ELSEVIER

Physica A 265 (1999) 78–84

PHYSICA A

Effect of replication on the third base of codons

S. Cebrat[a], M.R. Dudek[b,*], A. Gierlik[a], M. Kowalczuk[a], P. Mackiewicz[a]

[a] *Institute of Microbiology, University of Wrocław, ul. Przybyszewskiego 63/77 54-148 Wrocław, Poland*
[b] *Institute of Theoretical Physics, University of Wrocław, pl. Maxa Borna 9 50-204 Wrocław, Poland*

Received 31 July 1998

Abstract

We have analyzed third position in codons and have observed strong long-range correlations along DNA sequence. We have shown that the correlations are caused mostly by asymmetric replication. In the analysis, we have used a DNA walk (spider analysis Cebrat et al., Microbial Comparative Genomics 2(4) (1997) 259–268) in two-dimensional space [A–T,G–C]. The particular case of the *E.coli* sequence has been studied in detail. © 1999 Elsevier Science B.V. All rights reserved.

PACS: 87.10.+e; 87.15.-v

Keywords: DNA walk; Long-range correlations; Replication

1. Introduction

For the past few years the question about the existence of statistical long-range base–base correlations in some DNA sequences [1] has been widely discussed. Long-range correlations were observed first in 1992 by three groups, Li et al. [2,3], Peng et al. [4] and Voss [5]. It had been suggested in [2–4] that the non-coding sequences display long-range correlations while coding sequences have only short-range correlations. The observation was later confirmed several times by various methods, e.g. by Peng et al. [6], Bernaola-Galván et al. [7], Arnéodo et al. [8], or recently in papers, e.g. by Buldyrev et al. [9] and by Arnéodo et al. [10], but the nature of their generation and causes are still obscure. Moreover, there are methods of statistical analysis, e.g. in paper by Azbel [11], which show that there are no long-range correlations in DNA.

In the paper by Arnéodo et al. [10] there is an additional message that although coding sequences have no long-range correlations the same method of analysis (wavelet

[*] Corresponding author. Fax: +48-71-214-454; e-mail: mdudek@mirek.ift.uni.wroc.pl.

PII: S0378-4371(98)00657-8

transform) applied to the third position of each codon shows long-range base–base correlations. The result coincides with our finding strong asymmetry between sense and antisense strands of coding sequences [12,1,13,14]. In particular, in paper [14] we show how the results of statistical analysis of DNA in terms of DNA walks are influenced by triplet structure of coding sequences.

In the following, we show that replication, which is asymmetric, is responsible for introducing strong trends in the third base of codons and in consequence causes the long-range base–base correlations.

2. Coding trends

There is a complementation rule specific for double-strand DNA that the number of adenines (A) is equal to the number of thymines (T), and the number of guanines (G) is equal to the number of cytosines (C). If we count the numbers for a single DNA strand, we observe a small deviation from the rule, but still the number of A balances well the number of T, and the number of G balances well the number of C. The smaller the region of DNA strand, the more significant becomes the misfit of compensation. The most unbalanced regions are those represented by protein coding sequences, because of abundance of some nucleotides in specific positions of codons [12–14]. However, even very large subregions of DNA strand of a size of a few million nucleotides can have strong misfit of compensation. For example, in *Escherichia coli* genome (*E.coli*, 4.6 M bases long – downloaded 10 May 1997 from http://genom4.aist-nara.acjp) the global overplus of A is equal to $\sim 0.03\%$ of the total number of nucleotides. At the same time, the overplus of A is equal to $\sim 0.1\%$ of the total number of nucleotides in one half of the genome, the fragment from the TER sequence (the Terminator of Replication) to the ORI sequence (the Origin of Replication). *E.coli* has a circular genome which can be divided in a natural way into two segments, ORI-TER and TER-ORI. The reason for the non-compensation is that the process of replication is asymmetric and not equivalent for the two DNA strands. In procaryotic genomes the ORI and TER determine the switching mode of replication for each DNA strand: the mode when DNA strand is synthesized continuously (leading strand), and the mode when DNA strand is synthesized in fragments (lagging strand) which are then ligated. Enzymatic mechanisms involved in the synthesis of the two strands are different. This results in different mutational pressures which lead to compositional bias between the leading and lagging strand. If ORI-TER of one strand is leading, the complementary fragment of the other strand is lagging. This causes the purine/pyrimidine bias also in non-coding regions [15,16]. The DFA method applied for the purine-pyrimidine DNA walks [6,7] representing *E.coli* genome shows the existence of compositionally homogeneous patches of nucleotides along DNA strands.

The observation by Arnéodo [10] that the third position in codons has the same degree of long-range correlations as non-coding sequences can be explained with the help of replication mechanisms. This has also been suggested by us in paper [14]

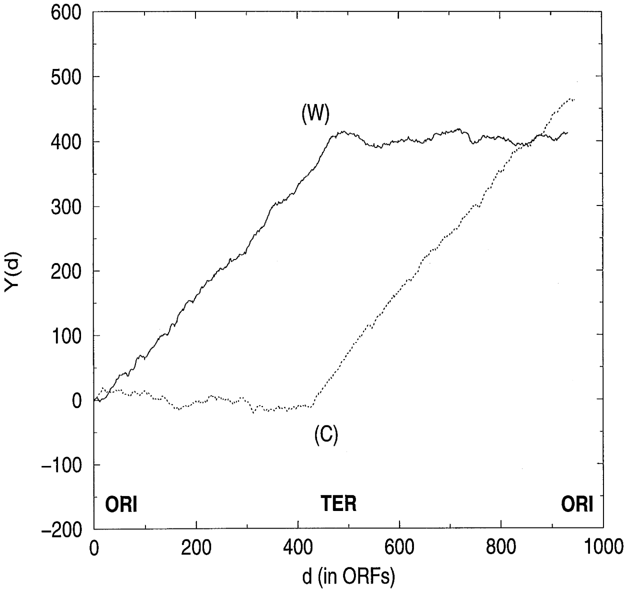

Fig. 1. DNA walks for position 3 in codons of nonoverlapping ORFs (longer or equal to 150 codons) belonging to strand (W) and (C). Walker step is equal to ±S. ORI represents the beginning of *E.coli* genome (4.6 M bases long) and TER is 2.3 M bases apart.

(Fig. 1 in [14]). In the following, we will discuss replication trends at third position in codons for *E.coli.*

The third position in codons is degenerated. Often the base at the position is not significant for the sense of the codon and therefore we can expect that the third positions of codons accumulate silent mutations.

Investigating the coding sequences of DNA (e.g., [1,14]), we have shown that their statistical analysis should be done separately for three subsequences representing bases in position 1, 2, and 3 of codons. We have used a DNA walk analysis, originating from the paper by Berthelsen et al. [17], for each of the three subsequences to show strong correlations present in coding sequences. In this case the walker makes unit shifts in two-dimensional space (T–A,C–G) depending on the type of nucleotide visited. The shifts are: (0,1) for G, (1,0) for A, $(0,-1)$ for C, and $(-1,0)$ for T. The resulting DNA walk, called by us spider [1] has three "legs" representing position 1,2 and 3 in codons. The vector determined by the origin and the end of the spider leg, expressed in terms of the vector length and its angle with (T–A) axis, appears to be a sufficient parameter for (*open reading frame*) (ORF) discrimination [1,12,13]. We have observed that spiders of typical genes use a restricted range of the leg angles and the legs corresponding to the bases in the first and second positions in codons are very elongated in space (T–A,C–G), whereas the third leg is much shorter, suggesting more random base composition. In practice, during sequence analysis, one considers only long ORFs (having more than 100 codons) to be coding [18], thinking of the

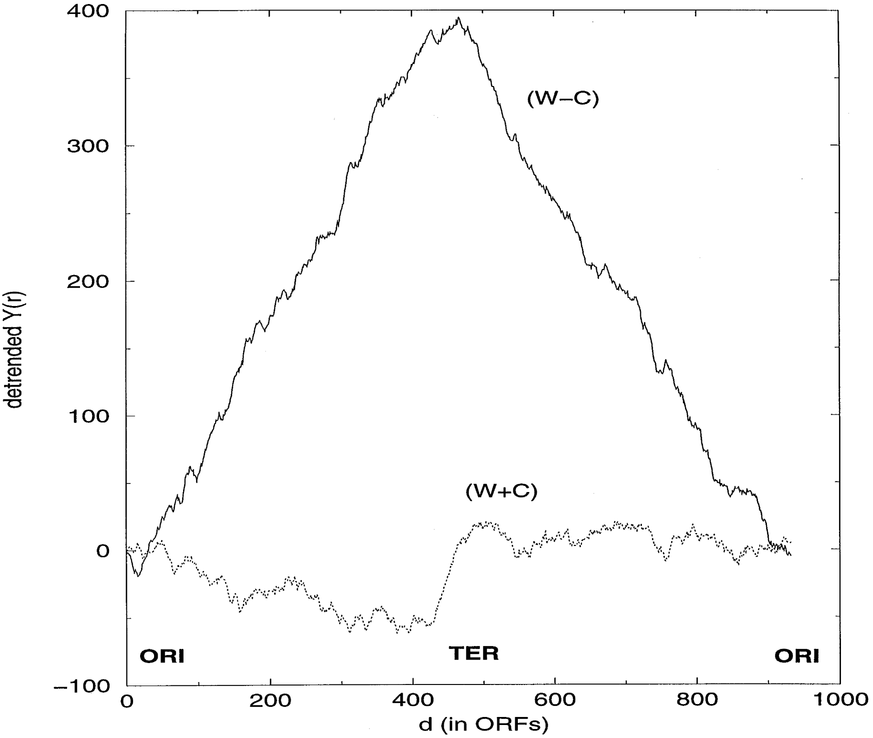

Fig. 2. The global linear trends of two DNA walks from Fig. 1 have been removed and the sum (W+C) and difference (W−C) of the detrended walks have been plotted.

short ones as purely random. We will analyze ORFs and therefore it is reasonable to weight them with respect to their length. What is more, we will restrict our analysis only to non-overlapping ORFs, longer or equal to 150 codons. This ensures that our statistics will not suffer from the overlaps and that ORFs almost certainly represent genes. Below, we examine $N_G - N_C$ value for ORFs, where N_G and N_C are numbers of G and C in ORFs at third position of codons (third spider leg) for which we expect to find long-range correlations discussed by Arnéodo et al. [10]. To this aim, we associate the following variable with each ORF:

$$S = (N_G - N_C)/\sqrt{N} \tag{1}$$

weighted by square root of the number N of codons determining the length of the ORF. $|S| > 1$ indicates the existence of strong trends. One can use an alternative formula for S

$$S = \frac{R}{\sqrt{N}} \sin(\phi)\,, \tag{2}$$

where $\sin(\phi) = (N_G - N_C)/R$, ϕ is the angle made by the vector determined by the origin and the end of the third leg of spider representing ORF with the (T–A) axis, and R is length of the vector.

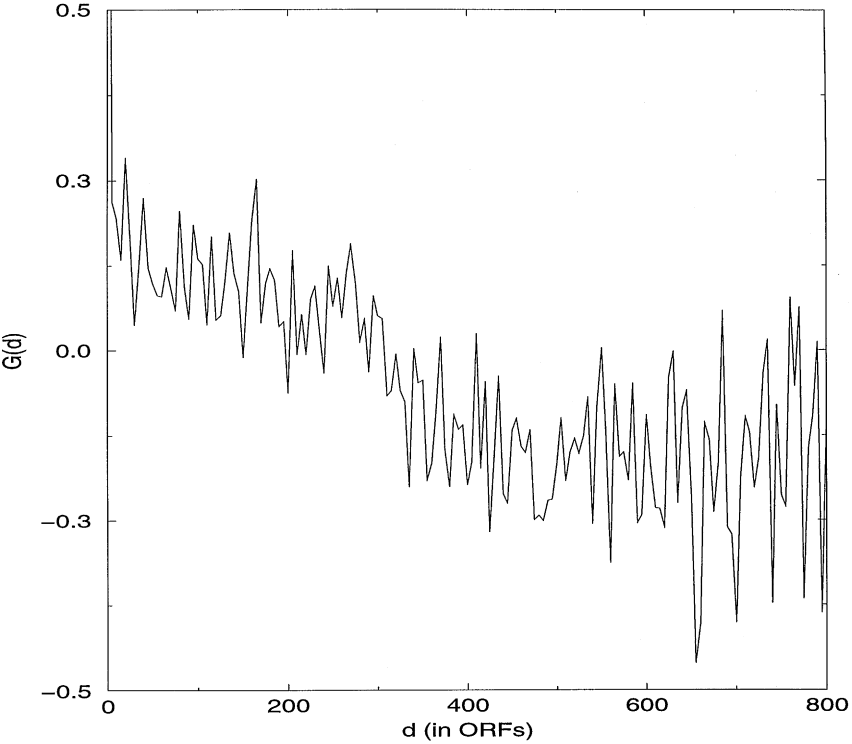

Fig. 3. Two-point correlation function $G(d)$ versus distance (d) (measured in ORFs) for S calculated for bases at third position in codons. The ORFs appear in natural order of strand W, they are non-overlapping and 150 or more codons long.

We have defined a walk restricted to the third position in codons of ORFs in such a way that while scanning succeeding ORFs in a DNA strand, the walker goes “up” or “down” by a value $\pm|S|$ depending on the sign of S. In Fig. 1 are presented the walks for ORFs of Watson (W) and Crick (C) strands. One can notice the points, where DNA walks change their directions. These points represent terminator of replication, where the switch between leading and lagging role of DNA occurs. Thus, it is possible to conclude that the process of replication is responsible for this asymmetry. If this conclusion is correct, asymmetries for the two complementary strands caused by replication process should posses opposite signs, i.e., they should compensate each other if added, and should double the effect, if subtracted. To show this, we have subtracted the global linear trends in the walks (W) and (C) from Fig. 1. Next, we have plotted the sum (W+C) and difference (W−C) of the detrended DNA walks (Fig. 2). The results confirm our expectation of asymmetry of replication, i.e. the enhancement of trends for (W−C) and compensation of them for (W+C). The trends cannot be caused by transcription mechanisms, because the transcription effect should be the same for coding sequences in both DNA strands, and should cumulate when added.

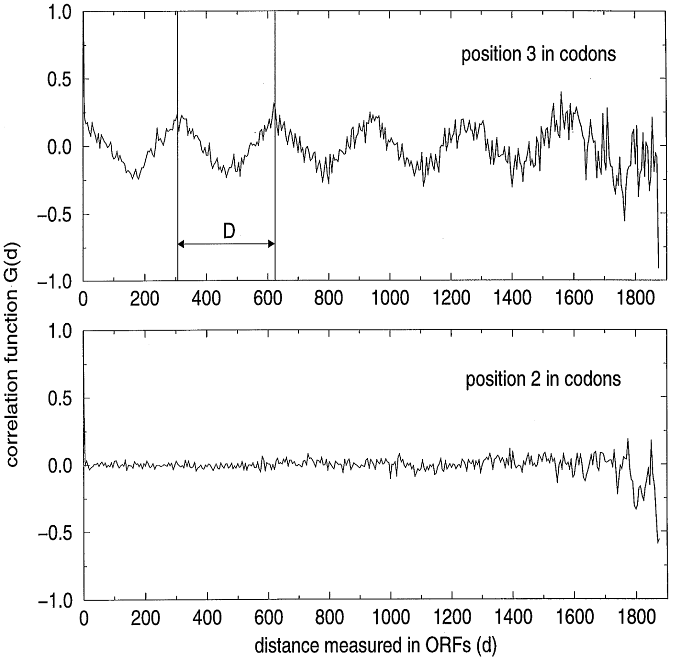

Fig. 4. Two-point correlation function $G(d)$ versus distance (d) (measured in ORFs) for sequence of ORFs ordered phase after phase (here, phases (1), (2), (3), (6), (5), (4)). Distance D has the length of one phase. S was calculated for bases at third position in codons (the upper part) and second position in codons (lower part). The ORFs are non-overlapping and are 150 or more codons long.

The trends at third position in codons can be examined with the help of the two-point correlation function $G(d)$ defined as follows:

$$G(d) = \langle S(r)S(r+d)\rangle - \langle S(r)\rangle^2 , \tag{3}$$

where d is distance between two ORFs (measured in ORFs number in between). It confirms the existence of the overall trends in ORI-TER and TER-ORI fragments. It can be observed in Fig. 3, where the correlation function $G(d)$ plotted versus distance d has non-vanishing tail of the length range comparable with the length of the ORI-TER fragment.

One might say that the replication trends at third position in codons are negligible. It is not the case. To show this we have used six-phase representation of DNA [14,18–20] (in this case each DNA strand is represented by three reading frames) and we have read all ORFs under consideration in their phases. Next, we spliced the phases (e.g., in the following order: phases (1), (2), (3), (6), (5), (4)) and calculated $G(d)$ (Eq. (3)) for the resulting sequence. Thus, we obtained strong long-range correlations between them. It is evident from Fig. 4 (upper part), where the periodic trends are clearly visible with the period of length of chromosome. This can be compared with the corresponding $G(d)$ calculated at the second position of codons (lower part of

Fig. 4), where we have not observed long-range correlations. We should remember that the second position in codons determines the property of coded amino acid and most mutations at these positions are not acceptable.

3. Conclusions

We have shown that the finding by Arnéodo et al. [10] that third position in codons has the same degree of long-range correlations as non-coding sequences can be explained by mutational pressure introduced by replication mechanisms. There are different mutational pressures imposed on leading and lagging DNA strands. Therefore, some nucleotides are not compensated by their counterparts in the same DNA strand. The small overplus of them produces long trends between ORI and TER, which mark start and stop for replication process. The trends are not caused by transcription process – they are witness to silent mutations at the third position of codons, and similar trends can be found in the non-coding sequences (this has been shown in paper by Cebrat et al. [14]).

References

[1] S. Cebrat, M.R. Dudek, P. Mackiewicz, M. Kowalczuk, M. Fita, Microbial Comparative Genomics 2 (4) (1997) 259–268.

[2] W. Li, International J. Bifurcation Chaos 2 (1992) 137–154.

[3] W. Li, K. Kaneko, Europhys. Lett. 17 (1992) 655–660.

[4] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature 356 (1992) 168–170.

[5] R. Voss, Phys. Rev. Lett. 68 (1992) 3805–3808.

[6] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M. Simons, H.E. Stanley, Physica A 221 (1995) 180–192.

[7] P. Bernaola-Galván, R. Román-Roldán, J.L. Oliver, Phys. Rev. E 53 (1996) 5181–5189.

[8] A. Arnéodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. 74 (1995) 3293–3296.

[9] S.V. Buldyrev, N.V. Dokholyan, A.L. Goldberger, S. Havlin, C.-Peng, H.E. Stanley, G.M. Viswanathan, Physica A 249 (1998) 430–438.

[10] A. Arnéodo, Y. d'Aubenton-Carafa, B. Audit, E. Bacry, J.F. Muzy, C. Thermes, European Phys. J. B 1 (1998) 259–263.

[11] M.Ya. Azbel, Phys. Rev. Lett. 75 (1995) 168–171.

[12] S. Cebrat, M.R. Dudek, A. Rogowska, J. Appl. Gen. 38 (1997) 1–9.

[13] S. Cebrat, M.R. Dudek, P. Mackiewicz, Theory Bioscienc. 117 (1998) 78–89.

[14] S. Cebrat, M.R. Dudek, European Phys. J. B (1998).

[15] C.-I. Wu, N. Maeda, Nature 327 (1987) 169–170.

[16] M.P. Francino, H. Ochman, Trends Genet. 13 (1997) 240–245.

[17] Ch.L. Berthelsen, J.A. Glazier, M.H. Skolnick, Phys. Rev. A 45 (1992) 8902–8913.

[18] B. Dujon, Trends Genet. 12 (1996) 263–270.

[19] S. Cebrat, M.R. Dudek, Trends Genet. 12 (1996) 12.

[20] B. Dujon, A. Goffeau, Trends Genet. 12, Poster (1996).