GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Comparison of period-3 correlation amplitudes in genomic DNA sequences

J.C.O. Guerra[1], P. Licinio[2] and P.C.P. Andrade[1]

[1]Instituto de Física, Universidade Federal de Uberlândia,
Uberlândia, MG, Brasil
[2]Departamento de Física, Universidade Federal de Minas Gerais,
Belo Horizonte, MG, Brasil

Corresponding author: J.C.O. Guerra
E-mail: jcog@infis.ufu.br

Genet. Mol. Res. 12 (4): 4459-4467 (2013)
Received December 14, 2012
Accepted August 7, 2013
Published October 10, 2013
DOI http://dx.doi.org/10.4238/2013.October.10.11

ABSTRACT. Period-3 oscillations in genome composition can be detected through correlation functions. Since these oscillations are closely related to the genetic code structure, we developed methods for quantitative comparison of genomic and exonic oscillation amplitudes and decay. In contrast to genomic correlations, exonic period-3 oscillation amplitudes are persistent. A model postulating an uncorrelated distribution of exons in the genome has been applied to the analysis of *Escherichia coli* K12, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens* genomic decay, allowing for a quantitative discussion of genome organization.

Key words: Nucleotide correlation; Irreducible correlation; Exons; Period-3 modulation

INTRODUCTION

Codon triplets carry a strong chemical bias into protein-encoding DNA sequences. Among the origins of 3-phase modulated composition trends is a greater tolerance for point mutations at the third base due to its strong amino acid coding degeneracy. Codon bias is readily observed as a period-3 oscillation in correlation functions of exonic or genomic origin and is seldom observed in intron correlations (Guerra and Licinio, 2010). Recently, we proposed an approach for describing general properties of DNA sequences in the most compact or irreducible form (Licinio and Guerra, 2007) and showed that binary correlations between nucleotides could be calculated as a combination of a maximum of six irreducible correlations along genomic DNA sequences. In addition, a procedure for extracting period-3 modulations and amplitudes out of a binary correlation was developed and applied to the *Drosophila melanogaster* genome (Guerra and Licinio, 2010). In that work, a relationship was found between the period-3 amplitudes calculated for the genomic and exonic DNA. The relationship showed that exons are dispersed along each chromosome in a phase-uncorrelated manner. We extended this quantitative analysis to other species (*Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens*) in order to gain comparative insight into how exons are organized in different genomes.

MATERIAL AND METHODS

In a recent study, we considered the normalized correlation function for the nucleotides *B, B'* (= {*A*; *T*; *C*; *G*}) separated by the downstream distance *k* (Guerra and Licinio, 2010):

$$C_{BB'}(k) = \frac{1}{N-k}\sum_{i=1}^{N-k} \rho_B(i)\,\rho_{B'}(i+k) \equiv \langle \rho_B(0)\,\rho_{B'}(k)\rangle \qquad \text{(Equation 1)}$$

where the local compositions $\rho_B(i)$ and $\rho_{B'}(i+k)$ are given as 0 or 1:

$$\rho_B(i) = \begin{cases} 1 & \Leftrightarrow \rho(i) = B \\ 0 & \Leftrightarrow \rho(i) \neq B \end{cases} \qquad \text{(Equation 2)}$$

The 16 possible dinucleotide correlation functions are related to one another and can be expressed in terms of nine irreducible correlation functions by using irreducible coordinates for the local nucleotide compositions *x*, *y*, *z* (Licinio and Caligiorne, 2004):

$$x(i) = \begin{cases} 1 \Leftrightarrow \rho(i) = \{A, G\} \\ -1 \Leftrightarrow \rho(i) = \{C, T\} \end{cases};\quad y(i) = \begin{cases} 1 \Leftrightarrow \rho(i) = \{A, C\} \\ -1 \Leftrightarrow \rho(i) = \{G, T\} \end{cases};\quad z(i) = \begin{cases} 1 \Leftrightarrow \rho(i) = \{A, T\} \\ -1 \Leftrightarrow \rho(i) = \{C, G\} \end{cases} \qquad \text{(Equation 3)}$$

The irreducible correlations are thus calculated as

$$C_{xy}(k) = \frac{1}{N-k}\sum_{i=1}^{N-k} x(i)\,y(i+k) \equiv \langle x(0)\,y(k)\rangle \qquad \text{(Equation 4)}$$

Consideration of statistical genome strand symmetry further reduces the correlation function set to the consideration of solely the following six irreducible correlations: {$C_{xx}(k)$;

$C_{yy}(k)$; $C_{zz}(k)$; $C_{xy}(k)$ (= $C_{yx}(k)$); $C_{xz}(k)$ (= - $C_{zx}(k)$) and $C_{zy}(k)$ (= - $C_{yz}(k)$)}. From this set, the 16 possible dinucleotide correlations can be properly obtained as

$$\begin{aligned} C_{AA}(k) = C_{TT}(k) &= \frac{1}{16}\left[1 + 2\langle z\rangle + C_{xx}(k) + C_{yy}(k) + C_{zz}(k) + 2C_{xy}(k)\right] \\ C_{AT}(k) &= \frac{1}{16}\left[1 + 2\langle z\rangle - C_{xx}(k) - C_{yy}(k) + C_{zz}(k) - 2C_{xy}(k) + 2C_{xz}(k) + 2C_{yz}(k)\right] \\ &\cdots \end{aligned} \qquad \text{(Equation 5)}$$

The *z*-coordinate discriminates strong-weak binding (AT-CG) and displays the strongest fluctuations and drift in composition along the genome. In this work, only the strong-weak self-correlation $C_{zz}(k)$ was analyzed. This work also addressed the period-3 modulations that can be detected from the binary correlation functions. To extract the period-3 modulation, we propose the following algorithm: the correlation function is smoothed by a 3-average running window that erases oscillations with period 3. The period-3 modulation $\Delta C^3_{ZZ}(k)$ is the difference between the natural irreducible correlation $C_{ZZ}(k)$ and the smoothed correlation $C^3_{ZZ}(k)$, as given by:

$$\Delta C^3_{zz}(k) \equiv C_{zz}(k) - C^3_{zz}(k); \quad k \geq 2 \qquad \text{(Equation 6)}$$

with

$$C^3_{zz}(k) \equiv \frac{1}{3}\left[C_{zz}(k-1) + C_{zz}(k) + C_{zz}(k+1)\right]; \quad k \geq 2 \qquad \text{(Equation 7)}$$

Self-correlation functions, such as the correlation $C_{ZZ}(k)$, should be pair functions and oscillate as cosines, with maxima at $k = 3n$, where n is an integer. In a general way, one observes the following profile for the self-correlation modulations:

$$\Delta C_{zz}^{3}(k) = A_{zz}^{3}(k)\cos\left(\frac{2\pi k}{3}\right), \quad k \geq 2 \qquad \text{(Equation 8)}$$

Thus, the period-3 modulation amplitude for the irreducible self-correlation $C_{ZZ}(k)$ can be obtained from inversion of equation (8):

$$A_{zz}^{3}(k) = \frac{\Delta C_{zz}^{3}(k)}{\cos\left(\frac{2\pi k}{3}\right)}, \quad k \geq 2 \qquad \text{(Equation 9)}$$

RESULTS

We applied this analysis to the genomes of the following species: *E. coli* K12 strains, *S. cerevisiae, C. elegans*, *D. melanogaster* (except for the smaller Y and MT chromosomes), and *H. sapiens*. In the case of *H. sapiens*, *S. cerevisiae*, and *C. elegans*, we did not include the organelle genomes. Table 1 lists the total genome length for each species.

Table 1. Total genome length, length of the coding sequences analyzed and corresponding number of exons (or CDSs).

Species	Genome length (Mbp)	Total coding length analyzed (Mbp)	Number of exons or CDSs
Escherichia coli K12	4.64	3.95	4,050 CDSs
Saccharomyces cerevisiae	12.07	3.36	3,541 CDSs
Caenorhabditis elegans	100.27	28.3	213,155 exons
Drosophila melanogaster	120.29	34.12	70,237 exons
Homo sapiens	2,861.33	3.69 (~8%)	13,890 exons

The genome sequences for *E. coli* K12, *S. cerevisiae, C. elegans*, and *D. melanogaster* were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov/). The sequences for the 24 chromosomes of *H. sapiens* were downloaded from the UCSC database (http://genome.ucsc.edu). For each species, the reading was performed along the euchromatic arms for all chromosomes (except, obviously, for the circular DNA genome of *E. coli* K12). In order to understand the role played by exons in each genome, we used currently available coding sequences. *D. melanogaster* exon sequences were downloaded from FlyBase (http://flybase.org/). *C. elegans* exon sequences were downloaded from WormBase (http://www.wormbase.org/). *H. sapiens* exon sequences and *S. cerevisiae* and *E. coli* K12 CDSs were downloaded from NCBI. Table 1 lists the number of exons or CDSs and the total coding length analyzed for each species.

Figure 1 shows genomic self-correlation C_{zz} (k) up to $k = 50$. Except for the very first points (up to $k \approx 6$), they oscillate in phase with a period of 3, although this is not so pronounced in *H. sapiens*. The mean level is higher for *C. elegans*, reflecting a long-range AT-CG unbalance.

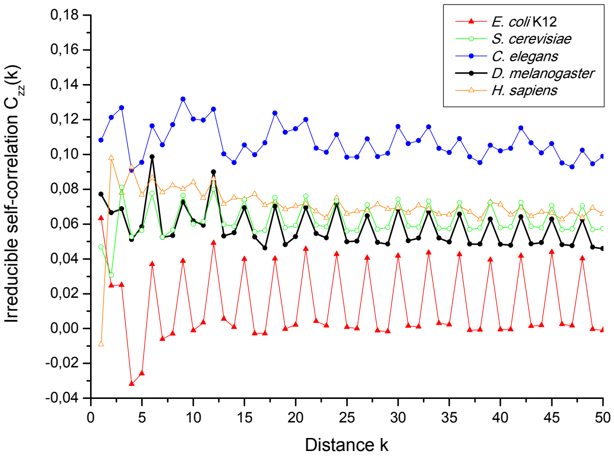

Figure 1. Short range irreducible self-correlation $C_{zz,G}$ (k) for the genomes of *Escherichia coli* K12, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. Correlations are seen to oscillate in-phase as cosines with period-3.

DISCUSSION

Modulation amplitudes for period-3 oscillations were calculated as described above. Figure 2 shows the genomic modulation amplitudes for short-range values of *k* up to 70. Modulation amplitudes relate to the fraction of the genome devoted to protein coding and are accordingly stronger for the compact *E. coli* K12 genome. However, for *H. sapiens*, coding sequences account for only about 1.5% of the genome. The modulation displays a barely coherent small amplitude and, due to noise, it fluctuates often to negative values. Only after averaging does *H. sapiens* modulation stand as positive.

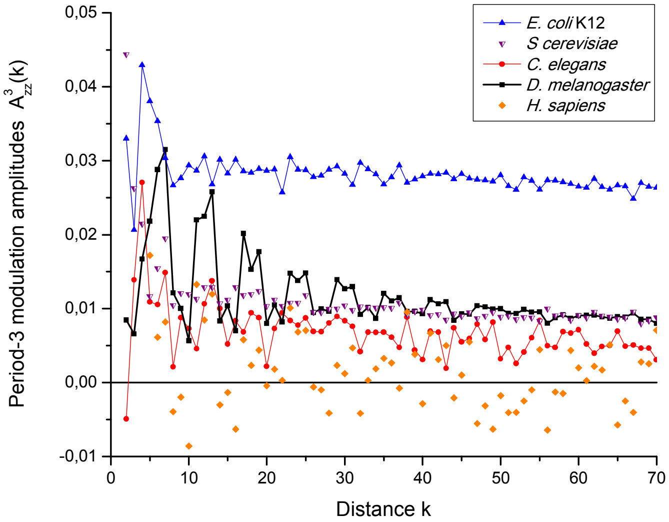

Figure 2. Short range genomic period-3 modulation amplitudes $A^3_{ZZ,G}$ (k) for the genomes of *Escherichia coli* K12, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*.

For *D. melanogaster*, *H. sapiens*, and *C. elegans* (in minor proportion), there is some oscillation at short range. The reason for this can be found in the repetitive elements in these genomes and may contribute to period-3 oscillations (see below). In fact, ~ 50% of the genomic DNA in *H. sapiens* is comprised of repetitive DNA, while *D. melanogaster* and *C. elegans* genomes are comprised of ~12 to 13% repetitive sequences (Taft et. al, 2007). Note how the amplitudes decay slowly in Figure 2. The long-range decay has been plotted in Figure 3, which shows that the period-3 oscillations in genomic correlations persist to distances of about 1000 bases, at which point they begin to decay more rapidly.

Recently, we showed that period-3 modulation amplitudes for the self-correlation C_{ZZ} (k) decay quasi-logarithmically for the *D. melanogaster* genome (Guerra and Licinio, 2010). Nevertheless this is not the case for the ensemble investigated here for which no general decay form could be found.

Asymptotic short-range (k ~ 10) values of period-3 amplitudes are listed in Table 2. The fact that only ~1.5% of the *H. sapiens* genome corresponds to exons has been thought to be an impediment to detection of period-3 oscillations in early correlation or Fourier analyses. With the techniques used here, however, these oscillations clearly detach out of data noise displaying low, yet measurable amplitudes.

Sources for 3-base periodicities are primarily the coding segments of the genome. In order to investigate the exon contribution to genome oscillations, we made similar calculations for the exons (or CDSs) available for these species. Figure 4 shows the exonic period-3 amplitudes for the five species. Contrary to the fast decay in genomic period-3 amplitudes, exonic period-3 amplitudes remain strongly correlated over 1000 bases.

As with genomic data, asymptotic short-range values of exonic period-3 amplitudes were also estimated and listed in Table 2.

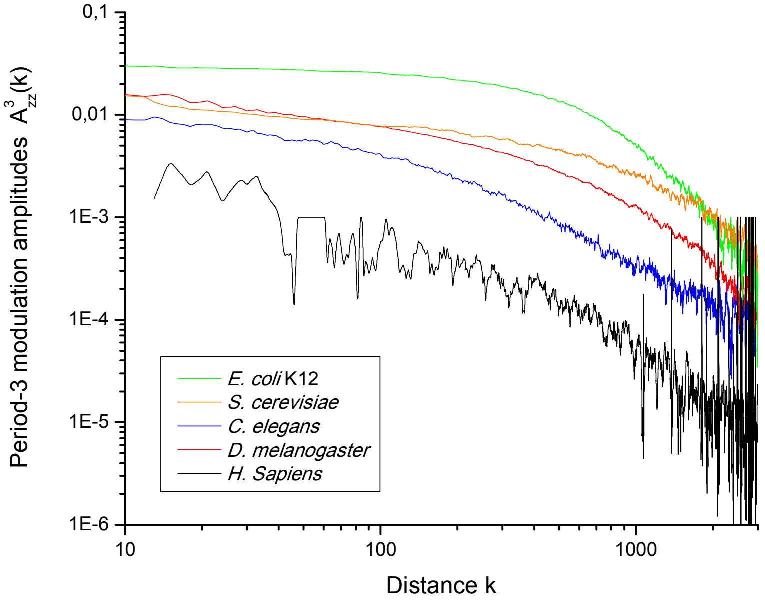

Figure 3. Long range genomic period-3 modulation amplitudes $A^3_{ZZ,G}$ (k) for the genomes of *Escherichia coli* K12, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. Both vertical and horizontal scales are logarithm. Data are smoothed by a 21-average.

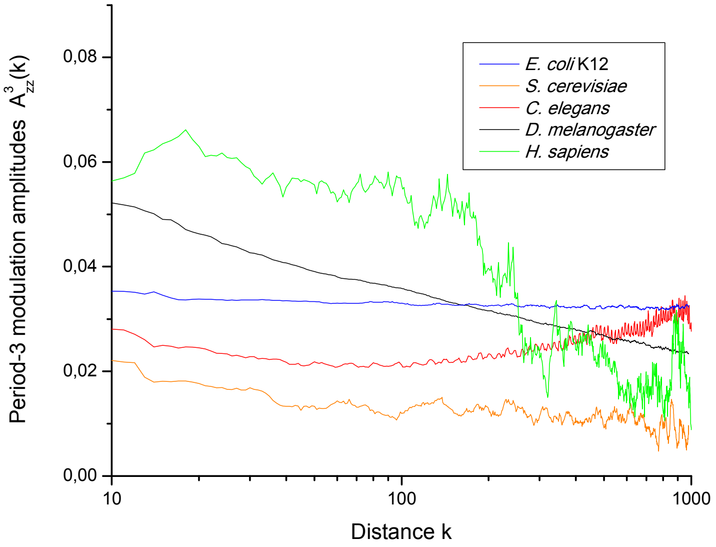

Figure 4. Long range exonic period-3 modulation amplitudes $A^3_{ZZ,E}$ (k) for the genomes of *Escherichia coli* K12, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. The horizontal scale is logarithm. Data are smoothed by a 21-average.

The exon contribution to genomic period-3 correlations can be computed considering only internal exon correlations, i.e., by neglecting exon-intron and exon-exon correlations - in other words, assuming the exons are phase-uncorrelated. In this case, the genomic period-3 amplitudes can be calculated from the exonic period-3 amplitudes using the following relation:

$$A^3_{zz,G}(k) = \frac{1}{N_G} \sum_{i=k+1}^{l_{MAX}} (i-k) h(i) A^3_{zz,E}(k) \qquad \text{(Equation 10)}$$

where N_G is the total genome length, *h(i)* is the number of occurrences of exons of length *i*, while l_{MAX} is the greatest exon length. The weight (*i* - *k*) in the sum acknowledges the number of pair correlations at distance *k* strictly internal to exons of size *i*. At short range (*k* = 1), genomic and exonic amplitudes become simply related as

$$A^3_{zz,GENOMIC} \cong \frac{N_E}{N_G} A^3_{zz,EXONIC} \qquad \text{(Equation 11)}$$

where the total exon content in the genome (number of base pairs) is $N_E = \sum_{i=1}^{l_{MAX}} ih(i)$, while $N_E >> \sum_{i=2}^{l_{MAX}} h(i)$, the total number of exons. Relation (11) shows that the coding content of a genome can be estimated from suitable sequence samplings given estimates for both genomic and exonic period-3 amplitudes. Exonic content evaluated according to the proposed methodology is also listed in Table 2, together with estimates from the literature (Blattner et al., 1997; Misra et al., 2002; Taft et al., 2007; *C. elegans* Sequencing Consortium, 1998). The agreement seems reasonable, although clearly not exact. Except for the very compact *E. coli* genome, Equation (11) estimates slightly higher exon content than the pertinent literature estimates. Some excess could be anticipated in view of the presence of fossil or inactive exons and other repetitive elements in the genomes. For *H. sapiens*, about 8% of exons were used but the main hindrance to precision was the intrinsic low 1.5% coding content in the genome, whose signal is superposed with strong noise from the remaining 98.5% of the non-coding genome.

Table 2. Short range (*k* ~ 10) asymptotic values of the genomic and exonic period-3 modulation amplitudes.

Species	$A^3_{ZZ,GENOMIC}$	$A^3_{ZZ,EXONIC}$	N_E/N_G (est.)	N_E/N_G (lit.)
Escherichia coli K12	0.029	0.035	0.83	0.88
Saccharomyces cerevisiae	0.015	0.021	0.71	0.70
Caenorhabditis elegans	0.009	0.028	0.32	0.27
Drosophila melanogaster	0.014	0.052	0.27	0.24
Homo sapiens	0.002	0.064	0.03	0.015

The amplitude ratio is displayed in the third column as an estimate of exonic content in the genomes (Equation 11). The exonic content as found in the literature is also shown in the last column.

In order to apply Equation 10, we should have on hand the exon length distribution for each species. However, we do not need to have data for all the exons; only a sampling with statistical significance is required. Thus, Figure 5 compares the period-3 modulation amplitudes calculated for the whole genome with the uncorrelated exon contribution as described above, for the five species being analyzed here. There is a striking agreement for the genomes of *C. elegans* and *D. melanogaster*. The long-range genomic correlation for *E. coli* decays faster than the model. This suggests that neighboring CDSs could be preferentially out-of-phase, instead of simply uncorrelated in this compact genome. The opposite tendency has been obtained for *S. cerevisiae*. The genomic period-3 modulation amplitude in *H. sapiens* looks too pathological, especially at short range, possibly due to a strong contribution of repetitive elements among sparse exons. However, long-range correlations can still be modeled by an uncorrelated exon distribution.

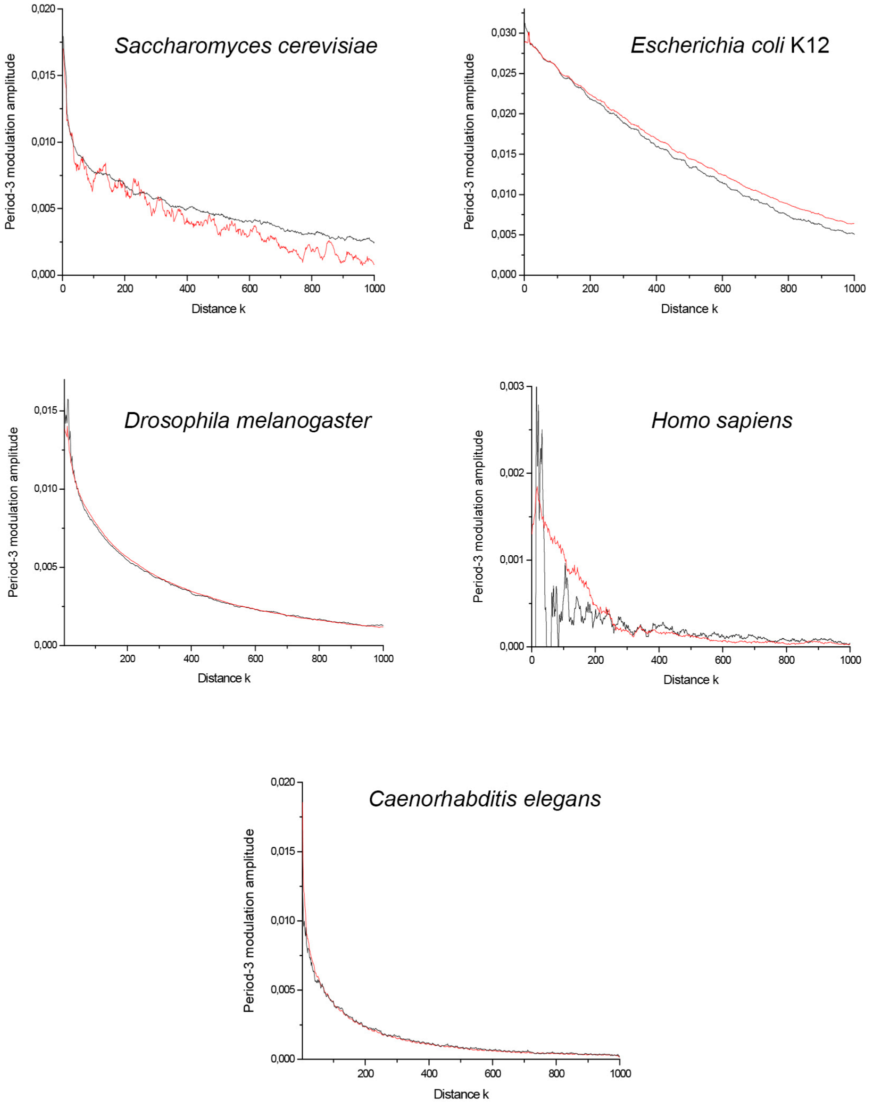

Figure 5. Genomic period-3 modulation amplitudes for the self-correlation C_{ZZ} (*k*) (black lines) compared with uncorrelated exon contribution from the model of Equation 10 (red lines). Data are smoothed by a 21-average.

CONCLUSION

Period-3 oscillations in genome compositions can be detected through correlation functions. Since these oscillations are closely related to the genetic code structure, we developed methods for quantitative comparison of genomic and exonic oscillation amplitudes. Contrary to genomic correlations, exonic period-3 oscillation amplitudes are persistent. A model postulating an uncorrelated distribution of exons in the genome could precisely account for *C. elegans* and *D. melanogaster* genomic decay. Deviations in other genomes have been discussed in light of this model. Estimates of exonic content given by this model slightly exceed other estimates from extensive sequence analysis found in the literature, a feature associated with the presence of fossil or inactive exons and other repetitive elements.

ACKNOWLEDGMENTS

We thank the Brazilian agencies CNPq and FAPEMIG for financial support.

REFERENCES

Blattner FR, Plunkett G, III, Bloch CA, Perna NT, et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453-1462.

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

Guerra JC and Licinio P (2010). The role played by exons in genomic DNA sequence correlations. *J Theor Biol* 264: 830-837.

Licinio P and Caligiorne RB (2004). Inference of phylogenetic distances from DNA-Walk divergences. *Phys. A Stat. Mech. Appl.* 341: 471-481.

Licinio P and Guerra JC (2007). Irreducible representation for nucleotide sequence physical properties and self-consistency of nearest-neighbor dimer sets. *Biophys. J.* 92: 2000-2006.

Misra S, Crosby MA, Mungall CJ, Matthews BB, et al. (2002). Annotation of the *Drosophila* melanogaster euchromatic genome: a systematic review. *Genome Biol.* 3: RESEARCH0083.

Taft RJ, Pheasant M and Mattick JS (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29: 288-299.