Citation:
Maria Kowalczuk, Dorota Mackiewicz, Paweł Mackiewicz, Natalia Polak, Kamila Smolarczyk, Joanna Kriaga, Mirosław Roman Dudek, Stanisław Cebrat (2005)
Mutational and selective mechanisms shaping large-scale organization and evolution of chromosomes. *Current Topics in Genetics* **1** pp. 87-101
http://www.researchtrends.net/tia/abstract.asp?in=0&vn=1&tid=45&aid=2113&pub=2005&type=3

Mutational and selective mechanisms shaping large-scale organization and evolution of chromosomes

Maria Kowalczuk[1], Dorota Mackiewicz[1], Pawel Mackiewicz[1], Natalia Polak[1], Kamila Smolarczyk[1], Joanna Banaszak[1], Miroslaw R. Dudek[2], Stanislaw Cebrat[1]

[1] Institute of Genetics and Microbiology, University of Wroclaw, Przybyszewskiego 63/77, 51-148 Wroclaw, Poland

[2] Institute of Physics, University of Zielona Gora, ul. Szafrana 4A, 65-516 Zielona Góra, Poland

Corresponding author:

Stanislaw Cebrat

Institute of Genetics and Microbiology

University of Wroclaw

Przybyszewskiego 63/77

51-148 Wroclaw

Poland

Tel. +48 71 3756 303

Fax +46 71 3252 151

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

Short running title:

Mutational and selective mechanisms shaping organization and evolution of chromosomes

Abstract

Sequencing projects provide a lot of information on genome structure, organization and evolution. Whole-genome analyses reveal large biases in nucleotide composition of genes and chromosomes, coding density and gene distribution, which are not easy to explain in the cause-effect manner. Usually these problems are approached from either neutralist or selectionist point of view. The present work is a review of phenomena of the whole-genome scale and their possible explanations. The major process which shapes organization of bacterial chromosomes is DNA replication, while in eukaryotic chromosomes it is mostly transcription and isochore structure. Asymmetric structure of chromosomes related with mechanisms of replication and inner asymmetry of genes related with protein coding function influence frequency and kinds of rearrangements and the rate of gene evolution. The mutational and selection pressures are tuned to optimize the costs of evolution. The key role in this process is played by the genetic code, which is universal for the whole living world and to which both mutational and selective pressures have had to adapt.

Keywords: mutation pressure, selection pressure, genome evolution, replication, isochore

Introduction

The genome is more than the sum of its parts, and studies of whole genomic sequences allow for finding general properties and laws governing the genetic information. However, there is little consensus about their interpretation. The present work attempts to present the on-going discussion about the origin and significance of genomic-scale phenomena observed in prokaryotic and eukaryotic genomes. The genome sequence is shaped by two main, sometimes opposing forces: mutation and selection pressures. Mutation pressure is usually assumed to be random, but in fact most mutations are introduced to DNA when it is single-stranded and thus most vulnerable, during replication and transcription, which are highly regulated processes. Therefore the mutation pressure allowed by these processes must have evolved under strong selection. In protein coding sequences the possibility of accepting mutations is connected with degeneracy of the genetic code. These factors introduce specific biases to nucleotide composition of genes and their location on chromosomes, and influence their rearrangements and evolution, which is discussed below.

Replication of the bacterial chromosome

Most bacterial chromosomes are circular and possess only one origin of replication (ori), from which two replication forks start in opposite directions to meet at the other extreme of the chromosome in the terminus of replication region (ter) (Fig 1). Because each DNA strand can be replicated only in one direction, from 5’ to 3’, and Watson and Crick strands are antiparallel,

replication forks are asymmetric – the strand replicated in the same direction as replication forks movement is synthesized continuously (the leading strand) while the strand replicated in the opposite direction to replication forks movement is synthesized from Okazaki fragments (the lagging strand). Each of the strands is synthesized by a different DNA polymerase subunit, and in some species, e.g. in *Escherichia coli*, these subunits are identical [1] while e.g. in *Bacillus subtilis* the leading and lagging strands are synthesized by polymerases encoded by different genes [2,3]. Nevertheless, replication forks have to be asymmetric because the leading strand complex has to be more processive to stay associated with the template during replication, while the lagging strand complex has to dissociate more easily [4].

As a consequence, Watson strand can be divided into two halves – the one replicated as leading strand and the one replicated as lagging strand (Fig 1). The complementary Crick strand can be divided respectively into lagging and leading strands. It is important to consider gene location respective to DNA strand and chromosome topology. If the sense strand of a gene (corresponding to mRNA) is the leading strand, the gene is designated as located on the leading strand, analogously the gene located on the lagging strand. Also, location of genes can be viewed in respect to the distance from the origin of replication: the genes located near ori are designated as proximal, the ones located near ter – as distal (Fig 1).

Replication-associated asymmetry in bacterial genomes

The mode of replication is reflected in nucleotide composition of chromosomes. The numbers of guanine and cytosine and of adenine and thymine are equal in double stranded DNA. In the absence of strand bias the number of complementary nucleotides within a single strand also should be equal [5]. It is true for whole chromosomes, but when one analyses leading and lagging strands separately, there are large differences in the numbers of complementary nucleotides, which has been defined as DNA asymmetry [6,7]. Sequences located on the leading strand are usually rich in guanine and thymine, while sequences from the lagging strand are rich in adenine and cytosine. It is a universal feature of bacterial genomes [6,8-14]. The asymmetry can be observed both in intergenic and protein coding sequences. It is reflected also in the codon composition and amino acid composition of proteins [15-17].

DNA asymmetry is best shown by DNA walks (Fig 2, 4) [12]. In this way one can see the local trends in the nucleotide composition whereas the global trend in the whole sequence is eliminated. In most bacterial species there is a sharp change of the cumulated local trend in the ori and ter regions (Fig 2, 4). In bacterial genomes the asymmetry is so strong that it enables precise location of the origin of replication based only on sequence analysis [18-21]. In archaeal genomes location of the origin of replication is usually unknown, the asymmetry is usually absent or not very

pronounced but if present, it may be applied to identification of replication origin (see [22] and references therein).

Most researchers find the cause of replication-associated asymmetry in the mutational pressure associated with the replication process. It is indicated by the change of compositional trends exactly in the origin and terminus of replication, and by strong asymmetry both in third positions in codons in protein coding sequences and in intergenic sequences [10,12,14,23]. The change of trends in ori and ter regions means the change of the mode of replication from leading to lagging and vice versa. Third positions in codons and intergenic sequences are under weaker selection pressure than first and second positions in codons, so they should better reflect mutational pressure.

Another kind of DNA walks ("spiders") illustrates the results of asymmetric mutational and selection pressures exerted on genes (Fig 3). In this walk nucleotides in each position in the codon are analysed separately, so the walker performs three walks for each gene. The dark lines show walks on gene sequences from the data base, and the grey lines show walks on the same genes but after evolution under asymmetric mutational pressure described by [24]. The first positions in codons are rich in guanine and adenine, which is typical for protein coding sequences. The second positions are rich in cytosine, and the third positions are rich in nucleotides typical for the strand: the leading strand genes are rich in G and T, and the lagging strand genes – in A and C. Genes after evolution without any selection lose the triplet structure, all the three walks are similar. Now the genes have nucleotide composition generated by pure mutation pressure, when there is no selection. Compositional trends for leading and lagging strands are opposite.

[25] formulated cytosine deamination theory to explain the asymmetry between leading and lagging strands of bacterial genomes. During synthesis of the lagging strand its template – the strand replicated in the previous cycle as leading – remains single stranded, thus more susceptible to mutations. Probably in single stranded DNA methylated cytosines are more frequently deaminated which leads to their substitutions by thymines [26-28]. Cytosine deamination would explain the observation that transitions of GC pairs into AT pairs are dominant mutations in *E. coli* [29]. Also mispairing of thymine with guanine has been observed during synthesis of the leading strand [30]. These processes enrich the leading strand in guanine and thymine, while the lagging strand – in adenine and cytosine. Actually, the mutational pressure associated with replication found for bacterial genomes confirms the cytosine deamination theory by indicating very high rate of the $C \rightarrow T$ substitution [24,31]. Universality of these biases in bacteria suggests that the processes of DNA replication are responsible for them. So far only one exception has been found in the case of the *Streptomyces coelicolor* genome in which the lagging strand, not the leading one, is richer in guanine than cytosine [32].

The occurrence of asymmetric substitutions has been also explained by asymmetric structure and function of replication forks. Experimental analyses however yield contradictory results on the level of mistakes occurring during the synthesis of the leading and lagging strands. [33] analyzed frequency of deletion of a palindromic fragment and found that it was much greater on the lagging strand, probably because during replication a longer stretch of the template is uncovered and loops are more easily formed. [34] analyzed frequency of deletion of one nucleotide and substitutions in genes placed on a plasmid in different orientations in respect to replication fork movement. They observed increased frequency of mutations on the lagging strand, and explained it by greater susceptibility to mutations of the lagging strand replication machinery. However, [4] observed a smaller frequency of mutations on the lagging strand, which they explained by the greater processivity of the DNA polymerase complex which replicates the leading strand. It results in greater possibility of repair of mispaired bases in the lagging strand, because the polymerase complex has to dissociate from the template more often. However, they analyzed mutations in the lacZ operon which in *E. coli* is located on the lagging strand, so its translocation to the leading strand must cause increase in mutation rate [35].

Since the coding strand remains single and uncovered during transcription, and thus more susceptible to mutations, asymmetry in nucleotide composition of genes and chromosomes has been also explained by the mutational pressure associated with the process of transcription [26-37,8]. However, in that case the asymmetry would be observed only in the genomes where genes are very unevenly distributed on leading and lagging strands or differ in expression levels, and it would not be apparent in intergenic sequences.

Asymmetric location of genes on differently replicated DNA strands

Protein coding sequences have their own asymmetry between the sense and antisense strands, resulting from the protein coding function and optimization of translation process [38-45]. Nucleotide composition and asymmetry are different for each position in codons (see Fig. 3), so biased distribution of genes on chromosome should introduce asymmetry in the scale of the whole chromosome.

In almost all bacterial species more genes are encoded on the leading strand than on the lagging one. In some species, e.g. Mycoplasma, the bias is so pronounced that it is the main cause of the asymmetry observed between differently replicated strands [15]. Interestingly, in the species in which two different DNA polymerase subunits replicate the leading and lagging strands (gram positive bacteria, Mycoplasma) the number of genes located on the leading strand (on average 78%) is much greater than in species in which both replication complexes are identical (on average 58% of genes are located on the leading strand) [46].

These differences can be explained by different selective mechanisms. The most frequently evoked one is selection against head-on collisions between DNA and RNA polymerases. Genes located on the leading strand are transcribed in the same direction as the movement of the replication forks. However, the speed of transcription is much slower than that of replication, as it has to be adjusted to the speed of translation. When the DNA polymerase complex encounters the transcription complex on the leading strand, it has to slow down, but usually neither replication nor transcription are aborted. On the lagging strand replication complex collides head-on with transcription complex, which results in replication arrest, aborted transcription and dissociation of partial transcript [47,48]. First observations indicated that highly expressed genes are located preferentially on the leading strand [49] which seemed to confirm the hypothesis of polymerase collision avoidance [50]. However, further research showed that the differences in the number of genes on leading and lagging strands are not greatest in the fast growing species in which collisions between polymerases would be the most deleterious [46]. Also the number of highly expressed genes is not high enough to account for the gene bias observed on the leading strand [51]. It was observed that in most bacterial genomes the conserved essential genes group preferentially on the leading strand independently of whether they are highly expressed or not [52,53]. Genes coding for ribosomal proteins are usually located on the leading strand [10,11,37]. In case of a collision of polymerases transcription on the leading strand can be completed, while on the lagging strand it is usually aborted and the incomplete transcript can be translated into an incomplete peptide, which may be deleterious for essential functions [51]. If such a truncated peptide is part of a protein complex, it may lead to its inactivation resulting in a dominant negative phenotype [54]. Thus harmfulness of collisions depends on the function of the transcribed gene and not on collision frequency [52].

Bias in regions of bacterial chromosomes proximal and distal in respect to ori

Regions of the genome located on reciprocal poles of the chromosome in respect to ori and ter differ in nucleotide composition. In a lot of genomes regions proximal to ori are rich in G+C, and distal to ori are rich in A+T, especially in third positions in codons [55,12,56]. It is very well visible in DNA walks on the [G+C] graph (Fig 4). The proximal-distal bias may result from mutations caused by different availability of nucleotide precursors needed for replication – at the beginning of the replication process they are abundant, and at the end they may become scarce which may lead to specific bias in nucleotide composition of the sequences located near ter [57,58]. The observed surplus of A+T near ter may also result from the presence of specific sites binding proteins which take part in replication termination and from forming specific tertiary structures which facilitate parting of chromosome copies after replication [56]. Also other repair processes in ter proximity not

based on homologous recombination may preferentially introduce A and T [55,56]. Not taking into account the G+C bias along the chromosome may result in an erroneous conclusion that a lot of sequences located in the ter region have been acquired via lateral transfer [59].

Proximal and distal trends have also been observed in distribution of genes along the chromosome, which results from the mechanism of replication. Fast growing bacteria, such as *E. coli*, can divide every 20 minutes, while they need about twice as much time to replicate the chromosome. Therefore the next replication round begins before the previous one ends, which leads to 4 or even 8 times more copies of genes located near ori because of the presence of additional replication forks [60]. Thus the region proximal to ori should be rich in genes whose products are needed in large quantities. Actually, in fast growing bacteria like *E. coli* or *B. subtilis*, a significant surplus of highly expressed genes is observed in the region of chromosome proximal to ori, e.g. genes connected with translation – rRNA or ribosomal proteins [61,62,51]. Location of highly expressed genes near ori is also important for the organization and division of the cell. After replication start ori regions rapidly move apart to opposite poles of the cell, while ter region is located at the cell center [63]. The motor force of the RNA polymerase could pull apart ori regions during transcription of genes located near ori [64]. It facilitates chromosome segregation and division, especially that highly expressed genes group around ori [65].

Proximally to terminus of replication adaptive genes are located, genes acquired via horizontal transfer and transposones [66,67,68]. It may be connected with increased recombination rate caused by the presence of prophages in that region, e.g. in *E. coli* [69]. It is thought that in that region occur recombination hotspots enabling incorporation of foreign DNA into the chromosome [70]. Genes from lateral transfer do not require high expression and large number of copies, therefore they may be located preferentially near ter [51].

Evolution of genes in the asymmetric bacterial genomes

Different mutational pressures acting on different parts of the genome together with biased location of genes on chromosomes influence gene evolution. Comparison of orthologous sequences (homologous in closely related genomes) has shown for many bacterial species that mean divergence of genes located on the leading strand is statistically significantly lower than the mean divergence of genes located on the lagging strand [35]. The difference may be caused by the lower rate of evolution of the essential genes, which are preferentially located on the leading strand. Leading and lagging strands have their own compositional asymmetry, also the sense and antisense strands of a gene are asymmetric in nucleotide composition. Thus gene orientation may influence its susceptibility to mutations. In the sense strand of genes guanine is preferred (see e.g. [38,41,44,43,10]). Therefore genes located on the leading strand are less susceptible to C→T

transitions which prevail on the leading strand [25,24]. Genes from the lagging strand, whose antisense strands are rich in cytosine, are more susceptible to C→T transitions (Fig 5). Thus the difference in divergence between genes from leading and lagging strands may result from different adaptation of genes to the mutational pressures acting on their strands. It is assumed that the longer the given gene remains on one strand the better its adaptation. In this way the number of mutations can be minimized and gene function preserved. A gene relocated from one strand to the other (or simply inversed) is under a different mutational pressure which causes increase in divergence. The orthologs which changed the strand in their evolutionary history (and now are located on different strands in analyzed genomes) have the highest divergence compared to orthologs located on leading or lagging strands in all analyzed genomes [71,31,35,72]. The effect of the new mutational pressure is very strong and after changing the strand genes quickly adapt their nucleotide composition to the composition of the new strand [16,71,31]. The increase in accumulation of mutations after gene inversion can contribute to the loss of function and elimination of the gene. This effect is especially deleterious for conserved genes from the leading strand transferred to the lagging strand [73]. However, it is possible that some gene inversions are connected with duplications which may lead to generation of paralogous sequences and redundant information [74]. Duplicated sequences probably evolve differently than one copy genes. They are free from selection and can cumulate mutations [75-77] and generate pseudogenes, which are very common in bacterial genomes [78-83]. Since when searching for orthologs it is difficult to completely eliminate this kind of sequences, it can partly explain the strangely high divergence values for genes which changed the strand [72]. Duplicated genes can also evolve new functions [84-86]. Thus they can accelerate evolution of bacteria and ensure fast adaptation to the changing environment.

A gene which remains on a given strand for a long time stays under directional mutational pressure which may be disadvantageous for some genes. To preserve a given nucleotide composition (intermediate between the pressures acting on the two strands), it may be advantageous for some genes to change the strand (inverse) with a given frequency to maintain proper composition. Certainly it causes increase in divergence, but also increase in survival probability, as confirmed by computer simulations of gene evolution [87,88]. The results of simulations agree with the observed higher divergence of the orthologs located on differently replicated strands in compared genomes and with the weak conservation of gene location on chromosome even in closely related bacterial species [89-94]. The phenomenon of differentiated rate of evolution may cause problems in estimating phylogenetic relationships if they are based on analysis of genes located on differently replicated strands [35,72].

There is also a relationship between the rate of gene evolution and their location in respect to ori and ter. It was noted during analysis of the rate of synonymous substitutions in *E. coli* and

Salmonella typhimurium that it is two times higher near ter than in the proximity of ori [57]. However, when taking into account the differences in codon usage and gene expression levels it was found that the differences in substitution rates connected with distance to ori makes only 5% of the total variance in the substitution level [95] and the genes with the highest substitution rate group in a relatively small, A+T rich region near ter [56]. This effect probably does not result from the greater probability of occurrence of defects through conversions and recombinations in the genes located near ori because of their high copy number, as it was suggested before [57,96,97] because a significant relationship was shown between the distance from ori and the number of transversions, which is difficult to explain by the effect of these repair processes which do not differentiate between types of substitutions [95]. It is possible that the different rate of mutations is connected with the change in enzymatic activity in different regions of the chromosome, e.g. DNA polymerase has a tendency to dissociate from the synthesised strand during replication and its reassociation may generate mistakes [98,99]. The increased mutation rate near ter may also result from the existence of single stranded regions typical of stalled replication forks which increases susceptibility of these regions to mutations and recombinations [100]. Also different processes happening during disconnecting of the chromosome dimers after replication and repair processes which are more susceptible to mistakes occurring near ter [57,56].

Gene rearrangements and the structure of the bacterial chromosomes

All the phenomena described above are related with rearrangements shaping chromosome structure and organization. The rate of rearrangements depends on the number of repeats facilitating chromosome recombinations [101]. The greatest asymmetry in nucleotide composition has been observed in obligatory intracellular bacteria [51] which are devoid of repeats and thus more stable [102,103].

A comparison of relative location on chromosome of closest orthologs in closely related species has shown that a lot of genes have been relocated without changing the distance from the origin and terminus of replication (Fig 6). Such symmetric translocations around the origin of replication have been repeatedly found in bacterial genomes [104-110]. The main cause of this phenomenon is thought to be the process of replication itself, as the symmetrically proceeding replication forks increase the probability of reciprocal translocation or transposition between the newly replicated DNA strands [106]. The possibility of physical contact of replication forks during replication [111,112] seems to confirm this hypothesis. However, it is not the only explanation of the prevalence of symmetric rearrangements around ori and ter. The key role may be played by selection [73]. Since the distance from ori and ter determines the relative number of gene copies in the cell, it is important for highly and lowly expressed genes to be located in the proper distance

from ori and ter [61,113-117]. The observed rearrangements keep this distance. Another cause of the frequent occurrence of this kind of rearrangements could be selection pressure on the equal length of the replichores (halves of the chromosome replicated simultaneously) [115] because it guarantees shortest replication time, and only symmetric rearrangements do not change the length of the replichores (Fig 6).

Interestingly, the prevalence of symmetric rearrangements has been observed for the genes which during evolution have not changed the strand, and are located on the same strand in the analysed genomes [73]. Inversion of a gene causes the change of mutational pressure which increases mutation rate. It leads not only to amino acid substitutions but also to change in codon usage which is especially important for highly expressed genes [118-120]. Thus it seems that symmetric rearrangements should be preferred for this group of genes.

The influence of asymmetry on the restraints on relocation of genes between the strands is confirmed by the negative correlation between the fraction of paralogs located on differently replicated strands in the genome and the magnitude of asymmetry [121]. Although one could assume that the high value of asymmetry in the genome is a result of decreased frequency of rearrangements between the strands, it seems unlikely because of the ubiquity of gene rearrangements in bacterial genomes [90,122,123,104,106]. The disappearance of conservation of gene location on chromosome with phylogenetic distance between species suggests that the gene is not always eliminated after changing the strand. The presence of orthologs located on different strands may result from inversion connected with duplication. It is confirmed by the random location of points on dot matrices when relative location of all orthologs from different strands is compared [73]. Most of the orthologs have very high divergence and it is possible that relocated were additional copies of genes which as paralogs can be free from strong selection pressure. In this case the increase in mutation rate caused by the change of mutational pressure resulting from changing the strand would not be deleterious for these genes.

The essential genes located on the leading strand are more vulnerable to collisions between DNA and RNA polymerases, thus the effect of changing the strand should be more pronounced in the direction leading – lagging than lagging – leading. The analysis of orthologs which changed the strand in closely related genomes have shown that relocation from lagging to leading strand is relatively more common [124]. Also analyses of orthologs in many genomes have shown that the orthologs located on leading strands and common for many genomes are present even after adding to the analysis a distant archaeal genome, while orthologs located on lagging strands disappear very quickly [121].

Until recently it was assumed that nucleotide composition of bacterial genes is quite uniform, while the observed deviations are usually a result of lateral transfer. However, now we know that

the differentiation within a genome may be significant. It concerns differently replicated strands: the leading and lagging one, and in a lot of species also regions located proximally and distally to origin of replication. Such organization of chromosomes is reflected in the specific distribution of genes on chromosome – on the leading strand essential genes dominate, and in the ori region – highly expressed genes. For such polarization of chromosomes a lot of phenomena connected with replication are responsible, especially mutational pressure, repair mechanisms and recombinations, and also selection pressure.

Replication-associated asymmetry in eukaryotic genomes

Eukaryotic chromosomes are much larger than prokaryotic ones, and possess multiple origins of replication, termed ARSes (Autonomously Replicated Sequences). Not all of these origins are active during each replication cycle, and their firing is not precisely synchronized. Termination of replication occurs where the replication forks meet, not in any predetermined point. Thus the leading and lagging roles of strands may change. Eukaryotic replicons are much shorter than prokaryotic ones, ranging from 10 to 300 kb [125,126], also Okazaki fragments are 10 times shorter in eukaryotes than in *E. coli* [51], so the replication-associated mutational pressure is expected to be much weaker in eukaryotes than in prokaryotes [127]. When analyzing (G-C) and (A-T) or GC skew and AT skew, one can observe asymmetry in yeast and human chromosomes [9,128,129], but as the precise location and firing patterns for ARSes are not known, it is not clear if the observed trends are connected with replication-associated mutational pressure. Even finding exact location of a replication origin present in a few primate genomes did not allow for detecting replication-related asymmetry [130]. Thus it is not easy to find asymmetry resulting from replication in eukaryotic genomes. The difficulties might result also from overlapping biases introduced by transcription-associated mutational pressure, recombination, and selection pressure [127].

Chromosome ends are sequence fragments with predetermined leading or lagging roles, from the last ARS to the telomere. Indeed, asymmetry connected with replication has been found in these sequences [9,128]. Another approach was to examine sites under weak selection (introns and four fold degenerate sites in codons) in neighboring genes from Watson and Crick strands, which are located most probably on differently replicated strands [127]. The detected GC and AT skews proved that there is asymmetry resulting from replication mechanism and it is possible to differentiate it from asymmetry introduced during transcription.

Transcription-associated asymmetry in eukaryotic genomes

Transcription increases single-strand deamination [131] and favors transcription-coupled repair [132], leading to pronounced asymmetries. [133] analyzed substitution rates by comparing a

sequence containing 9 genes, orthologous in 9 mammal species. They found excess of G over C and T over A resulting from excess of purine transitions and deficit of pyrimidine transitions in transcribed regions. They explained it as a byproduct of transcription-coupled repair acting on the mismatched base pairs that result from uncorrected DNA polymerase substitution errors during DNA replication [133]. [134] analyzed intron sequences in the whole set of human genes and included the role of transversions in creating the observed GC and AT skews. The same group [135] later analyzed introns in mammalian, invertebrate and plant genomes and discovered two different biases. Intron ends show selection-driven bias associated with splicing. This bias is greatest at intron extremities and decreases to zero in the internal region. Internal regions of introns exhibit a different, constant bias resulting from transcription-associated mutation pressure and repair mechanisms. However, transcription-associated skews differ between mammals and invertebrates, suggesting different mutation patterns or repair mechanisms [135].

Isochores in mammalian and bird genomes

When analyzing structure of mammalian and bird chromosomes, the most obvious compositional feature are isochores, very long (over 300 kb) stretches of DNA of similar GC content (see [136] for an excellent review). G+C rich parts of the genome are abundant in genes and in short interspersed repetitive DNA elements (SINES), contrary to G+C poor regions, where there are fewer genes and long interspersed repetitive DNA elements (LINES) prevail. G+C rich regions are actively transcribed, and correspond to an open chromatine structure [137], are associated with CpG islands, and housekeeping genes are preferentially located there [138]. These regions are characterized also by higher recombination frequency and early replication.

Isochores were first observed by the group of Giorgio Bernardi in the early seventies [139], but up till now their presence and persistence in genomes has not been adequately explained by either mutational or selective mechanisms. Since the nucleotide composition of genes, especially third positions in codons, is correlated with the GC level of the whole isochore, it seems that mutation bias must play a role in formation of isochores. [140] observed that as different regions of chromosome are replicated early or late, they must have different mutation patterns because levels of free nucleotides vary in the cell cycle which influences the pattern of base misincorporation. Other mutational causes could be variation in the efficiency of DNA repair [141], or cytosine deamination [142]. On the other hand, because isochores are typical of the homeotherms – birds and animals, and have been conserved on large evolutionary distances, Giorgio Bernardi has argued that they are a consequence of natural selection (see [143] for review). If one assumes that the threshold of optimal GC values increased with increasing body temperature from cold- to warm-blooded vertebrates, the observed increase in GC levels in the genomes of mammals and birds could be

explained by negative selection [144]. Intragenic analysis of synonymous positions of quartet codons showed that there is negative selection on maintaining GC-rich genes, while the changes in GC-poor genes are mostly neutral [145,146].

The role of the genetic code in the interplay between mutation and selection pressures

The composition of genomes is shaped by mutation and selection pressures. Selection pressure can choose among the results of mutation pressure, but it seems that also mutation pressure is not random and it is tuned to the requirements of selection. The most frequent mutations, nucleotide substitutions, are the least deleterious ones. It is a result of the structure of the genetic code and its degeneracy – most amino acids are encoded by several different codons. The degeneracy of the genetic code seems to be related to the GC content of codons. The codons with the same combination of G and/or C in the first two positions code for the same amino acid independently of what is in the third position. It can be explained by the chemistry of DNA: G and C are connected by a triple hydrogen bond which is strong enough to ensure coding by the first two positions in the codon, independently of which nucleotide is in the third position, while A and T are connected by two hydrogen bonds and in codons containing A and/or T in the first two positions the third position plays a role in the codon-anticodon interaction [147,148]. Assignment of codons to amino acids is also not random: codons for the five most hydrophobic amino acids (phenylalanine, isoleucine, methionine, valine, and leucine) have T in second positions and differ only in the first positions, while codons for the six most hydrophilic amino acids have adenine in the second positions [149-151]. Amino acids encoded by triplets with C in second positions have intermediate values of hydrophobicity [151]. Codons for amino acids with similar chemical properties are close in the table of the genetic code: codons for acidic amino acids (aspartic acid and glutamic acid) differ only in third positions and their amino derivatives (asparagine and glutamine) differ only in first positions; codons for the three basic amino acids (lysine, arginine and histidine) also differ only in single positions; similarly for codons for aromatic amino acids (phenylalanine, tyrosine and tryptophane) and amino acids containing hydrophilic groups (serine, threonine and tyrosine). Thus, substitutions changing the meaning of a codon often lead to changing an amino acid for a similar one.

The genetic code is universal for both prokaryotes and eukaryotes, the whole living world. It has been observed that the prevalence of amino acids in proteins is correlated with the number of codons representing a given amino acid in the genetic code [152], and that the least stable codons correspond to the least represented amino acids which in turn are the best preserved by selection [153,154]. One could assume that the optimal genetic code is the one which minimizes the number of amino acid subsitutions resulting from nucleotide substitutions. Such condition enables

quantifying the quality of the genetic code. Computer simulations have shown that the existing genetic code is much more resistant to point mutations than randomly generated codes [155]. A genome evolution model based on the *Borrelia burgdorferi* genome has shown that nucleotide composition of genes, mutational pressure and the genetic code are optimized to give relatively low level of gene elimination by selection while maintaining high sequence diversity (Table 1) [156]. A change in the level of degeneration of the genetic code led to increase in the number of genes "killed" in the simulation. Also, using symmetric matrix of nucleotide substitutions or symmetrization of amino acid composition of the evolving sequence caused increase in the rate of gene elimination (Table 1). Thus the present code seems quite optimal, although it is possible to find a code which would be better adapted to a given combination of mutational and selection pressures. However, a code fitted to the genes from the leading strand would probably increase mutation rate for the genes from the lagging strand. Even if we could find a code better for both the genes from the leading and lagging strands from one genome, it is unlikely that it would be better for any other genome, not to mention the whole biosphere. The observed asymmetries, mutation and selection pressures vary among genomes, and the cases where the code has changed are few and under limited conditions – it seems that they have had to adapt to the universal code [156]. Thus the question about a "better" code for the whole biosphere seems irrelevant because organisms have already adapted to the existing code.

References

[1] Yuzhakov, A., Turner, J., and O'Donnel, M. 1996, Cell, 86, 877.
[2] Bruck, I., and O'Donnel, M. 2000, J. Biol. Chem., 275, 28791.
[3] Dervyn, E., Suski, C., Daniel, R., Bruand, C., Chapuis, J., Errington, J., Jannière, L., and Ehrlich, S. D. 2001, Science, 294, 1716.
[4] Fijałkowska, I. J., Jonczyk, P., Maliszewska-Tkaczyk, M., Bialoskorska, M., and Schaaper, R. M. 1998, Proc. Natl. Acad. Sci. USA, 95, 10020.
[5] Lobry, J. R. 1995, J. Mol. Evol., 40, 326. Erratum 41: 680.
[6] Lobry, J. R. 1996, Mol. Biol. Evol., 13, 660.
[7] Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M. R., and Cebrat, S. 2001, J. Appl. Genet., 42(4), 553.
[8] Freeman, J. M., Plasterer, T. N., Smith, T. F., and Mohr, S. C. 1998, Science, 279, 1827.
[9] Grigoriev, A. 1998, Nucl. Acids Res., 26, 2286.
[10] McLean, M. J., Wolfe, K. H., and Devine, K. M. 1998, J. Mol. Evol., 47, 691.
[11] Mrazek, J., and Karlin S. 1998, Proc. Natl. Acad. Sci. USA, 95, 3720.
[12] Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M. R, and Cebrat, S. 1999, Genome Res., 9, 409.
[13] Rocha, E. P. C., Danchin A., and Viari, A. 1999, Mol. Microbiol., 32, 11.
[14] Tillier, E. R. M., and Collins R. A. 2000, J. Mol. Evol., 50, 249.
[15] Perrière, G., Lobry, J. R., and Thioulouse, J. 1996, CABIOS 12, 519.
[16] Lafay, B., Lloyd, A. T., McLean, M. J., Devine, K. M., Sharp, P. M., and Wolfe, K.H. 1999, Nucleic Acids Res., 27, 1642.

[17] Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M. R, and Cebrat, S. 1999., J. Appl. Genet., 40(1), 1.
[18] Lobry, J. R. 1996, Science, 272, 745.
[19] Picardeau, M., Lobry, J. R., and Hinnebusch, B. J. 1999 Mol. Microbiol., 32, 437.
[20] Zawilak, A., Cebrat, S., Mackiewicz, P., Król-Hulewicz, A., Jakimowicz, D., Messer, W., Goscinak G., and Zakrzewska-Czerwinska, J. 2001, Nucleic Acids Res., 29, 2251.
[21] Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M. R., and Cebrat, S. 2004, Nucl. Acids. Res., 32, 3781.
[22] Zhang, R., and Zhang, C. T. 2004, Archaea 1,5,1 (http://archaea.ws/archive/pdf/volume1/issue5/1-Zhang.pdf)
[23] Lobry, J. R., and Sueoka, N. 2002, Gen. Biol., 3:RESEARCH0058.
[24] Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M. R., and Cebrat, S. 2001, BMC Evol. Biol., 1, 13.
[25] Frank, A. C., Lobry, J. R. 1999, Gene, 238, 65.
[26] Lindahl, T., and Nyberg, B. 1974, Biochemistry, 13, 3405.
[27] Frederico, L. A., Kunkel, T. A., and Shaw B. R. 1990, Biochemistry, 29, 2532.
[28] Lutsenko, E., and Bhagwat, A. S. 1999, Mutat. Res., 437, 11.
[29] Echols, H., and Goodman, M. F. 1991, Annu. Rev. Biochem., 60, 477.
[30] Gawel, D., Maliszewska-Tkaczyk, M., Jonczyk, P., Schaaper, R. M., and Fijalkowska, I. J. 2002, Mutat. Res., 501, 129.
[31] Rocha, E. P. C., and Danchin, A. 2001, Mol. Biol. Evol., 18(9), 1789.
[32] Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C. H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M. A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B. G., Parkhill, J., and Hopwood, D. A. 2002, Nature, 417(6885), 141.
[33] Trinh, T. Q., and Sinden, R. R. 1991, Nature, 352, 544.
[34] Iwaki, T., Kawamura, A., Ishino, Y., Kohno, K., Kano, Y., Goshima, N., Yara, M., Furusawa, M., Doi, H., and Imamoto, F. 1996, Mol. Gen. Genet., 251, 657.
[35] Szczepanik, D., Mackiewicz, P., Kowalczuk, M., Gierlik, A., Nowicka, A., Dudek, M. R., and Cebrat, S. 2001, J. Mol. Evol., 52, 426.
[36] Francino, M. P., Chao, L., Riley, M. A., and Ochman, H. 1996, Science, 272, 107.
[37] Francino, M. P., and Ochman, H. 1997, Trends Genet., 13, 240.
[38] Shepherd, J. C. 1981, Proc. Natl. Acad. Sci. USA, 78, 1596.
[39] Smithies, O., Engels, W. R., Devereux, J. R., Slightom, J. L. and Shen, S. H. 1981, Cell, 26, 345.
[40] Wong, J. T., and Cedergren, R. 1986, Eur. J. Biochem., 159, 175.
[41] Karlin, S., and Burge, C. 1995, Trends Genet., 11, 283.
[42] Cebrat, S., Dudek, M. R., Mackiewicz, P., Kowalczuk, M., and Fita, M. 1997, Micr & Comp. Genomics, 2(4), 259.
[43] Cebrat, S., Dudek, M. R., and Mackiewicz, P. 1998, Theory Bioscienc., 117, 78.
[44] Gutierrez, G., Marquez, L., and Martin, A. 1996. Nucleic Acids Res., 24, 2525.
[45] Lagunez-Otero, J., and Trifonov, E. N. 1992, J. Biomol. Struct. Dyn., 10, 455.
[46] Rocha, E. P. C. 2002, Trends Microbiol., 10(9), 393.
[47] French, S. 1992, Science, 258, 1362.
[48] Deshpande, A. M., and Newlon, C. S. 1996, Science, 272, 1030.
[49] Ellwood, M., and Nomura, M. 1982, J. Bacteriol., 149, 458.
[50] Brewer, B. J. 1988, Cell, 53, 679.
[51] Rocha, E. P. C. 2004, Microbiology, 150, 1609.
[52] Rocha, E. P. C., and Danchin, A. 2003, Nat. Genet., 34, 377.

[53] Rocha, E. P. C., and Danchin, A. 2003, Nucleic Acids Res., 31, 6570.
[54] Pakula, A. A., and Sauer, R. T. 1989, Annu. Rev. Genet., 23, 289.
[55] Deschavanne, P., and Filipski, J. 1995, Nucleic Acids Res., 23, 1350.
[56] Daubin, V., and Perrière, G. 2003, Mol. Biol. Evol., 20(4), 471.
[57] Sharp, P. M., Shields, D. C., Wolfe, K. W., and Li, W.-H. 1989, Science, 246, 808.
[58] Rocha, E. P. C., and Danchin, A. 2002, Trends Genet., 18(6), 291.
[59] Guindon, S., and Perrière, G. 2001, Mol. Biol. Evol., 18(9), 1838.
[60] Chandler, M. G., and Pritchard, R. H. 1974, Mol. Gen. Genet., 138, 522.
[61] Louarn, J. M., Bouche, J. P., Legendre, F., Louarn, J., and Patte, J. 1985, FEMS Microbiol. Rev., 26, 533.
[62] Schmid, M. B., and Roth, J. R. 1987, J. Bacteriol. 169, 2872.
[63] Errington, J., Daniel, R. A., and Scheffers, D. J. 2003, Microbiol. Mol. Biol. Rev., 67, 52.
[64] Dworkin, J., and Losick, R. 2002, Proc. Natl. Acad. Sci. USA, 99, 14089.
[65] Rocha, E. P. C., Fralick, J., Vediyappan, G., Danchin, A., and Norris, V. 2003, Mol. Microbiol., 49(4), 895.
[66] Corre, J., Cornet, F., Patte, J., and Louarn, J. M. 1997, Genetics, 147, 979.
[67] Lawrence, J. G., and Ochman, H. 1998, Proc. Natl. Acad. Sci. USA, 95, 9413.
[68] 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, Nature, 390, 249.
[69] Corre, J., Patte, J., and Louarn, J. M. 2000, Genetics, 154, 39.
[70] Danchin, A. 2003, Curr. Issues Mol. Biol., 5(2), 37.
[71] Tillier, E. R. M., and Collins, R. A. 2000, J. Mol. Evol., 51, 459.
[72] Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Dudkiewicz, M., Dude,k M. R., and Cebrat, S. 2003, J. Appl. Genet., 44(4), 561.
[73] Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., and Cebrat, S., 2001, Gen. Biol., 2(12), 1004.1.
[74] Brookfield, J. F. 1997, Adv Genet., 36, 137.
[75] Lynch, M., and Conery, J. S. 2000, Science, 10, 1151.
[76] Conery, J. S., and Lynch, M. 2001, Pac. Symp. Biocomput., 167.
[77] Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. 2002, Genome Biol., 3(2),research0008.
[78] 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., and Kurland, C. G. 1998, Nature, 396, 133.
[79] Andersson, J. O., and Andersson, S. G. 2001, Mol. Biol. Evol., 18, 829.
[80] Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Ganier, T., Churcher, C., Harris, D. et al. 2001, Nature, 409, 1007.
[81] Mira, A., Ochman, H., and Moran, N. A. 2001, Trends Genet., 17, 589.
[82] Homma, K., Fukuchi, S., Kawabata, T., Ota, M., and Nishikawa, K. 2002, Gene, 294, 25.
[83] Liu, Y., Harrison, P. M., Kunin, V., and Gerstein, M. 2004, Genome Biol., 5, R64.
[84] Ohno, S. 1970, Evolution by gene duplication, George Allen and Unwin, London.
[85] Barker, W. C., and Dayhoff, M. O. 1980, Bio-Science, 30, 593.
[86] Li, W. H. 1997, Molecular Evolution, Sunderland, MA:Sinauer Associates.
[87] Mackiewicz, P., Dudkiewicz, M., Kowalczuk, M., Mackiewicz, D., Banaszak, J., Polak, N., Smolarczyk, K., Nowicka, A., Dudek, M. R., and Cebrat, S. 2004, Lecture Notes in Computer Science, 3039, 687.
[88] Dudkiewicz, M., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Nowicka, A., Polak, N., Smolarczyk, K., Banaszak, J., Dudek, M. R., and Cebrat, S. 2004, Biosystems (in press)
[89] Mushegian, A. R., and Koonin, E. V. 1996, Trends Genet., 12, 289.
[90] Kolsto, A. B. 1997, Mol. Microbiol., 24, 241.
[91] Watanabe, H., Mori, H., Itoh, T., and Gojobori, T. 1997, J. Mol. Evol.. 44. (Suppl. 1), S57.

[92] Bellgard, M. I., Itoh, T., Watanabe, H., Imanishi, T., and Gojobori, T. 1999, Ann. N. Y. Acad. Sci., 18, 293.
[93] Itoh, T., Takemoto, K., Mori, H. and Gojobori, T. 1999, Mol. Biol. Evol., 16, 332.
[94] Hughes, D. 2000, Genome Biol., 1(6), REVIEWS0006.
[95] Mira, A., and Ochman, H. 2002, Mol. Biol. Evol., 19, 1350.
[96] Sharp, P. M. 1991, J. Mol. Evol., 33, 23.
[97] Birky, C. W. Jr., and Walsh, J. B. 1992, Genetics, 130, 677.
[98] Goodman, M. F. 2000, Trends Biochem. Sci., 25, 189.
[99] Courcelle, J., and Hanawalt P. C. 2001, Proc. Natl. Acad. Sci. USA, 98, 8196.
[100] Bierne, H., Ehrlich, S. D., and Michel, B. 1997, EMBO J., 16, 3332.
[101] Rocha, E. P. C. 2003, Trends. Genet., 19, 600.
[102] Frank, A. C., Amiri, H., and Andersson, S. G. 2002, Genetica, 115, 1.
[103] Achaz, G., Coissac, E., Netter, P., and Rocha, E. P. C. 2003, Genetics, 164, 1279.
[104] Eisen, J. A., Heidelberg, J. F., White, O., and Salzberg, S. L. 2000, Genome Biol., 1(6),research0011.
[105] Read, T. D., Brunham, R. C., Shen, C., Gill, S. R., Heidelberg, J. F., White, O., Hickey, E. K., Peterson, J., Umayam, L. A., Utterback, T. et al. 2000, Nucleic Acids Res., 28, 1397.
[106] Tillier, E. R. M., and Collins, R. A. 2000, Nat. Genet., 26, 195.
[107] Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, et al. 2000, Nucleic Acids Res., 28, 4317.
[108] Suyama, M., and Bork, P. 2001, Trends Genet., 17, 10.
[109] Moran, N. A., and Mira, A. 2001, Genome Biol.2(12), RESEARCH0054.
[110] Zivanovic, Y., Lopez, P., Philippe, H., and Forterre, P. 2002, Nucleic Acids Res., 30, 1902.
[111] Lemon, K. P., and Grossman, A. D. 1998, Science, 28, 1516.
[112] Newport, J., and Yan, H. 1996, Curr. Opin. Cell Biol., 8, 365.
[113] Segall, A., Maha,n M. J., and Roth, J. R. 1988, Science, 241, 1314.
[114] Liu, S. L., and Sanderson, K. E. 1995, Proc. Natl. Acad. Sci. USA, 92, 1018.
[115] Liu, S. L., and Sanderson, K. E. 1996, Proc. Natl. Acad. Sci. USA, 93, 10303.
[116] Wu, L. J., and Errington, J. 2002, EMBO J., 21, 4001.
[117] Campo, N., Dias, M. J., Daveran-Mingot, M. L., Ritzenthaler, P., and Le Bourgeois, P. 2004, Mol. Microbiol., 52, 511.
[118] Ikemura, T. 1981, J. Mol. Biol., 151, 389.
[119] Gouy, M., and Gautier, C. 1982, Nucleic Acids Res., 10, 7055.
[120] Sharp, P. M., and Li, W. H. 1987, Nucleic Acids Res., 15, 1281.
[121] Mackiewicz, D., Mackiewicz, P., Kowalczuk, M., Dudkiewicz, M., Dudek, M. R., and Cebrat, S. 2003, Acta Microbiologica Polonica, 52(3), 245.
[122] Koonin, E. V., and Galperin, M. Y. 1997, Curr. Opin. Genet. Dev., 7, 757.
[123] Tamames, J., Ouzounis, C., Casari, G., and Valencia, A. 1997, J. Mol. Evol., 44, 66.
[124] Mackiewicz, P., Szczepanik, D., Gierlik, A., Kowalczuk, M., Nowicka, A., Dudkiewicz, M., Dudek, M. R., and Cebrat, S. 2001, J. Mol. Evol., 53(6), 615.
[125] Brewer, B. J., and Fangman, W. L. 1993, Science, 262, 1728.
[126] Hyrien, O., Marheineke, K., and Goldar, A. 2003, Bioessays, 25, 116.
[127] Niu, D. K., Lin, K., and Zhang, D.-Y. 2003, J. Mol. Evol., 57, 325.
[128] Gierlik, A., Kowalczuk, M., Mackiewicz, P., Dudek, M. R., and Cebrat, S. 2000, J. Theor. Biol., 202, 305.
[129] Shioiri, C., and Takahata, N. 2001, J. Mol. Evol., 53(4-5), 364.
[130] Francino, M. P., and Ochman, H. 2000, Mol. Biol. Evol., 17(3), 416.
[131] Beletskii, A., and Bhagwat, A. S. 1996, Proc. Natl. Acad. Sci. USA, 93, 13919.
[132] Svejstrup, J.Q. 2002, Nat. Rev. Mol. Cell. Biol., 3(1), 21.
[133] Green, P., Ewing, B., Miller, W., Thomas, P. J., and Green, E. D., NISC Comparative Sequencing Program, 2003, Nat. Genet., 33, 514.

[134] Touchon, M., Nicolay, S., Arneodo, A., d'Aubenton-Carafa, Y., and Thermes, C. 2003, FEBS Letters, 555, 579.
[135] Touchon, M., Arneodo, A., d'Aubenton-Carafa, Y., and Thermes, C. 2004, Nucl. Acids Res., 32(17), 4969.
[136] Eyre-Walker, A., and Hurst, L. D. 2001, Nat. Rev., 2, 549.
[137] Kerem, B. S., Goiten, R., Diamond, G., Cedar, H., and Marcus, M. 1984, Cell, 38, 493.
[138] Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. 1992, Genomics, 13, 1095.
[139] Filipski, J., Thiery, J. P., and Bernardi, G. 1973, J. Mol. Biol., 80, 177.
[140] Wolfe, K. H., Sharp, P. M., and Li, W.-H. 1989, Nature, 337, 283.
[141] Filipski, J. 1987, FEBS Lett., 217, 184.
[142] Fryxell, K., and Zuckerlandl, E. 2000, Mol. Biol. Evol., 17, 1371.
[143] Bernardi, G. 2000, Gene, 241, 3.
[144] Bernardi, G. 1993, Mol. Biol. Evol., 10, 186.
[145] Cacciò, S., Zoubak, S., D'Onofrio, G., and Bernardi, G. 1995, J. Mol. Evol., 40, 280.
[146] Zoubak, S., D'Onofrio, G., Cacciò, S., and Bernardi, G. 1995, J. Mol. Evol., 40, 293.
[147] Lagerkvist, U. 1978, Proc. Natl. Acad. Sci. USA ,75, 1759.
[148] Lagerkvist, U. 1980, Amwer. Scient., 68, 192.
[149] Woese, C. R. 1965, Proc. Natl. Acad. Sci. USA, 54, 71.
[150] Volkenstein, M. V. 1966, Biochim. Biophys. Acta, 119, 421.
[151] Woese, C. R., Dugre, W. C., Dugre, S. A., Kondo, M. et al. 1966, Cold Spring Harb. Symp. Quant.Biol. 31, 723.
[152] King, J. L., and Jukes, T. H. 1969, Science, 164, 788.
[153] Nowicka, A., Mackiewicz, P., Dudkiewicz, M., Mackiewicz, D., Kowalczuk, M., Cebrat, S., and Dudek, M. R. 2003, Lecture Notes in Computer Science, 2658, 650.
[154] Nowicka, A., Mackiewicz, P., Dudkiewicz, M., Mackiewicz, D., Kowalczuk, M., Banaszak, J., Cebrat, S., and Dudek, M. R. 2004, Applied Bioinformatics, 3(1), 31.
[155] Gilis, D., Massar, S., Cerf, N. J., and Rooman, M. 2001, Genome Biol., 2(11), 49.1.
[156] Dudkiewicz, M., Mackiewicz, P., Nowicka, A., Kowalczuk, M., Mackiewicz, D., Polak, N., Smolarczyk, K., Banaszak, J., Dudek, M. R., and Cebrat, S. 2004, Fut. Gener. Comp. Sys. (in press).

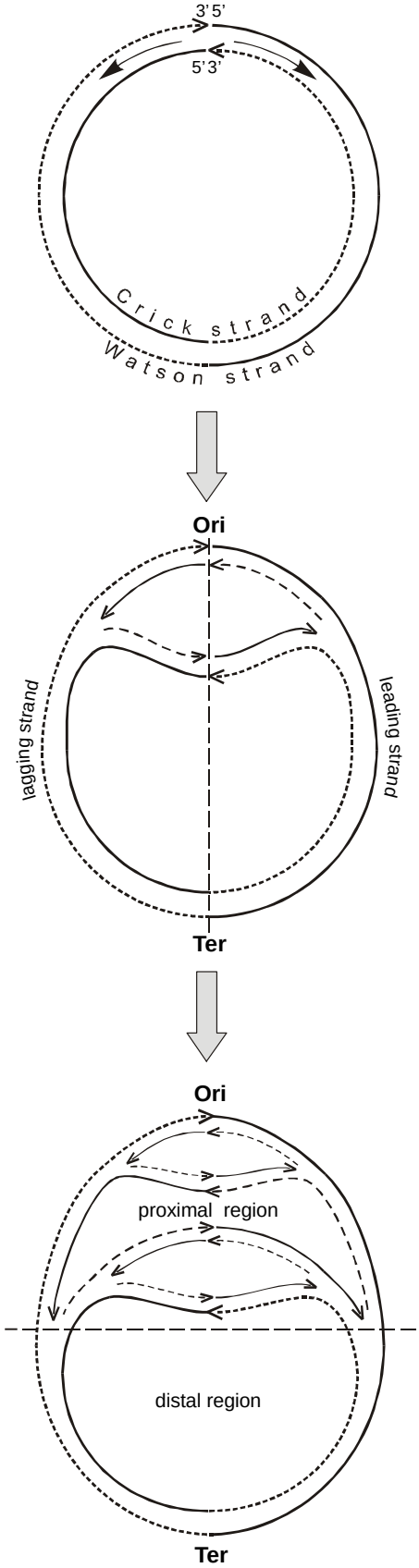

Fig 1. Organization and replication of the bacterial chromosome. Top picture: bacterial chromosome before replication. The inner circle represents Crick strand, the outer one – Watson strand. Black arrows show direction of replication forks movement. Middle picture: the beginning of replication, which starts from the ori region to finish in the ter region, and formation of replication bubble. Solid lines denote strands replicated in the previous replication round as leading, dashed lines – as lagging. Bottom picture: in the log phase of bacterial growth the next replication round may begin before the previous replication round is over. Division of the chromosome into regions proximal and distal in respect to origin of replication (ori).

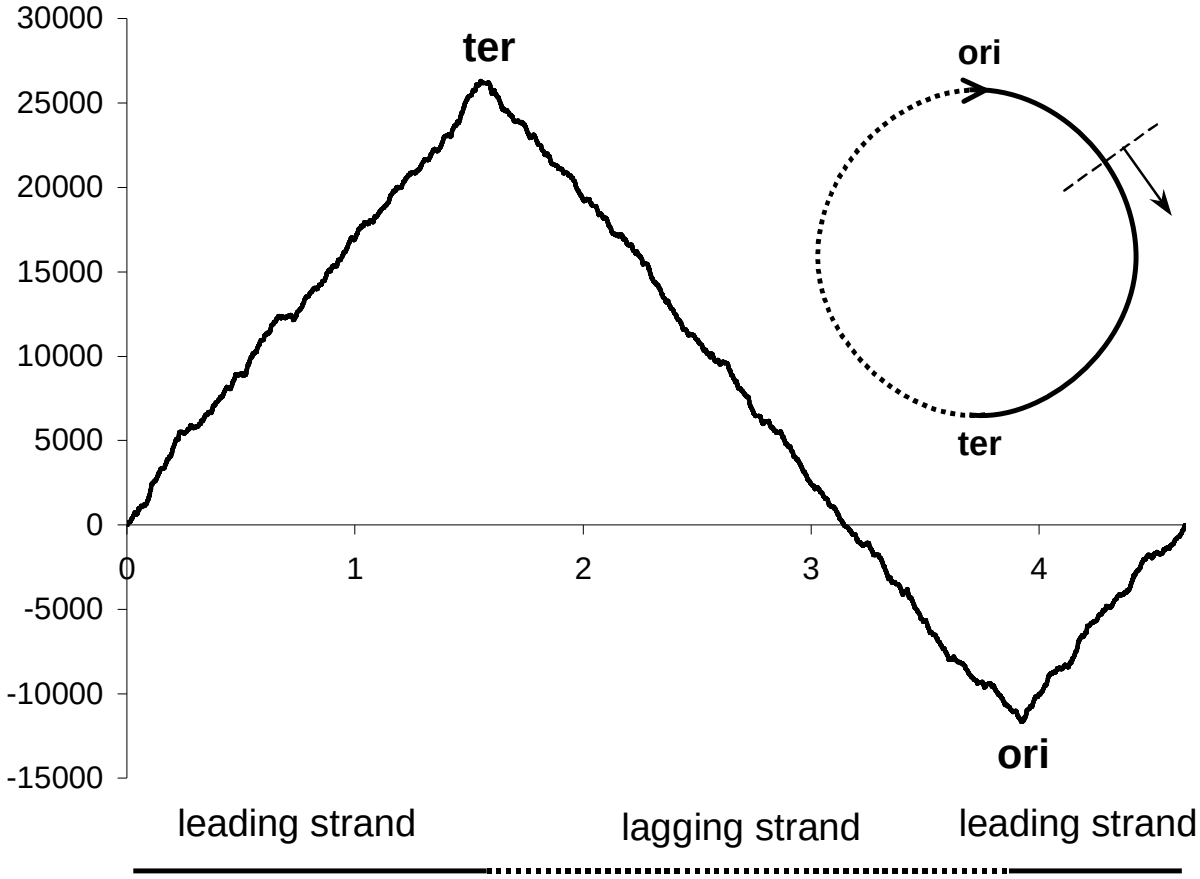

Fig 2. Analysis of asymmetry in the *E. coli* chromosome by DNA walks. Virtual walker goes along the Watson strand of the chromosome, and moves up and forward when it meets a guanine, down and forward when it meets a cytosine, and only forward when it meets an adenine or a thymine. Thus the X axis represents location on chromosome in Mbp and the Y axis measure of asymmetry. In this way we can analyse the difference between complementary nucleotides [G-C] along the chromosome. Also, the global trend of the whole sequence was detracted so the walk begins and ends in Y=0 which emphasizes local trends in asymmetry. The extrema of asymmetry connected with the change of strand from leading to lagging or vice versa correspond to the region of replication initiation (ori) and termination (ter). Dashed line with an arrow on the chromosome scheme (insertion) shows the beginning of the DNA walk.

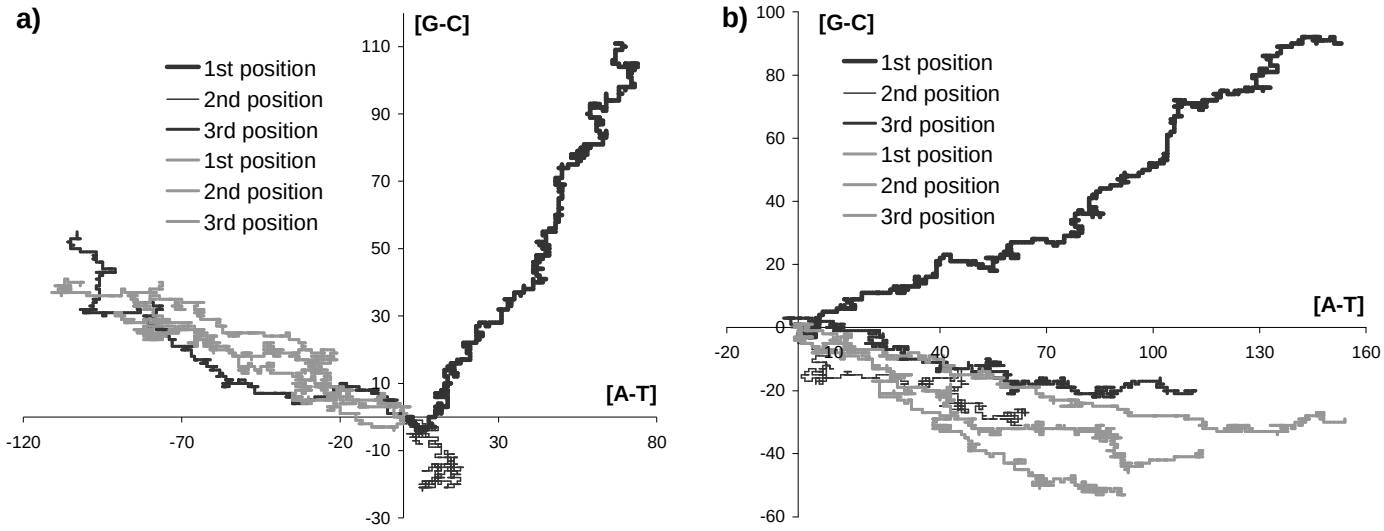

Fig 3. Analysis of two genes from the *Borrelia burgdorferi* genome located on the leading strand – BB0020 (a) and on the lagging strand – BB0040 (b) by "spider" DNA walks. Nucleotides from each position in the codon were analysed separately, so there are three walks for each gene (black lines). Grey lines represent walks on the same genes but after evolution under mutational pressure connected with replication process characteristic for this geneome (Kowalczuk et al. 2001b), without any selection pressure. The virtual walker went up when it encountered guanine, down for cytosine, left for adenine and right for thymine.

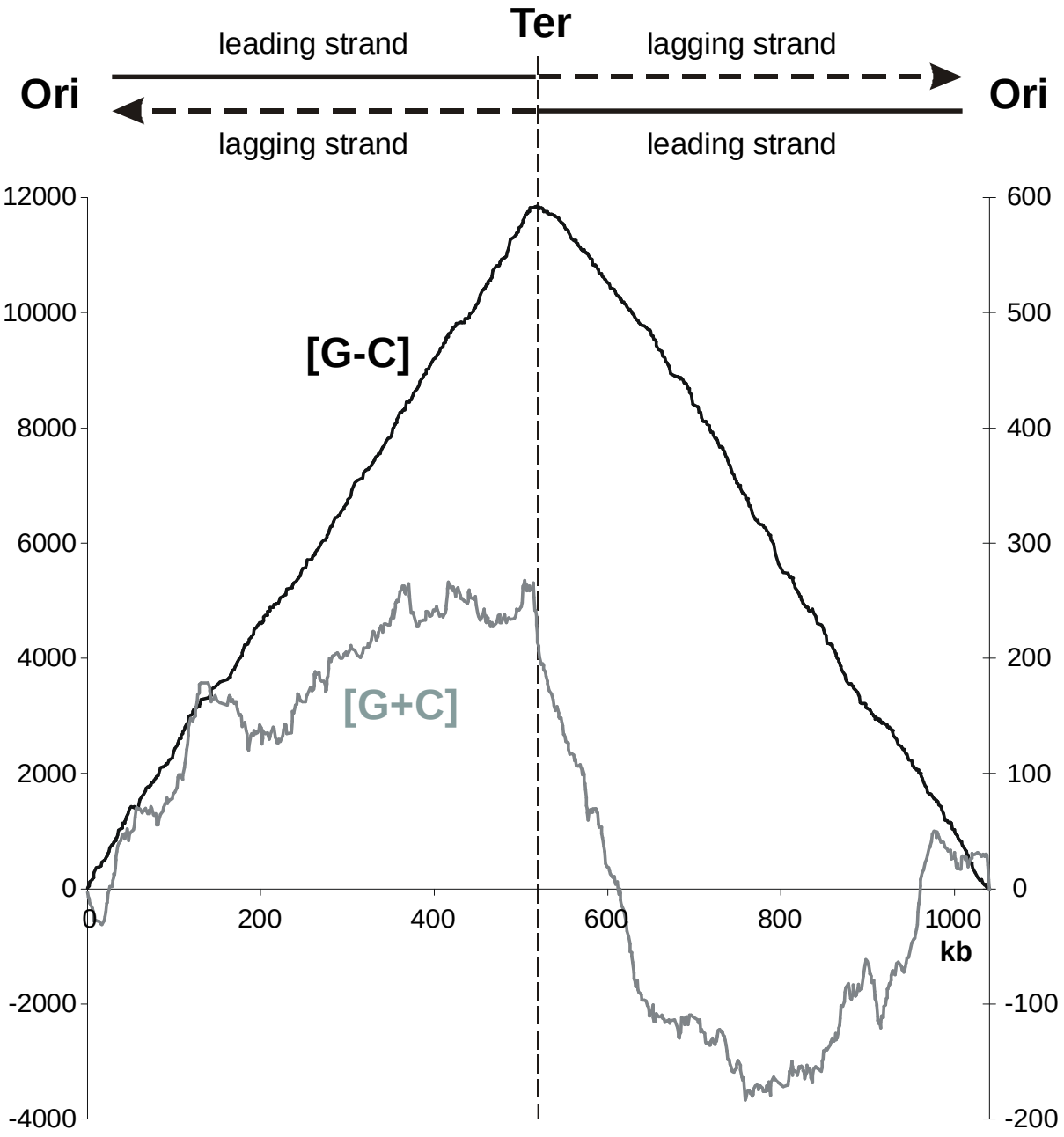

Fig 4. Analysis of asymmetry in the *Chlamydia trachomatis* chromosome by DNA walks. The walks were performed as discribed in caption for Fig 2, except that here only third positions in codons in protein coding sequences were analysed.

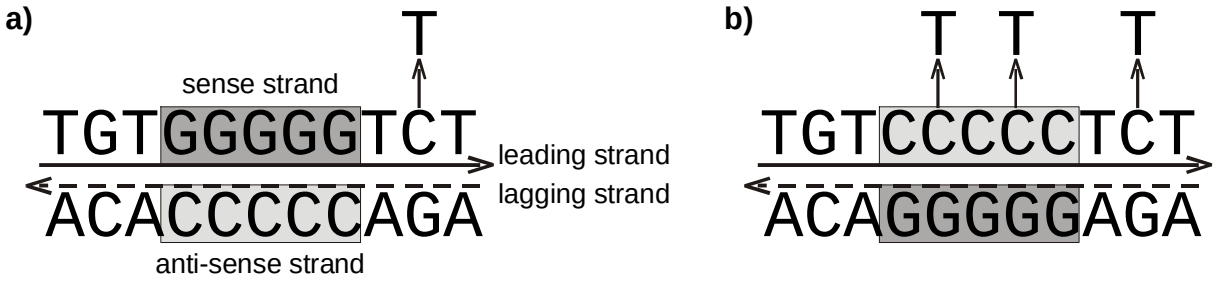

Fig 5. Schematic representation of sense and antisense strands of protein coding sequence (shadowed boxes) on leading and lagging strands of the chromosome. The sense strand of protein coding sequence is rich in guanine, thus genes located on the leading strand (a) are less susceptible to C→T transitions than genes located on the lagging strand (b).

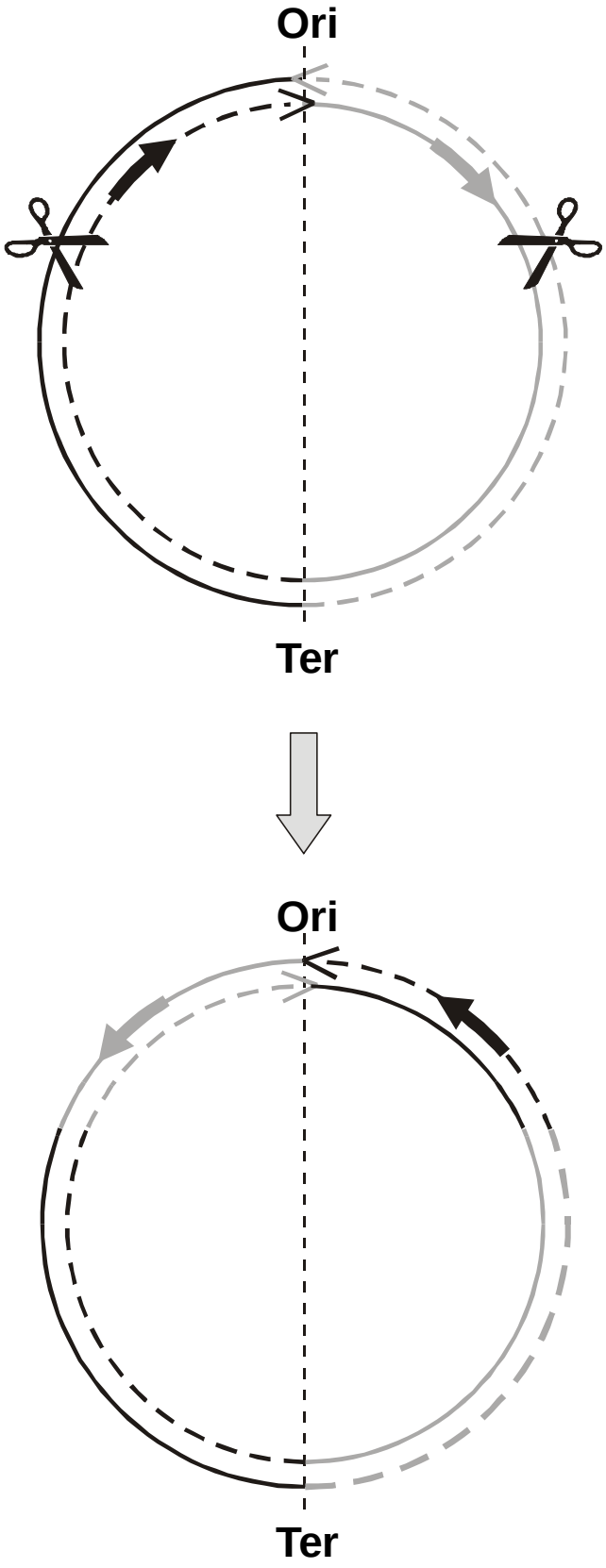

Fig 6. A symmetrical inversion encompassing the origin of replication. Such inversions do not change the distance of the genes to the origin and terminus of replication and their location on leading and lagging strands. Also, the lengths of replichores remain equal.

Table 1. Computer simulations of DNA sequence evolution

DNA strand	**Number of eliminated genes**			
	Standard simulation	One-parameter substitution matrix	Uniform amino acid composition	Redefined meanings of codons
Leading	84	92	222	97
Lagging	30	69	122	40

Results of computer simulations of evolution of the *Borrelia burgdorferi* chromosome. In the standard simulation the chromosome evolved under the actual mutational pressure associated with replication (Kowalczuk et al. 2001b), and the amino acid composition and the genetic code were as in Nature. A gene was eliminated and replaced by its allele from a simultaneously evolving chromosome when it accumulated amino acid substitutions above an assumed threshold. In the next simulation as mutational pressure one parameter substitution matrix was used, all other parameters remained unchanged. Uniform amino acid composition means that the fraction of each amino acid in the sequence was the same, but codon preferences and lengths of genes were as in the real *B. burgdorferi* genome. In the last simulation blocks of codons coding for one amino acid were redefined to code for another amino acid which had the same level of degeneracy, i.e. if an amino acid was encoded by four codons, these codons could be ascribed to another amino acid also coded by a four codon block.