BioSystems 107 (2012) 142–144

Contents lists available at SciVerse ScienceDirect

BioSystems

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

On the origin of three base periodicity in genomes

Kushal Shah*, Annangarachari Krishnamachari[1]

School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India

ARTICLE INFO

Article history:
Received 30 September 2011
Received in revised form 1 November 2011
Accepted 4 November 2011

Keywords:
Correlation
Periodicity
Nucleotide distribution

ABSTRACT

Genomes of almost all organisms have been found to exhibit several periodicities, the most prominent one is the three base periodicity. It is more pronounced in the gene coding regions and has been exploited to identify the segments of a genome that code for a protein. The reason for this three base periodicity in the gene-coding region has been attributed to inhomogeneous nucleotide compositions in the three codon positions. However, this reason cannot explain the three base periodicity present at the level of the whole genome where the codon concept is not applicable. Even though the distribution of each nucleotide is uniform at the positions 0(mod 3), 1(mod 3) and 2(mod 3) when the whole genome data is considered, our analysis reveals that the three base periodicity is arising because of higher correlations among the nucleotides separated by three bases.

1. Introduction

Correlation-based methods (Li et al., 1994; Voss, 1992; Hod and Keshet, 2004) and Fourier transform techniques (Voss, 1992; Tiwari et al., 1997) have become important tools in the analysis of a wide variety of systems including genomic DNA. Though these two tools can be used independently, they have a simple relationship between them. The Fourier transform of the correlation function is the power spectrum of the original sequence.

Several important results have been found for genomic DNA using correlation-based tools. One of them is the presence of long-range correlations in the DNA sequence of organisms. Also, not only is there long-range correlation, it also obeys a power-law with the exponent close to one (Voss, 1992). Another important result that has been found using these techniques is the presence of certain periodicities (mainly at 3, 10.5, 200, 400 base pairs) in the genomes of a wide variety of organisms (Trifonov, 1998; Herzel et al., 1999). Among these several periodicities found, the most prominent one is at three base pairs. Since this three base-pair periodicity is much more pronounced in the gene-coding region compared to the non-coding region, it has been exploited to identify gene-coding regions in several bacterial genomes (Tiwari et al., 1997). For the gene-coding regions, it has been shown that this three base-pair periodicity arises due to the inhomogeneous nucleotide distributions in the three codon positions (Datta and Asif, 2005; Yin and Yau, 2005). However, their argument does not explain the occurrence of three base-pair periodicity even when the whole genome is taken into consideration. Nucleotide distributions across the whole genome are uniform in the three codon positions. In this paper, we show that the nucleotides separated by three positions along the length of the genome have a higher correlation strength than those separated by one or two positions.

2. Materials and methods

The auto-correlation function, $C(k)$, of a discrete sequence, $\{a_i, i = 1, 2, \ldots, N\}$, is defined as (Beauchamp and Yuen, 1979; Cavicchi, 2000)

$$C(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} (a_j - \mu)(a_{j+k} - \mu) \tag{1}$$

where each a_i takes a value from $\{+1, -1\}$, so that $\mu = \langle a_i \rangle = 0$.

To use Eq. (1), we need to convert the DNA sequence into a sequence of bits. Since a DNA sequence is made up of four bases, we can generate a string of bits by assigning a value of +1 to every occurrence of G and -1 to all other positions (similarly for A, T, C). Thus, a given DNA sequence gives rise to four different bit strings and four different values of correlation strength (i.e. auto-correlation values) corresponding to each of the four bases, A, T, G, and C. In this paper, we have considered only correlations due to the Guanine (G) residue since it has shown to play an important role in determining the 3-base periodicity in the gene-coding region (Yin and Yau, 2005).

The value for $C(k)$ ranges from 0 to 1 and is independent of the length of the sequence. Lower value of $C(k)$ corresponds to lower correlation strength and vice-versa. The value of $C(k)$ for a typical random sequence will be zero and for a highly correlated sequence will approach unity.

Genomes of all the species used in this paper were downloaded from NCBI FTP website.

3. Results and discussion

Fig. 1 shows the values of $C(k)$ for the genome of *Escherichia coli* (containing both gene-coding and non-coding segments), which is

* Corresponding author. Tel.: +91 11 2673 8703; fax: +91 11 2674 1586.
E-mail addresses: kkshah@mail.jnu.ac.in, atmabodha@gmail.com (K. Shah), chari@mail.jnu.ac.in (A. Krishnamachari).
[1] Tel.: +91 11 2673 8703; fax: +91 11 2674 1586.

0303-2647/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2011.11.006

Table 1
Values of percentage of Guanine base and its correlation for several organisms. f_G is the fraction of Guanine base among all nucleotides present in the given genome. f_{Gi} is the fraction of the i(mod 3)th positions of the genome ($i = 0, 1, 2$) having Guanine base. C_G is the average of all the $C(k)$ values ($C_G = \langle C(k) \rangle$). As can be seen, higher value of f_G results in a lower value of C_G. Also, those organisms which have close values of f_G tend to have similar values for C_G. However, the values of $\langle C_k \rangle_{k=0(mod)3}$ has been found to be higher than that of $\langle C_k \rangle_{k=1(mod)3}$ and $\langle C_k \rangle_{k=2(mod)3}$ (in this case, the averaging was done for $k \in \{1, 2, \ldots, 99\}$).

Accession no.	Length	Name	f_G	f_{G1}	f_{G2}	f_{G0}	C_G	$\langle C_k \rangle_{k=0(mod)3}$	$\langle C_k \rangle_{k=1(mod)3}$	$\langle C_k \rangle_{k=2(mod)3}$
NC_000909	1,664,970	*Methanocaldococcus jannaschii*	0.1589	0.1598	0.1587	0.1583	0.4615	0.4941	0.4695	0.4684
NC_001135	316,617	*Saccharomyces cerevisiae* CHR III	0.1884	0.1909	0.1856	0.1886	0.3880	0.4033	0.3903	0.3896
NC_000964	4,215,606	*Bacillus subtilis*	0.2171	0.2186	0.2167	0.2159	0.3039	0.3383	0.3187	0.3179
NC_006582	4,307,996	*Bacillus clausii*	0.2231	0.2210	0.2234	0.2249	0.3007	0.3245	0.3039	0.3030
NC_000917	2,178,400	*Archaeoglobus fulgidus*	0.2438	0.2440	0.2430	0.2442	0.2571	0.2929	0.2639	0.2640
NC_000913	4,639,675	*Escherichia coli*	0.2537	0.2553	0.2526	0.2532	0.2434	0.2597	0.2402	0.2398

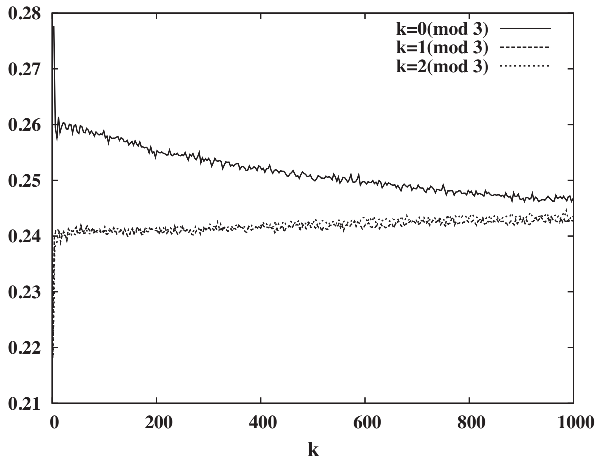

Fig. 1. Plot of $C(k)$ for the genome of *E. coli*. As can be clearly seen from the plot, the correlations for $k = 0$(mod 3) is higher than the correlations for other values of k.

one of the most widely studied bacterial model organism. As can be seen, the correlations of nucleotide for positions separated by 0(mod 3) base pairs is higher than that for positions separated by 1(mod 3) or 2(mod 3) base pairs. The same result was also found for the GroEL gene of *E. coli*, as shown in Fig. 2. Similar results were found for other organisms as well and that is summarized in Table 1.

In Table 1, f_G is the fraction of Guanine base among all nucleotides present in the given genome. f_{Gi} is the fraction of the i(mod 3)th positions of the genome ($i = 0, 1, 2$) having Guanine base and it is found that $f_{G1} \approx f_{G2} \approx f_{G0}$ for all the genomes considered in this paper. This is different from the case of individual genes where f_{G1} is typically greater than f_{G2} and f_{G0} (Yin and Yau, 2005). This shows that the reason for three base periodicity found in entire genome is different from that for individual genes.

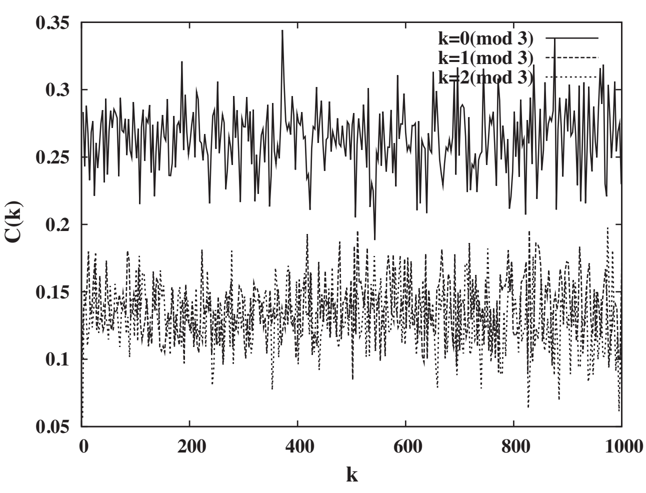

Fig. 2. Plot of $C(k)$ for the GroEL gene of *E. coli*. As can be clearly seen from the plot, the correlations for $k = 0$(mod 3) is higher than the correlations for other values of k.

In Table 1, C_G is the average of all the $C(k)$ values ($C_G = \langle C(k) \rangle$). As can be seen, higher value of f_G results in a lower value of C_G (with an approximately linear scaling). Also, those organisms which have similar values of f_G tend to have similar values for C_G. This is primarily because the Fourier transform of most genomes has been shown to obey a power-law with exponent close to 1 (Voss, 1992). Thus, the value of the correlation depends more or less on the percentage of a given nucleotide and not its precise positional distribution. Although this phenomenon is found to hold for the entire genome, deviations exist when a sub-sequence of a given genome is considered. As can be seen in Table 1, the values of $\langle C_k \rangle_{k=0(mod)3}$ are found to be higher than that of $\langle C_k \rangle_{k=1(mod)3}$ and $\langle C_k \rangle_{k=2(mod)3}$ (in this case, the averaging was done for $k \in \{1, 2, \ldots, 99\}$), even though $f_{G1} \approx f_{G2} \approx f_{G0}$. This shows that the Guanine base positions separated by three bases are more correlated than those separated by one or two bases, thereby leading to a 3 base periodicity across the entire genome of various organisms. Unlike the case of gene-coding regions, this difference in correlations is not captured by measuring the occurrence frequency of G in the three positions. Thus, the reason for the presence of period 3 across the entire genome is different from that of individual genes.

The higher correlations reported in this paper are not due to the codon system (information organization) of gene-coding regions. The peak in the Fourier transform for gene-coding regions is due to the inhomogeneous distribution of nucleotides at the three codon positions (Yin and Yau, 2005). If we extend this argument to the genome level, then we do expect a similar inhomogeneous distribution. But the results obtained in this paper show otherwise. The nucleotide distribution is uniform when the whole genome is taken. Surprisingly, we see this homogeneous distribution in bacteria, archaea as well as yeast, i.e. across the phyla implying that it is not a species specific characteristic. The higher correlations observed in this paper may be due to biological reasons (currently unknown) which needs further investigation.

Acknowledgement

KS would like to thank SCIS-JNU, the Center of Excellence of the Department of Biotechnology (Government of India) for financial support.

References

Beauchamp, K.G., Yuen, C.K., 1979. Digital Methods for Signal Analysis. Geroge Allen and Unwin, London.

Cavicchi, T.J., 2000. Digital Signal Processing. John Wiley & Sons, New York.

Datta, S., Asif, A., 2005. A fast DFT based gene prediction algorithm for identification of protein coding regions. In: Proceedings of ICASSP, pp. 113–116.

Herzel, H., Weiss, O., Trifonov, E.N., 1999. 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics 15, 187–193.

Li, W., Marr, T.G., Kaneko, K., 1994. Understanding long-range correlations in DNA sequences. Physica D 75, 392–416.

Hod, S., Keshet, U., 2004. Phase transition in random walks with long-range correlations. Phys. Rev. E 70, 015104(R).

Tiwari, S., Ramachandran, S., Bhattacharya, A., et al., 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Comput. Appl. Biosci. 13, 263–270.

Trifonov, E.N., 1998. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. Physica A 249, 511.

Voss, R.F., 1992. Evolution of long range fractal correlations and 1/*f* noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.

Yin, C., Yau, S., 2005. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. J. Comp. Biol. 12, 1153.