionally masking 5 kbp upstream of the start-of-gene (TSS) and 2 kbp downstream of the end-of-gene (TES), or (ii) we masked only around the TSS by adding flanks of 500 bp upstream of the TSS and 100 bp downstream of the TSS. For small genomes (insects, worm, and S. cerevisiae): (iii) we either followed the TSS masking (narrowing the upstream masking to 100 bp) as in (ii); or (iv) we followed the TSS masking again, but with the additional masking of each protein coding segment, with symmetric flanks 50 bp upstream and downstream. The criterion adopted for choosing chromosomes for the study of the distribution of orphan CGIs is to include the chromosome with the most extended linearity for the standard threshold choices G&G&F or T&J, and several other chromosomes, with a medium E value, taken at random.

We have also checked the existence of power-law-like distributions for the set of epigenetically determined CGIs (Illingworth et al., 2010) not intersecting with TSSs and having a minimal distance of 100 bp from a TSS of the RefSeq annotation. We name these Islands "non-TSS". Alternatively, we have formed the subset of Illingworth and co-workers' original CGIs, which have a minimal distance of 5 kbp upstream or 2 kbp downstream of any RefSeq TSS and we name them "TSS-unrelated" in the present text. This is done in order to test the distribution of CGIs that are unlikely to have any role in the transcriptional regulation of protein coding genes at medium-distances (and probably spatially linked). This filtering is analogous to the case (i) of the previous paragraph, dealing with the study of masking of threshold-based CGIs. In the framework of the present study, the two classes of epigenetically determined CGIs defined above are considered as orphan.

2.6. Simulations using the segmental duplication – CGI loss model

Examples of simulations are given in Fig. 5(a–c). Initially, 1000 markers (representing CGIs) are randomly inserted in a sequence 2 Mbp long. In (a), 173 segmental duplication events occurred, with lengths sampled from a uniform distribution not exceeding in size the 5% of the actual length of the simulated sequence. After each such event a number of markers (CGIs) equal to 90% of the number of the duplicated ones are eliminated. The length reached by the simulated chromosome was ~195 Mbp (steps i and ii of the model, see in the Section 4). In (b) the segmental duplications were 87 (~50% of those in (a)) and the final sequence length ~22 Mbp. In (c) the number of segmental duplications is as in (a) and the finally reached length ~119 Mbp. Here, 256 additional events of non-duplicated CGIs eliminations are also allowed (step iii of the model). These eliminations represent cases of loss of function of ancestral CGIs with a subsequent progressive decomposition. For genomic and for model-generated size distributions, circles and diamonds are used respectively, throughout. Simulations presented in Fig. 5 are representative of a large number of numerical experiments conducted.

Scripts in Perl and programs in FORTRAN used for parsing genomic data files, computing the presented size distributions and simulating the evolutionary model are available upon request.

3. Results

3.1. Chromosomal distributions of CpG-Islands in various genomes

In Fig. 1 we present examples of the complementary cumulative inter-CGI distances’ size distributions, in double logarithmic scale, for whole chromosomes from the 11 considered genomes. The G&G or T&J thresholds values have been used here. More plots are included in “Supplementary Plots”. In Table 1 the following information is included: (i) mean values of linearity extent E (M.V.) for all studied chromosomes from each organism; (ii) mean values of linearity Extent E for the five more extended power-law incidences (M.V.-5), serving as an additional estimator of the intensity of the appearance of the power-law-like pattern per organism. In Supplementary Table all chromosomes are presented. In these Tables we have included along with the standard parameter sets, the ones we called “alternative” in sub-Section 2.4.

The principal result of this study is that power-law-like distributions of inter-CGI distances are observed in several cases with linearity extent E > 2, which in two instances surpass the three orders of magnitude (in dog chr. 31 and human chr. 21, see the Supplementary Table).

We have also performed gene masking and subsequently we studied the spacers’ size distribution for orphan CGIs, in some chromosomes for each organism, as described in detail in the Methods. In Fig. 2, examples of these distributions are presented, for the same chromosomes as in Fig. 1. In Table 2 mean values are given for the orphan CGI spacers’ size distribution in the same form with Table 1, while in Supplementary Table all the studied cases are included. Power-law-like distributions of orphan CGIs are wide-spread, while, in some instances of masked genomes, we observe increase of the extent of linearity when orphan islands are considered instead of the full populations.

Additionally, for the human and mouse genomes, chromosomal distributions for the islands determined by Illingworth et al. (2010) based on methylation data are studied and the power-law-like pattern is also found. When a masking procedure is applied in the epigenetically determined CGI collection (for details see Methods, sub-Section 2.5) the picture remains qualitatively the same. We observe that power-law-like distributions are again formed, like the ones found in the orphan CGI populations derived on the basis of compositional criteria, although the extent of linearity is somewhat reduced. In Fig. 3, for one chromosome of each organism, the complete sets of epigenetically determined CGI are presented (Fig. 3a and d), along with the corresponding CGI subsets remaining after gene-masking (non-TSS CGIs; Fig. 3b and e), as well as the fraction of the TSS-unrelated CGIs (Fig. 3c and f). In Table 2, summary information about non-gene-related epigenetically determined CGIs is also included. The full set of epigenetically determined CGI populations and their orphan subsets are presented in the Supplementary Table. See also the Supplementary Plots.

We have also attempted a search for the determination of the form of distribution of distances between CGIs using

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Mean values of linearity extent E (M.V.) for all studied chromosomes from each organism; (ii) mean values of linearity extent E for the five more extended power-law incidences (M.V.-5). Standard and alternative threshold values are considered, see sub-Section 2.4.</td>
</tr>
<tr>
<td>G&amp;G/F</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Large genomes</td>
</tr>
<tr>
<td>Bos taurus</td>
</tr>
<tr>
<td>Canis familiaris</td>
</tr>
<tr>
<td>Danio rerio</td>
</tr>
<tr>
<td>Gallus gallus</td>
</tr>
<tr>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Monodelphis domestica</td>
</tr>
<tr>
<td>Mus musculus</td>
</tr>
<tr>
<td>Small genomes</td>
</tr>
<tr>
<td>Apis mellifera</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Tsiagkas, G., et al., Orphan and gene related CpG Islands follow power-law-like distributions in several genomes; Evidence of function-related and taxonomy-related modes of distribution. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j. compbiolchem.2014.08.013
3.2. An evolutionary scenario reproducing the observed power-law-like distributions based on segmental duplications and CGI loss

Segmental duplication events occurred continuously in the evolutionary past of virtually all eukaryotes (De Grassi et al., 2008; Gardiner-Garden and Frommer, G-G&F (L_{min} = 200 bp, GC > 50%, CpGo/e > 0.6) or the Takai and Jones, T&J (L_{min} = 500 bp, G+C > 55%, CpGo/e > 0.65) sets of thresholds have been used. Genomic curves are accompanied in each plot by 10 curves of surrogate data (continuous lines), corresponding to randomly distributed markers. The linear segments are inferred by linear regression.

Fig. 1. Inter-CGI spacers’ complementary cumulative size distributions in whole chromosomes. Linearity in log-log scale (a power-law-like pattern) is observed. Logarithms are to base 10. Each sub-plot (a–f) corresponds to a chromosome of one of the examined genomes, for a specific choice of thresholds in the detection of CpG–Islands. Either the Gardiner–Garden and Frommer, G-G&F (L_{min} = 200 bp, GC > 50%, CpGo/e > 0.6) or the Takai and Jones, T&J (L_{min} = 500 bp, G+C > 55%, CpGo/e > 0.65) sets of thresholds have been used. Genomic curves are accompanied in each plot by 10 curves of surrogate data (continuous lines), corresponding to randomly distributed markers. The linear segments are inferred by linear regression.

Kehrer-Sawatzki and Cooper, 2008; Kirsch et al., 2008; McLysaght et al., 2002; Shapira and Finnerty, 1986). It has been shown that at least 10% of the non-repetitive human genome consists of identifiable (i.e. relatively recent) segmental duplications (Bailey et al., 2002). Additionally, most extant taxa have experienced paleopolyploidy during their evolution (i.e. duplication of the whole genome and subsequent reduction to diploidy), see e.g. Adams and Wendel (2005), Gibson and Spring (2000), Sémon and Wolfe (2007) and references therein. It is estimated that 50% of all genes in a genome are expected to duplicate giving an “offspring” at least once on time scales of 35–350 million years, not taking into account events of polyploidization (Lynch and Conery, 2000). Segmental duplication and polyploidisation generate copies of some or all the genes of an organism, but also of other functional

Fig. 2. Same cases of chromosomal distributions as in Fig. 1, sub-plot (a–f), but here excluding gene-related CGIs. Several alternative ways of retaining only orphan CGIs have been used. Again, power-law-like linearity is observed, often more extended than in the corresponding complete population. Surrogate data curves and linear regression are similar to Fig. 1. Logarithms are to base 10.
genomic localizations, like CGIs. In the literature, there are reported three possibilities for the evolution of the duplicated genes (Adams and Wendel, 2005; De Grassi et al., 2008; Lynch and Conery, 2000): In some cases one member of the gene pair adopts a new function while the other remains unchanged. In other cases, the two copies continue to survive in the genome sharing the multiple functions of the initial gene between them. However, as all authors agree, see e.g. Adams and Wendel (2005), Kasahara (2007), Lynch and Conery (2000), Sémon and Wolfe (2007), the fate of most duplicated genes is that one copy is silenced, losing the ability to be transcribed, and then disintegrates progressively by random mutations, while it is also exposed to the possibility of excision due to recombination driven eliminations. The above fate of duplicated genes is normal to have its parallel to the fate of CGIs accompanying genes and being involved in their regulatory mechanism. We can consider that CGIs with yet unknown roles, regarded as orphan, when duplicated often become superfluous and stop being under purifying selection. These orphan CGIs will be gradually decomposed and lost, with the accelerated rate imposed by the rapid turnover of the Cpg dinucleotide (Nacimiento and Crowell, 2000). The existence of another source of CGI loss can be supported by current findings of comparative genomics. Several cases of syntenic genes of different organisms have been shown to be under differential control of CGIs, with one gene being under the regulatory control of a CGI, while its syntenic counterpart in another genome appears to have been emancipated from the action of its ancestral CGI, which has been subsequently decomposed (for details see in the following sub-section).

Moreover, sequencing and analysis of the human genome has shown that segmental duplications and gene loss are widespread phenomena of genome remodeling during evolution. An example of relatively recent gene loss is that ~60% of the olfactory receptors in the human genome (~1000 genes and pseudogenes in total) have disrupted ORFs and appear to be pseudogenes, which is consistent with findings suggesting massive functional gene loss in the last 10 Myr (IHGSC, 2001). Even if there are not available quantitative data for the CGI corresponding populations, we may expect that loss of CGIs due to one of the above mentioned reasons, often connected to the loss of (protein-coding or not) genes due to changes of the needs of a given species is also a widespread phenomenon in genomic evolution.

The implementation of a “genomic duplication–CGI loss” model is based on the following genomic phenomena: occasional gene (and associate CGI) silencing and subsequent degradation, orphan CGIs becoming redundant, emancipation of genes from their associate CGI, segmental duplications occurring during genomic evolution and (in some cases) complete genome duplications. This model includes events of the types:

i. Segmental duplications of extended regions of chromosomes.
   This step may include as limiting case whole genome duplications, although not explicitly considered in the present implementation.

ii. Random eliminations of a number of CGIs which is lower or equal to the number of the duplicated ones.

iii. Occasionally, additional eliminations of non-duplicated CGIs.

iv. Insertions of sequences increasing the total chromosomal length (these could be transposable elements, retroviruses, microsatellite expansions etc).

v. Deletions of sequence stretches (which usually are under weak or no purifying selection).

The proposed evolutionary scenario reproduces power-law-like inter-CGI distances size distributions. Moreover, this property is proven numerically to be robust to quantitative modifications of all the involved types of molecular events. Only events i and ii are indispensable for the appearance of the power-law-like pattern, given that the population of the genomic elements we study are (principally) functional and thus preserved by purifying selection. In a completely different genomic framework—i.e. in the study of non-conserved elements, e.g. transposable elements (TE) or microsatellites, event types iii and iv (in that context, eliminations of repeat copies of the studied TE family, and insertions of TE families more recent than the studied one) are required instead of i

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Values of linearity extent E as: (i) Mean values (M.V.); (ii) mean values for the five more extended power-law incidences cases (M.V.-5), for several choices of gene-unrelated (orphan) sets of CGIs.</td>
</tr>
<tr>
<td>Masked TSS (-500 bp, +100 bp) for large genomes, (-100 bp, +100 bp) for small genomes</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Large genomes</td>
</tr>
<tr>
<td>Bos taurus</td>
</tr>
<tr>
<td>Canis familiaris</td>
</tr>
<tr>
<td>Danio rerio</td>
</tr>
<tr>
<td>Gallus gallus</td>
</tr>
<tr>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Monodelphis domestica</td>
</tr>
<tr>
<td>Mus musculus</td>
</tr>
<tr>
<td>Small genomes</td>
</tr>
<tr>
<td>Aphis mellifera</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>G-G&amp;F</td>
</tr>
<tr>
<td>Large genomes</td>
</tr>
<tr>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Mus musculus</td>
</tr>
<tr>
<td>Small genomes</td>
</tr>
<tr>
<td>Aphis mellifera</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
and ii (see Sellis et al., 2007; Klimopoulos et al., 2012). Events iii, iv and v are numerically shown to be dispensable for the emergence of power-laws in the computer simulations studied herein. Events of type iii, although not essential for the emergence of power-law-type linearity, are important, as several authors bring evidence about loss of CGIs over evolutionary time due to de novo methylation (see e.g. Antequera and Bird, 1993). The involvement of this type of events in the formation of the observed linearities is indicated by the relatively higher linearities observed in mouse where higher island elimination rates are observed, as the above authors remark (see also discussion in the next sub-section). The inclusion of events type iv tests the robustness of the model,

**Fig. 3.** Inter-CGI spacers’ complementary cumulative size distributions in whole chromosomes for epigenetically determined CGIs from the human and mouse genome. Several cases of non-TSS or TSS-unrelated (orphan) islands are considered. Surrogate data curves and linear regression are similar to Fig. 1. Logarithms are to base 10.

because for many organisms important parts of the genome are generated by repeat proliferation. Events of type v represent either deletion of sequence regions usually due to unequal recombination or gradual shrinkage by a balance of indel (insertion/deletion) events favoring decrease of the sequence length.

The model presented herein is formally similar to another model introduced previously, in order to account for the appearance of a power-law-like pattern in the chromosomal distribution of genes (Sellis and Almirantis, 2009). Both are based on an analytically solvable model introduced by Takayasu et al. (1991) for the appearance of power-law size distributions in aggregative growth of particles in physicochemical systems.

Notice that the model presented herein, conceived to describe the genomic dynamics of CGIs, is not analytically solvable and thus no universal exponents (slopes for the linear segment in log–log scale) may be obtained. This is verified by all our computer simulations. Figure 4 demonstrates this clearly. We have plotted the complementary cumulative size distributions of inter-CGI spacers for several whole chromosomes in two representative organisms: \textit{A. mellifera} and \textit{B. taurus}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig4.pdf}
\caption{Plots of inter-CGI spacers' complementary cumulative size distributions in whole chromosomes. Here non-standard threshold sets are used, for more details see the supplementary file "grid". Surrogate data curves and linear regression are similar to Fig. 1. Logarithms are to base 10.}
\end{figure}

simulations and is in accordance with the variety of slope values met in the study of genomic CGI distributions. Thus, our data deviate from the typical power-law not only because they always have the linearity in log–log scale truncated at a lower and an upper cut-off (in fact, this is a feature common to all cases of naturally occurring “power-laws”) but mainly because they lack any universal exponent (slope). This is the principal reason why we call the pattern we have found “power-law-like” throughout this article. For a recent in depth view of the requirements for having a power-law, see Stumpf and Porter (2012). These authors state that these requirements include a statistically sound power-law (extended linearity in log–log scale with indications of convergence to a universal exponent) and a concrete underlying theory to support it. In our case we clearly show that the log–log linearities we observe in our genomic data (often being quite extended) lack a universal exponent (slope). On the other hand, this deviation from universality is a characteristic feature shared between the genomic inter-CGI distributional patterns we describe herein and the simulations of the proposed evolutionary model. This feature, along with the common dependence on the evolutionary parameters shared between model and genomic distributions, as found in several instances, corroborates the hypothesis that the evolutionary dynamics described by this model is at the origin of the observed genomic patterns.

In the simulation of Fig. 5a, only events of the types i and ii are included (segmental duplications followed by CGI loss). We see that, as observed in real chromosomes, power-law-like curves (linear in log–log scale) are formed in the inter-CGI spacers’ size distributions. A decrease in the number of allowed segmental duplications, depicted in Fig. 5b, decreases the extent of the linearity. If we include additional elimination of CGI (events of type iii), as intuitively expected, the linearity is considerably increased (see Fig. 5c). These results are compatible with extended power-law linearity found, in cases where an extended genomic remodeling in the recent past of the organism is plausible (perhaps, the case of the human genome) with many genes inactivated (e.g. the case of human olfactory receptors) and several genes having their regulatory pattern changed. We may infer that gene inactivation and reassignment of the roles of genes (which may have caused their emancipation from an associate CGI) led several CGIs to cease being under purifying selection and thus they progressively decomposed.

4. Discussion

4.1. Analysis of the observed power-law-like chromosomal CGI distributions. Taxonomy-related evidence

The property of the formulated simple model to produce power-law-like distributions is an indication that it could be extended to include more detailed modeling of conservation through evolutionary time, in line with the view that functionality is a necessary condition for CpG islands to form stable power-laws. The similarity of the power-law-like pattern emerging in simulations of this model with the genomic distributions of inter-CGI distances cannot be considered, however, as an absolute proof that this model is at the origin of the genomic pattern. Moreover, it is probable that the observed distributions are the result of the proposed mechanism in synergy with other aspects of genomic dynamics. More specifically, the expansion–modification model proposed by Li (1991), shown to generate long-range correlations in nucleotide sequences might have significantly contributed to power-law-like distributional patterns of several genomic components (repeats, protein coding segments, CGIs), especially in small genomes. For a related discussion see Klimopoulos et al. (2012).

In the previous paragraph we consider the possibility that the genomic power-law-like size distributions of CGIs might be an indication of functionality. The power-law inter-repeat size distributions of distances between TEs could be seen as a counter-example (Klimopoulos et al., 2012). However, as extensively discussed therein, TEs are long-lived genomic objects, under no purifying selection, recognizable in mammalian genomes up to
200My, (Ostertag and Kazazian, 2001) thus allowing the interplay of event types iii and iv of the proposed model. These are events not requiring evolutionary conservation of the involved genomic elements in order to generate power-law-like distributions. Contrary to TEs, Cpg Islands retain their identity in the evolutionary time scale (i.e. they stay above given thresholds for CG% and for CpgO/e only for relatively short times after ceasing to be under purification selection. This is due to the mutability of Cpg after being methylated, which makes its genomic turnover rate to be the highest of any other dinucleotide. CpG is, in fact, reported to disappear from a given position at least one order of magnitude faster than any other dinucleotide in the eukaryotic genome; see Nachman and Crowell (2000) and Antequera (2003).

An immediate observation stemming from our study is that the standard threshold sets (G-G&F and T&J) may produce power-law-like distributions not only in human, mouse and other mammalian genomes, from which they are principally derived, but also in a variety of other genomes. From log–log linearities found herein one might infer that an important proportion of the CGIs detected in all examined genomes using these thresholds are functional, on the basis of the argument presented above: i.e. that log–log linearity implies functionality for CGI populations. In H. sapiens, the maximum of the linearity in log–log scale is indeed observed for the standard threshold sets or for valuing in their close vicinity, with no signs of skewed deviations to any direction in the parameter (threshold) space for all analyzed chromosomes. The situation changes, to some extent, for several of the other studied genomes as may be revealed by an inspection of the data presented in the supplementary file “grid”. Especially in mouse, where CGIs are known to be eroded, i.e. they are decreased in number while the “surviving” ones are shorter and with lower CG% and CpgO/e values (Antequera and Bird, 1993; Matsuo et al, 1993) we observe that the maximum linear length is systematically positioned at lower threshold combinations than the standard ones. Human and mouse, and in fact other vertebrate genomes too, according to Antequera and Bird (1993), are losing Cpg Islands over evolutionary time due to de novo methylation in the germ line followed by Cpg loss through mutation. More specifically, type iii events of the proposed model occur according to the cited literature with a particularly higher rate in rodents. On the other hand, by simple inspection of Table 1 we see that the mouse genome presents for G-G&F the highest mean extent, i.e. highest mean value of log–log linearity (M.V.) of all examined genomes, this being a corroboration of the proposed model. Another point compatible with the model is that this optimum is reached for the relaxed (G-G&F) and not for the stringent (T&J) threshold set, as the erosion of mouse CGIs probably makes the relaxed thresholds more suitable for filtering in most of the functional CGIs. Note that this result does not contradict the earlier finding of Han and Zhao (2009) that the T&J threshold set is particularly suitable for identifying promoter-associated CGIs in human and probably other genomes as well. As already mentioned, in our study we attempt to identify the distributional patterns of CGI populations of different functionalities and not only the gene-transcription related, often focusing on their non-gene-related subset (orphan islands).

In other genomes the optimum (maximum linear extent) is positioned at lower value combinations, while sporadically high values of thresholds may perform better (see data in the “grid”). The existence of several organisms with higher power law linearities for lower thresholds and shorter minimum length (i.e. when $L_{\text{min}} = 200$ performs better than $L_{\text{min}} = 500$) is compatible with the finding that “many known functional CGIs are shorter than commonly assumed – the extreme example being Xenopus CGIs” (Hackenberg et al., 2006).

On the other hand, the observation of Han and Zhao (2008) that in fish genomes stringent thresholds (T&J) gave the best results on functional CGIs determination is compatible with our finding that in D. rerio, the lengthier log–log linear extents (the only ones above 2) are found in the higher threshold value region. The prevalence of the higher thresholds in the formation of the most extended linear segments in log–log scale suggests that in fish genomes, functional islands in general might be compositionally moved toward high CG % and CpgO/e values.

Another point where findings presented herein comply with findings of other research groups following a different approach is our observation that chromosome III of Saccharomyces cerevisiae is the one presenting the globally maximal extent of linear segment (E = 1.85) and the only one which exhibits linearity for alternative threshold set (altG-G&F and altT&J). This is in accordance with the findings of Li et al. (1998) and Bradnam et al. (1999), see also Sharp and Lloyd (1993), about chromosome III being atypical in many regards: It presents large-scale compositional (C+G content) inhomogeneity and high extent of compositional mosaicism of genes found only in this chromosome. It is also worthy to note that the two chromosomes (I and XI) reported by Bradnam et al. (1999) to be next to III in their degree of variation of silent-site C+G content are the ones with the next higher extents of linearity (1.6 and 1.5 respectively). These authors consider as the more likely reason for the particularities of chromosome III that it contains the mating type loci thus being under selective pressure and protecting it from transpositions and keeping it intact in evolutionary time. On the other hand, slowing down of the transposition dynamics, according to our proposed model, also has as effect the protection of the power-law-like pattern from random interruptions of the inter-CGI spacers thus maximizing linearity in log–log scale. This has also been shown numerically through computer simulations allowing random transpositions, where a decrease of the linearity extent in log–log scale was observed (figure not included herein).

4.2. Gene-related vs. orphan CGIs

The distributions of coding segments and of genes in a variety of genomes are shown to follow a power-law-like spatial distribution, long-rangeness and fractality, by using several different approaches (Sellis and Almirantis, 2009; Athanasopoulou et al., 2010). In order to exclude the possibility that the power-law-like spatial distribution of CGIs is a mere consequence of a power-law distribution of genes, given the systematic connection of many CGIs with promoters of protein-coding genes, we extended our analysis to the subset of CGIs characterized as orphan, i.e. not spatially connected to genes. Moreover, the separated study of orphan CGIs allows for the investigation of different modes of chromosomal distribution and thus of different functionalities of this class of genomic element compared to entire CGI populations. The observation that power-law-like linearity persists (and is sporadically more extended) in orphan CGI distributions is indicative of distinct functionalities for this type of genomic localizations.

The result of Han et al. (2008) that the dog genome hosts the highest percentage of gene-unrelated functional CGIs if compared to any studied genome is compatible, in view of the proposed model, with our result (see Table 2) that dog genome, for both standard threshold choices, represents the global optimum of linearity extent, when full gene and flanks masking is applied, to all examined genomes. The same optimum is reached if we examine the M.V.-5 values. Along the same lines is the observation that this genome is characterized by the maximum of increase of linearity when full gene and flanks masking is applied either for G-G&F or for T&J thresholds (if masked and unmasked genomes are compared). Considering mean values from all chromosomes as depicted in Tables 1 and 2, we note that this increase is much higher for the G-G&F than for the T&J threshold set (0.67 vs. 0.45 respectively). This...
In this work, we investigated the distributional features of CpG-island-containing gene segments in terms of patterns of DNA methylation, in order to associate DNA methylation with functional properties. We consider the presence of CpG islands as a functional constraint on DNA methylation, as CpG islands are known to be associated with gene expression. We hypothesized that CpG islands act as a constraint on DNA methylation, which is consistent with the observation that CpG islands are associated with gene expression. We also propose that the distribution of DNA methylation in CpG islands is a function of the DNA methylation of the surrounding DNA, which is consistent with the observation that DNA methylation is conserved between species.

Appendix A: Supplementary data

We would like to acknowledge Professors Savas J. Adams and George E. Church for their support and advice. We thank T.J. Tsiagkas for his help with the preparation of the manuscript.


