J. Appl. Genet. 40(1), 1999, pp. 1-14

Asymmetry of nucleotide composition of prokaryotic chromosomes

Paweł MACKIEWICZ[1], Agnieszka GIERLIK[1], Maria KOWALCZUK[1], Mirosław R. DUDEK[2], Stanisław CEBRAT[1]

[1]Institute of Microbiology, Wrocław University, Wrocław, Poland
[2]Institute of Theoretical Physics, Wrocław University, Wrocław, Poland

Abstract. We have analysed the causes of asymmetry in nucleotide composition of DNA complementary strands of prokaryotic chromosomes. Analysing DNA walks we have separated the effect of replication-associated processes from the effect introduced by transcription and coding functions. The asymmetry introduced by replication switches its polarity at the origin and at the terminus of replication, which is observed in both noncoding and coding sequences and varies with respect to positions in codons. Coding functions introduce very strong trends into protein coding ORFs, which are specific for each nucleotide position in the codon. Using detrended DNA walks we have eliminated the effect of coding density and we were able to distinguish between mutational pressure associated with replication and compositional bias for genes proximal and distal to the origin of replication.

Key words: asymmetrical replication, *Bacillus subtilis*, *Borellia burgdorferii*, DNA asymmetry, DNA walks, *Escherichia coli*, *Treponema pallidum*.

Introduction

A random DNA sequence should not exhibit any statistically significant compositional bias between the two complementary strands. Nevertheless, there are some processes which do not treat the two strands of natural DNA molecule equally. One of these processes is replication. The main cause of unequal fidelity of leading and lagging strand replication is still not clear. It is controversial if replication of only one or both strands is discontinuous (OKAZAKI et al. 1968, KORNBERG, BAKER 1992, WANG, CHEN 1992, 1994). Nevertheless, the topology of the replication fork itself requires the involvement of different enzymatic

Received: November 1998.
Correspondence: S. CEBRAT, Institute of Microbiology, Wrocław University, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland, E-mail: cebrat@angband.microb.uni.wroc.pl.

mechanisms in replication of each DNA strand (KUNKEL 1992, WAGA, STILLMAN 1994). Besides the above-mentioned mechanisms, differences in processivity of leading and lagging DNA strands may be responsible for differential accuracy of DNA replication of these two strands (FIJALKOWSKA et al. 1998). Thus, both strands are exposed to different mutational pressures and compositional bias has been found between them as a result (LOBRY 1996a,b, BLATTNER et al. 1997, GRIGORIEV 1998, MRAZEK, KARLIN 1998).

A different kind of asymmetry in DNA strands is generated by their coding functions. Some authors argue that it is the transcription mechanism itself which introduces the asymmetry into DNA strands (FRANCINO et al. 1996, FRANCINO, OCHMAN 1997, FREEMAN et al. 1998). Transcription mechanism is asymmetrical because only one strand serves as a source of information during transcription and the other one is, as a single strand, exposed to mutational pressure. The non-transcribed strand is four times more prone to deamination than the transcribed one, which leads to substitutions C→U or deamination of 5-methylcytosine to thymine (BELETSKII, BHAGWAT 1996). Nevertheless, the selection pressure imposed on the amino acid composition of the coded peptides and the codon bias connected with the efficiency of translation can also introduce the compositional bias into coding sequences (IKEMURA 1982, SHARP et al. 1993). It seems difficult to distinguish between the mutational pressure of transcription mechanism itself and this selection.

Since the trends introduced by coding sequences are very strong and different for each position in codons (CEBRAT et al. 1997), we have analysed walks performed separately for the first, second and third position in codons. In addition, we have performed detrended walks by elimination of the average trends to show the effect of gene position on the chromosome. This allowed us to separate the effect of replication from other effects introducing asymmetry into DNA complementary strands.

Material and methods

All the results presented in the paper were obtained by an analysis of the *Escherichia coli* genome which was downloaded from http://www. genetics.wisc.edu. on December 1, 1997. After retrieval, the data have not been updated.

DNA compositional bias was described by DNA walks. To describe the bias in the whole chromosome sequence we have performed the G↔C and A↔T DNA walks. In the G↔C DNA walk, the walker moved along the chromosome sequence (co-ordinates of axis X) and it moved one unit up (co-ordinates of axis Y) when the visited nucleotide was G, or down when the visited nucleotide was C. In the A↔T DNA walks the walker moved up when the visited nucleotide was A and down when the visited nucleotide was T.

When analysing Open Reading Frames (ORFs) the walker starts its walk at the start codon of the first ORF on the analysed strand. If the first codon positions were analysed, the walker visited only first positions of codons, jumping every third nucleotide and moving up or down as in previously described walks. After checking one ORF the walker jumped to the start of the next ORF of the same strand until the last ORF of the analysed strand was checked. The values on the X-axis represent the positions of start codons on the scale of the whole chromosome. The other codon positions were analysed similarly.

To analyse intergenic sequences we have performed the same kind of walks for sequences outside ORFs longer than 70 codons. In this case the walker checked each nucleotide position (we assumed that there was no reading frame or triplet structure outside ORFs).

To show specific trends for each position in codons we have performed walks in two-dimensional space (Figure 3). We have spliced all ORFs longer than 150 codons of one strand and performed three independent walks on such a sequence, one for each position in codons. Every jump of a walker is associated with a unit shift in the two-dimensional space 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. Hence, each DNA walk represents the "history" of nucleotide composition of the first, the second or the third position of codons along the DNA sequence. This walk is a modification of two-dimensional Berthelsen walk (BERTHELSEN et al. 1992). Note that in this kind of DNA walks neither axis represents chromosome co-ordinates.

Since the trends introduced by coding functions into specific nucleotide positions in codons are very strong and mask a possible asymmetry of strands which could be a result of mutational pressure, we have performed a kind of "detrended walks". In this kind of walks each movement of the walker was corrected by a factor which allows the walker to finish the walk at value 0. For example, during the analysis of nucleotide composition of specific positions in codons, the frequency of a particular nucleotide at the analysed positions was counted. Next, the value of the walker jump for a given ORF was counted from the equation:

$$J = [N] - (\overline{F_N} \times L),$$

where J – the value of the walker jump for the ORF, N – the number of molecules of the analysed nucleotide (A,T,G or C) in the analysed positions of the ORF, $\overline{F_N}$ – the frequency of occurrence of this nucleotide at the examined positions in all ORFs in the genome, and L – the length of the ORF (in codons).

If the intergenic sequences were analysed, F was the frequency of a nucleotide in the whole set of intergenic sequences and L was the length of the visited sequence in nucleotides.

Analogous detrended walks have been done for G↔C and A↔T walks.

When walks for two strands were added, the walker visited non-overlapping ORFs of both strands as they appear in the chromosome, scanned them in

the proper reading frame and moved according to the result of scanning. When walks for the Crick strand were subtracted from the walks for the Watson strand, the value of walker jump for each ORF lying in the Crick strand was multiplied by (–1).

Results

Asymmetry of the whole chromosome

In Figure 1 DNA walks for the whole Watson strand of the *E. coli* chromosome are shown. The walker starts at the origin of replication, moves along the strand in 5'→3' direction and ends its analysis at the same point (the analysed chromosome is circular). Note that the same DNA walk performed for the whole Crick strand in 3'→5' direction (not shown) gives the exact mirror picture of the DNA walk for the Watson strand. The DNA walk in Figure 1 shows specific asymmetry in G/C composition. The first part of the plot represents the leading DNA strand. At the maximum – the terminus of replication – the role of the Watson strand is switched from leading to lagging. The second switch is at the origin of replication.

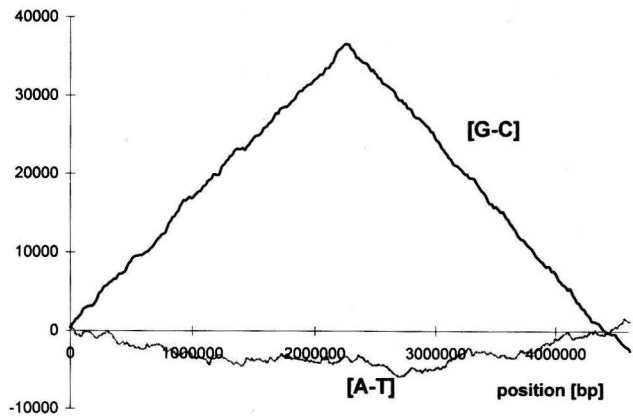

Figure 1. DNA walk on the W strand of the *E. coli* genome. In [G-C] walk the walker moved one unit up when the visited nucleotide was G, and down when the visited nucleotide was C. In [A-T] walk the walker moved up or down when the visited nucleotide was A or T, respectively.

Asymmetry of coding and noncoding sequences

To resolve the problem whether replication or transcription is responsible for the observed asymmetry, we have performed DNA walks for intergenic sequences and for non-overlapping ORFs longer than 150 codons spliced into one

sequence (assuming that most of these ORFs are coding). As intergenic (noncoding) sequences we have considered sequences which are outside ORFs longer than 70 codons (Figure 2). It has been shown previously (CEBRAT et al. 1997) that coding sequences show strong asymmetry between coding and noncoding strands. This asymmetry masks the asymmetry introduced by other mechanisms almost totally (Figure 2a). Asymmetry seen in the intergenic sequences (Figure 2b) corresponds to the asymmetry of the whole chromosome (Figure 1). To show the compositional bias of each position in codons we have

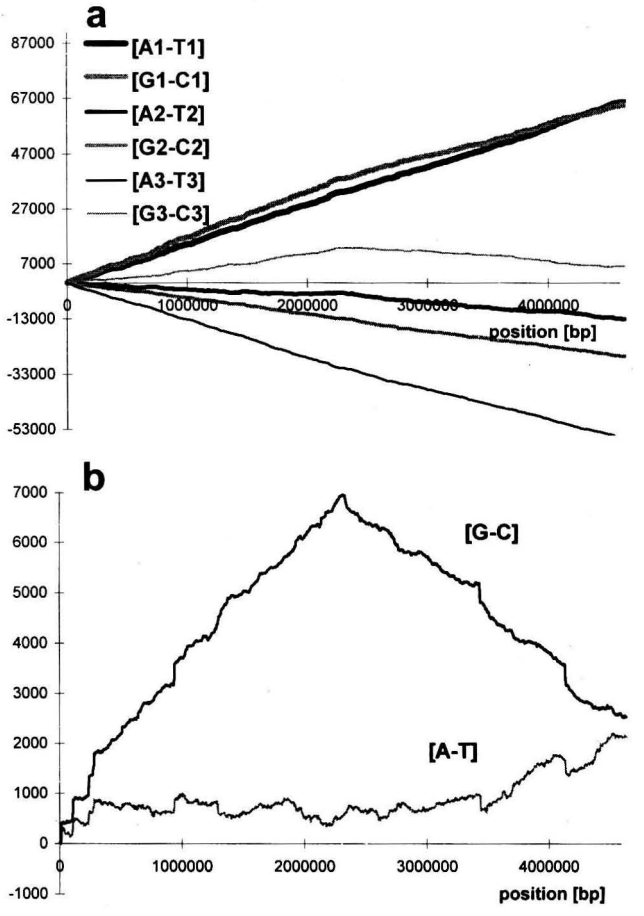

Figure 2. DNA walks on the W strand of *E. coli* sequences: a – ORFs longer than 150 codons; b – intergenic sequences, i.e. sequences outside all ORFs longer than 70 codon. Walks were done as described in Figure 1 but in the case of ORFs walkers analysed each codon position separately. Scale of X-axis corresponds to real co-ordinates of the sequence on the chromosome.

performed three DNA walks on Watson strands of all ORFs, in two-dimensional space (Figure 3). The results suggest that the walks representing the first and the second positions are straight and are not sensitive to position on the chromosome. The trend in the third position changes along the chromosome (note that neither scale represents position on the chromosome and walks are performed on consecutive ORFs lying on one strand). To find trends depending on position on the chromosome we have used detrended walks. We have performed the G↔C

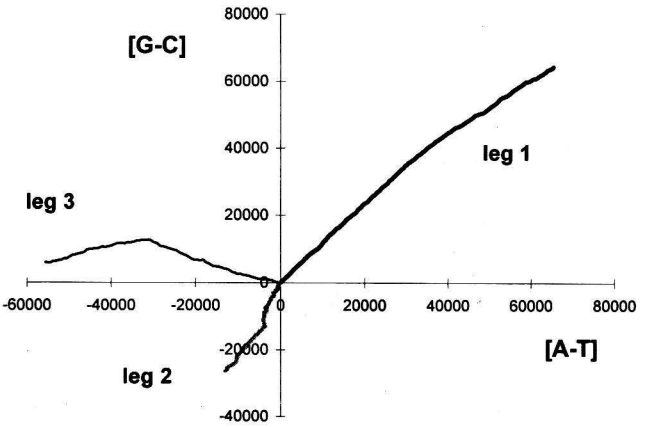

Figure 3. Three two-dimensional DNA walks (a spider) for spliced ORFs longer than 150 codons of the *E. coli* W strand. Walkers analysed first, second and third positions in codons separately (leg *1*, *2* and *3*, respectively). Walkers changed their positions on the plot by (0,1) for G, (1,0) for A, (0,–1) for C and (–1,0) for T. Note that in this kind of walks, neither axis represents the co-ordinates of the analysed sequence.

and A↔T walks for each codon position separately in ORFs on the Watson strand (Figure 4). To compare the walks we have presented all data on the same scale. It can be seen that the asymmetry is different for each codon position. The one for the third codon position resembles the asymmetry for intergenic sequences.

Distinguishing between replication and transcription mutational pressure

Since the mechanisms of replication of leading and lagging strands are different and the DNA walks change their directions at the leading/lagging switches, it is possible to conclude that replication-associated processes are responsible for this asymmetry. If this conclusion is correct, asymmetries for two complementary strands should be of reciprocal sign and compensate each other if added, while subtraction of the values of asymmetry should double the effect (note that the analysed spliced ORFs from Watson strands do not overlap any ORF from Crick

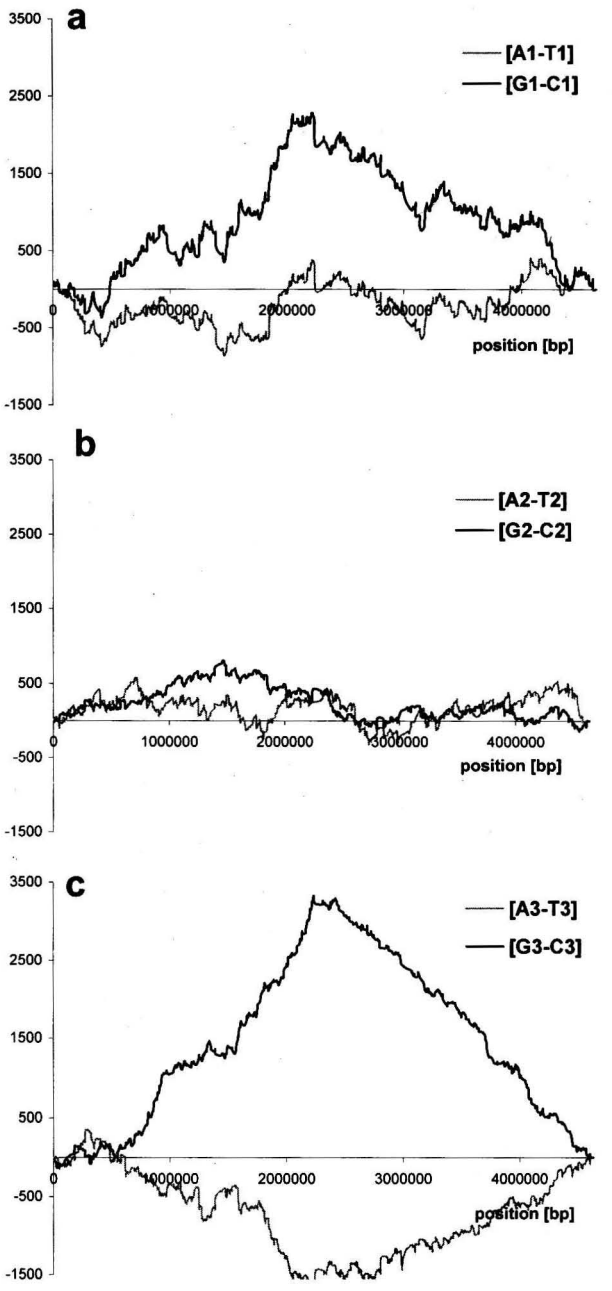

Figure 4. Detrended DNA walks performed for the same sequences as in Figure 2a. In these walks the steps of walkers up were the same as in Figure 2 but the steps down were multiplied by a parameter allowing the walk to finish at the position y = 0. The scale of X-axis corresponds to real co-ordinates of the sequence on the chromosome; a – DNA walks on the first positions in codons; b – the second positions; c – the third positions.

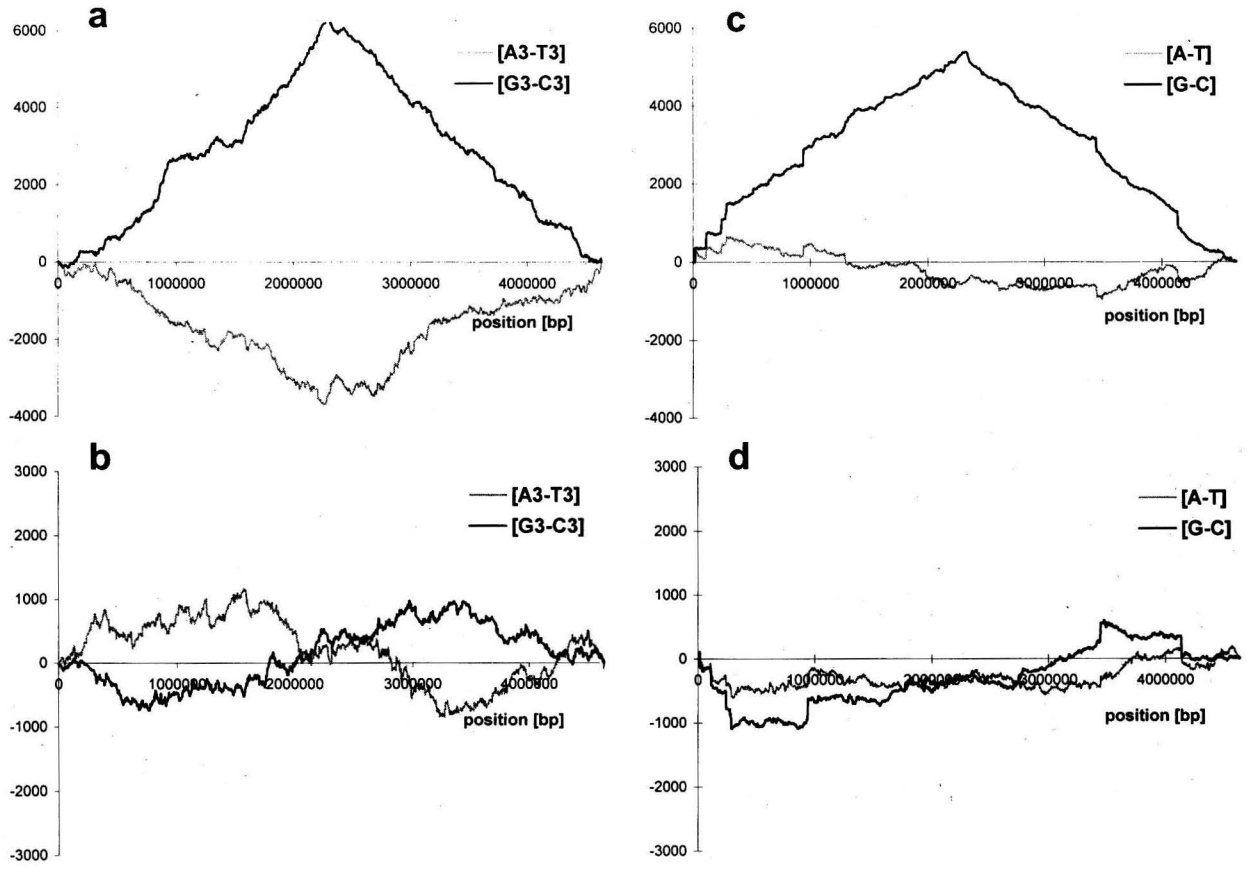

Figure 5. The results of addition (b, d) and subtraction (a, c) of detrended DNA walks for the third codon positions (a, b) and for intergenic sequences (c, d). In the case of addition, the values of walkers' steps have the same sign for both strands, in the case of subtractions, the values of walkers' steps on the C strand were multiplied by (–1) and added to the values of DNA walk on strand W. To perform these transforms for intergenic sequences, the consecutive sequences were numbered and uneven sequences were considered as lying on strand W while even sequences were considered as lying on strand C.

strands). If the asymmetry is introduced by transcription mechanisms (and/or other selection pressures for coding functions), their effect should be the same for sequences of both strands and the results of addition should cumulate the effect of transcription. The results of such transformations of DNA walks for the third positions in codons of ORFs longer than 150 codons are shown in Figure 5. The results indicate that the introduced bias in composition of the third positions is caused mainly by replication-associated processes but the effect of transcription mechanisms (we prefer to say coding constraint) can also be noticed. Figure 5a (the effect of subtraction) represents the bias introduced by replication associated processes which depend on the leading or lagging role of DNA strands. Trends shown in Figure 5b (the effect of addition) are introduced by coding functions and are independent of the leading/lagging role of DNA strands. They rather follow the trends connected with proximal or distal location of genes. To prove that our reasoning is correct we have performed two DNA walks on intergenic sequences. Since there is no sense in performing DNA walks on both strands of the same sequence because they are complementary and have to give exactly mirror pictures, we have spliced every second intergenic sequence and performed DNA walks on this sequence and on the sequence obtained by splicing the rest of intergenic sequences. The transformations of the resulting walks are shown in Figures 5c and d. As it was expected, the result of subtraction doubles the effect of asymmetry and the sum of two walks is close to zero (no transcription effect).

Preferences in nucleotide substitutions

It could be concluded from the above results that both mechanisms introduce specific trends into prokaryotic chromosomes, especially in G/C ratio. The problem of the primary cause of this G/C asymmetry has been discussed previously. A widely accepted explanation is that methylated cytosine is transformed to thymidine by deamination. In such a case the diminishing C should be associated to growing T in the same strand and correlated with [A-T] values. Some authors argue that if the [T-A] value does not follow the [G-C] value it means that C→T transitions are balanced by T→A transversions (FRANCINO et al. 1996, FRANCINO, OCHMAN 1997). However, there may be another explanation – the higher G/C ratio is not caused by the disappearance of C but by the preferences of substitutions by G over any other nucleotide. To check this hypothesis we have performed detrended walks for each of the four bases separately for coding and noncoding sequences. In Figure 6 such walks for strand W are shown. These walks prove that simple C→T transition cannot be the only cause of the observed asymmetry in G/C and A/T ratios. In noncoding sequences (Figure 6d) as well as in the third codon positions (Figure 6c) the accumulation of G in the leading strand is accompanied by the accumulation of T. One of the explanation could be that it is the result of T→A transversions as in the hypothesis of FRANCINO et al. (1996), and FRANCINO and OCHMAN (1997).

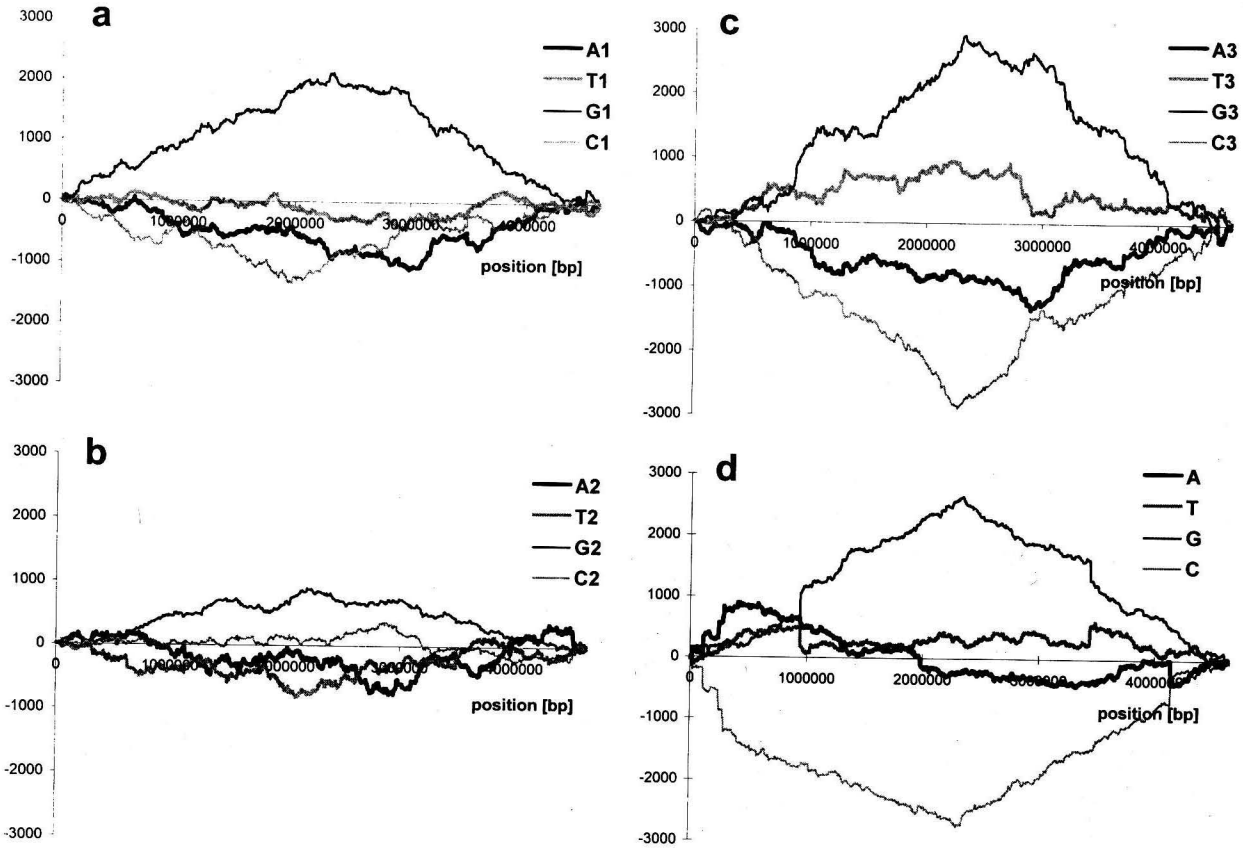

Figure 6. The results of subtractions of detrended DNA walks for different codon positions of ORFs and for noncoding sequences of the *E. coli* genome. When the walker analysed the frequency of a specific nucleotide in the sequence, it moved one unit up when it met this very nucleotide, otherwise it stepped down and the size of step down was of such a size that the whole walk ended at y = 0. The DNA walks for strand C were then subtracted from the respective walks done for strand W; a – DNA walks for the first positions in codons ORFs, b – DNA walks for the second positions in codons ORFs, c – DNA walks for the third positions in codons ORFs, d – DNA walks for intergenic sequences.

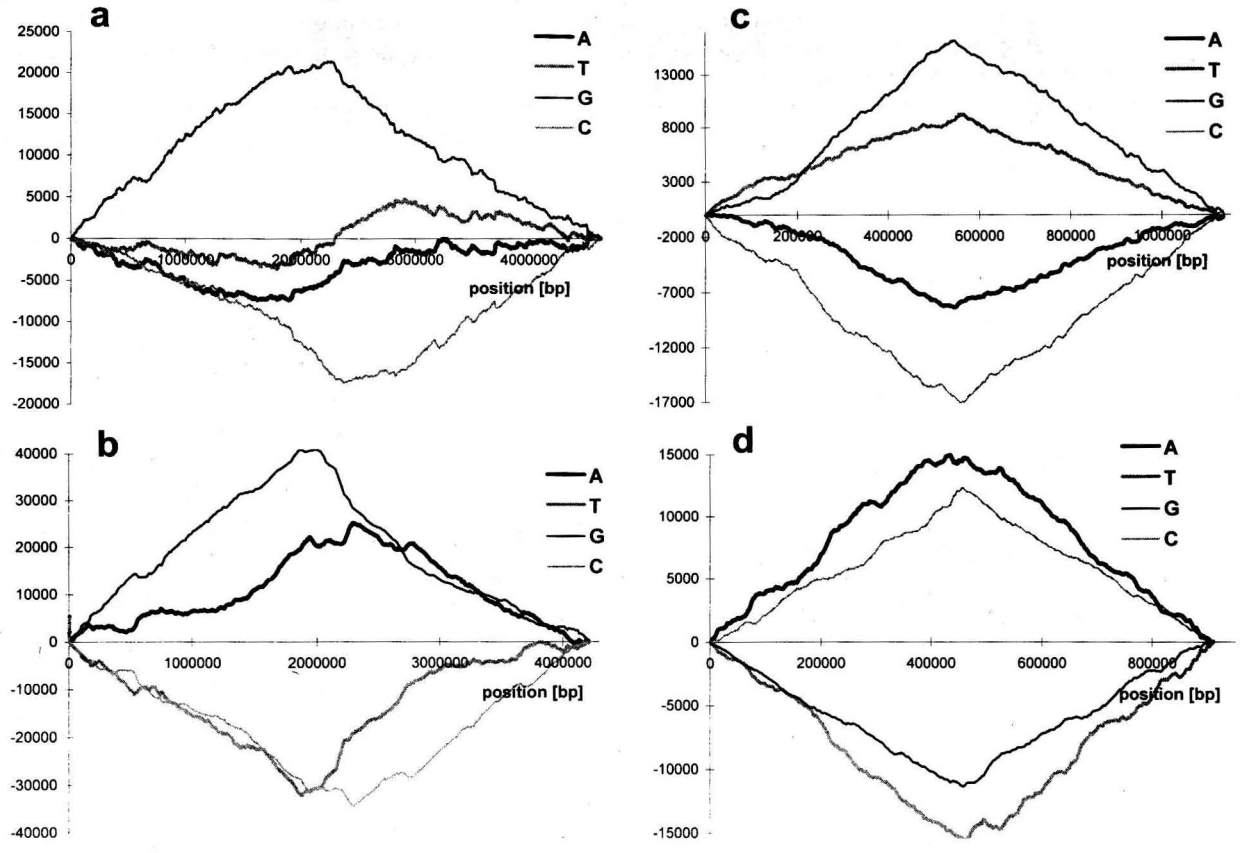

Figure 7. DNA walks on strand W of the following genomic sequences: a – *Escherichia coli,* b – *Bacillus subtilis*, c – *Treponema pallidum* and d – *Borellia burgdorferii.* When the walker analysed the frequency of a specific nucleotide in the sequence, it moved one unit up when it met this very nucleotide, otherwise it stepped down and the size of step down was of such a size that the whole walk ended at y = 0.

One can argue that the observed differences could be the results of different coding capacities of leading and lagging strands. An analysis of the method shows that the addition of detrended DNA walks done on the scale of the chromosome eliminates the influence of coding density on the results. This transformation of DNA walks eliminates the replication-associated mutational pressure, leaving asymmetries caused by other mechanisms. We have proved it in simulations of mutation in chromosomes (data not presented).

We have found asymmetry in many other prokaryotic genomes whose sequences have already been identified. The pictures of these asymmetries vary. In different genomes, different preferences are observed in nucleotide substitutions by replication mechanisms (Figure 7).

Discussion

DNA walks have shown that there are two different causes of strand asymmetry in prokaryotic chromosomes. One is the influence of replication-associated processes which are different for leading and lagging DNA strands. The other one is the coding role of genes. Protein coding sequences locally introduce very strong asymmetry which is specific for each codon position. This asymmetry is compensated locally by specific location of genes on both strands of chromosome which has been shown elsewhere (CEBRAT, DUDEK 1996,1998). Since trends introduced by coding sequences are similar for both strands, the resulting asymmetry can be eliminated by subtracting coding trends. In the *E. coli* genome the asymmetry specific to the first two codon positions depends on position on the chromosome to a small extent, nevertheless the first position accumulates some surplus of G in the leading strand, which seems to be the effect of replication processes. Asymmetry introduced by the third position depends on the position on chromosome and follows the trend introduced into intergenic sequences by replication. Furthermore, in the third codon position in *E. coli* or even in the second codon position in *Bacillus subtilis* it is possible to distinguish between the asymmetry introduced by replication processes and coding functions. Some authors claim that transcription-coupled repair mechanisms introduce the asymmetry into DNA strands and remove the most frequent types of DNA damage, which results in abundance of pyrimidines in the transcribed strand (FREEMAN et al. 1998). If this is true, the effect of this mechanism should be the most pronounced in the third codon positions which are the most prone to accumulate mutations. Thus, the third positions should become the most purine rich. In fact this position is the least purine rich.

Such a strong asymmetry in nucleotide composition of prokaryotic chromosomes raises another problem – what is the primary force establishing and maintaining the asymmetry in coding sequences – mutational pressure changing the gene sequences, or selection and recombination which shifts the genes to chromosome regions where they fit the structural requirements of DNA better. Genes

translocated with invertion into another location on the same replichore should be unstable, prone to accumulate mutations at a high rate. This could explain the relative conservatism in the chromosomal map structure of eubacterial genomes (WILKINS 1988).

Conclusions

Nucleotide composition of complementary strands of eubacterial chromosomes is highly asymmetric. The asymmetry is introduced by replication-associated mechanisms, as well as by transcription and/or by coding functions of sequences. The asymmetry introduced by replication switches its polarity at the origin and at the terminus of replication, which is observed in both noncoding and coding sequences and varies with respect to positions in codons. Coding functions introduce very strong trends into protein coding ORFs, which are specific to each nucleotide position in the codon. Detrended DNA walks show transcription-associated compositional bias for genes proximal and distal to the origin of replication.

Acknowledgements. This work was supported by the State Committee for Scientific Research, Grant No. 6PO4A 030 14.

REFERENCES

BELETSKII A., BHAGWAT A.S. (1996). Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 93: 13919-13924.

BERTHELSEN Ch.L., GLAZIER J.A., SKOLNICK M.H. (1992). Global fractal dimension of human DNA sequences treated as pseudorandom walks. Phys. Rev. A45: 8902-8913.

BLATTNER F.R., PLUNKETT III G., BLOCH C.A., PERNA N.T., BURLAND V., RILEY M., COLLADO-VIDES J., GLASNER J.D., RODE Ch.K., MAYHEW G.F., GREGOR J., DAVIS N.W., KIRKPATRICK H.A., GOEDEN M.A., ROSE D.J., MAU B., SHAO Y. (1997). The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1462.

CEBRAT S., DUDEK M.R. (1996). Symmetry in chromosome fractal organization and DNA domain structure. Proceedings of the 8th Joint EPS-APS Int. Conference on Physics Computing '96 (Borchards P., Bubak M., Maksymowicz A., eds.). Academic Computer Center, CYFRONET-KRAKÓW: 371 -374.

CEBRAT S., DUDEK M.R., MACKIEWICZ P., KOWALCZUK M., FITA M. (1997). Asymmetry of coding versus non-coding strands in coding sequences of different genomes. Microbial & Comparative Genomics 2(4): 259 - 268.

CEBRAT S., DUDEK M.R. (1998). The effect of DNA phase structure on DNA walks. Euro. Phys. J. B 3: 271-276.

FIJALKOWSKA I.J., JONCZYK P., MALISZEWSKA-TKACZYK M., BIALOSKORSKA M., SCHAAPER R.M. (1998). Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. Proc. Natl. Acad. Sci. USA 95: 10020-10025.

FRANCINO M.P., CHAO L., RILEY M.A., OCHMAN H. (1996). Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272: 107-109.

FRANCINO M.P., OCHMAN H. (1997). Strand asymmetries in DNA evolution. Trends Genet. 13(6): 240-245.

FREEMAN J.M., PLASTERER T.N., SMITH T.F., MOHR S.C. (1998). Patterns of genome organization in bacteria. Science 279: 1827.

GRIGORIEV A. (1998). Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26(10): 2286-2290.

IKEMURA T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J. Mol. Biol. 158: 573-597.

KORNBERG A., BAKER T.A. (1992). DNA Replication. Freeman, New York.

KUNKEL T.A. (1992). Biological asymmetries and the fidelity of eukaryotic DNA replication. Bioessays 14: 303-308.

LOBRY J.R. (1996a). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13(5): 660-665.

LOBRY J.R. (1996b). Origin of replication of *Mycoplasma genitalium*. Science 272: 745-746.

MRÁZEK J., KARLIN S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA 95: 3720-3725.

OKAZAKI R., OKAZAKI T., SAKABE K., SUGIMOTO K., SUGINO A. (1968). Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. Proc. Natl. Acad. Sci. U.S.A. 59(2): 598-605.

SHARP P.M., STENICO M., PEDE J.F., LLOYD A.T. (1993). Codon usage: mutational bias, translational selection or both? Biochem. Soc. Trans. 21: 835-841.

WAGA S., STILLMAN B. (1994). Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication *in vitro*. Nature 369: 207-212.

WANG T.C., CHEN S.H. (1992). Similar-sized daughter-strand gaps are produced in the leading and lagging strands of DNA in UV-irradiated *E. coli* uvrA cells. Biochem. Biophys. Res. Commun. 184(3): 1496-1503.

WANG T.C., CHEN S.H. (1994). DNA fragments contain equal amounts of lagging-strand and leading-strand sequences. Biochem. Biophys. Res. Commun. 198(3): 844-849.

WILKINS B. M. 1988. Organization and plasticity of enterobacterial genomes. Soc. Appl. Bacteriol. Symp. Ser.17: 51-69.