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

A method for extracting period~10 modulations from DNA sequence correlations applied to the *Drosophila melanogaster* genome

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

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

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

Genet. Mol. Res. 11 (3): 2835-2846 (2012)
Received June 1, 2012
Accepted August 2, 2012
Published August 24, 2012
DOI http://dx.doi.org/10.4238/2012.August.24.8

ABSTRACT. Nucleosome DNA packaging and positioning within the *Drosophila melanogaster* genome imposes a weak modulation, with a period of about 10 bp in the genomic composition correlations. We present formalism for extracting such modulations from an irreducible set of six correlation functions calculated along the *D. melanogaster* genome. These modulations were seen to be stronger for the irreducible self-correlation $C_{zz}(k)$ (strong-weak binding). Using an FFT procedure, we show that the period~10 modulation extracted from such self-correlation is viewed to be an oscillation with period~10.9 overmodulated by an oscillation with period~153. This behavior of the modulation reflects the organization of the eukaryotic genomic DNA. But, since the period~10 modulation dies for $k \sim 150$, the constraints imposed by the nucleosome arrangement over the nucleosome sequence composition must be weak, provided that such constraints are the sources for the modulations.

Key words: Nucleotide correlations; Irreducible correlations; Period~10 modulation; Nucleosome sequence; Linker DNA

INTRODUCTION

Today, with the sequencing of the genomic DNA of an increasing number of species, we see a vigorous production of scientific papers in the area of statistical genomics and bioinformatics. Before the human genome project era, in the sixties, studies concerned mostly short-range binary correlations (Josse et al., 1961) or base density heterogeneity (isochoric distribution) in digested DNA segments (Sueoka, 1959). Statistical regularities were primarily used to detect coding regions (Shulman, 1981) and other functional sites. Correlation measures of DNA sequences have been widely studied with the objective of understanding how the genomes of various species are organized. Fourier analysis (Yin and Yau, 2007), dinucleotide frequencies (Kogan and Trifonov, 2005; Cohanim et al., 2006a,b; Kogan et al., 2006), random-walk variances (Peng et al., 1992; Buldyrev et al., 1995; Te Boekhorst et al., 2008), and wavelet analysis (Arneodo et al., 1998) are some examples of related statistical measures. Herzel and co-workers used binary correlation functions and constructed a dependence matrix that would count all the statistical dependences between all nucleotides (Herzel and Groβe, 1995, 1997; Herzel et al., 1998). In applying this covariance matrix to human chromosomal regions, for example, long-range correlations in human DNA have been observed, correlated to G + C distributions or isochoric structure of genomes (Carpena et al., 2007; Oliver et al., 2008), as well as chromatin structure (Audit et al., 2001); periodicities of 3 and 10-11 bases have been verified in yeast and bacterial DNA. Many authors have considered dinucleotide frequency distributions rather than nucleotide correlations (Kogan and Trifonov, 2005; Cohanim et al., 2006a,b; Kogan et al., 2006). In a recent study, we were confronted with the problem of calculating correlations along DNA sequences using a representation of the four-nucleotide set as a tetrahedron in 3-D (Silverman and Linsker, 1986; Coward, 1997; Licinio and Caligiorne, 2004; Licinio and Guerra, 2007). We concluded that the only 10 kinds of 5'3'-nucleotide pair correlations to be considered (AA = TT, AT, TA; CC = GG, CG, GC; AG = CT; AC = GT; TG = CA, and TC = GA) are subject to composition closure relationships and can be deduced from an irreducible set of 6 basic correlation functions (Guerra and Licinio, 2010). We examined composition correlations for the *Drosophila melanogaster* genome and focused on the problem of describing the period-3 modulations directly related to the presence of exons and coding regions owing to the codon structure (Shulman et al., 1981; Fickett, 1982). It was verified that the period-3 modulation amplitudes are highest for the irreducible self-correlation between strong or weak nucleotides. A relationship between exonic and genomic period-3 amplitudes allowed the conclusion that exons are dispersed along each chromosome in a phase uncorrelated manner, in the *D. melanogaster* genome.

In this study, we extended the methodology initially presented for extracting period-3 modulations to deal with period~10 modulations. The period~10 has been widely investigated by the scientific community. Widom (1996), for example, applied Fourier transform analysis and separated out the strong period-3 modulation into a single peak so that other spectral regions could be analyzed. Widom (1996) also observed one peak at 10-11 bp in two eukaryotic genomes and, in particular, for the *Caenorhabditis elegans* genome, such spectral component becomes the dominant feature. The author's conclusions suggest that the nucleosome positioning directed by the adjacent DNA sequence has significance, not only at promoters, but also at locations throughout genes. However, in contrast, Lowary and Widom (1997) concluded that at least 95% of genomic DNA sequences

have an affinity for the histone octamer in the nucleosome structure, comparable to that of random sequences. In this way, genomic DNA would contribute little to its own packaging. Thus, to make the two studies coincide, the authors suggested that the signals that favor nucleosome packaging are sparsely but uniformly distributed along the entire genome or are concentrated in a small subset of the genome. Yuan et al. (2005) developed a hidden Markov model (HMM) to determine nucleosome/linker boundaries. The proposed model calculates the probability that a given probe in the array corresponds to a nucleosomal DNA, delocalized nucleosome, or linker DNA, and identifies the most likely nucleosome positions. Nucleosomes could occupy multiple positions in ensemble measurements, because there would be little thermodynamic preference by the histone octamer for most of the genomic DNA. However, the results of their study showed that 65 to 69% of nucleosomal DNA were found in well-positioned nucleosomes. In another way, delocalized nucleosomes were inhomogeneously distributed. Later, Segal et al. (2006) isolated nucleosome-bound sequences at high resolution from yeast and used these sequences in a new computational approach to construct and validate experimentally a nucleosome-DNA interaction model, and to predict the genome-wide organization of nucleosomes. Their results showed that ~50% of nucleosome organization *in vivo* can be explained only by preferences for sequences from nucleosomes. Besides, their results indicate that the nucleosome depletions observed at coding and intergenic regions are attributed in part to unstable nucleosomes encoded in these regions. Also, the distribution of 'pairwise' distances between positions of the highly stable nucleosomes showed significant correlations, persisting over at least six adjacent nucleosomes with an average nucleosome repeat length of 177 bp. Thus, the yeast genome would encode not only the preferred positions of individual nucleosomes but also the highest structural levels of chromatin organization directly. Cohanim et al. (2006a), using dinucleotide frequency measures, observed periodic oscillations of AA and TT dinucleotides; such oscillations generate two patterns, one referred to as the nucleosome DNA pattern and other that would correspond to the curved DNA (which would also participate in nucleosome formation).

METHODOLOGY AND FORMALISM

In a recent study, we defined composition correlations for DNA double strands according to Guerra and Licinio (2010):

$$C_{BB'}(k) = \frac{1}{16}\left\{1 + (B_z + B_z')\langle z\rangle + (B_x \quad B_y \quad B_z)\begin{bmatrix} C_{xx}(k) & C_{xy}(k) & C_{xz}(k) \\ C_{xy}(k) & C_{yy}(k) & C_{yz}(k) \\ -C_{xz}(k) & -C_{yz}(k) & C_{zz}(k) \end{bmatrix}\begin{pmatrix} B_x' \\ B_y' \\ B_z' \end{pmatrix}\right\}, \quad \text{(Equation 1)}$$

where B_i is the *i-th* component of the nucleotide-state vector, which can be one of the four:

$$|A\rangle = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, |T\rangle = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, |C\rangle = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}, \text{or}, |G\rangle = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}.$$

Equation 1 considers the symmetries of complementary pairing between the chains. In Equation 1, we applied the definition:

$$C_{xy}(k) = \frac{1}{N-k}\sum_{i=1}^{N-k} x(i)y(i+k) \quad \text{(Equation 2)}$$

for the irreducible cross-correlation $C_{xy}(k)$ and extended it for the others. In Equation 2, *x(i)* is the *x*-component of the nucleotide-state vector associated with the nucleotide occupying the *i-th* site along the sequence, and *y(i + k)* is the *y*-component of the nucleotide-state related to the nucleotide at the position *i + k*. The calculation of any irreducible correlation or even of the nucleotide correlation 1 is performed along anyone of the chains. Application of the Equation 1 to the 16 possible correlations between nucleotides (AA, AT, and so on) provides

$$\begin{aligned} C_{AA}(k) = C_{TT}(k) &= \frac{1}{16}\left[1 + 2\langle z\rangle + C_{xx}(k) + C_{yy}(k) + C_{zz}(k) + 2C_{xz}(k)\right] \\ C_{CC}(k) = C_{GG}(k) &= \frac{1}{16}\left[1 - 2\langle z\rangle + C_{xx}(k) + C_{yy}(k) + C_{zz}(k) - 2C_{xz}(k)\right] \\ C_{AC}(k) = C_{GT}(k) &= \frac{1}{16}\left[1 - C_{xx}(k) + C_{yy}(k) - C_{zz}(k) - 2C_{yz}(k)\right] \\ C_{AG}(k) = C_{CT}(k) &= \frac{1}{16}\left[1 + C_{xx}(k) - C_{yy}(k) - C_{zz}(k) - 2C_{xz}(k)\right] \end{aligned} \quad \text{(Equation 3)}$$

where, $C_{AC}(k) = C_{GT}(k)$, $C_{AG}(k) = C_{CT}(k)$, and so on, reducing the number of nucleotide correlations from 16 to 10. These 10 possible nucleotide correlations, in another way, are expanded in terms of 6 irreducible correlation functions, namely, $C_{xx}(k)$, $C_{yy}(k)$, $C_{zz}(k)$, $C_{xy}(k)$, $C_{xz}(k)$, and $C_{yz}(k)$.

All the irreducible correlations show a strong period-3 modulation, except the irreducible cross-correlation $C_{xy}(k)$, for the *D. melanogaster* genome. A careful analysis about the period-3 was then developed; in particular, we defined that the period-3 modulation $\Delta C^3_{ij}(k)$ is the difference between the natural irreducible correlation $C_{ij}(k)$ and the smoothed correlation $C^3_{ij}(k)$ [$C_{ij}(k)$ can represent any one of the 6 irreducible correlations]:

$$\Delta C^3_{ij}(k) \equiv C_{ij}(k) - C^3_{ij}(k) \quad \text{(Equation 4)}$$

with

$$C^3_{ij}(k) \equiv \frac{1}{3}\left[C_{ij}(k-1) + C_{ij}(k) + C_{ij}(k+1)\right];\ \ k \geq 2$$

where the smoothed correlation $C^3_{ij}(k)$ is a 3-average correlation at position *k*. Obviously, such smoothed correlation will lack oscillations with period-3. For *D. melanogaster*, we determined that the period-3 modulation amplitudes are the highest for the irreducible self-correlation between strong or weak nucleotides [$C_{zz}(k)$]. Thus, we defined the period-3 modulation amplitude as $A^3_{zz}(k) = \Delta C^3_{ij}(k) / \cos(2\pi k / 3)$ and correlated the genomic amplitude to the exonic amplitude using the distribution of lengths of exons encountered along the entire genome of *D. melanogaster*. The strong period-3 signal is isolated from other signals, which are weaker, on smoothing the irreducible self-correlation $C_{zz}(k)$. Therefore, proceeding to an analysis of the smoothed correlation $C^3_{zz}(k)$, we note that there are modulations with an apparent period~10 that extend for values of *k* up to *~150*. Figure 1 presents the graph of the smoothed correlation $C^3_{zz}(k)$ for *Drosophila*.

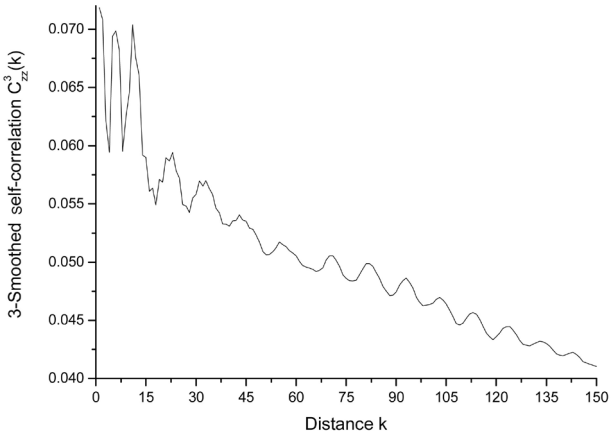

Figure 1. Short-range 3-smoothed self-correlation $C^3_{zz}(k)$ as a function of k. Note that such correlation oscillates with apparent period~10.

The graph of Figure 1 plots the 3-smoothed self-correlation $C^3_{zz}(k)$ for distances k up to $k = 150$. We see that an oscillation with apparent period~10 extends for values of k up to $k \sim 150$. Thus, a period~10 modulation is a second spectral component of self-correlation $C^3_{zz}(k)$.

To filter the period~10 modulation from the smoothed correlation $C^3_{ij}(k)$, we proposed to execute a similar procedure used to obtain the period-3 modulations. Consequently, we define the period~10 modulation $\Delta C^{10}_{ij}(k)$ as the difference between the smoothed correlation $C^3_{ij}(k)$ and the newly smoothed correlation $C^{10}_{ij}(k)$:

$$\Delta C^{10}_{ij}(k) \equiv C^3_{ij}(k) - C^{10}_{ij}(k); \quad k > 5 \qquad \text{(Equation 5)}$$

with

$$C^{10}_{ij}(k) \equiv \frac{1}{11}\sum_{j=-5}^{5} C^3_{ij}(k+j);\ k > 5$$

Note that we defined the smoothed correlation $C^{10}_{ij}(k)$ as the 11-average 3-smoothed correlation at position k. Obviously, such smoothed correlation will lack oscillations with period~10.

APPLICATION, RESULTS AND DISCUSSION

We performed the calculation of period~10 modulations for each of the six irreducible correlations for the entire genome of *D. melanogaster*. Figure 2 presents two graphs, one containing such oscillations for the irreducible self-correlations, and the other containing the modulations for the irreducible cross-correlations.

As we can observe from the graphs in Figure 2, all the irreducible correlations present some oscillatory behavior. Cross-correlations tend to present a weaker oscillation, while the irreducible self-correlation $C_{zz}(k)$ again shows the highest amplitude. Thus, we can readily conclude that, if the signal of period~10 favors the nucleosome positioning and packaging, then self-correla-

tion $C_{zz}(k)$ is the principal component of the patterns related to the curved DNA in the nucleosome structure. Henceforth, we can contrast our approach with that of other authors, such as Cohanim et al., (2006a) in the sense of that we derived the period~10 directly from nucleotide correlations. The argument to study frequencies of dinucleotides, or probability distributions of dinucleotides, across the genome, and to extract from them period~10 modulations, is that the dinucleotide step is the first instance to examine sequence-dependent properties of DNA sequences. One example is the sequence-dependent mechanics of DNA bending, which is considered essential for histone-DNA association. However, as we can observe in the graphs of Figure 2, a correlation between nucleotides at the nucleosomal level shows a period~10 modulation.

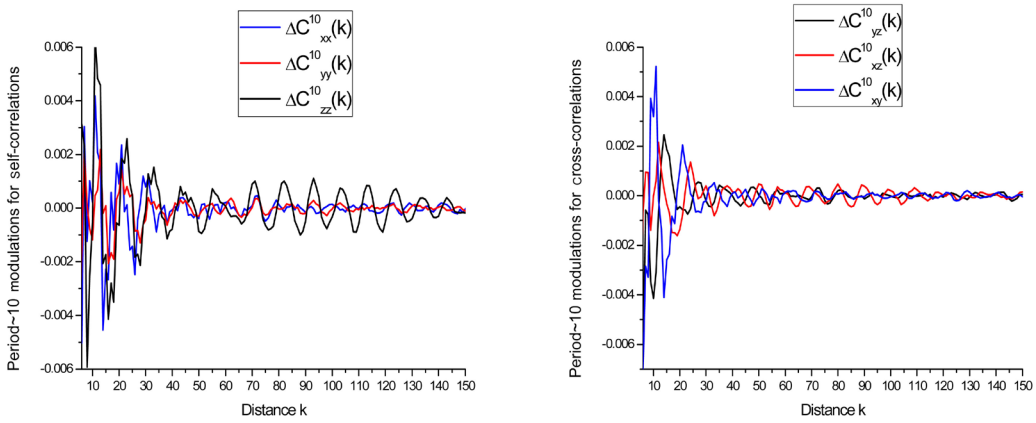

Figure 2. Period~10 modulations for the 6 irreducible correlations. Self-correlation irreducible $C_{zz}(k)$ presents period~10 modulation with the most important amplitude.

It can also argue for the contribution of the different domains of the genome of *D. melanogaster* for the period~10 modulation. In this way, we developed the approach shown in Equations 1-5 for exonic, intronic and intergenic sequences available for *D. melanogaster*. We were especially interested in the extraction of the period~10 modulation for the self-correlation $C_{zz}(k)$. Such modulation for each one of the referred domains is plotted in the graph of Figure 3.

Figure 3 shows that period~10 modulation has a similar amplitude for the different domains. This means that all the *D. melanogaster* genome encodes an intrinsic nucleosome organization; that is, the nucleosome positioning directed by the adjacent DNA sequence has significance, at all the locations throughout the genome, in agreement with results obtained by Widom (1996). Although the amplitude of the period~10 modulation is small (that is, much less than 1), the signals that favor nucleosome packaging are uniformly distributed along the entire genome of *Drosophila*. In fact, Figure 3 shows that it is present in exons, introns and intergenic sequences of the *D. melanogaster* genome.

As shown in Figure 3, at short range, the amplitude of the oscillations decays until $k \sim 60$, and from $k \sim 60$, the oscillations behave in a manner such that their amplitude slowly increases, reaches a maximum, and decreases also slowly. Close to $k \sim 40$, the amplitude of the oscillations drops more rapidly, becoming unimportant by $k \sim 150$.

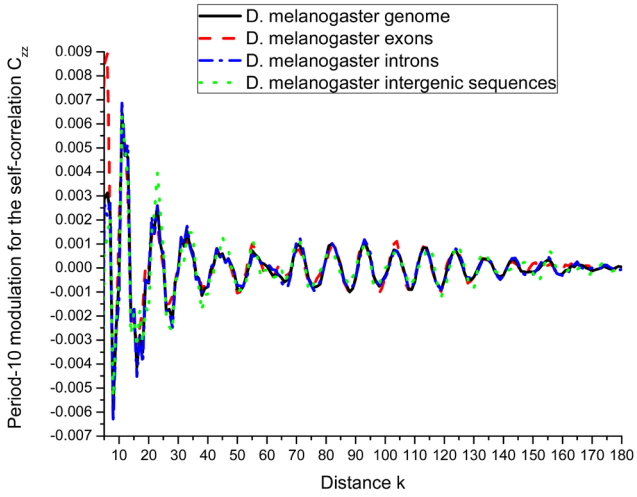

Figure 3. Period~10 modulation extracted from the self-correlation $C_{zz}(k)$ calculated for exons, introns, and, intergenic sequences of *Drosophila melanogaster*. The genomic period~10 modulation is also shown for comparison.

It is known that eukaryotic genomic DNA exists as highly compacted nucleosome arrays called chromatin. Each nucleosome contains a stretch of DNA whose length is of 147 bp, which is sharply bent and tightly wrapped around a histone protein octamer. This sharp bending occurs at every DNA helical repeat (~10 bp). Such bends would be facilitated by specific dinucleotides. In another way, neighbor nucleosomes are separated from each other by stretches of unwrapped linker DNA whose lengths vary from ~10 to ~50 bp.

A careful analysis of the graphs in Figures 2 and 3 also reveals that the period~10 modulation $\Delta C^{10}_{zz}(k)$ can be divided into two characteristic components: the first, due to one pair of base pairs located in the same nucleosomal structure (and separated by a distance k), and, the second, due to one pair of base pairs in which one of the two base pairs is located in one nucleosome sequence, and the other is located in a linker DNA (also separated by the same distance k). We shall call the first component the intranucleosome pattern, and the second the linker-nucleosome pattern.

The intranucleosome pattern is due to the fact the nucleosome positioning and packaging dictates the preferences of the nucleosome DNA sequence for certain specific dinucleotides. Since Figures 2 and 3 show that A/T (or C/G) base pairs have a tendency of repeating at each helical repeat (note that in the graphs in Figures 2 or 3, the distances between maxima or minima are on the order of 10 bp) and A+T-rich regions are preferentially separated by ~5 bp from C+G-rich regions (note that in the same graphs, the distances between adjacent maxima and minima are on the order of 5 bp), the nucleosomal DNA then faces inwards towards the histone octamer through the dinucleotides AA, AT, and TA. Besides, the nucleosomal DNA will face outward through the dinucleotides CC, CG, and GC. Such composition pattern for the nucleosome DNA sequence facilitates its bending around the histone protein octamer.

The linker-nucleosome pattern, in turn, contributes to the period~10 modulation $\Delta C^{10}_{zz}(k)$ because nucleosome-free regions are characterized as being constituted by stretches of poly(dA-dT), which confers to them a certain rigidity (Yuan et al., 2005). Thus, the compo-

sition along such sequences is preferentially constant. Therefore, when we account for the correlation between bases in which one of the two bases is located in one nucleosome sequence and the other is located in a linker DNA, the result will be a modulation similar to that of the intranucleosome pattern (but, with a phase difference between them).

Correlations between nucleotides located in neighboring nucleosomes do not contribute significantly to the period~10 modulation $\Delta C^{10}_{zz}(k)$. In fact, Figures 2 or 3 show that, from *k ~150*, the amplitude of the period~10 modulation $\Delta C^{10}_{zz}(k)$ becomes unimportant. Thus, the contribution to the period~10 modulation $\Delta C^{10}_{zz}(k)$ by the correlation between nucleotides located in two particular neighboring nucleosomes will generally be phase-lagged in relation to the contribution by the correlation between nucleotides located in the same nucleosomal structure (the nucleosome pattern). This implies, therefore, that the pattern of composition of nucleosome DNA is not strictly the same for all nucleosome sequences, such that there must be phase differences between different composition patterns. Finally, all the contributions to the period~10 modulation $\Delta C^{10}_{zz}(k)$ by the correlations between nucleotides located in neighboring nucleosomes throughout the entire genome of *D. melanogaster*, when summed, must interfere destructively at least partially.

Now, any irreducible correlation $C_{ij}(k)$, especially the self-correlation $C_{zz}(k)$, can be expanded as a sum of the 10-smoothed correlation plus the period-3 and period~10 modulations. In fact, we have

$$C_{zz}(k) = C_{zz}(k) - C^{3}_{zz}(k) + C^{3}_{zz}(k) - C^{10}_{zz}(k) + C^{10}_{zz}(k), \qquad \text{(Equation 6)}$$

and since $\Delta C^{3}_{zz}(k) = C_{zz}(k) - C^{3}_{zz}(k)$, and, $\Delta C^{10}_{zz}(k) = C^{3}_{zz}(k) - C^{10}_{zz}(k)$, then

$$C_{zz}(k) = C^{10}_{zz}(k) + \Delta C^{3}_{zz}(k) + \Delta C^{10}_{zz}(k) \qquad \text{(Equation 7)}$$

As we discussed earlier, the period-3 modulation $\Delta C^{3}_{zz}(k)$ relates to the coding segments along the genome (in the present context, the *D. melanogaster* genome). In another way, the modulation $\Delta C^{10}_{zz}(k)$ relates to the nucleosomal pattern. Since such modulation is believed to contain information about the nucleosome positioning and packaging, we proceeded to analyze it more carefully. The first step consisted in applying to it the fast Fourier transform (FFT). The graph in Figure 4 plots the FFT magnitude versus the inverse of the distance *k*.

Figure 4 shows that there are two very pronounced peaks for the FFT magnitude. The first peak occurs for k^{-1}~0.086 and the second for k^{-1} ~0.098. This agrees with the experimental observation that nucleosome positioning and packaging induces a periodicity of 10 bp in nucleosomal DNA, which facilitates its bending. In fact, the observed peaks in Figure 4 occur correspondingly for *k* ~11.6 and *k* ~10.2. This indicates that the nucleosome pattern can be divided into two oscillatory signals, one with a period of 11.6 and other with a period of 10.2. Their magnitudes are similar. There is also a very discrete peak for k^{-1} ~0.0131, which is on the order of the difference between the peaks observed for k^{-1} ~0.086 and k^{-1} ~0.098. Since we have two modulations with similar amplitudes and very close frequencies, this results in a beat that consists in a modulation with period~10.9 overmodulated by another modulation with period~153 (the double of the inverse of 0.0131). We then proposed to develop a fitting for the period~10 modulation $\Delta C^{10}_{zz}(k)$. Such fitting was developed for the range $20 < k < 150$. The test curve, basically, was a product of two cosine functions. In fact, since the FFT procedure resulted in two very close and sharp peaks, we waited to determine if the behavior of

the period~10 modulation would be similar to that of the proposed test curve. Figure 5 shows a graph that plots the period~10 modulation $\Delta C^{10}_{zz}(k)$ as a function of k for *D. melanogaster*, and a proposed fitting.

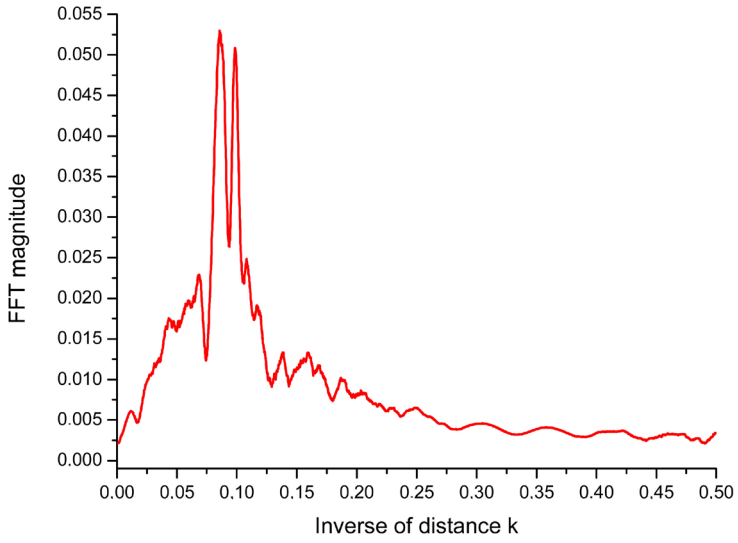

Figure 4. Fast Fourier transform (FFT) magnitude for the period~10 modulation $\Delta C^{10}_{zz}(k)$ for *Drosophila melanogaster*.

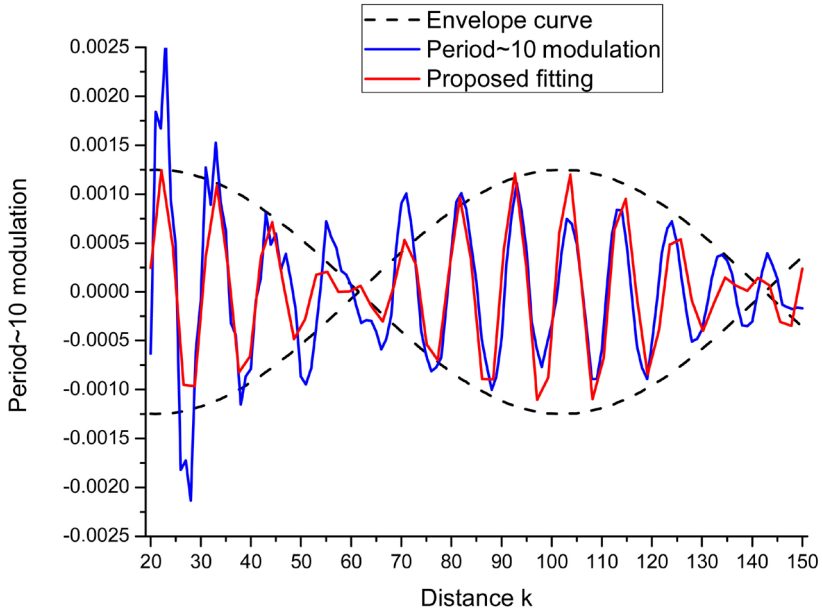

Figure 5. The period~10 modulation $\Delta C^{10}_{zz}(k)$ is a period~10.8 modulation overmodulated by a period~157 modulation. The period~157 modulation is an envelope inside where the period~10.8 modulation oscillates.

The proposed fitting was of the type

$$\Delta C_{zz}^{10}(k) = A cos\left[\frac{2\pi(k-k_0)}{T_0}\right] cos\left[\frac{2\pi(k-k_1)}{T_1}\right] \quad \text{(Equation 8)}$$

where

$A = 0.00125 \pm 0.00007$
$k_0 = 20 \pm 3$
$T_0 = 157 \pm 6$
$k_1 = 0.8 \pm 0.2$
$T_1 = 10.80 \pm 0.03$

The results obtained for the periods, in special, are in good agreement with the results obtained in the FFT procedure. Then, the period~10 modulation $\Delta C^{10}_{zz}(k)$ can be seen as an oscillation with period~10.9 overmodulated by an oscillation with period~153. This last can be regarded as an envelope curve inside where the first oscillates, as Figure 5 shows. Equation 8 can be interpreted as a result of beats between 2 cosines and written in an alternative form as:

$$\Delta C_{zz}^{10}(k) = A_1 cos\left[\frac{2\pi(k-k_1)}{T_1}\right] + A_2 cos\left[\frac{2\pi(k-k_2)}{T_2}\right] \quad \text{(Equation 9)}$$

where now

$A_1 = (6.7 \pm 0.5) \times 10^{-4}$
$k_1 = 2.1 \pm 0.3$
$T_1 = 10.09 \pm 0.03$
$A_2 = (5.8 \pm 0.5) \times 10^{-4}$
$k_2 = -0.6 \pm 0.4$
$T_2 = 11.59 \pm 0.05$

Equations 8 and 9 illustrate the fact that the period~10 modulation $\Delta C^{10}_{zz}(k)$ can be divided into two oscillatory signals, as discussed after Figure 3. Such oscillatory components have similar amplitudes, but slightly different frequencies. One component was identified as the intranucleosome pattern, and the other, as the linker-nucleosome pattern. However, both Equations 8 and 9 do not consider the decaying of the period~10 modulation amplitude observed for $k < 60$. However, in terms of frequencies (or periods), the fitting contained in Equations 8 and 9 produced results in good concordance with those obtained by the FFT procedure shown in Figure 4. Besides, the approach contained in Equation 9 provides information with respect to the role played by the linker DNA in the genomic correlation. In fact, the second term in Equation 9 must be identified as the linker-nucleosome pattern,

$$\Delta C_{zz}^{10^{l-n}}(k) = A_2 cos\left[\frac{2\pi(k-k_2)}{T_2}\right] \quad \text{(Equation 10)}$$

and, the first as the intranucleosome pattern,

$$\Delta C_{zz}^{10^{intra}}(k) = A_1 cos\left[\frac{2\pi(k-k_1)}{T_1}\right] \quad \text{(Equation 11)}$$

The linker-nucleosome pattern 10 is delayed in relation to the intranucleosome pattern 11. When the correlation along the entire genome is taken into account and the period~10 modulation is calculated, for pairs of base pairs separated one from other by a distance *k*, there can be contributions of pairs located in the same nucleosomal structure or pairs where one of the two base pairs is located in a linker DNA, if this distance *k* is sufficiently large. Thus, since the linker DNA does not contribute to the period~10 modulation $\Delta C^{10}_{zz}(k)$, we can conclude that the linker DNA introduces a phase difference in relation to the original intranucleosome pattern. If there were no linker DNAs or nucleosome-free regions, the period~10 modulation would be given by Equation 11. However, the existence of linker DNAs or nucleosome-free regions introduces the additional oscillatory signal in Equation 10, which when added to the signal in Equation 11 creates an overmodulation of period~153, which will modulate the period~10.9 oscillations.

CONCLUSION

In this study, we developed a formalism for calculating period~10 modulations extracted from nucleotide correlations calculated along DNA sequences. We found that the period~10 modulation amplitudes are the highest for the irreducible self-correlation $C_{zz}(k)$ (strong-weak binding), whether calculated along the genomic DNA or calculated for exons, introns or intergenic sequences, for the *D. melanogaster* genome. Since the principal sources for the period~10 modulations are the constraints on the composition of the nucleosome sequences, it was feasible to divide the period~10 modulation into two characteristic components: the intranucleosome and linker-nucleosome patterns (also determined by Cohanim et al. 2006a). Due to the fact that the linker DNAs are preferentially all adenine or all thymine segments or poly(dA-dT) segments, one of the two patterns is viewed as phase-lagged in relation to the other, and the phase difference found was ~3. However, the constraints imposed by the nucleosome arrangement over the nucleosome sequence composition are not strong since the period~10 modulation dies for *k ~150*.

ACKNOWLEDGMENTS

Research supported by the Brazilian agencies CNPq and FAPEMIG. We want to make a special thanks to Professor José Roberto Tozoni, from the Institute of Physics of the Federal University of Uberlândia, for the suggestions given, which we believe have been very important to this study.

REFERENCES

Arneodo A, D'Aubenton-Carafa Y, Audit B, Bacry E, et al. (1998). What can we learn with wavelets about DNA sequences? *Physica A* 249: 439-448.

Audit B, Thermes C, Vaillant C, D'Aubenton-Carafa Y, et al. (2001). Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys. Rev. Lett.* 86: 2471-2474.

Boekhorst R, Abnizova I and Nehaniv C (2008). Discriminating coding, non-coding and regulatory regions using rescaled

range and detrended fluctuation analysis. *Biosystems* 91: 183-194.
Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, et al. (1995). Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* 51: 5084-5091.
Carpena P, Bernaola-Galvan P, Coronado AV, Hackenberg M, et al. (2007). Identifying characteristic scales in the human genome. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 75: 032903.
Cohanim AB, Kashi Y and Trifonov EN (2006a). Three sequence rules for chromatin. *J. Biomol. Struct. Dyn.* 23: 559-566.
Cohanim AB, Trifonov EN and Kashi Y (2006b). Specific selection pressure at the third codon positions: contribution to 10- to 11-base periodicity in prokaryotic genomes. *J. Mol. Evol.* 63: 393-400.
Coward E (1997). Equivalence of two Fourier methods for biological sequences. *J. Math. Biol.* 36: 64-70.
Fickett JW (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* 10: 5303-5318.
Guerra JC and Licinio P (2010). The role played by exons in genomic DNA sequence correlations. *J. Theor. Biol.* 264: 830-837.
Herzel H and Groβe I (1995). Measuring correlations in symbol sequences. *Physica A* 216: 518-542.
Herzel H and Groβe I (1997). Correlations in DNA sequences: the role of protein coding segments. *Phys. Rev. E.* 55: 800-810.
Herzel H, Trifonov EN, Weiss O and Grobe I (1998). Interpreting correlations in biosequences. *Physica A* 249: 449-459.
Josse J, Kaiser AD and Kornberg A (1961). Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. Biol. Chem.* 236: 864-875.
Kogan S and Trifonov EN (2005). Gene splice sites correlate with nucleosome positions. *Gene* 352: 57-62.
Kogan SB, Kato M, Kiyama R and Trifonov EN (2006). Sequence structure of human nucleosome DNA. *J. Biomol. Struct. Dyn.* 24: 43-48.
Licinio P and Caligiorne RB (2004). Inference of phylogenetic distances from DNA-Walk divergences. *Physica 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.
Lowary PT and Widom J (1997). Nucleosome packaging and nucleosome positioning of genomic DNA. *Proc. Natl. Acad. Sci. U. S. A.* 94: 1183-1188.
Oliver JL, Bernaola-Galvan P, Hackenberg M and Carpena P (2008). Phylogenetic distribution of large-scale genome patchiness. *BMC Evol. Biol.* 8: 107.
Peng CK, Buldyrev SV, Goldberger AL, Havlin S, et al. (1992). Long-range correlations in nucleotide sequences. *Nature* 356: 168-170.
Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, et al. (2006). A genomic code for nucleosome positioning. *Nature* 442: 772-778.
Shulman MJ, Steinberg CM and Westmoreland N (1981). The coding function of nucleotide sequences can be discerned by statistical analysis. *J. Theor. Biol.* 88: 409-420.
Silverman BD and Linsker R (1986). A measure of DNA periodicity. *J. Theor. Biol.* 118: 295-300.
Sueoka N (1959). A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. *Proc. Natl. Acad. Sci. U. S. A.* 45: 1480-1490.
Widom J (1996). Short-range order in two eukaryotic genomes: relation to chromosome structure. *J. Mol. Biol.* 259: 579-588.
Yin C and Yau SS (2007). Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J. Theor. Biol.* 247: 687-694.
Yuan GC, Liu YJ, Dion MF, Slack MD, et al. (2005). Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309: 626-630.