

DNA Asymmetry and the Replicational Mutational Pressure

Maria Kowalczyk¹, Paweł Mackiewicz¹, Dorota Mackiewicz¹, Aleksandra Nowicka¹, Małgorzata Dudkiewicz¹, Mirosław R. Dudek², Stanisław Cebrat^{1#}

¹Institute of Microbiology, Wrocław University, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland,

tel.:+48-71-3247-303

fax:+48-71-3252-151

e-mail: cebrat@microb.uni.wroc.pl

<http://smORFland.microb.uni.wroc.pl>

²Institute of Physics, Zielona Góra Pedagogical University

To whom all correspondence should be addressed.

Short running title: Mutational pressure

Abstract

The mode of replication and organisation of bacterial genomes impose asymmetry on their nucleotide composition. The asymmetry is seen in coding and non-coding sequences and is reflected in the amino acid composition of proteins. The mechanisms generating asymmetry include: unequal mutation rates connected with replication and transcription, selection forces positioning genes and signal sequences nonrandomly in the genome, and protein coding constraints on coding sequences. There are different methods of visualising and measuring the asymmetry. Some of them can assess the contribution of individual mechanisms to the observed asymmetry and those have been described in greater detail. Asymmetric mutational and selection pressures differentiate the rates of evolution of genes on leading and lagging strands. The genes relocated to the opposite strand have to adapt to a different mutational pressure or are eliminated. Translocations from leading to lagging strands are more often selected against than from lagging to leading strands. Comparison of intergenic sequences that have lost the coding function to the original genes enables finding the frequencies of the twelve substitution rates in sequences free from selection. In the absence of selection, the half-time of substitution of a given type of nucleotide is linearly correlated with the fraction of that nucleotide in the sequence.

Key words

DNA asymmetry, mutational pressure, selection pressure, DNA walk, GC skew, AT skew, rate of evolution.

Glossary

antisense strand = coding strand = template strand = transcribed strand of a gene

sense strand = noncoding strand = non-template strand = nontranscribed strand of a gene

AT skew = $(A-T)/(A+T)$

GC skew = $(G-C)/(G+C)$

CAI = Codon Adaptation Index, i.e. codon usage for optimal translation rate

CDS = coding sequence

non-CDS = noncoding sequence

lagging DNA strand = strand synthesised from Okazaki fragments

leading DNA strand = strand replicated continuously

ORF = open reading frame

PR2 = parity rules type 2, i.e. $[A]=[T]$ and $[G]=[C]$ are true for a single DNA strand

Definition of DNA asymmetry

The specific structure of the double-stranded DNA molecule implies many of its genetic and chemical features. One of the most important features is the complementarity of the two DNA strands, where the number of adenines is exactly the same as the number of thymines, while the number of guanines is exactly the same as the number of cytosines. These are the Chargaff's rules (CHARGAFF 1950) which helped WATSON and CRICK (1953) to describe the structure of the double helix and to find out that the Chargaff's rules are deterministic. If we assume that there are no mutational or selection pressures which influence the composition of the two DNA strands, the rules: $[A]=[T]$ and $[G]=[C]$ should be in force not only for double-stranded DNA but also for each of the two strands. These rules for a single DNA strand are stochastic instead of deterministic, and are called parity rules type 2, in short PR2 (LOBRY 1995). Deviation from PR2 means that the two DNA strands are under different mutational or selection pressures or both, which leads to asymmetric substitution patterns and DNA asymmetry.

Finding DNA asymmetry

PR2 are valid for random DNA molecules as well as for whole chromosomes. Let us count the number of each of the four types of nucleotides in the Watson strand of the *Borrelia burgdorferi* genome and construct a single-stranded "leading" and "lagging" DNA sequences by drawing the nucleotides randomly from the pool in which the frequency of each nucleotide is the same as in the *B. burgdorferi* genome. In Table 1 the composition of such a random DNA sequence (an example of one

computer simulation) is shown. There are no significant differences in the nucleotide composition of these two sequences. It is not the case when one looks at the halves of the real Watson strand replicated as leading and lagging (see Table 1). The Chargaff's parity rules are valid but PR2 are not: the differences between the numbers of complementary nucleotides in each strand are significant.

Deviations from PR2 and differences in composition between the leading and the lagging strands were observed in many eubacterial genomes (LOBRY 1996a, 1996b, BLATTNER et al. 1997, KUNST et al. 1997, FRASER et al. 1997, 1998, ANDERSON et al. 1998, FREEMAN et al. 1998, MRAZEK, KARLIN 1998, GRIGORIEV 1998, MCLEAN et al. 1998, SALZBERG et al. 1998, MACKIEWICZ et al. 1999a, 1999b, TILLIER, COLLINS 2000a) and are still detected in newly sequenced genomes. The asymmetry was also detected in many viruses (DANIELS et al. 1983, FILIPSKI 1990, MRAZEK, KARLIN 1998, GRIGORIEV 1999). Generally it was found that the leading strand is rich in guanine and thymine, and the lagging strand, in cytosine and adenine.

The asymmetry is observed even at the level of codons and amino acids (PERRIERE et al. 1996, MCINERNEY 1998, LAFAY et al. 1999, MACKIEWICZ et al. 1999b, ROCHA et al. 1999a, ROMERO et al. 2000). This kind of asymmetry generally does not exist in archaeal genomes (MRAZEK, KARLIN 1998, GRIGORIEV 1998, MCLEAN et al. 1998) and was found only in some of them (SALZBERG et al. 1998, LOPEZ et al. 1999, 2000). Analyses of eukaryotic genomes do not show the asymmetry on a large scale (MRAZEK, KARLIN 1998, GRIGORIEV 1998, GIERLIK et al. 2000), although some specific asymmetry in subtelomeric regions of yeast chromosomes was observed (GIERLIK et al. 2000, see also GRIGORIEV 1998).

The asymmetry is so strong that it can come to assistance in experimental searches for the origin and terminus of replication (QIN et al. 1999, PICARDEAU et al. 1999, 2000, ZAWILAK et al. 2001).

Mechanisms generating asymmetry

A substitution in one DNA strand is reflected by a change in the complementary strand. However, to understand the origin and meaning of asymmetry, it is important to find where, and how the primary changes occur which lead to different substitutions in different regions of the genome.

Mechanisms that could introduce asymmetry into DNA strands have been discussed many times (see for review: FRANCINO, OCHMAN 1997, MRAZEK, KARLIN 1998, FRANK, LOBRY 1999, KARLIN 1999, TILLIER, COLLINS 2000a). Nucleotide composition of a sequence is shaped by two different and sometimes opposite forces: mutational and selection pressures. Generally, they include mutational pressures on DNA during replication and transcription, selection forces positioning genes and signal sequences nonrandomly in the genome, and protein coding constraints on coding sequences. In this paper we would like to present and shortly comment on the most important causes of asymmetry accepted to date.

Replication-associated mutational pressure

An important structural feature of the DNA molecule is that the two strands are antiparallel. Together with the properties of replication mechanisms, it has very significant genetic implications. DNA strands can be synthesised only in one direction: from the 5' to 3' end. Because the strands are antiparallel and replication forks move along the maternal double strand molecule, the two new strands have to be synthesised by different mechanisms and different replication-associated mutational pressures may influence their nucleotide composition. After many generations, parity rules type 2 should not be in force in such DNA molecules (see Table 1, the composition of the replichores of the *B. burgdorferi* genome).

Synthesis of one strand, called the leading strand, is continuous, while synthesis of the other strand, called the lagging strand, has short intermediates named Okazaki fragments (OKAZAKI et al. 1968). Differences between the synthesis of the strands have been reviewed by FRANK and LOBRY (1999). In *Escherichia coli* both DNA strands are synthesised by symmetric core enzymes (Pol III holoenzyme, BAKER, WICKNER 1992, MARIANS 1992, YUZHAKOV et al. 1996), and therefore base incorporation and proof-reading should be the same for both strands. However, the enzyme complexes differ in processivity (tendency to remain on a single template, MARIANS 1992). The leading strand complex needs to be more processive to remain on the template, while the lagging strand complex needs to dissociate more often, which facilitates excision of a mismatch by some cellular exonuclease (FIJALKOWSKA, SCHAAPER 1996). Thus, the lagging strand synthesis should be more faithful. Also, the strands may differ in stepwise progression speed and mismatch repair mechanisms (RADMAN 1998). The lagging strand polymerase should synthesise DNA faster to compensate for the time of its recycling, so more errors may be committed in the process, but on the other hand the discontinuous replication provides nicks in DNA, which are required by mismatch repair, so the lagging strand repair could be more efficient. Experimental analyses of the relative fidelity of the leading and lagging strand replication have given contradictory results (e.g. compare IWAKI et al. 1996 and FIJALKOWSKA et al. 1998). Generally, in experiments lagging strands seem more prone to mutations (e.g. TRINH, SINDEN 1991, BASIC-ZANINOVIC et al. 1992, VEAUTE, FUCHS 1993; ROBERTS et al. 1994, THOMAS et al. 1996). However, these results should be carefully considered, because the experiments were performed in specific conditions, e.g. the strains used in the studies were deficient in proof-reading or mismatch repair.

A theory (named the cytosine deamination theory) that explains the influence of replication-associated mutational pressure on asymmetry was presented by FRANK and LOBRY (1999). During replication, stretches of the template for the newly synthesised lagging strand are temporarily single-stranded. In this state the template is more exposed to damage and mutations (similarly to the sense strand during transcription). The most frequent mutation is deamination of cytosine and its homologue 5-methylcytosine to uracil (ECHOLS, GOODMAN 1991, LINDAHL 1993, KREUTZER, ESSIGMANN 1998),

which leads in consequence to C->T transition. It was found that cytosine deaminates 140 times faster in single-stranded DNA than in double-stranded (FREDERICO et al. 1990). This transition explains the excess of guanine and thymine in the leading strand, and adenine and cytosine in the lagging strand; similarly for the less common A->G transition resulting from deamination of adenine to hypoxanthine (LINDAHL 1993). The deamination theory gives an especially convincing explanation of asymmetry in mitochondrial (TANAKA, OZAWA 1994, REYES et al. 1998) and viral genomes (GRIGORIEV 1998, 1999).

Transcription-associated mutational pressure

The genetic information stored in DNA can also be read only in one direction: from the 5' to 3' end. However, there are six possible reading frames and the transcribed strand of a gene may be located in the same or opposite direction to the replication fork movement (on the leading or lagging strand). Because genes are not distributed uniformly in the genome, and coding and non-coding strands are treated differently by transcription, PR2 may be violated.

A potential cause of asymmetry may also be deamination of methylated cytosines which leads to thymines. Some authors have claimed that this type of substitution differentiates sense and antisense strands of coding sequences, and that transcription mechanisms introduce the asymmetry into DNA strands (FRANCINO et al. 1996, FRANCINO, OCHMAN 1997, FREEMAN et al. 1998). During transcription a part of the nontranscribed DNA strand is exposed and more prone to deamination (BELETSKII, BHAGWAT 1996), while the other strand is protected by the enzymatic transcription complex and by transcription-coupled repair that preferentially repairs pyrimidine dimers (MELLON, HANAWALT 1989, HANAWALT 1991). Some experiments have proved that the frequency of mutations introduced into the non-transcribed DNA strand is higher than into the transcribed one (FRANCINO et al. 1996).

Unequal distribution of genes and oligomers on the chromosome

Transcription-associated mutational pressure alone does not distinguish between the leading and lagging strand. However, if highly transcribed genes are preferably located on one strand, the bias between strands should be generated in coding sequences and in the intergenic regions that are partly transcribed.

There are preferences for transcribing DNA strands in the direction of replication, rather than in the inverse direction, possibly to avoid collisions between replication and transcription complexes (BREWER 1988). In *Mycoplasma genitalium*, *M. pneumoniae* and *Bacillus subtilis* over 75% of genes are located on the leading strands (FRASER et al. 1995, HIMMELREICH et al. 1996, KUNST et al. 1997). Hence, transcription-associated mutational pressure may contribute to the leading/lagging strand asymmetry. However, in the *E. coli* genome the bias is relatively low and only 54% of genes are located on the leading strand (BLATTNER et al. 1997), so it cannot account for the asymmetry observed in that genome.

It has also been found that in prokaryotic genomes usually the majority of highly expressed genes are located on the leading strand. Those genes use a small subset of specific codons (GOUY, GAUTIER 1982, SHARP, LI 1987), which may contribute to the asymmetry observed. TILLIER and COLLINS (2000a) tried to estimate the contribution of transcription to asymmetry by analysing Codon Adaptation Index values of genes. They found that genes with the highest and lowest CAI did not account for the correlation of base composition skew with replication orientation, and these skews were not completely explained by the selection for highly expressed genes on the leading strand. However, these analyses assumed that CAI value (which actually measures translation level/intensity) corresponded to transcription level/intensity, which may not be true. Only experimental research can provide evidence that genes on the leading strand are more intensively transcribed than genes on the lagging strand.

Leading- and lagging-strand-specific codon usage has been observed in *Borrelia burgdorferi*, *Treponema pallidum* and *Chlamydia trachomatis* (MCINERNEY 1998; LAFAY et al. 1999, ROCHA et al. 1999a, ROMERO et al. 2000). However, because the presumably highly expressed genes in those genomes do not differ in codon usage from other genes located on the same strand, some authors conclude that the two different patterns are the result of replication-associated mutational pressure and not selection, and codon usage is strand-specific and not correlated with the level of expression (MCINERNEY 1998; LAFAY et al. 1999).

Oligomers that are over-represented on one of the strands could contribute to DNA asymmetry. For example, Chi sequence 5' GCTGGTGG 3', which is a recombinational hot-spot, is located preferentially on the leading strand of the *E. coli* genome (BLATTNER et al. 1997). SALZBERG et al. (1998) observed skewed distribution of some oligomers on leading and lagging strands. In *Borrelia burgdorferi*, *Treponema pallidum*, *E. coli*, *Bacillus subtilis* and other genomes 7-, 8-, and 9-mers are statistically significantly skewed and are helpful in finding the origin and terminus of replication. The nucleotide composition of these oligomers is correlated with the most abundant codons in those genomes, although they do not occur preferentially within coding regions. Their asymmetry is much stronger than the inequality in the distribution of coding sequences. Their function is unknown but they are expected to play a role as biological signals in replication and transcription (ROCHA et al. 1998), so their distribution should be subject to selection. However, Chi-sites make up only 0.25 % of the *E. coli* genome and are not likely to be an important source for global base composition asymmetry (FRANK, LOBRY 1999), which is further supported by analyses of several genomes by TILLIER and COLLINS (2000a). Removal of all skewed octamer sequences from the *E. coli* and *Haemophilus influenzae* genomes gave reduced asymmetry but did not eliminate it. Apparently skewed oligomers are not the main source of asymmetry.

Protein coding constraints on coding sequences

Coding for proteins requires a specific nucleotide composition. It has been long known that coding strands of genes are rich in purines (e.g. SHEPHERD 1981, SMITHIES et al. 1981, KARLIN, BURGE 1995, FRANCINO et al. 1996, CEBRAT et al. 1997a, FREEMAN et al. 1998). DNA sequences which code for proteins have a triplet structure. Each position in the codon has specific preferences in nucleotide composition (WONG, CEDERGREN 1986, ZHANG, ZHANG 1991, GUTIERREZ et al. 1996, MRAZEK, KARLIN 1998, CEBRAT et al. 1997b, 1998, MCLEAN et al. 1998, WANG 1998), which suggests that it plays a unique role and remains under a specific selection pressure. Generally, the first codon positions of protein coding sequences are rich in adenine and guanine and the second are rich in adenine and cytosine. The asymmetry between coding and non-coding strands of genes is so strong that it can be used to successfully discriminate between coding and non-coding sequences (CEBRAT et al. 1997a, 1997b, 1998).

There are many mechanisms which contribute to the specific composition of genes. The common presence of purines in the sense strand is favoured by evolution because they are less prone to mutations than pyrimidines (especially dimers) (HUTCHINSON 1996). During transcription the sense strand is more exposed than the antisense strand which is preferably repaired by removal of pyrimidine dimers (MELLON, HANAWALT 1989, HANAWALT 1991). Therefore selection should increase the purine content of the coding strand (FREEMAN et al. 1998, FRANK, LOBRY 1999).

Furthermore, the base composition of the first and second positions in codons reflects the high usage of acidic amino acids coded by GAN (KARLIN, MRAZEK 1996) and Gly, Ala and Val (KARLIN et al. 1992). The second codon position determines the polarity of the encoded amino acid and its change may have a detrimental effect on the protein.

Periodical codon composition pattern $(GCU)_n$ plays a role in mRNA-rRNA interaction during translation in the ribosome (TRIFONOV 1987, 1992, LAGUNEZ-OTERO, TRIFONOV, 1992). Guanines in the first codon positions interact with periodically distributed cytosines in rRNA and ensure the correct reading frame during translation.

Third codon positions are degenerated and most substitutions in them are silent. These substitutions are not necessarily neutral. They may change the rate of translation of the product (IKEMURA 1981, BENNETZEN, HALL 1982, SHARP, COWE 1991). However, selection on third positions in codons is the weakest and the effect of mutational pressure should be observed in them.

If coding sequences are in the same number on both leading and lagging strands, their compositional bias should be cancelled out. Otherwise, if they are not randomly distributed on the chromosome, they can contribute to the global asymmetry (asymmetry of the whole chromosome).

Relative contribution of different factors to DNA asymmetry

Although there are many different and sometimes contradictory hypotheses and opinions about the influence of asymmetry, it is possible to draw some conclusions. The impact of uneven gene

distribution on global asymmetry is different in various genomes. TILLIER, COLLINS (2000a) have assessed the relative contribution of gene orientation in many genomes to base composition asymmetry. In some genomes the influence of gene bias is opposite to that resulting from mutational pressure (MCLEAN et al. 1998, TILLIER, COLLINS 2000a). Highly expressed genes and signal sequences contribute to the bias only to a very small extent. The replication-associated mutational pressure is the most significant factor of the observed asymmetry. Some authors (CEBRAT et al. 1999, MACKIEWICZ et al. 1999a, 1999b, 1999c, TILLIER, COLLINS 2000a) have filtered by different methods the influence of replication from other mechanisms. Although the degree of influence of transcription-associated mutational pressure still remains open, it seems that it is weaker than the influence of replication (FRANK, LOBRY 1999). The time of single-stranded state is shorter for the coding strand than for the lagging strand template. Deaminations in the sense strand cause only premutagenic lesions that have to wait for the next round of replication to become fixed and during this time can be repaired. The uracil resulting from deaminations of cytosine occurring in the lagging strand template is almost immediately paired with an incoming adenine in the synthesis of the lagging strand.

Furthermore, the influence of transcription on asymmetry may be consistent with the influence of replication-associated mutational pressure, because deamination of cytosine occurs both in the lagging strand template during replication, and during transcription in the coding strands, which are preferably located on the leading strand.

Methods of analysing DNA asymmetry

Deviations from PR2 are usually analysed in terms of an excess of the number of guanines relative to cytosines or adenines relative to thymines. The bias is measured by GC and AT skews, $(G-C)/(G+C)$ and $(A-T)/(A+T)$, respectively. The method of analysing GC and AT skews with a sliding window (LOBRY 1996a) is helpful to detect replication origin in some prokaryotic genomes but the results are often difficult to interpret. If a small-sized window is used, strong fluctuations obscure the asymmetry (MRAZEK, KARLIN 1998), if the window is large, the trends in nucleotide composition are diminished (MCLEAN et al. 1998). Cumulative skew diagrams or plots of numerically integrated skew of GRIGORIEV (1998) and TILLIER and COLLINS (2000a) eliminate fluctuations and give a much clearer picture. FREEMAN et al. (1998) performed cumulative diagrams of purine (A+T versus T+C), keto (G+T versus A+C) and coding strand excess.

ROCHA et al. (1999a,b) applied a statistical linear discriminant function to assess strand asymmetry at the level of nucleotides, codons and amino acids.

A method which differentiates the influence of replication processes and transcriptional/translational forces on genomic sequences has been proposed by TILLIER and COLLINS (2000a). Their ANOVA analyses quantify and measure statistical significance of individual effects of replication and gene direction on GC and AT skews. The skews were measured in each codon position of CDS and in non-CDS separately. They found that the effect of replication orientation is independent

of the effects of transcriptional or translational processes and in fact can be of the opposite sign. They also found that AT and GC skews in non-CDS are similar in size and sign to the skews seen with replication orientation at the third positions in codons.

DNA walks

A lot of information can be obtained by analysing sequences with so-called DNA walks (CEBRAT, DUDEK, 1998). In one version of a DNA walk, which is very similar to cumulative diagrams, the walker moves in two-dimensional space (A - T, or G - C) depending on the type of nucleotides analysed (MACKIEWICZ et al. 1999a). The shifts are: (1,1) for G and for A, and (1,-1) for C and for T. Thus, the plot shows for each position on the chromosome (X-axis co-ordinates) the relative abundance of A over T or G over C (Y-axis co-ordinates). The result of the DNA walk for the whole *Treponema pallidum* genome is shown in Fig. 1a. Extrema in the plots represent the positions of the origin (minimum) and terminus (maximum) of replication, where the role of DNA strands changes from the leading to lagging or vice versa. The walk has been normalised by the length of the analysed sequence to compare asymmetry in different sequences and genomes. The walk on intergenic sequences of *T. pallidum* (Fig. 1b) looks similar, contrary to walks on the three codon positions of CDS (Fig. 2a-c). The latter sequences show very strong trends connected with coding for proteins which mask other trends. These walks can be corrected by a factor which allows the walker to finish the walk at y=0. In Fig. 2d-f such detrended walks are shown, where the distinct asymmetry differentiating leading and lagging sequences is revealed. It is the strongest in third codon positions, which are under the weakest selection pressure. Numbers on the Y-axis indicate local relative cumulative abundance of A over T and G over C in the analysed sequence independently of the whole base composition of analysed sequence and its length. The same method can be used for analysing the chromosome structure in respect to many other parameters i.e. nucleotide, codon and amino acid composition of proteins, G/C relations or purines to pyrimidines (MACKIEWICZ et al. 1999b). Examples of DNA walks of this type can be found at our web site: <http://smORFland.microb.uni.wroc.pl>.

Subtraction and Addition of DNA walks

Transformations of detrended DNA walks enable distinguishing between the effects which treat the two complementary strands unequally (subtraction of DNA walks) and the effects which have the same influence on both strands (addition of DNA walks) (CEBRAT et al. 1999, MACKIEWICZ et al. 1999a,b,c). To subtract walks, separate analyses of the two DNA strands are performed: the Watson (W) strand, the one whose sequence is in GenBank, and the Crick (C) strand, complementary to W. The asymmetry introduced by replication-associated mechanisms into the leading and lagging DNA strands is of reciprocal sign. To enhance the picture of replication-associated asymmetry, one needs to subtract the values of walks performed on Watson strands from walks on non-complementary sequences from Crick strands. The asymmetry is seen not only on the scale of the whole chromosome,

but also in each codon position in open reading frames (ORFs, Fig. 3a-c). It is very strong in intergenic sequences, which clearly shows that it is a result of replication and not transcription-related processes. It is the weakest in second codon positions because they are under the strongest selection pressure. Interestingly, in some genomes, e.g. *B. burgdorferi*, it is greater in third codon positions than in intergenic sequences (MACKIEWICZ et al. 1999c). A possible explanation is that intergenic sequences change their orientation more often than genes, since they are under no selective constraints. Genes remain on the leading or lagging strand longer, so asymmetric mutations have time to accumulate.

When detrended DNA walks on ORFs situated on the W strand done on the scale of chromosome are added to DNA walks performed on ORFs from the C strand, the reciprocal values of replication-associated asymmetry compensate each other and disappear, leaving the effect of asymmetry introduced by other mechanisms. The addition of DNA walks on ORFs from both DNA strands leaves also the asymmetries which are the result of the same, unbalanced composition of linked genes from both complementary DNA strands. The asymmetry not associated with replication is seen most clearly in the *B. subtilis* genome (Fig. 3d-e). Some nucleotides in the second and especially in the third codon positions are more common on both strands in proximal parts of the chromosome (near origin of replication), and others in distal parts (near the terminus of replication). The trends in the second codon position concern adenine and thymine distribution and are connected with preferential location of genes coding for hydrophobic or hydrophilic proteins (MACKIEWICZ et al. 1999b). The strong trends in the third codon positions may reflect the effect of mutational pressure connected with variation of nucleotide precursor pools during replication and/or differences in DNA repair modes in those regions (DESCHAVANNE, FILIPSKI 1995).

Spiders

In another version of DNA walk, the walker analyses all four nucleotides during one walk. The method was used in different variants by MIZRAJI, NINIO 1985, GATES 1986, BERTHELSEN et al. 1992, LOBRY 1996b, and CEBRAT et al. 1997b, 1998. The shifts are: (0,1) for G, (1,0) for A, (0,-1) for C and (-1,0) for T. The walks can be used to picture each codon position in a gene separately (Fig. 4a). The whole graph is called a spider and the walks are called spider legs. The nucleotide preferences in each position in the codon are very apparent, because each leg goes in a different direction. Intergenic sequences (Fig. 4b) have no trends. The parameters of the walks, like their angles and lengths of vectors, have been successfully used to discriminate between protein coding and non-coding sequences and to estimate the total number of protein coding genes in genomes (CEBRAT et al. 1997b, CEBRAT et al. 1998; KOWALCZUK et al. 1999a). The same kind of walk can be done for spliced genes from a genome. The walks are normalised to compare asymmetry in sequences of different length. Fig. 5a-b shows these walks performed on leading and lagging strand genes from the *B. burgdorferi* genome. The walks on first and second positions in codons are similar for both strands, but third codon positions show opposite trends. Selection pressure on third positions in codons is the weakest and they

show asymmetry introduced by replication-associated mutational pressure, which is of opposite sign on leading and lagging strands (MACKIEWICZ et al. 1999c).

Rate of evolution on leading and lagging strands

Comparisons of orthologs from closely related genomes

Replication-associated mutational pressure is strong enough to influence gene evolution and rearrangements. A way to observe the influence of replication direction on gene evolution is to compare pairs of orthologs from closely related genomes. LAFAY et al. (1999) compared codon and amino acid usage between leading and lagging strand genes of *B. burgdorferi* and *T. pallidum*. Despite species-specific G+C content and chromosome structure and organisation, they found similar G-T versus A-C biases between the leading and lagging strands in these two species. The biases were found at the level of nucleotides, codons, and amino acids. The orthologs that have switched strands have adapted their codon and amino acid usage to their new strand and have the same codon usage as the genes of the new strand.

TILLIER and COLLINS (2000c), who compared orthologs from *Chlamydia trachomatis* and *Chlamydia pneumoniae*, also observed that the genes that switched the strand have acquired the skew of their current strand. Comparison of amino acid similarity and identity between the orthologs showed that the switched genes were on average more diverged than the nonswitched ones. Changing the replication direction significantly changed the amino acid sequence and affected evolution of these sequences. Thus the substitutions resulting from mutational pressure are not neutral.

SZCZEPANIK et al. (2001) have measured differences in the rate of divergence between genes lying on the leading strand, lagging strand, and genes which changed their positions on chromosome during evolution. Analyses have been performed on 12,645 orthologs derived from 11 eubacterial genomes showing evident compositional asymmetry between leading and lagging strands. In almost all cases the distances between genomes measured by the divergence of orthologs from the lagging strand are statistically significantly larger than the distances counted on the basis of the leading strand orthologs. Apparently the orthologs situated on lagging strands diverge quicker than the orthologs situated on leading strands. This phenomenon can be explained either by a higher mutation rate on the lagging strand, or by stronger selection on the more conserved genes located on the leading strand (SZCZEPANIK et al. 2001). For closely related genomes the rate of divergence between the orthologs located on different strands is even greater than that of the lagging strand orthologs. The genes which have switched the strand recently are under a greater mutational pressure and diverge very quickly. The differences in the rate of divergence are significant enough to affect the structure of phylogenetic trees constructed on the basis of leading and lagging strand orthologs (SZCZEPANIK et al. 2001). Different mutational pressures on the two DNA strands group genes into slower and faster evolving groups. It may play an important role in adaptation to the quickly changing environment.

Rearrangements in genomes

MACKIEWICZ et al. (2001a) have found a method to determine which of two orthologs located on different strands has actually been relocated. Two pairs of highly asymmetric genomes were analysed, *C. trachomatis* vs. *C. pneumoniae*, and *B. burgdorferi* vs. *T. pallidum*. CG and AT skews were measured for each analysed gene as well as mean values and standard deviations for all leading and lagging strand genes in each genome. The gene whose GC and AT skews were more distant from the mean for its current strand was considered switched. The authors have found that genes have been relatively more often transferred from lagging to leading DNA strands than vice versa. That may be because the more conserved genes from the leading strand can tolerate fewer substitutions which change their amino-acid composition and codon usage when affected by a higher mutational pressure after inversion. Highly expressed genes seem to be more sensitive to discrimination control through codon usage (i.e. IKEMURA 1981, GOUY, GAUTIER 1982, SHARP, LI 1987). Moreover, the possible collisions between transcription and replication complexes may be more deleterious for highly expressed genes switched from the leading to lagging strand (MCINERNEY 1998).

The most specific rearrangements occur around the origin of replication (SUYAMA, BORK 2001, EISEN et al. 2000, READ et al. 2000, TILLIER, COLLINS 2000b). In closely related genomes, many orthologs coding for the same function remain at the same distance and orientation to the origin or terminus of replication, but they can be positioned on either of the two replichores. This property gives a specific picture when the positions of genes in one genome are plotted against the positions of their homologs in a closely related genome. TILLIER and COLLINS (2000b) have argued that the structure of replication forks, which are hot-spots of recombination, is responsible for that picture. However, the strand and distance from the origin of replication may be as well conserved by selection (MACKIEWICZ et al. 2001b). Firstly, the distance from the origin of replication determines copy number of a gene in bacteria whose generation time is shorter than replication period. In those cells the newly replicated origins initiate the next round of replication before the end of the previous round. Thus, in the cell there are several copies of genes proximal to the origin. Highly and lowly expressed genes should be located in optimal distances from the origin (LIU, SANDERSON 1995, 1996). Secondly, transfer of a gene to the opposite strand increases mutational pressure on that gene, as mentioned above, and thus should be selected against. Thirdly, there is a trend to keep both replichores the same size (LIU, SANDERSON 1996), possibly because that ensures the shortest time of replication of the genome.

Deciphering the mutational pressure

It is impossible to find out which substitutions generate the observed asymmetry just by measuring the asymmetry itself. The asymmetric DNA composition can be realised by an infinite number of combinations of frequencies of the twelve possible nucleotide substitutions. Protein coding sequences are not only under the mutational pressure typical for their location, but are also subject to

selection for function, which mercilessly eliminates all undesirable substitutions. However, the intergenic sequences which are remnants of duplications of genes should accumulate all mutations. Their comparison with the original genes should reveal the influence of mutational pressure. This approach was adopted by KOWALCZUK et al. (2001a). About thirty intergenic sequences were found in the *B. burgdorferi* genome, which read from the leading strand had some homology to ORFs from that genome's database. They were aligned and all the observed differences between the ORFs and their homologues were assumed to result from substitutions in the intergenic sequences. In that way the frequencies of all substitutions were found and a table of mutations for the leading strand was constructed (Table 2). C->T transition is the most frequent, as predicted by deamination theory (see above). The obtained table of substitution frequencies was tested in computer simulations. When a random, equimolar ($[A]=[T]=[G]=[C]$) DNA sequence was put under such a mutational pressure, after a sufficient number of generations it had the same asymmetry as the third codon positions of the leading strand ORFs of the real *B. burgdorferi* genome. When the *B. burgdorferi* ORFs were put under the pressure of the table, the nucleotide composition of third codon positions did not change. Also, precise corrections could be counted, which should be introduced into experimentally found differences in nucleotide sequences to find the real numbers of substitutions during the divergence time (KOWALCZUK et al. 2001b). The correction proposed by KIMURA (1980) does not take into account different rates of mutation of each nucleotide. Now it is possible to count the corrections very precisely and for a wider degree of homology. Furthermore, it should be possible to separate the effects of mutational and selection pressures.

The table of substitutions has an interesting feature: the time when a half of nucleotides of a given type are substituted by other nucleotides is linearly correlated with the fraction of the analysed type of nucleotide in the sequence. It seems to be a property of the pure mutational pressure. A table of substitution rates for sequences under strong selection (ZHANG 1999) gave no correlation. Also, a computer-generated table of substitutions by KOWALCZUK et al. (1999b) imposed asymmetry on DNA sequence but there was no correlation between fractions of nucleotides in that sequence and the rate of their substitution. However, correlation was found for substitution matrices which were obtained for sequences free from selection pressure by TAMURA (1992), ADACHI, HASEGAWA (1996), LI et al. (1984), YANG (1994), and GOJOBORI et al. (1982). Apparently the mutational pressure in a genome is highly correlated with its nucleotide composition in such a way that the higher substitution turnover of a nucleotide determines the lower fraction of this nucleotide in the sequence.

Conclusions

In prokaryotic genomes, location of a gene on the chromosome influences its nucleotide composition, which is reflected by its codon composition and amino acid composition of the protein. In different regions of the genome, mutational and selection pressures act differently, often oppositely. The skews introduced by particular mechanisms generating asymmetry are universal, but relative

contribution of various mechanisms is species-specific. The basic division between these forces lies between the parts of the chromosome replicated as leading and lagging strands. Consequently rates of divergence for the genes from the two strands are not equal. Translocation of a gene to the opposite strand results in sudden increase in the number of mutations and may be detrimental, especially if it was transferred into the lagging strand. If the gene is to survive, it has to adapt to local DNA composition.

Methods of analysis available to date enable assessment of relative contribution of mutational and selection forces to the observed asymmetry. The frequency of each kind of substitution can be found for sequences free from selection. This will enable measuring phylogenetic distances more precisely, and constructing better phylogenetic trees.

Acknowledgements

This work was supported by the State Committee for Scientific Research, grant numbers 6 P04A 025-18 and 6 P04A 016 20. Additionally, P.M. was supported by the Foundation for Polish Science.

References

- ADACHI J., HASEGAWA M. (1996). Tempo and mode of synonymous substitutions in mitochondrial DNA of primates. *Mol. Biol. Evol.* 13: 200-208.
- ANDERSSON S.G., ZOMORODIPOUR A., ANDERSSON J.O., SICHERITZ-PONTEN T., ALSMARK U.C., PODOWSKI R.M., NASLUND A.K., ERIKSSON A.S., WINKLER H.H., KURLAND C.G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396:133-140
- BAKER T.A., WICKNER S.H. (1992). Genetics and enzymology of DNA replication in *Escherichia coli*. *Annu. Rev. Genet.* 26: 447-477.
- BASIC-ZANINOVIC T., PALOMBO F., BIGNAMI M, DOGLIOTTI E. (1992). Fidelity of replication of the leading and the lagging DNA strands opposite N-methyl-N-nitrosourea-induced DNA damage in human cells. *Nucleic Acids Res* 20: 6543-6548.
- 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.
- BENNETZEN J.H., HALL B.D. (1982). Codon selection in yeast. *J. Biol. Chem.* 257: 3026-3031.
- BERTHELSEN Ch.L., GLAZIER J.A., SKOLNICK M.H. (1992). Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Phys. Rev. A* 45: 8902-8913.
- BLATTNER F.R., PLUNKETT G. 3rd, BLOCH C.A., PERNA N.T., BURLAND V., RILEY M., COLLADO-VIDES J., GLASNER J.D., RODE C.K., MAYHEW G.F., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453-1462.
- BREWER B.J. (1988). When polymerases collide: Replication and the transcriptional organisation of the *E. coli* chromosome. *Cell* 53: 679-686.
- CEBRAT S., DUDEK M.R. (1998). The effect of DNA phase structure on DNA walks. *Eur. Phys. J.* 3: 271-276.
- CEBRAT S., DUDEK M.R., GIERLIK A., KOWALCZUK M., MACKIEWICZ P. (1999). Effect of replication on the third base of codons. *Physica A* 265: 78-94.
- CEBRAT S., DUDEK M.R., MACKIEWICZ P. (1998). Sequence asymmetry as a parameter indicating coding sequences in *Saccharomyces cerevisiae* genome. *Theory Bioscienc.* 117: 78-89.

- CEBRAT S., DUDEK M.R., ROGOWSKA A. (1997a). Asymmetry in nucleotide composition of sense and antisense strands as a parameter for discriminating open reading frames as protein coding sequences. *J. Appl. Genet.* 38: 1-9.
- CEBRAT S., DUDEK M.R., MACKIEWICZ P., KOWALCZUK M., FITA M. (1997b). Asymmetry of coding versus non-coding strands in coding sequences of different genomes. *Microb. & Comp. Genom.* 2: 259-268.
- CHARGAFF E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* 6: 201-240.
- DANIELS D.L., SANGER F. COULSON A.R. (1983). Features of bacteriophage lambda: analysis of the complete nucleotide sequence. *Cold Spring Harbor Symp. on Quantum Biology Vol.* 47: 1009-1024.
- DESCHAVANNE P., FILIPSKI J. (1995). Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. *Nucleic Acids Res.* 23:1350-1353.
- ECHOLS H., GOODMAN M.F. (1991). Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* 60: 477-511.
- EISEN, J.A. HEIDELBERG JF, WHITE O, SALZBERG SL. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biology* 1: 0011.1-0011.9.
- 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.
- FIJALKOWSKA I.J., SCHAAPER R.M. (1996). Mutants in the Exo I motif of *Escherichia coli* dnaQ: defective proofreading and inviability due to error catastrophe. *Proc. Natl. Acad. Sci. USA* 93: 2856-2861.
- FILIPSKI J. (1990). Evolution of DNA sequences. Contributions of mutational bias and selection to the origin of chromosomal compartments. In: Obe G. (ed.) *Advances in mutagenesis research* 2, pp. 1-54. Springer Verlag, Berlin.
- 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: 240-245.
- FRANK A.C., LOBRY J.R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238: 65-77.
- FRASER C.M., CASJENS S., HUANG W.M., SUTTON G.G., CLAYTON R., LATHIGRA R., WHITE O., KETCHUM K.A., DODSON R., HICKEY E.K. et al. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390: 580-586
- FRASER C.M., GOCAYNE J.D., WHITE O., ADAMS M.D., CLAYTON R.A., FLEISCHMANN R.D., BULT C.J., KERLAVAGE A.R., SUTTON G.G., KELLEY J.M. et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397-403.
- FRASER C.M., NORRIS S.J., WEINSTOCK G.M., WHITE O., SUTTON G.G., DODSON R., GWINN M., HICKEY E.K., CLAYTON R., KETCHUM K.A. et al. (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* 281: 375-388
- FREDERICO L.A., KUNKEL T.A., SHAW B.R. (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rare constants and the activation energy. *Biochemistry* 29: 2532-2537.
- FREEMAN J.M., PLASTERER T.N., SMITH T.F., MOHR S.C. (1998). Patterns of genome organisation in bacteria. *Science* 279: 1827.
- GATES M.A. (1986). A simple way to look at DNA. *J. Theor. Biol.* 119: 281-300.
- GIERLIK A., KOWALCZUK M., MACKIEWICZ P., DUDEK M.R., CEBRAT S. (2000). Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome?. *J. Theor. Biol.* 202: 305-314.
- GOJOBORI T., LI W-H, GRAUR D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* 18: 360-369.
- GOUY M., GAUTIER C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10: 7055-7074.
- GRIGORIEV A. (1998). Analysing genomes with cumulative skew diagrams. *Nucleic Acids Res.* 26: 2286-2290.

- GRIGORIEV A. (1999). Strand-specific compositional asymmetries in double-stranded DNA viruses. *Viol. Res.* 60:1-19.
- GUTIERREZ G., MARQUEZ L., MARTIN A. (1996). Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translation efficiency. *Nucleic Acids Res.* 24: 2525-2528.
- HANAWALT P.C. (1991). Heterogeneity of DNA repair at the gene level. *Mutat. Res.* 247: 203-211.
- HIMMELREICH R., HILBERT H., PLAGENS H., PIRKL E., LI B.C., HERRMANN R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 24: 4420-4449.
- HUTCHINSON F. (1996). Mutagenesis. In: Neidhardt F.C. (red.) *Escherichia coli* and *Salmonella*. Cellular and molecular biology. Asm. Press, Washington D.C.: 749-763
- IKEMURA T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein sequence: a proposal for a synonymous codon choice that is optimal for the *E. coli* translation system. *J. Mol. Biol.* 151: 389-409.
- IWAKI T., KAWAMURA A., ISHINO Y., KOHNO K., KANO Y., GOSHIMA N., YARA M., FURUSAWA M., DOI H., IMAMOTO F. (1996). Preferential replication-dependent mutagenesis in the lagging DNA strand in *Escherichia coli*. *Mol. Gen. Genet.* 251: 657-664.
- KARLIN S. (1999). Bacterial DNA strand compositional asymmetry. *Trends Microbiol.* 8:305-308.
- KARLIN S., BURGE C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 11: 283-290.
- KARLIN S., MRAZEK J. (1996). What drives codon choices in human genome? *J. Mol. Biol.* 262: 459-472.
- KARLIN S., BLAISDELL B.E., BUCHER P. (1992). Quantile distributions of amino acid usage in protein classes. *Protein Eng.* 5: 729-738.
- KIMURA M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16: 111-120.
- KOWALCZUK M., MACKIEWICZ P., GIERLIK A., DUDEK M.R., CEBRAT S. (1999a). Total number of coding open reading frames in the yeast genome. *Yeast* 15: 1031-1034.
- KOWALCZUK M., GIERLIK A., MACKIEWICZ P., CEBRAT S., DUDEK M.R. (1999b). Optimization of gene sequences under constant mutational pressure and selection. *Phys. A* 273: 116-131.
- KOWALCZUK M., MACKIEWICZ P., SZCZEPANIK D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001a). Precise genome evolution law of turnover of nucleotides under mutational pressure. Submitted to *BMC Evolutionary Biology*
- KOWALCZUK M., MACKIEWICZ P., SZCZEPANIK D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001b). Multiple base substitution corrections in DNA sequence evolution. *Int. J. Modern Phys. C* 12(7), in press.
- KREUTZER D.A., ESSIGMANN J.M. (1998). Oxidized, deaminated cytosines are a source of C->T transitions *in vivo*. *Proc. Natl. Acad. Sci. USA* 95: 3578-3582.
- KUNST F., OGASAWARA N., MOSZER I., ALBERTINI A.M., ALLONI G., AZEVEDO V., BERTERO M.G., BESSIERES P., BOLOTIN A., BORCHERT S., et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390: 249-256.
- LAFAY B., LLOYD A.T., MCLEAN M.J., DEVINE K.M., SHARP P.M., WOLFE K.H. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.* 27: 1642-1649.
- LAGUNEZ-OTERO J., TRIFONOV E.N. (1992). mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. *J. Biomol. Struct. Dyn.* 10: 455-464.
- LI W-H, WU C.I., LUO C-C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitution in pseudogenes and its evolutionary implications. *J. Mol. Evol.* 12: 58-71.
- LINDAHL T. (1993). Instability and decay of the primary structure of DNA. *Nature* 362: 709-715.
- LIU S.L., SANDERSON K.E. (1995). Rearrangements in the genome of the bacterium *Salmonella typhi*. *Proc. Natl. Acad. Sci. USA* 92: 1018-1022.

- LIU S.L., SANDERSON K.E. (1996). Highly plastic chromosomal organization in *Salmonella typhi*. Proc. Natl. Acad. Sci. USA 93: 10303-10308.
- LOBRY J.R. (1995). Properties of a general model of DNA evolution under no-strand bias conditions. J. Mol. Evol. 40: 326-330. Erratum 41:680.
- LOBRY J.R. (1996a). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13: 660-665.
- LOBRY J.R. (1996b). A simple vectorial representation of DNA sequence for the detection of replication origins in bacteria. Biochimie 78: 323-326.
- LOPEZ P., FORTERRE P., LE GUYADER H., PHILIPPE H. (2000). Origin of replication of *Thermotoga maritima*. Trends. Genet. 16: 59-60.
- LOPEZ P., PHILIPPE H., MYLLYKALLIO H., FORTERRE P. (1999). Identification of putative chromosomal origins of replication in Archaea. Mol. Microbiol. 32: 881-891.
- MACKIEWICZ P., GIERLIK A., KOWALCZUK M., DUDEK M.R., CEBRAT S. (1999a). Asymmetry of nucleotide composition of prokaryotic chromosomes. J. Appl. Genet. 40: 1-14.
- MACKIEWICZ P., GIERLIK A., KOWALCZUK M., DUDEK M.R., CEBRAT S. (1999b). How does replication-associated mutational pressure influence amino acid composition of proteins? Gen. Res. 9: 409-416.
- MACKIEWICZ P., GIERLIK A., KOWALCZUK M., SZCZEPANIK D., DUDEK M.R., CEBRAT S. (1999c). Mechanisms generating long-range correlation in nucleotide composition of the *Borrelia burgdorferi* genome. Phys. A 273: 103-115.
- MACKIEWICZ P., SZCZEPANIK D., GIERLIK A., KOWALCZUK M., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001a). The differential killing of genes by inversions in prokaryotic genomes. J. Mol. Evol. 53, in press.
- MACKIEWICZ P., SZCZEPANIK D., KOWALCZUK M., CEBRAT S. (2001b). Flip-flop around the origin and terminus of replication in prokaryotic genomes. Submitted to Genome Biology.
- MARIANS K.J. (1992). Prokaryotic DNA replication. Annu. Rev. Biochem. 61: 673-719.
- MCINERNEY J.O. (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. USA 95: 10698-10703.
- MCLEAN M.J., Wolfe K.H., DEVINE K.M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 47: 691-696.
- MELLON I., HANAWALT P.C. (1989). Induction of the *Escherichia coli* lactose operon selectively increases repair of its transcribed DNA strand. Nature 342: 95-98.
- MIZRAJI E. I NINIO J. (1985). Graphical coding of nucleic acids sequences. Biochimie 67: 445-448.
- MRAZEK 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 synthesised chains. Proc. Natl. Acad. Sci. USA 59(2): 598-605.
- PERRIERE G., LOBRY J.R. I THIOULOUSE J. (1996). Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acids sequences. Comput. Appl. Biosci. 12: 519-524.
- PICARDEAU M., LOBRY J.R., HINNEBUSCH B.J. (1999). Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome. Mol. Microbiol. 32: 437-445.
- PICARDEAU M., LOBRY J.R., HINNEBUSCH B.J. (2000). Analyzing DNA strand compositional asymmetry to identify candidate replication origins of *Borrelia burgdorferi* linear and circular plasmids. Gen. Res. 10: 1594-1604.
- QIN M.H., MADIRAJU M.V., RAJAGOPALAN M. (1999). Characterization of the functional replication origin of *Mycobacterium tuberculosis*. Gene 233: 121-130.
- RADMAN M. (1998). DNA replication: one strand may be more equal. Proc. Natl. Acad. Sci. USA 95: 9718-9719.

- READ T.D., BRUNHAM R.C., SHEN C., GILL S.R., HEIDELBERG J.F., WHITE O., HICKEY E.K., PETERSON J., UTTERBACK T., BERRY K. et al. (2000). Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* 28: 1397-1406.
- REYES A., GISSI C., PESOLE G., SACCONI C. (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.* 15: 957-966.
- ROBERTS J.D., IZUTA S., THOMAS D.C., KUNKEL T.A. (1994). Mismatch-, site-, and strand-specific error rates during simian virus 40 origin-dependent replication *in vitro* with excess deoxythymidine triphosphate. *J. Biol. Chem.* 269: 1711-1717.
- ROCHA E.P.C., DANCHIN A., VIARI A. (1999a). Universal replication biases in bacteria. *Mol. Microbiol.* 32: 11-16.
- ROCHA E.P.C., DANCHIN A., VIARI A. (1999b). Bacterial DNA strand compositional asymmetry: Response. *Trends Microbiol.* 7: 308.
- ROCHA E.P.C., VIARI A., DANCHIN A. (1998). Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. *Nucleic Acids Res.* 26: 2971-2980.
- ROMERO H., ZAVALA A., MUSTO H. (2000). Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucl. Acids Res.* 28: 2084-2090.
- SALZBERG S.L., SALZBERG A.J., KERLAVAGE A.R. TOMB J-F. (1998). Skewed oligomers and origins of replication. *Gene* 217: 57-67.
- SHARP P.M. I COWE E. (1991), Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* 7: 657-678.
- SHARP P.M., LI W.-H. (1987). The codon adaptation index: a measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Res.* 15: 1281-1295.
- SHEPHERD J.C. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci. USA* 78: 1596-1600.
- SMITHIES O., ENGELS W.R., DEVEREUX J.R., SLIGHTOM J.L. I SHEN S.H. (1981). Base substitutions, length differences, and DNA strand asymmetries in the human G γ and A γ fetal globin gene region. *Cell* 26: 345-353.
- SUYAMA, M., BORK, P. (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* 17, 10-13.
- SZCZEPANIK D., MACKIEWICZ P., KOWALCZUK M., GIERLIK A., NOWICKA A., DUDEK M.R., CEBRAT S. (2001). Evolution rates of genes on leading and lagging DNA strands. *J. Mol. Evol.* 52, 426-433.
- TAMURA K. (1992). The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. *Mol. Biol. Evol.* 9: 814-825.
- TANAKA M., OZAWA T. (1994). Strand asymmetry in human mitochondrial DNA mutations. *Genomics* 22: 327-335.
- THOMAS D.C., SVOBODA D.L., VOS J.M., KUNKEL T.A. (1996). Strand specificity of mutagenic bypass replication of DNA containing psoralen monoadducts in a human cell extract. *Mol. Cell. Biol.* 16: 2537-2544.
- TILLIER E.R.M., COLLINS R.A. (2000a). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* 50: 249-257.
- TILLIER E.R.M., COLLINS R.A. (2000b). Genome rearrangement by replication-directed translocation. *Nat. Genet.* 26: 195-197.
- TILLIER E.R.M., COLLINS R.A. (2000c). Replication orientation affects the rate and direction of bacterial gene evolution. *J. Mol. Evol.* 51: 459-463.
- TRIFONOV E.N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. *J. Mol. Biol.* 194: 643-652.
- TRIFONOV E.N. (1992). Recognition of correct reading frame by the ribosome. *Biochimie* 74: 357-362.
- TRINH T.Q., SINDEN R.R. (1991). Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli*. *Nature* 352: 544-547.

- VEAUTE X., FUCHS R.P.P. (1993). Greater susceptibility to mutations in lagging strand of DNA replication in *Escherichia coli* than in leading strand. *Science* 261: 598-600.
- WANG J. (1998). The base contents of A, C, G, or U for three codon positions and the total coding sequences show positive correlation. *J. Biomol. Struct. Dyn.* 16: 51-57.
- WATSON J.D., CRICK F.C.H. (1953). A structure for deoxyribose nucleic acid. *Nature* 171: 737-738.
- WONG J.T. I CEDERGREN R. (1986). Natural selection versus primitive gene structure as determinant of codon usage. *Eur. J. Biochem.* 159: 175-180.
- YANG Z. (1994). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39: 105-111.
- YUZHAKOV A., TURNER J., O'DONNELL M. (1996). Replisome assembly reveals the basis for asymmetric function in leading and lagging strand replication. *Cell* 86: 877-886.
- ZAWILAK A., CEBRAT S., MACKIEWICZ P., KRÓL-HULEWICZ A., JAKIMOWICZ D., MESSER W., GOSCINAK G., ZAKRZEWSKA-CZERWINSKA J. (2001). Identification of a putative chromosomal replication origin from *Helicobacter pylori* and its interaction with the initiator protein DnaA. *Nucleic Acids Res.* 29: 2251-2259.
- ZHANG J. (1999). Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.* 16: 868-875.
- ZHANG C.T., ZHANG R. (1991). Analysis of distribution of bases in codon in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.* 19: 6313-6317.

Table 1

Numbers of nucleotides in the leading and lagging part of the Watson strand of the *Borrelia burgdorferi* genome. The random strand was obtained by drawing nucleotides randomly from a pool of all nucleotides from the Watson strand.

Strand type	Number of bases in strand				Strand length
	A	T	G	C	
Random leading Watson strand	160,481	162,812	64,209	64,946	452,448
Random lagging Watson strand	162,216	165,081	65,453	65,526	458,276
Real leading Watson strand	145,921	178,068	75,741	52,718	452,448
Real lagging Watson strand	177,186	149,128	53,911	78,051	458,276

Table 2

Frequencies of substitutions in the leading strand of the *Borrelia burgdorferi* genome. All frequencies sum up to 1.

		To:			
		A	T	G	C
From:	A	-	0.103	0.067	0.023
	T	0.065	-	0.035	0.035
	G	0.164	0.116	-	0.015
	C	0.070	0.261	0.047	-

Figure captions

Fig. 1. The DNA walks performed on the whole chromosome (a) and intergenic sequences (b) of the *Treponema pallidum* genome. Scale on the X-axis corresponds to the real co-ordinates of the chromosome in base pairs. Numbers on the Y-axis indicate cumulative abundance of A over T and G over C normalised by the length of the analysed sequence.

Fig. 2. The DNA walks performed on the three codon positions of CDS from the Watson strand of the *Treponema pallidum* genome. (a-c) normal DNA walks, as described for Fig. 1; (d-f) detrended DNA walks. Scale on the X-axis corresponds to the real co-ordinates of the chromosome in base pairs. Numbers on the Y-axis indicate cumulative abundance (or local relative cumulative abundance for detrended DNA walks) of A over T and G over C, normalised by the length of the analysed sequence.

Fig. 3. Subtraction (a-c) and addition (d-f) of DNA walks performed on nucleotides in three codon positions of CDS of the *Bacillus subtilis* genome. Scale on the X-axis corresponds to real co-ordinates of the chromosome in base pairs. Numbers on the Y-axis demonstrate the local cumulative deviations from the average composition normalised by the length of the analysed sequence.

Fig. 4. DNA walks (spiders) performed on *Borrelia burgdorferi* sequences; (a) coding sequence lying on the leading strand (gene BB0005; tryptophanyl-tRNA synthetase), (b) intergenic sequence. For detailed description see text.

Fig. 5. DNA walks (spiders) performed on the sequence of all genes from the leading (a) and lagging (b) strand of the *Borrelia burgdorferi* genome spliced together.

Figure 1.

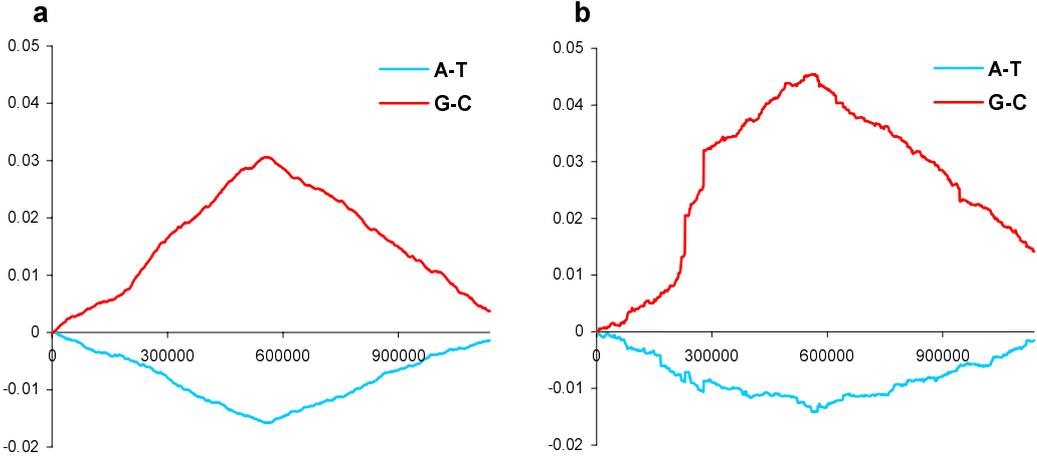


Figure 2.

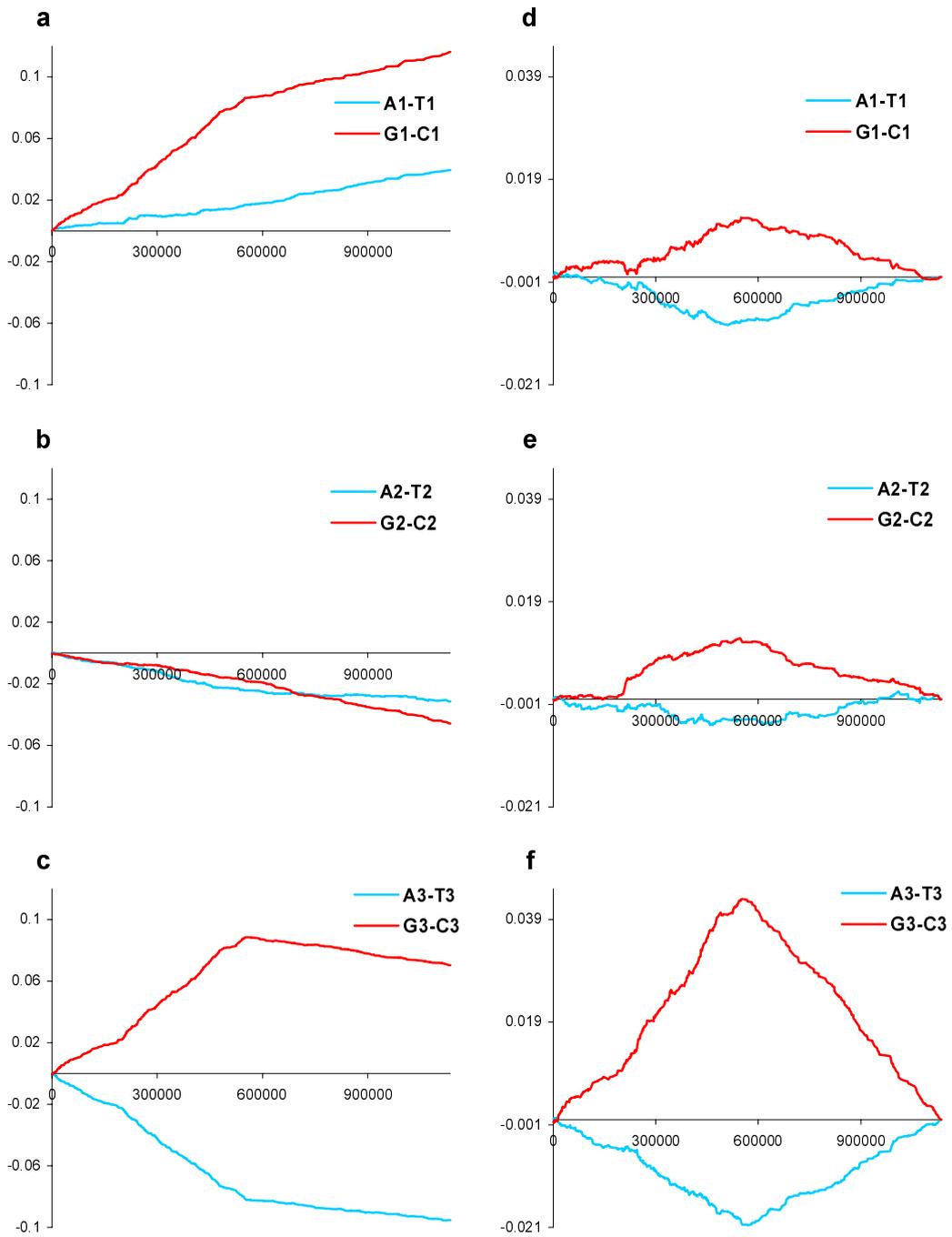


Figure 3.

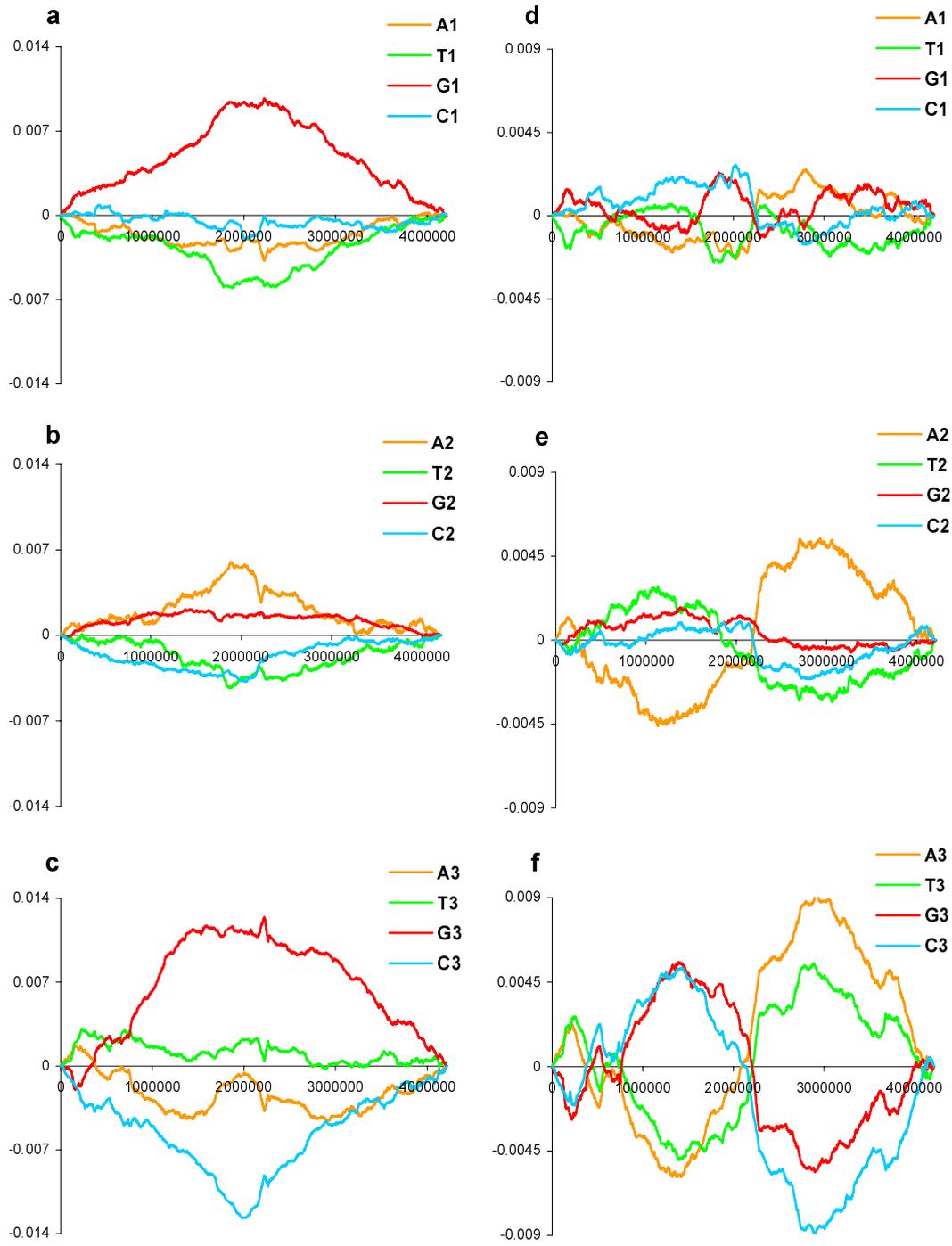


Figure 4.

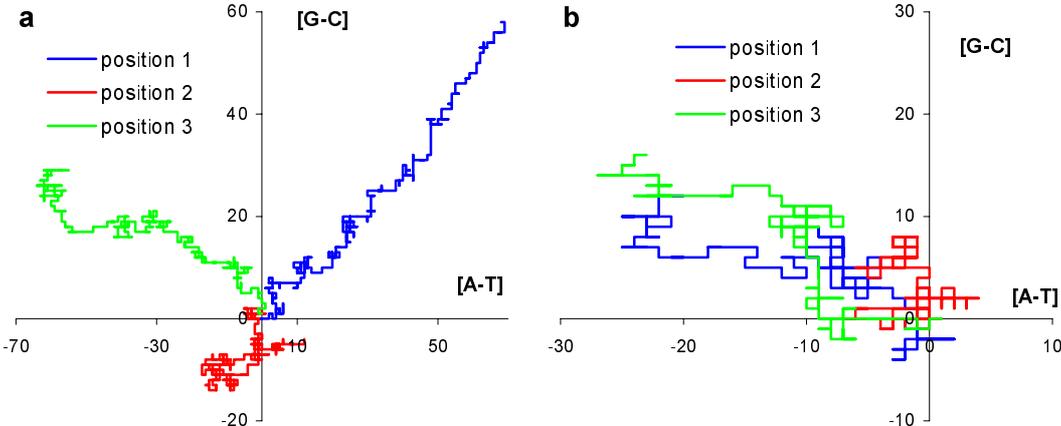


Figure 5.

