Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

The *Caenorhabditis elegans* genome: a multifractal analysis

P.E. Vélez[1], L.E. Garreta[2], E. Martínez[3], N. Díaz[3], S. Amador[3], I. Tischer[2], J.M. Gutiérrez[4] and P.A. Moreno[2]

[1]Departamento de Biología, Universidad del Cauca, Popayán, Colombia
[2]Escuela de Ingeniería de Sistemas y Computación,
Universidad del Valle, Santiago de Cali, Colombia
[3]Escuela de Ingeniería de Sistemas y Computación,
Universidad del Cauca, Popayán, Colombia
[4]Instituto de Física de Cantabria,
Universidad de Cantabria-CSIC, Santander, Spain

Corresponding author: P.A. Moreno
E-mail: pedro.moreno@correounivalle.edu.co

Genet. Mol. Res. 9 (2): 949-965 (2010)
Received January 10, 2010
Accepted February 25, 2010
Published May 25, 2010
DOI 10.4238/vol9-2gmr756

ABSTRACT. The *Caenorhabditis elegans* genome has several regular and irregular characteristics in its nucleotide composition; these are observed within and between chromosomes. To study these particularities, we carried out a multifractal analysis, which requires a large number of exponents to characterize scaling properties. We looked for a relationship between the genetic information content of the chromosomes and multifractal parameters and found less multifractality compared to the human genome. Differences in multifractality among chromosomes and in regions of chromosomes, and two group averages of chromosome regions were observed. All these differences were mainly dependent on differences in the contents of repetitive DNA. Based on these properties, we propose a nonlinear model for the structure of the *C. elegans* genome, with some biological implications. These results suggest that examining

 ©FUNPEC-RP www.funpecrp.com.br

differences in multifractality is a viable approach for measuring local variations of genomic information contents along chromosomes. This approach could be extended to other genomes in order to characterize structural and functional regions of chromosomes.

Key words: *Caenorhabditis elegans* genome; Multifractal analysis; Repetitive DNA sequences; Chaos game representation

INTRODUCTION

Several tens of Eukaryotic genomes have already been sequenced. These genomes have revealed a complex structure (Lewin, 2008) that is highly regionalized, which can be studied by methods that allow measuring how the information content is fragmented.

Information theory has been a very useful conceptual framework to study the information content along a symbol (or signal) sequence. Its beginnings can be traced to discussions about entropy of information. These concepts were rigorously formulated in other terms by Shannon (1948) in the theory of information for communication. During the 60's Mandelbrot (1982) established fractal geometry, as a new geometry able to measure the irregularity of nature. The fractal geometry paradigm introduces several ways to measure this information content by calculating the fractal dimension, an exponent derived from a power law, which gives us an idea of the fragmentation level or information content of a complex phenomenon. Fractal geometry is a useful approach for searching for self-similar properties in biological structures and processes (Goldberger, 1996). These self-similarities are organized following scaling or power laws, which emerge from the spatial-temporal, extended interaction of the different system components (Voss, 1992). These systems have a complex structure, which result from deterministic as well as random events, leading to an apparently disordered distribution. Thus, power laws are powerful tools to quantify the scaling properties of information contents.

Fractal geometry has been a useful approach for addressing several problems related to coding and non-coding DNA sequences, phylogenetic relationships, and for searching for explanations of regularities observed in molecular data bases. These studies were initially focused on long-range correlation properties in DNA sequences. Other surveys applying power laws have been reported (for a detailed review, see Yu et al., 2001).

Sometimes, one exponent may not be enough to characterize a complex phenomenon; so that more exponents are required. Multifractal formalism allows for this. It is applied when many fractal subsets with different scaling properties (with a large number of exponents or fractal dimensions) coexist simultaneously. As a result, when a spectrum of singularities of multifractal measure is generated, the scaling behavior of the symbol frequencies of a sequence can be quantified.

Multifractal analysis has been implemented to improve characterization of spatial inhomogeneity of both theoretical and experimental fractal patterns (Hentschel and Procaccia, 1983; Peitgen et al., 1992) in physics, chemistry (Stanley and Meakin, 1988), economy, and ecology. For example, it has been applied to study the turbulence phenomenon, time series analysis, and financial modeling. It has also been useful in studying different kinds of problems at DNA and protein sequence levels, under two modalities: 1) by using subintervals in a

 ©FUNPEC-RP www.funpecrp.com.br

one-dimensional (1-D) space to represent substrings and 2) by using a 2-D space in the chaos game representation context (Jeffrey, 1990). For a detailed review using the chaos game representation, see Yu et al. (2004).

The first modality was applied in pre-genomics times, to study DNA sequences by making spectral and multifractal analyses of measurements (Berthelsen et al., 1994). It was then used to reconstruct phylogeny from mitochondrial DNA (Glazier et al., 1995). Subsequently, in post-genomics times, multifractal analysis was used in discriminate analyses of complete bacterial genomes (Yu et al., 2001) and to distinguish coding and non-coding sequences in DNA sequences (Zhou et al., 2005). It has been useful in bacterial analysis to construct phylogenetic trees (Yu et al., 2003) and clustering of protein structures due to hydrophobic free energy and solvent accessibility of proteins (Yu et al., 2006). Recently, multifractal analysis has shown its utility in predicting the existence of promoters for RNA polymerase II (Yang et al., 2008).

In the second modality, multifractality analysis applied to chaos game representation patterns was first used on DNA discrete sequences (Gutiérrez et al., 1998); this was extended to a genomic level in a preliminary study (Moreno et al., 2000). Any influence of coding and non-coding content was observed in the correlation length of a given sequence (Gutiérrez et al., 2001). Bacteria classification based on the detailed Hydrophobic-Polar model in proteins revealed the utility of this approach (Yu et al., 2004).

In short, most of these approaches have practical applications for classifying microorganisms, distinguishing coding and non-coding sequences, studying proteins, and for promoter prediction. Much less attention has been paid to their use to characterize structural genetics information content. To apply multifractal formalism to this problem, we chose the *Caenorhabditis elegans* genome.

For a long time, *C. elegans* has been an excellent model for various types of studies (Wood, 1987; Brenner and Miller, 2002). Its genome was the first sequenced eukaryotic genome. Its molecular structure is particularly homogenous along the chromosome sequences (Anonymous, 1998), due to the presence of several regular characteristics. For example, with an average size of 3000 bp, genes are usually small and close to 50% of introns are extremely small, being about 47 bp long (Spieth and Lawson, 2006). Also, several bioinformatics studies have identified a remarkable periodicity for AA/TT dinucleotides, with periodicity at each 10 bp (hyperperiodicity; Moreno-Herrero et al., 2006); DNA periodicity in exons is determined by codon usage frequencies (Eskesen et al., 2004). All these characteristics contribute to producing homogenous nucleotide topography along the *C. elegans* genome.

Sequencing of the *C. elegans* genome revealed a regional organization of chromosomes, given by a distinctive distribution of some repeats and highly conserved genes (Anonymous, 1998). That is, the *C. elegans* genome also has some irregular characteristics. For example, gene density is a little higher in the central regions of the chromosomes than in the chromosome arms, and the vast majority of several kinds of DNA repeats are located in the chromosome arms rather than in the central regions (Figure 1, Panel A1). Additionally, mutation (Wolfe et al., 1989) and recombination (Wood, 1987) rates are more frequent in the chromosome arms, and (G + C) contents are slightly higher in the chromosome arms than in their central regions.

Given these regular and irregular characteristics, we decided to determine if some of these distributions can be analyzed by a multifractal formalism.

MATERIAL AND METHODS

Data bases, sequences, and multifractal approaches

Chromosome 21 from *Homo sapiens* [hs_ref chr21 build 36.1, contig NT_011512.10 (from 28.6 Mbp we took ~15 Mbp)] and *C. elegans* NC_003279-84.fna, data bases were downloaded from the NCBI web site. Four approaches were used to measure multifractality in the *C. elegans* genome. 1) In comparison with the human chromosome. The multifractal spectra for chromosome 21 of *H. sapiens* and chromosome I of *C. elegans* were studied. 2) By chromosome. Here the DNA sequences of five allosomes (I-V) and one sexual chromosome (X) were analyzed. 3) By chromosome region. We took a straightforward approach: the complete length of each chromosome was divided by 3. These regions were called L (left), C (center), and R (right), and the resulting 18 DNA sequences were analyzed by multifractal analysis (See Figure 1, Panel A1). 4) By group averages. Here the molecular and multifractal parameters for 6 C regions of the *C. elegans* chromosomes were averaged and compared with the averages of the 12 L and R regions of the chromosomes.

Molecular parameters: contents of (G + C), DNA repeats, and genes

Using scripts in Python, the (G + C) and repeat TTAGGC contents and gene density were calculated for each chromosome and fragment of chromosomes of *C. elegans*. Likewise, the dinucleotide AA/TT contents were counted for each DNA fragment for the motifs: $AAN_{10}AA$ and $TTN_{10}TT$. Additionally, other DNA repeat (AAAAT, AAATT, TTTTC, and TTTTT) contents were examined in each chromosome fragment.

Chaos game representation of DNA sequences

According to Jeffrey (1990) and Yu et al. (2004), the chaos game representation for a given DNA sequence $s = s_1, \ldots. s_N$ is obtained as follows: the corners of the square [0, 1] x [0, 1] are labeled with A, C, G, and T. The first point of s is plotted halfway between the center of the square and the corner corresponding to the first nucleotide of s; the i-th point of s is then plotted halfway between the (i - 1)-th point and the corner corresponding to the i-th nucleotide. An example of a chaos game representation is shown in Figure 1, Panel A2.

Multifractal analysis and discrimination analysis

In spite of arguments for the sandbox method (Tél et al., 1989; Yu et al., 2004), we decided to use the box-counting method for our multifractal spectrum estimation, because it counts boxes in a straightforward manner and maintains statistics allowing dimension calculation (Peitgen et al., 1992). Given a grid of mesh size ε over the fractal, the number M_i of fractal points that fall in the i-th grid box is determined and related to the total number M_0. The multifractal spectrum D_q^{bc} based on the box-counting method (Mandelbrot, 1982; Halsey et al., 1986) is defined as:

$$D_q(\varepsilon) = \frac{\ln(\sum_i (M_i/M_0)^q)}{\ln(\varepsilon)} \frac{1}{q-1}, \qquad \text{(Equation 1)}$$

where q is an integer parameter. The multifractal dimension is defined as

$$D_q = \lim_{\varepsilon \to 0} D_q^{bc}(\varepsilon) \qquad \text{(Equation 2)}$$

A practical use of the box-counting method transforms Equation 1 into

$$\ln(M_i^q) = D_q^{bc}(\varepsilon)(q-1)\ln(\varepsilon) + (q-1)\ln(M_0^q) \qquad \text{(Equation 3)}$$

and performs a linear fit of $\ln(M_i^q)$ *vs* $(q - 1)\ln(\varepsilon)$, using sample data of M_i. The slope that is obtained is an approximation of the multifractal dimension D_q. From Equation 3 it can be inferred that a high D_q for positive q values emphasizes regions where the points are dense; for negative q values, it is associated with the structure and properties of sparse regions. Directly from the multifractal dimension D_q, the correlation exponent $\tau(q)$ is derived as $\tau(q) = (q - 1)D_q$, which will also be used in some of our analyses. The degree of multifractality, ΔD_q, is defined as the difference between maximum and minimum values of D_q: $\Delta D_q = D_{q\max} - D_{q\min}$ (Dewey and Strait, 1996; Ivanov et al., 1999). If ΔD_q is high, the multifractal spectrum is rich in information; for a small ΔD_q, the resulting dimension spectrum is poor in information.

In this study, the effects of the different DNA lengths on multifractality were not taken into account (for theory concerning this, see Fu et al. (2007)). Finally, the discrimination analyses were performed by using plots in 2-D and 3-D spaces, with combined molecular and multifractal parameters.

RESULTS

The box-counting method used here proved to be a good estimator for multifractal parameters in the *C. elegans* genome. It has been argued that the sandbox method is a more precise method than the box counting method (Yu et al., 2004). However, the box counting method is easy, has high computational efficiency and the program can be carried out for shapes with and without self-similarity (Peitgen et al., 1992), as those patterns shown for the chaos game representation. We used four approaches of multifractal analysis to examine the *C. elegans* genome.

1. Multifractal comparison between the *C. elegans* chromosomes and human chromosomes

The *C. elegans* multifractal spectra revealed marked multifractality reduction compared with human DNA sequences (see also Moreno et al., 2000). Here, we examined whether such a behavior is consistent at the chromosome level. Due to its intermediate repeat and gene densities, we selected human chromosome 21 as a landmark for comparison (Venter et al., 2001). Chromosome 21 is more multifractal (0.27 more) than the *C. elegans* chromosome I (Figure 1, Panels B1,B2; Table 1), suggesting that the genetics information along the *C. elegans* chromosome is more regularly ordered in comparison with the human genome. Human chromosome 21 is known to have a highly asymmetric complex structure that is aperiodic and rich in heterogeneous genetic information (Venter et al., 2001).

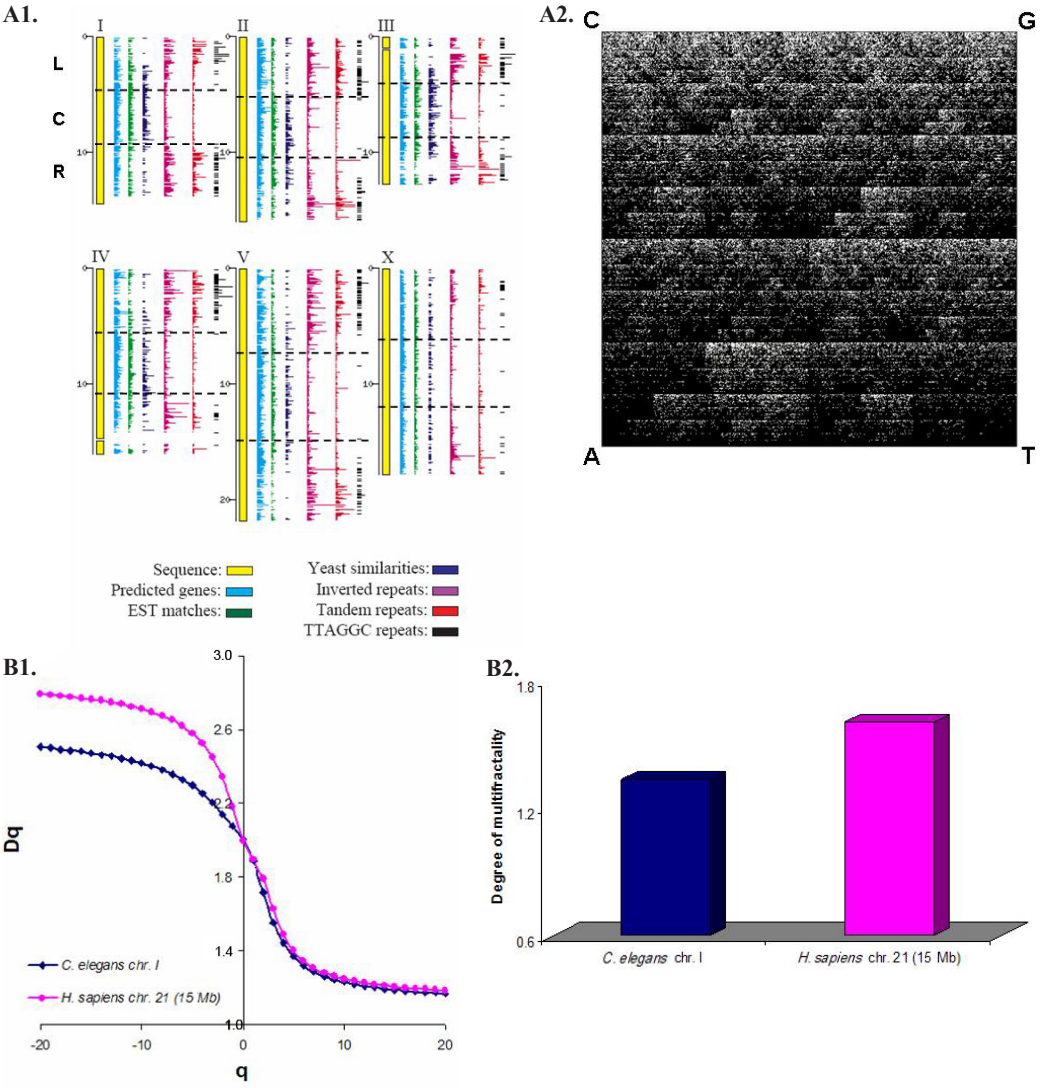


Figure 1. A1. Map of the *Caenorhabditis elegans* genome showing regular and irregular distributions in the contents of predicted genes and repetitive DNA (adapted from Figure 3 in Anonymous, 1998). The horizontal dashed lines show how each chromosome was divided into three regions (L = left; C = center; R = right). **A2.** Chaos game representation of a *C. elegans* chromosome I fragment (~500 kbp). Note the high concentration of (A + T) content in the sub-quadrants AA and TT. Multifractal comparative analysis: **B1.** Comparison between the multifractal spectra for *Caenorhabditis elegans* and *Homo sapiens*. **B2.** Multifractality for both chromosomes.

Table 1. Multifractal parameters for two eukaryote chromosomes.

	D-20	D-2	D-1	D1	D2	D20	ΔDq
C. elegans chr. I	2.5044	2.1416	2.0773	1.8894	1.7143	1.1679	1.3365
H. sapiens chr. 21 (15 Mb)	2.7909	2.3422	2.1795	1.8947	1.7918	1.1843	1.6066

D1 is called the information dimension and D2 the correlation dimension of the measurement.

2. Multifractal analysis by chromosome

We studied the multifractal behavior of each *C. elegans* chromosome. We obtained the generalized dimension spectrum for each chromosome and used it to examine genetic information content (Figure 1, Panel A1). Chromosome III of *C. elegans* has high multifractality, whereas chromosome X has the lowest multifractality, suggesting strong periodicity of the nucleotides along the chromosome X sequence, and a high aperiodicity on chromosome III (Figure 2A,B; Table 2A). Discrimination analysis placed chromosomes X and III of *C. elegans* at opposite coordinates along the linear regression (Figure 2C). Likewise a low density of genes and repeats was found on chromosome X when compared with the other chromosomes (Table 2B).

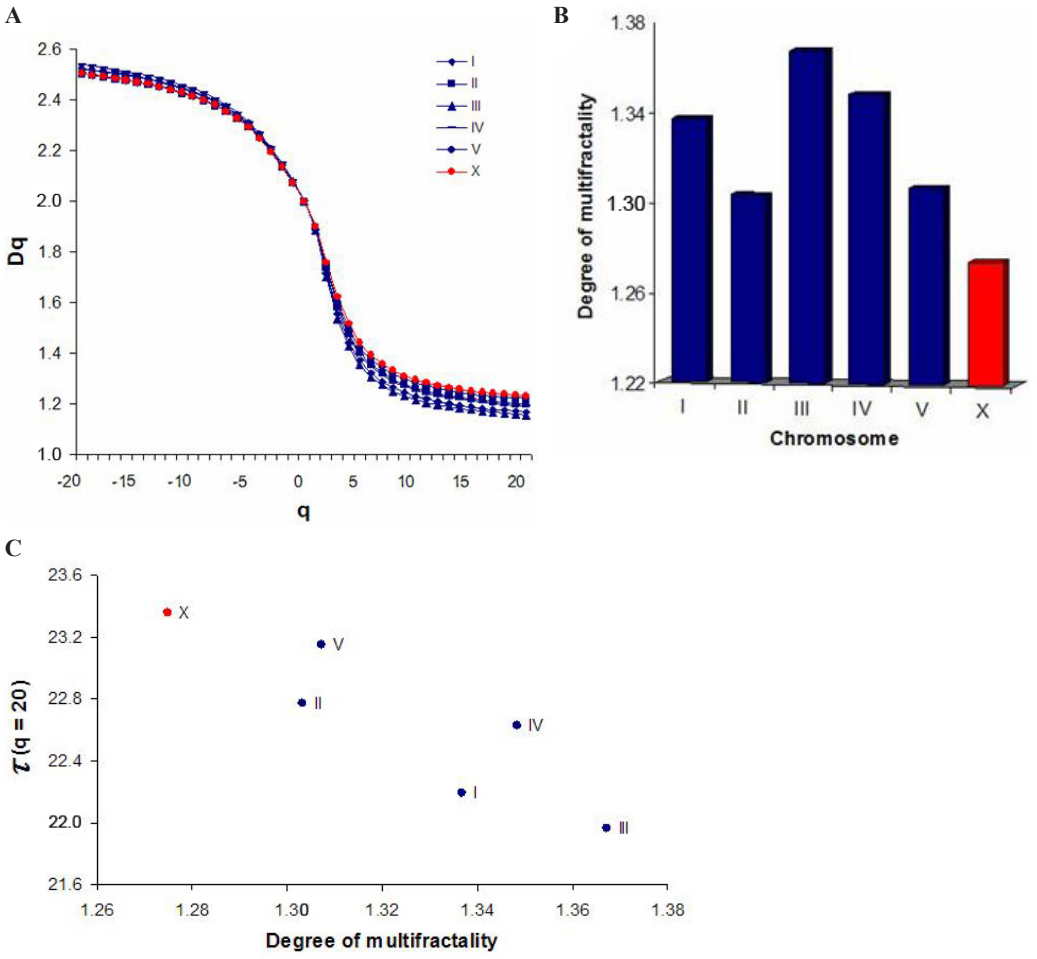


Figure 2. Multifractal analysis by chromosome. **A.** Dimension spectra for the *Caenorhabditis elegans* chromosomes. **B.** Degree of multifractality of Figure 2A. **C.** Discrimination method based on multifractal formalism in a distribution of two-dimensional points, R^2 ~0.8.

Table 2.

A. Multifractal parameters by chromosome.

Chr.	D-20	D-2	D-1	D1	D2	D20	ΔDq	τ(q = 20)
I	2.505	2.142	2.077	1.889	1.714	1.168	1.337	22.190
II	2.502	2.137	2.073	1.899	1.74	1.198	1.303	22.770
III	2.523	2.145	2.079	1.885	1.702	1.156	1.367	21.962
IV	2.54	2.151	2.082	1.885	1.718	1.191	1.348	22.631
V	2.526	2.14	2.075	1.897	1.743	1.218	1.307	23.149
X	2.504	2.134	2.073	1.901	1.76	1.229	1.275	23.358

B. Molecular parameters by chromosome.

Chr.	Size (Mb)	(G + C)	TTAGGC	Density (repeat/Mb)	Genes	Density (genes/Mb)
I	15.07	35.75	3742	248	3327	221
II	15.28	36.2	3466	227	3968	260
III	13.78	35.66	3195	232	3168	230
IV	17.49	34.6	2890	165	3706	212
V	20.92	35.43	3912	187	5392	258
X	17.72	35.2	2271	128	3276	185

3. Multifractal analysis by chromosome region

The most irregular characteristic exhibited by the *C. elegans* genome was examined. With the exception of chromosome X, this asymmetric distribution is strongest on chromosomes I-V. Hence, we divided each chromosome into three fragments of DNA sequences (Figure 1, Panel A1). A detailed analysis for each generalized dimension spectrum revealed that the C regions of each chromosome have somewhat less multifractality than the multifractal spectra of the L and R regions (Figure 3A, Table 3A). The scaling exponent $\tau(q)$ can reveal other aspects of chromosome structure. Monofractal behavior would correspond to a straight line for $\tau(q)$; for multifractal behavior, $\tau(q)$ is nonlinear. The differences between the curves for chromosome arms and C regions can be seen in Figure 3B. The changing curvature for the data for the chromosome arms indicates multifractality. In contrast, $\tau(q)$ tends to be linear for the C regions of the chromosomes, indicating partial loss of multifractality.

We hypothesized that these two kinds of behaviors were due to different repetitive DNA sequence contents. Counts of different kinds of repetitive DNA sequences confirmed a clear contribution of these sequences in increasing the multifractality degree in the L and R regions of *C. elegans* chromosomes (Table 3B). A similar result can be observed by calculating the percentage in size of seven types of repeats present in each chromosome fragment (Figure 3C, Table 3C). The densities of DNA repeats were also found to be correlated with the multifractal degree (Table 3B; Supplementary material). Discrimination methods based on distributions of 2-D and 3-D points show how C regions can easily be discriminated from L and R region points (Figure 3D,E). Plots (D-1, D1, D-2) also allow discriminating C regions from L and R regions (Supplementary Material).

 ©FUNPEC-RP www.funpecrp.com.br

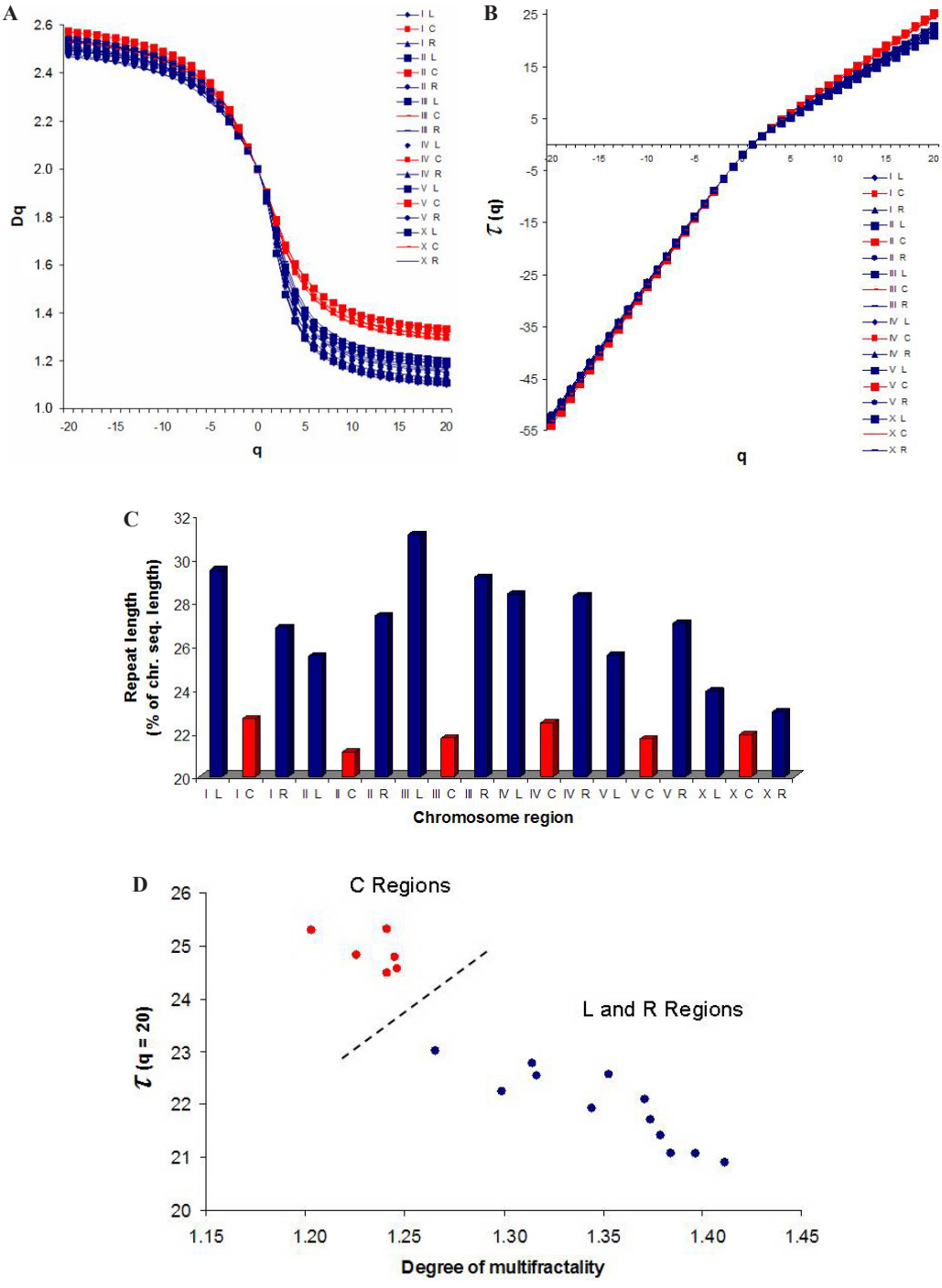


Figure 3. Continued on next page.

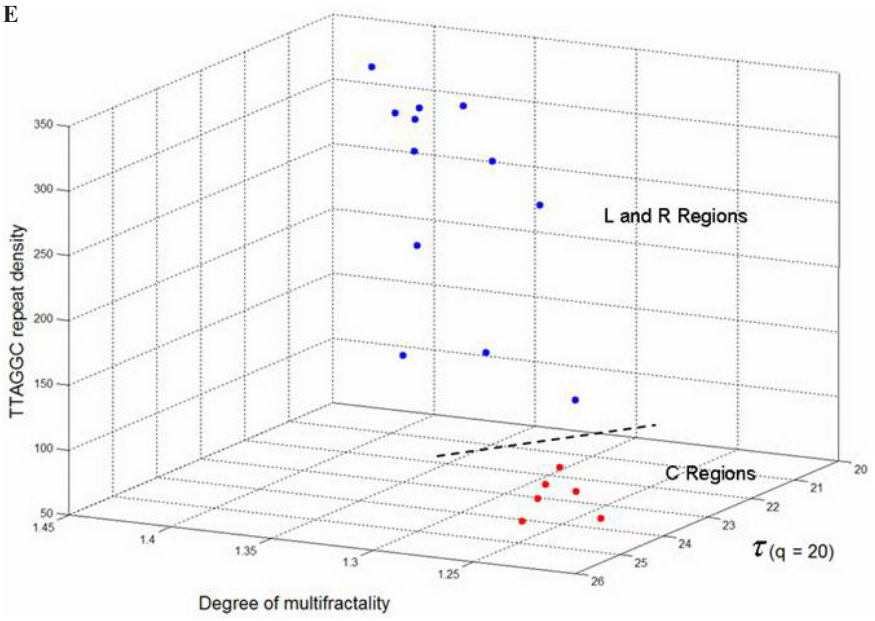


Figure 3. Multifractality by chromosome region. **A.** Generalized dimension spectrum for each chromosome region. **B.** Multifractal spectrum $\tau(q)$ for each chromosome region. **C.** Percentage of length for seven kinds of repeats. **D.** Discrimination method based on a distribution of two-dimensional points, $R^2 = 0.914$. Blue points, L and R chromosome regions; red points, chromosome C regions. This multifractal approach discriminates C regions from L and R chromosome regions. **E.** Discrimination method based on two multifractal parameters and one molecular parameter. Each chromosome fragment dataset is characterized by three quantities. The first quantity (x-axis) is the degree of multifractality for each chromosomic region. The second quantity (y-axis) is the density of TTAGGC content of the chromosomic fragments. The third quantity (z-axis) is the exponent value $\tau(q = 20)$, characterizing the scaling of the 20th moment. L = left; C = center; R = right.

Table 3A. Multifractal parameters by chromosome region.

Chr. region	D-20	D-2	D-1	D1	D2	D20	ΔDq	$\tau(q = 20)$
I L	2.512	2.151	2.084	1.872	1.653	1.100	1.412	20.907
I C	2.539	2.170	2.090	1.895	1.766	1.293	1.246	24.564
I R	2.498	2.142	2.078	1.889	1.704	1.154	1.345	21.923
II L	2.503	2.138	2.074	1.898	1.728	1.186	1.317	22.541
II C	2.534	2.168	2.088	1.902	1.786	1.331	1.203	25.290
II R	2.506	2.140	2.077	1.887	1.686	1.126	1.379	21.403
III L	2.494	2.151	2.085	1.868	1.649	1.109	1.384	21.076
III C	2.533	2.163	2.086	1.902	1.776	1.307	1.226	24.828
III R	2.505	2.151	2.084	1.873	1.660	1.109	1.397	21.068
IV L	2.516	2.144	2.080	1.882	1.688	1.142	1.374	21.699
IV C	2.550	2.174	2.092	1.894	1.766	1.305	1.245	24.787
IV R	2.533	2.165	2.092	1.870	1.684	1.162	1.371	22.082
V L	2.540	2.144	2.077	1.894	1.723	1.187	1.353	22.553
V C	2.574	2.172	2.090	1.899	1.782	1.333	1.241	25.325
V R	2.469	2.141	2.077	1.890	1.708	1.170	1.299	22.239
X L	2.513	2.139	2.075	1.898	1.742	1.199	1.314	22.773
X C	2.529	2.152	2.081	1.901	1.774	1.288	1.241	24.474
X R	2.477	2.141	2.076	1.900	1.754	1.211	1.266	23.005

L = left; C = center; R = right.

Table 3B. Number of repeats and genes per chromosome region.

Chr. region	Size (Mb)	(G + C)	TTAGGC repeat		AAN10AA repeat			TTN10TT repeat			Other kinds of repeats				Genes		Repeats + Genes	
			N	D	N	D	N-Ov.	N	D	N-Ov.	AAAAT	AAATT	TTTTC	TTTTT	N	D	N	D
I L	5.024	36.07	1654	329.2	124641	24808	63193	126753	25229	63932	55921	51170	37546	47808	935	186.1	1193333	79173.3
I C	5.024	34.95	451	89.8	103877	20676	59454	103134	20528	59314	38912	32751	28421	32159	1329	264.5		
I R	5.024	36.23	1637	325.8	114183	22727	59667	117460	23379	60634	50163	44806	34621	41938	1063	211.6		
II L	5.093	36.77	1514	297.3	110013	21600	59063	112143	22019	60444	47598	43512	32377	39396	1421	279	1131346	74044.3
II C	5.093	35.38	416	81.7	99535	19543	58576	96016	18852	57042	36488	30102	26360	29362	1486	291.8		
II R	5.093	36.46	1536	301.6	121726	23900	62169	121519	23860	61966	51863	48256	34320	43326	1061	208.3		
III L	4.594	35.42	1401	304.9	122495	26662	60976	118161	25718	59328	56724	51000	35583	44558	930	202.4	1137026	82492.9
III C	4.594	35.75	422	91.9	90061	19602	52138	91072	19822	52290	33580	28918	25549	28047	1312	285.6		
III R	4.594	35.81	1372	298.6	117787	25637	58870	114501	24922	58013	50257	47445	33055	41870	926	201.5		
IV L	5.831	35.75	1647	282.4	142846	24497	73738	140951	24172	73099	62164	56940	41765	50738	1199	205.6	1383016	79057.6
IV C	5.831	34.58	479	82.1	117630	20172	68663	118918	20393	69111	44374	37775	31760	37269	1522	261		
IV R	5.831	33.46	764	131.0	142940	24513	77472	146630	25145	79149	60292	53631	39272	50525	985	168.9		
V L	6.974	35.83	1575	225.8	153860	22062	84188	154165	22105	84139	64752	58576	43440	53227	1779	255.1	1542348	73718.1
V C	6.974	34.83	508	72.8	137317	19690	81552	136909	19631	81178	51137	42685	36948	41793	1928	276.5		
V R	6.974	35.63	1829	262.3	159845	22920	85562	161213	23116	86478	69785	65147	46444	55801	1685	241.6		
X L	5.906	35.47	905	153.2	127766	21632	70726	125146	21189	69943	50176	42574	33668	42777	1097	185.7	1215845	68618.7
X C	5.906	34.87	605	102.4	118391	20045	69304	118108	19997	68990	43379	35576	31133	37507	1193	202		
X R	5.906	35.27	761	128.8	122532	20746	69575	122211	20692	69346	46830	39803	31836	40885	986	166.9		

N: number. D: density. N-Ov.: non-overlapped. L = left; C = center; R = right. Size and (G + C) content are indicated.

©FUNPEC-RP www.funpecrp.com.br

Table 3C. Percentage in size of repetitive DNA for 18 chromosome fragments.

Chr. region	TTAGGC	AAN10AA	TTN10TT	AAAAT	AAATT	TTTTC	TTTTT	Percent
I L	9924	252772	255728	279605	255850	187730	239040	29.47
I C	2706	237816	237256	194560	163755	142105	160795	22.67
I R	9822	238668	242536	250815	224030	173105	209690	26.84
II L	9084	236252	241776	237990	217560	161885	196980	25.55
II C	2496	234304	228168	182440	150510	131800	146810	21.14
II R	9216	248676	247864	259315	241280	171600	216630	27.38
III L	8406	243904	237312	283620	255000	177915	222790	31.10
III C	2532	208552	209160	167900	144590	127745	140235	21.78
III R	8232	235480	232052	251285	237225	165275	209350	29.14
IV L	9882	294952	292396	310820	284700	208825	253690	28.39
IV C	2874	274652	276444	221870	188875	158800	186345	22.46
IV R	4584	309888	316596	301460	268155	196360	252625	28.29
V L	9450	336752	336556	323760	292880	217200	266135	25.56
V C	3048	326208	324712	255685	213425	184740	208965	21.75
V R	10974	342248	345912	348925	325735	232220	279005	27.03
X L	5430	282904	279772	250880	212870	168340	213885	23.94
X C	3630	277216	275960	216895	177880	155665	187535	21.92
X R	4566	278300	277384	234150	199015	159180	204425	22.98
Average								25.41

The percentages were calculated as follows: Percent = ∑(for 7 types of repeats*)100 / chromosome region size. *Number of repeats multiplied by repeat size. For the AA/TT dinucleotides, we used the number of non-overlapped repeats (in Table 3B) multiplied by 4.

4. Multifractal analysis by group averages

We also examined the multifractality by group averages of chromosome regions. For all grouped L and R regions there is a clear increase in multifractality in comparison with that displayed by all grouped central regions (Figure 4A, Table 4). For all grouped L and R regions, τ(q) is a nonlinear function, indicating that the nucleotide composition in the *C. elegans* genome is more aperiodic in the chromosome arms than in the grouped C regions (Figure 4B).

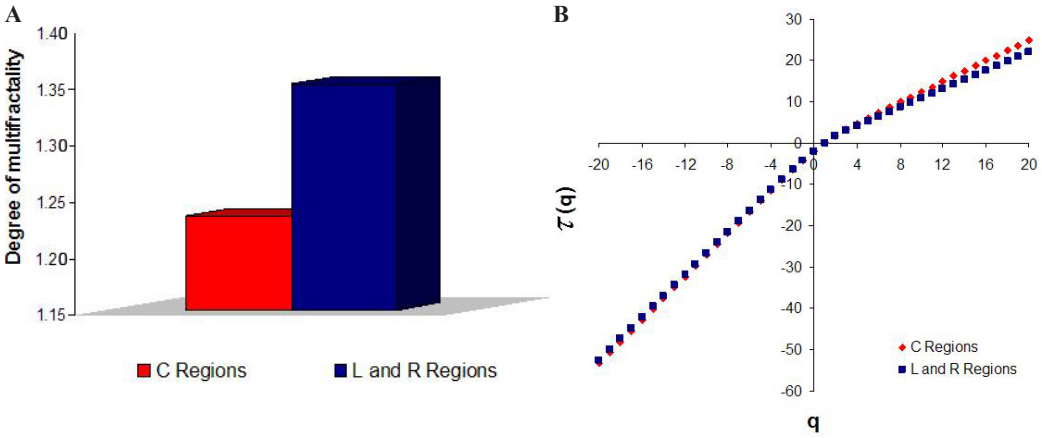


Figure 4. Multifractal analysis by group averages. **A.** Multifractal degree for averaged groups. **B.** Multifractal spectra $\tau(q)$ for the two group averages.

Table 4. Multifractal parameters by group averages of chromosome regions.

Chr. region	D-20	D-2	D-1	D1	D2	D20	ΔDq
I C	2.539	2.17	2.09	1.895	1.766	1.293	1.246
II C	2.534	2.168	2.088	1.902	1.786	1.331	1.203
III C	2.533	2.163	2.086	1.902	1.776	1.307	1.226
IV C	2.55	2.174	2.092	1.894	1.766	1.305	1.245
V C	2.574	2.172	2.09	1.899	1.782	1.333	1.241
X C	2.529	2.152	2.081	1.901	1.774	1.288	1.241
Average							1.234
I L	2.512	2.151	2.084	1.872	1.653	1.1	1.412
I R	2.498	2.142	2.078	1.889	1.704	1.154	1.345
II L	2.503	2.138	2.074	1.898	1.728	1.186	1.317
II R	2.506	2.14	2.077	1.887	1.686	1.126	1.379
III L	2.494	2.151	2.085	1.868	1.649	1.109	1.384
III R	2.505	2.151	2.084	1.873	1.66	1.109	1.397
IV L	2.516	2.144	2.08	1.882	1.688	1.142	1.374
IV R	2.533	2.165	2.092	1.87	1.684	1.162	1.371
V L	2.54	2.144	2.077	1.894	1.723	1.187	1.353
V R	2.469	2.141	2.077	1.89	1.708	1.17	1.299
X L	2.513	2.139	2.075	1.898	1.742	1.199	1.314
X R	2.477	2.141	2.076	1.9	1.754	1.211	1.266
Average							1.351

DISCUSSION

We discovered a strong relationship between the multifractal parameters and part of the genetic information coded by the *C. elegans* genome.

Initially, the nematode's chromosomes were found "poor" in genetic information when compared to the human chromosomes. For instance, chromosome I of *C. elegans* has a more periodic behavior than chromosome 21 of *H. sapiens* (Figure 1, Panel B1). This is due to numerous regularly arranged elements in the *C. elegans* genome (Eskesen et al., 2004; Moreno-Herrero, 2006; Spieth and Lawson, 2006; Figure 1, Panels A1, A2). In contrast, human chromosome 21 is richer in variable non-coding sequences (~98%; Venter et al., 2001) than the *C. elegans* chromosome (~74%; Anonymous, 1998).

Subsequently, we found that the nematode's autosomes are more multifractal than the sexual chromosome X, confirming what is known about the molecular structure of the X chromosome, which is particularly poor in genes and repeated DNA sequences (Figure 2, Table 2B; Anonymous, 1998). This difference can also be explained by determining the density by chromosome for genes and repeats from Table 3B (~79173, 74044, 82492, 79057, 73718, 68618 for chromosomes I, II, III, IV, V, X, respectively). Therefore, we suggest that chromosome X has a less stable nonlinear structure, which could generate greater genomic instability compared to the autosomes (especially chromosome III).

The nematode's chromosome arms were found to be more multifractal than the central regions of the chromosomes (Figure 3A,B; Table 3A). These results reveal that chromosome arms contain significant correlation structure; these correlations are lower in the central regions of chromosomes. These results also indicate that the observed multifractality is related to nonlinear distributions in the chromosome arms; this is particularly evident in regions where $q>>1$, which are densely populated with nucleotides (Figure 3A), have considerable repetitive DNA. Thus, the repetitive DNA contents located in the L and R regions account for the high aperiodicity and genetic variability of these regions. In other words, the multifractal character of these regions can vary greatly due to TTAGGC contents, inverted and tandem

repeats (Figure 1, Panel A1 and Figure 3C). We only measured 25% of repetitive element size. This seems to contribute to increasing the degree of multifractality. However, other repetitive DNA elements that were not taken into account might also be involved, as well as other genomic characteristics, for example, high mutation and recombination rates (Wolfe et al., 1989) and G + C contents that are mainly observed in the chromosome arms (Anonymous, 1998).

Similarly, group averages are more multifractal in chromosome arms than in central regions. This approach clearly separates the 18 fragments of chromosome regions into two averaged groups; central regions fall into one group and the L and R regions into the other (Figure 4A). For all chromosome arms, $\tau(q)$ is a nonlinear function (Figure 4B), indicating that the molecular structure of the chromosome arms has a multifractal behavior. For the central chromosome regions, $\tau(q)$ tends to be closer to linear behavior, especially for $\tau(q \geq 4)$, indicating partial loss of multifractality (Figure 4B). These results suggest that nucleotide fluctuations are less anti-correlated in central regions than in the chromosome arms. A similar behavior that will discriminate group averages has also been observed in another system (Ivanov et al., 1999). These changes in multifractality are also related to the density of group averages for genes and repeats, as can be calculated from Table 3B (~81583 for L and R regions and ~65386 for C regions).

Finally, we propose a nonlinear model for the structure of the *C. elegans* genome, with some biological implications. From a multifractal genomics perspective, the detection of a multifractal scaling in the *C. elegans* genome is of interest because it indicates that the chromosome molecular structure might be organized as a system operating far from equilibrium (Meneveau and Sreenivasan, 1987; Ivanov et al., 1999). This operation can be discussed from three points of view.

1. When the *C. elegans* genome is compared to the human chromosome, it seems to be less aperiodic. The somewhat lower multifractality of the *C. elegans* genome is due to its high structural homogeneity, due to many regular characteristics. Lowered non-linearity has been observed in many vital systems close to equilibrium (Goldberger, 1996; Burgos and Moreno-Tovar, 1996). For example, in the dynamics of the human heartbeat, the loss of multifractality is related to heart failure (Ivanov et al., 1999). That is, the system is operating dangerously close to the equilibrium. In consequence, the lifespan can be dramatically reduced. Thus, we suggest that the *C. elegans* genome might be operating closer to equilibrium than the human genome in order to maintain a simple life cycle.

2. When the *C. elegans* chromosomes are compared among themselves, less multifractality is observed for the X chromosome. That means the X chromosome is operating near equilibrium, which can be translated as major genetic instability. In *C. elegans*, sex is determined by the opposing actions of X and autosomal signaling elements (Gladden et al., 2007). Additionally, correct chromosome segregation depends upon three coupled processes during meiosis: chromosome pairing, synapses, and recombination (Garcia-Muse and Boulton, 2007; Zetka, 2009). Therefore, the instability of X might contribute to the molecular mechanism of sex determination during meiosis in a selective way (XX or X0). Thus, the X chromosome would be operating close to equilibrium in order to maintain its particular sexual dimorphism.

3. When the *C. elegans* genome is compared by chromosome regions, changes in multifractality are found that are related to the regional organization (in center and arms) exhibited by the chromosome. These behaviors are associated with changes in the contents of repetitive DNA. These results indicate that the chromosome arms are even more complex than previously suspected; since it is there that we found nonlinear properties that take the chromosome away from equilibrium. For example, part of the machinery that regulates several

physiological processes related to cell proliferation, genome stability, and aging are located in the telomeric regions (Im and Lee, 2005; Cheung and Deng, 2008) in the chromosome arms. Thus, the TTAGGC telomeric sequences would be operating far from equilibrium (perhaps as an attractor of mutations) to protect the genetic information coded by the entire chromosome.

All these biological arguments might explain why the *C. elegans* genome is organized in a nonlinear manner.

CONCLUSIONS

We report evidence for multifractality in the *C. elegans* genome. We demonstrated that this genome has a more periodic behavior than the human genome. We also found changes in multifractality based on other approaches. These changes in multifractality depend upon changes in contents and densities in repetitive DNA and genes. Based on these characteristics, a nonlinear model for the structure of the *C. elegans* genome is proposed. These findings shed light for understanding and quantifying the nonlinear structure organization of the *C. elegans* genome, which can be extended to other genomes, including the human genome.

ACKNOWLEDGMENTS

We gratefully acknowledge the Departamento Administrativo de Ciencia, Tecnología e Innovación - COLCIENCIAS of the Republic of Colombia, for supporting this project with Biotechnology Research Grant (#1103-12-16765). Thanks to Dr. Felipe Garcia for his comments and suggestions. Finally, thanks to the VRI of the University of Cauca for the publication financial support. The authors declare that they have no competing interests.

REFERENCES

Anonymous (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282: 2012-2018.

Berthelsen CL, Glazier JA and Raghavachari S (1994). Effective multifractal spectrum of a random walk. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* 49: 1860-1864.

Brenner S and Miller JH (Editors) (2002). Encyclopedia of Genetics. Introduction. Academics Presss, California.

Burgos JD and Moreno-Tovar P (1996). Zipf-scaling behavior in the immune system. *Biosystems* 39: 227-232.

Cheung AL and Deng W (2008). Telomere dysfunction, genome instability and cancer. *Front Biosci.* 13: 2075-2090.

Dewey TG and Strait BJ (1996). Multifractals, encoded walks and the ergodicity of protein sequences. *Pac. Symp. Biocomput.* 216-229.

Eskesen ST, Eskesen FN, Kinghorn B and Ruvinsky A (2004). Periodicity of DNA in exons. *BMC Mol. Biol.* 5: 12.

Fu W, Wang Y and Lu D (2007). Multifractal analysis of genomes sequences' CGR graph. *Sheng Wu Yi. Xue. Gong. Cheng Xue. Za Zhi.* 24: 522-525.

Garcia-Muse T and Boulton SJ (2007). Meiotic recombination in *Caenorhabditis elegans*. *Chromosome Res.* 15: 607-621.

Gladden JM, Farboud B and Meyer BJ (2007). Revisiting the X:A signal that specifies *Caenorhabditis elegans* sexual fate. *Genetics* 177: 1639-1654.

Glazier JA, Raghavachari S, Berthelsen CL and Skolnick MH (1995). Reconstructing phylogeny from the multifractal spectrum of mitochondrial DNA. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* 51: 2665-2668.

Goldberger AL (1996). Non-linear dynamics for clinicians: chaos theory, fractals, and complexity at the bedside. *Lancet* 347: 1312-1314.

Gutiérrez JM, Iglesias A, Rodríguez MA, Burgos JD, et al. (1998). Analyzing the Multifractal Structure of DNA Nucleotide Sequences. In: Chaos and Noise in Biology and Medicine. World Scientific Publishing Co., London, 315-319.

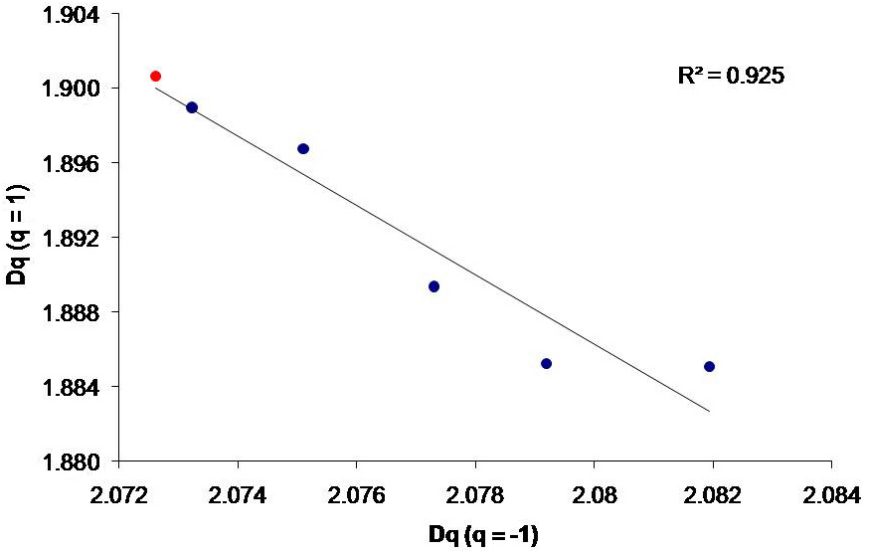
Gutiérrez JM, Rodríguez MA and Abramson G (2001). Multifractal analysis of DNA sequences using a novel chaos-game

representation. *Phys. Stat. Mech. Appl.* 300: 271-284.
Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, et al. (1986). Fractal measures and their singularities: The characterization of strange sets. *Phys. Rev. A* 33: 1141-1151.
Hentschel HGE and Procaccia I (1983). The infinite number of generalized dimensions of fractals and strange attractors. *Phys. Nonlinear Phenom.* 8: 435-444.
Im SH and Lee J (2005). PLP-1 binds nematode double-stranded telomeric DNA. *Mol. Cells* 20: 297-302.
Ivanov PC, Amaral LA, Goldberger AL, Havlin S, et al. (1999). Multifractality in human heartbeat dynamics. *Nature* 399: 461-465.
Jeffrey HJ (1990). Chaos game representation of gene structure. *Nucleic Acids Res.* 18: 2163-2170.
Lewin B (2008). Genes IX. Jones and Bartlett Publishers Inc., Missinauga.
Mandelbrot B (1982). La Geometría Fractal de la Naturaleza. Tusquets editores, Barcelona.
Meneveau C and Sreenivasan KR (1987). Simple multifractal cascade model for fully developed turbulence. *Phys. Rev. Lett.* 59: 1424-1427.
Moreno PA, Burgos JD, Vélez PE, Gutiérrez JM, et al. (2000). Multifractal Analysis of Complete Genomes. In: Proceedings of the 12th International Genome Sequencing and Analysis Conference, September 12-15, Miami Beach, 80-81.
Moreno-Herrero F, Seidel R, Johnson SM, Fire A, et al. (2006). Structural analysis of hyperperiodic DNA from *Caenorhabditis elegans*. *Nucleic Acids Res.* 34: 3057-3066.
Peitgen HO, Jürgen H and Saupe D (1992). Chaos and Fractals. New Frontiers of Science. Springer-Verlag, New York.
Shannon CE (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27: 379-423, 623-656.
Spieth J and Lawson D (2006). Overview of gene structure. *WormBook* 18: 1-10.
Stanley HE and Meakin P (1988). Multifractal phenomena in physics and chemistry. *Nature* 335: 405-409.
Tél T, Fülöp A and Vicsek T (1989). Determination of fractal dimensions for geometrical multifractals. *Phys. Stat. Theor. Phys.* 159: 155-166.
Venter JC, Adams MD, Myers EW, Li PW, et al. (2001). The sequence of the human genome. *Science* 291: 1304-1351.
Voss RF (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* 68: 3805-3808.
Wolfe KH, Sharp PM and Li WH (1989). Mutation rates differ among regions of the mammalian genome. *Nature* 337: 283-285.
Wood WB (1987). Introduction to *C. elegans* Biology. In: The Nematode *Caenorhabditis elegans* (Wood W, ed.). Cold Spring Harbor Laboratory Press, New York, 1-6.
Yang JY, Zhou Y, Yu ZG, Anh V, et al. (2008). Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. *BMC Bioinformatics* 9: 113.
Yu ZG, Anh V and Lau KS (2001). Measure representation and multifractal analysis of complete genomes. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 64: 031903.
Yu ZG, Anh V and Lau KS (2003). Multifractal and correlation analyses of protein sequences from complete genomes. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 68: 021913.
Yu ZG, Anh V and Lau KS (2004). Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *J. Theor. Biol.* 226: 341-348.
Yu ZG, Anh VV, Lau KS and Zhou LQ (2006). Clustering of protein structures using hydrophobic free energy and solvent accessibility of proteins. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 73: 031920.
Zetka M (2009). Homologue pairing, recombination and segregation in *Caenorhabditis elegans*. *Genome Dyn.* 5: 43-55.
Zhou LQ, Yu ZG, Deng JQ, Anh V, et al. (2005). A fractal method to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation. *J. Theor. Biol.* 232: 559-567.

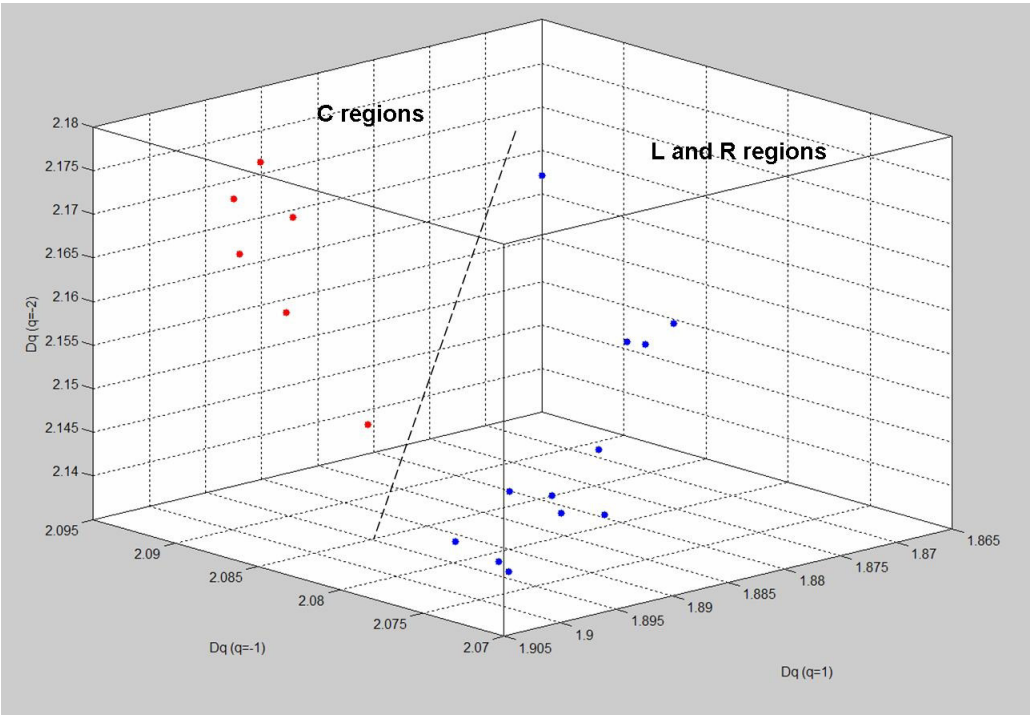
Description of additional data files

Additional file 1. Multifractal and molecular information.

This file has the information for all multifractal and molecular parameters calculated for our study.



Additional Figure 2C. Distribution of 2-D points (D-1, D1) of the *Caenorhabditis elegans* chromosomes.



Additional Figure 3E. Multifractality by chromosome region. Distribution of 3-D points (D-1, D1, D-2) of the *Caenorhabditis elegans* chromosome fragments.