ELSEVIER

Gene 238 (1999) 65–77

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

# Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms

A.C. Frank, J.R. Lobry *

*CNRS UMR 5558, Biométrie, Biologie Evolutive, Université Claude Bernard, 43 Bd. du 11-NOV-1918, F-69622 Villeurbanne, France*

Received 16 February 1999; received in revised form 1 June 1999; accepted 6 July 1999; Received by G. Bernardi

**Abstract**

In the absence of bias between the two DNA strands for mutation and selection, the base composition within each strand should be such that A=T and C=G (this state is called Parity Rule type 2, PR2). At a genome scale, i.e. when considering the base composition of a whole genome, PR2 is a good approximation, but there are local and systematic deviations. The question is whether these deviations are a consequence of an underlying bias in mutation or selection. We have tried to review published hypotheses to classify them within the mutational or selective group. This dichotomy is, however, too crude because there is at least one hypothesis based simultaneously upon mutation and selection. © 1999 Elsevier Science B.V. All rights reserved.

*Keywords:* Base composition; Codon usage; Gene orientation; Replication origin; Skew; Strand asymmetries

## 1. Introduction

Chargaff (1950) experimentally determined A=T and G=C equimolar frequencies when analysing both DNA strands together. Three years later, Watson and Crick (1953) determined the DNA secondary structure and stated the base-pairing rules that explain such frequencies. More surprisingly, these equalities are still observed within each strand (Lin and Chargaff, 1967). Under *no-strand-bias* conditions, when mutation and selection have equal effect on both strands, there are six possible substitution rates instead of 12, as stated in the Parity Rule type 1, PR1 (Sueoka, 1995). The rationale for this is as follows: since substitution rates are scored on one strand, a change such as T to C in a given strand results from either a T→C substitution on that strand or an A→G on the complementary strand. The type-2 parity rule, PR2 can be formally derived from PR1 (Lobry, 1995) to give the base frequencies within each strand at equilibrium: A=T and G=C. Moreover, convergence to PR2 is also expected when the substitution rates are not constant over time (Lobry and Lobry, 1999). Any deviation from PR2 implies asymmetric substitution: the result of different mutation rates, different selective pressures, or both, between the two strands of DNA. There are two principal ways of studying asymmetric substitution; phylogenetic reconstruction of base substitution and detection of deviations from PR2.

In the first method, asymmetries are detected by aligning homologous sequences, estimating the substitution matrix and comparing the frequencies of complementary changes. Wu and Maeda (1987) used this method to test for asymmetric substitution in a region of the β-globin complex of primates. They did detect asymmetries, but since origins and termini of replication of their data were not known, their results are not reliable, as shown by Bulmer (1991), who re-examined an adjacent region of the human β-globin complex. Francino et al. (1996) used the same method to search for asymmetric substitution in eubacteria. They found a difference between complementary changes C→T and G→A when scoring substitutions on the coding strand. The advantage of this method is that it directly detects the number and type of substitutions, but the access to a suitable data set is rather limited because of the difficulty of finding orthologous sequences with an adequate divergence time.

The second method (Lobry, 1996a) builds on the analysis of the DNA sequences for deviations from A=T and G=C frequencies. In 1990 such deviations in SV40 were interpreted as evidence for asymmetric mutation pressure because of a polarity switch at the origin

* Corresponding author. Tel.: +33-4-72431287; fax: +33-4-78892719.
*E-mail address:* lobry@biomserv.univ-lyon1.fr (J.R. Lobry)

0378-1119/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(99)00297-8

of replication (Filipski, 1990). GC AT skews and are measured for example as the quantity of (C−G)/(G+C) and (A−T)/(A+T) along a DNA sequence using a sliding window. Lobry (1996a) showed the existence of GC and AT skews in the genome of *Haemophilus influenzae* and in parts of the *Escherichia coli* and *Bacillus subtilis* genomes. There is a disadvantage to this method as it is indirect, but the increasing number of completely sequenced genomes allows an extensive analysis of the variation in nucleotide composition within and between genomes. Several recent studies used this method to analyse mitochondrial, viral and bacterial genomes (with emphasis on the latter) for compositional asymmetries, revealing systematic deviations from PR2 in these genomes.

In bacteria, the deviations switch sign at the origin and terminus of replication, such that the leading strand of replication is generally richer in G than in C, and in T than in A. Absolute values for AT skews tend to be lower than for GC skews. Grigoriev (1998) measured skews over all bases and found that the leading strand generally contains more G than C. Third codon position skews show a GT-rich leading strand in all eubacteria (Francino and Ochman, 1997; McLean et al., 1998) except in the *Mycoplasma* species, where it is CT-rich (McLean et al., 1998), and in *Synechocystis*, where no skew was detected (McLean et al., 1998; Mrázek and Karlin, 1998). Archaebacteria generally do not skew (Karlin et al., 1998; Mrázek and Karlin, 1998), except for *M. thermoautotrophicum*, where a weak skew has been detected (McLean et al., 1998; Rocha et al., 1999). Compositional studies of all genome positions in several bacteria report a correlation between purine and coding strand excess (Freeman et al., 1998), and an excess of keto bases (GT) over amino bases (AC) in the leading strand (Freeman et al., 1998; Perrière et al., 1996). Rocha et al. (1998) observed compositional asymmetries between the leading and lagging strand genes at the level of nucleotides, codons and amino acids. Additionally, a strand compositional asymmetry was confirmed in the complete genomes of *B. subtilis* (Kunst et al., 1997), *E. coli* (Blattner et al., 1997), *Rickettsia prowazekii* (Andersson et al., 1998) and *Treponema pallidum* (Fraser et al., 1998).

The deviation divides the chromosome into two segments that are homogenous for GC(AT) skews, called *chirochores* (Lobry, 1996a) in analogy with *isochores* (Bernardi, 1989), which are domains of mammalian chromosomes with homogenous GC content. Chirochores coincide with replichores (Blattner et al., 1997), so that skews switch sign at the origin and terminus of replication. This polarity switch allows for the confirmation of the origin of replication. The method was used to predict the origin in *Mycoplasma genitalium* (Lobry, 1996b), *R. prowazekii* (Andersson et al., 1998), *T. pallidum* (Fraser et al., 1998) and *Borrelia burgdorferi* (Fraser et al., 1997), where the origin could not be detected (because of a lack of consensus patterns) or had not yet been detected experimentally. There is now experimental evidence that the replication origin is located where it was predicted in *B. burgdorferi* (Picardeau et al., 1999).

The perhaps clearest skews are seen in mitochondria: studies of the nucleotide composition of mitochondrial genomes (Jermiin et al., 1995; Perna and Kocher, 1995; Reyes et al., 1998; Tanaka and Ozawa, 1994) all report patterns of asymmetric substitution.

An early study of the bacteriophage λ genome (Daniels et al., 1983) reveals base distribution skew in this molecule, but gives no biological interpretation for the skew. Recently, Mrázek and Karlin (1998) observed asymmetric substitution in some herpesviruses and in the phages λ and T7, and Grigoriev (1998) detected skew in adenovirus type 40.

During revision of this article, several additional publications appeared, showing the existence of strand asymmetries for instance in chloroplasts and ds DNA viruses (see Note added in proof).

Generally, there are two ways of looking at the evolutionary changes of nucleotide composition; the selectionist and the neutralist point of view. These hypotheses differ in the estimate of the role of selection on base substitution. The neutralist hypothesis assumes that the average composition of non-coding DNA depends on a bias of selectively neutral mutations which accumulate during evolution. As an example, there are two main theories that explain the origin of isochores. According to the selectionist hypothesis, isochores are the result of positive selection for GC content as an adaptation to the high body temperature in warm blooded vertebrates (Bernardi et al., 1985). GC content would thereby be the result of positive selection for the functional advantages of the GC content itself. The mutational hypothesis assumes that the compositional biases of mutagenic processes are different in structurally and functionally distinct segments of DNA (Sueoka, 1988, 1992). This hypothesis is based on directional mutation pressure, and must therefore be regarded as being neutral rather than selective. Similarly, selective and mutational theories can be developed for the origin of strand specific nucleotide composition (although the selective hypotheses do not assume that asymmetric substitution is positively selected for because of a functional advantage of the asymmetry itself). Even though the mechanisms creating such patterns are not fully understood, recent publications provide us with several plausible hypotheses, which have been partly summarised in two recent papers (Francino and Ochman, 1997; Mrázek and Karlin, 1998). The idea of this review is to investigate current hypotheses and classify them as mutational or selective. There is, however, at least one hypothesis that must be regarded as based on both mutation and selection.

## 2. Selective mechanisms

### *2.1. Bias on local scale*

In organisms in which a large proportion of the genome consists of coding sequence (prokaryotes, mitochondria, chloroplasts and viruses), selective bias acting on a local scale can potentially influence global nucleotide composition. Because of the low proportion of control sequences and different species of RNA that do not translate into proteins, only protein coding sequences will be considered here.

#### *2.1.1. Amino acid constraints*

The base composition of codon positions 1 and 2 is connected with the amino acid content of proteins. Thus constraints on amino acids (for protein structure and function) could induce PR2 violation in codon positions 1 and 2. Such a selective pressure should, however, be weak because the GC content has been shown to modulate the amino acid content (Lobry, 1997; Sueoka, 1961).

Amino acid preferences in bacteria often result in PR2 violation in positions 1 and 2. Codon position 1 of *E. coli* has A>T and G>C, while position 2 shows the opposite skew. In *B. subtilis* position 1 and 2 are G-rich and G-poor respectively, and they both show A>T (Nakamura et al., 1999). The enrichment of G in codon position 1 seem to be a universal phenomenon, codon position 2 is usually biased towards A or T, and is deficient in G (Frank and Makeev, 1997; Mrázek and Karlin, 1998; Trifonov, 1987). Mrázek and Karlin (1998) pointed out that this could reflect high usage of acidic amino acids encoded by codons GAN, and that the preference of G at site 1 is further amplified by frequent usage of glycine, alanine and valine, as seen in *E. coli* (Lobry and Gautier, 1994). A theory based on selection for amino acid usage must be related to specific constraints on proteins such as structure, function or localisation in the cytoplasm. The following is an example of a selective model for bias in amino acid usage: the global hydrophobicity of proteins is the main factor for variation in amino acid content of proteins in *E. coli* (Lobry and Gautier, 1994). This is due to a selective pressure to increase the content of hydrophobic amino acids (e.g. *Phe*, *Ile*, *Leu*, *Met*, *Val*, *Trp*, *Tyr*) for integral membrane proteins. Because of the nature of the genetic code, the result will be an asymmetric selective pressure between coding and non-coding strand with an excess of T over A in the second position. However, this group only makes up about 10% of the total amount of proteins (Lobry and Gautier, 1994), and is probably too small to have any impact on global composition patterns.

#### *2.1.2. Selection for function at nucleotide level*

It has been suggested that rather than being the trivial result of amino-acid preferences, periodical codon composition patterns [e.g. (G–non-G–N)] have a function in mRNA–rRNA interaction in the ribosome (Lagunez-Otero and Trifonov, 1992; Trifonov, 1987). The authors found that a preference for G in the first codon position and the lack of G in the second position is universal, and the pattern was proposed to be responsible for monitoring the correct reading frame during translation. Several sites with complementary C-periodical structure were found in the *E. coli* 16 S rRNA sequence, and the ribosome slippage was shown to be accompanied by a local disruption (frame-shift) of the periodical pattern, which supports such a hypothesis.

#### *2.1.3. Codon bias*

Although synonymous codon sites are free of selective constraints for amino-acid specification, composition at these sites may still be skewed because of bias in codon usage. Most organisms use a preferred set of codons, and selection acting on codon choice could create local asymmetries between coding and non-coding strands. It is well-known that the bias of codon usage in bacteria is related to the level of expression; highly expressed genes have a strong preference for a limited set of codons (Gouy and Gautier, 1982; Ikemura, 1981; Sharp and Matassi, 1994). Bias in synonymous codons is correlated with the relative abundance of specific tRNAs (Gouy and Gautier, 1982). Consequently, biased codon usage has been explained as the result of selection at translational level. Sueoka (1995) demonstrated that PR2 violation at the third position of synonymous codons in *E. coli* varies in an amino-acid-specific manner. He therefore considers it likely that the major cause of PR2 violation at these sites is the functional selection by tRNA abundance. On the other hand, McInerney (1998) showed that *B. burgdorferi* genes have two separate codon usages depending on whether the gene is transcribed on the leading or lagging strand of replication, although this is probably an exceptional case; the two spirochaetes *B. burgdorferi* and *T. pallidum* show unusually strong strand-specific skews in nucleotide composition. Moreover, Rocha et al. (1998) detected asymmetry in codon usage between strands in several bacteria that was distinct from the usual codon bias due to gene expression levels.

### *2.2. Bias on global scale*

The selectional mechanisms discussed so far would generate bias on a local scale, between coding and non-coding strands in genes. If genes were oriented randomly between the leading and the lagging strand, this bias would be cancelled out since the leading strand would contain approximately the same number of sense and antisense strands. However, the selective mechanisms discussed above could still provide plausible explanations for global strand asymmetry if the repartition of coding sequences between the lagging and leading strand is asymmetric.

### *2.2.1. Collisions between polymerases*

According to a hypothesis originally developed by Brewer (1988), a selective pressure is expected for concordance of orientation of replication and transcription, for the sense strand of genes to be the leading strand of replication. The effects of collisions between DNA and RNA polymerases could be minimised if genes are oriented so that RNA polymerase when transcribing them, moves in the same direction as a replication fork would move during replication. Opposite orientated collisions between the replication fork and the RNA polymerase complex are therefore expected to be counterselected. Highly expressed genes seem to show a higher degree of uneven repartition (Blattner et al., 1997; Brewer, 1988). For instance, for all genomes studied by McLean et al. (1998, table 1), a higher percentage of genes encoded on the leading strand was found when considering only genes coding for ribosomal proteins compared to all genes. This supports the existence of a selective pressure to maintain biased gene orientation.

Brewer (1988) proposed that the inverse orientation is very disadvantageous because of the possible lack of solution mechanism for head-on collisions, but Liu and Alberts (1995) proved the existence of such a mechanism in *E. coli*. Examining the consequences of a head-on collision, they found that RNA polymerase switches its template strand to the strand that has just been synthesised by the leading strand polymerase. This creates a brief pause in replication that is longer than the pause observed for a co-directional collision. Clearly, there is a disadvantage in this head-to-head orientation, but their results show that RNA polymerase *can* stay on the duplex regardless of the orientation of the collision, and thus both orientations should work. If co-orientation of transcription and replication were strongly evolutionary favoured, one would expect all genes to be encoded on the leading strand, which is not the case. The question is to what extent there exists a selective pressure to avoid head-on collisions, how this is manifested in different genomes, and if the resulting uneven repartition of genes is strong enough to contribute to global asymmetric substitution patterns.

### *2.2.2. The gene distribution in different genomes*

The complete sequence of *E. coli* (Blattner et al., 1997) revealed the number of genes oriented in the replication direction to be rather low. Although the repartition of coding sequences is slightly biased towards the leading strand, the global percentage is as low as 54%. In *M. genitalium* and *M. pneumoniae*, genes are strongly skewed towards the leading strand (Fraser et al., 1995; Himmelreich et al., 1996), and in *R. prowazekii* there is a weak asymmetry in the gene distribution (Andersson et al., 1998). In *H. influenzae* gene orientation shows the same mild asymmetry as in *E. coli* (Fleischmann et al., 1995), and in *M. tuberculosis* it is slightly uneven, 59% of genes are transcribed on the leading strand (Cole et al., 1998). Examination of orientation of 96 genes in *B. subtilis* (Zeigler and Dean, 1990) showed 91 to be oriented co-directional with replication, and the complete genome sequence of *B. subtilis* revealed the density of total coding sequence to be skewed by 75% towards the leading strand (Kunst et al., 1997). Thus three species (*M. genitalium*, *M. pneumoniae* and *B. subtilis*) have strongly skewed gene orientation, while the rest show an intermediate or mild skew. This information has also been confirmed and compiled by McLean et al. (1998, table 1), who examined the gene distribution for a number of microbial genomes.

Genes in human mitochondria are asymmetrically distributed between strands, with the L strand containing the sense strand of the rRNAs and most of the tRNAs and mRNAs (Andersson et al., 1981). A multiplicity of gene arrangements is found in metazoan mitochondrial DNA; in some mtDNAs all genes are transcribed from the same strand; in others, both strands encode genes (Boore et al., 1995). Bacteriophages ϕX174 (Sanger et al., 1977) and T7 (Dunn and Studier, 1983) have all their genes transcribed in the direction of replication.

Unlike bacteria, where DNA replication starts from a single origin, eucaryotes initiate DNA synthesis from numerous origins along the chromosomes. This makes it difficult to determine whether the directions of transcription and replication are non-randomly arranged as they are in prokaryotes.

To conclude, biased gene orientation seems to be a general phenomenon in prokaryotes, mitochondria and viruses but the degree of bias varies in different organisms. However, it cannot be excluded that uneven repartition of coding sequences contributes to the compositional bias, at least for some of the genomes studied. McLean et al. (1998, fig. 2) found that in some cases the direction of the skew for total DNA is opposite to the direction of skew for codon position 3, which demonstrates the impact of asymmetric gene distribution on the skew. Furthermore, both biased gene orientation and bias in codon usage are probably pronounced for highly expressed genes (see Section 2.1.3), which could perturb symmetries in base composition on a global scale, as proposed by Francino and Ochman (1997).

### *2.2.3. Uneven distribution of signal sequences as a result of biased selection*

Publications on complete genomic sequences of *E. coli* (Blattner et al., 1997), *B. burgdorferi* (Fraser et al., 1997), *B. subtilis* (Kunst et al., 1997) and *T. pallidum* (Fraser et al., 1998) report the existence of oligomers whose distribution is skewed. Such skewed oligomers could contribute to the global skew provided that the base composition within the oligomers is also skewed. The occurrence of over(under)-represented sequences may signify a phenomenon of positive(negative) selection, indicating important roles as biological signals. Recent studies (Karlin et al., 1996; Rocha et al., 1996) have focused on the occurrence of such sequences and proposed functional roles in replication, control of gene expression, etc., but they did not investigate their potential contribution to strand compositional bias.

A computer analysis carried out on a number of complete bacterial genomes proved skewed oligomers to be a general phenomenon (Salzberg et al., 1998). An oligomer was considered to be skewed if it occurred much more often on the leading strand than on the lagging (more often than its reversed complement in a given strand), and if it was present more frequently than predicted statistically for a random oligomer. The study reveals that most bacterial genomes contain a large number of skewed octamers, in particular *E. coli*, *B. burgdorferi*, *B. subtilis* and *T. pallidum* genomes, which all have hundreds of different skewed octamers (as well as other oligomers). A problem with this approach, however, is that coding and non-coding regions are treated at the same time, so the phenomenon could be a result of uneven repartition of genes between the two strands (see Section 2.2.2).

As long as the function of the skewed sequences is unknown, which is generally the case, it is impossible to develop selective theories for their skew. No biological significance was proposed for the skewed oligomers reported in the complete genomes of *B. burgdorferi* (Fraser et al., 1997) and *T. pallidum* (Fraser et al., 1998). For the most common *E. coli* octamers, Blattner et al. (1997) proposed roles that could explain their bias. These octamers in *E. coli* form a group containing the trimer CGT, which is implicated in the priming of Okazaki fragments (Yoda and Okazaki, 1991). Furthermore, the third most abundant octamer in *E. coli* is the recombinational hot-spot Chi ($\chi$), and none of the other octamers in the group differ from $\chi$ in a way that would inactivate $\chi$ activity. Therefore, the authors propose that the most common skewed oligomers all are $\chi$ sequences containing a primase-binding site. This arrangement would facilitate the recombination that follows the cleavage at the $\chi$ site by RecBCD, because Okazaki initiation at $\chi$ would be helpful in branch migration. Such a functionality could explain their skew. However, the occurrence of common octamers containing the CGT trimer could have a more trivial explanation since CTG happens to be by far the most abundant codon in *E. coli*. Similarly, over-represented codons are found in the most highly skewed octamers of *T. pallidum*.

Kuzminov (1995) proposed another role of the $\chi$ sequence in the recombinational repair of collapsed replication forks. Single-stranded interruption in the template could cause the collapse of the replication fork, and the repair of the fork would involve a RecBCD mediated unwinding starting from a $\chi$ site. Since nicks in DNA from the preceding round of replication should be unevenly distributed between strands, this could explain the preferential orientation of $\chi$-sites in the *E. coli* chromosome. However, the 24 most frequent octamers in *E. coli* together only make up 0.25% of the genome, and are therefore not likely to be the source for a global base composition skew.

Another theory for the function of skewed oligomers proposes their involvement in the post-replicative reconstruction of nucleoide structure (Cornet et al., 1996), at least for the terminus half of the chromosome.

## 3. Mutational mechanisms

Two important facts strongly suggest that strand asymmetries could be caused by mutational mechanisms. First, the violation of PR2 is pronounced at third codon positions and intergenic regions (Lobry, 1996a), where the selective pressure should be nearly neutral or at least weak. Second, the GC and AT deviations switch sign at origin and terminus of replication, which suggests a coupling with replication, repair or both.

### *3.1. Replication biases*

Mrázek and Karlin (1998) have proposed several replication related theories such as different mutation rates between leading and lagging strand, enzymological asymmetry and architectural asymmetry of the replication fork. If semi-conservative replication is responsible for strand asymmetries, it would be so because leading and lagging strands mutate at different rates, and this would be a consequence of the functional or architectural asymmetry of the replication fork. This fundamental asymmetry is introduced to the process of replication because of the antiparallel nature of the strands and the fact that DNA polymerase only synthesises DNA in the 5′–3′ direction. Thus, the leading strand can be replicated continuously, while the lagging strand must be synthesised in a series of Okazaki fragments. This has been observed in vitro but the situation is less clear in vivo. Because of the presence of DNA repair processes, experiments do not succeed in showing that leading strand replication is continuous (Wang and Smith, 1989). If the asymmetry of the replication fork causes strands to mutate at different rates, the question is at which biochemical step of DNA replication, editing or repair, the inequality occurs.

The high accuracy of replication is maintained by three fidelity mechanisms (Schaaper, 1993). In the insertion step, a correct or incorrect nucleotide is added at the growing strand. Exonucleolytic proofreading is the editing step, where an incorrectly inserted nucleotide is removed by an associated 3′–5′ exonuclease activity before elongation. Those errors that escape proofreading can be removed by mismatch repair, which selectively removes errors in the newly synthesised strand. The asymmetry of the replication fork suggests different protein requirements in DNA synthesis for the two strands. Since the incorporation rates differ among DNA polymerases in eucaryotes (Kunkel, 1992a,b), one could expect differences in mutation rates between strands if they use different polymerases during replication. Such an enzymological asymmetry could exist in eucaryotes but is unlikely in prokaryotes because the synthesis of both DNA strands in *E. coli* is carried out by the same polymerase, Pol III holoenzyme (Baker and Wickner, 1992; Marians, 1992). This suggests equal incorporation rate in both strands. The holoenzyme is, however, made up of two core enzymes that coordinate the simultaneous replication of the leading and the lagging strand. A functional asymmetry is imposed to the replication fork because a new Okazaki fragment must be regularly initiated. To accommodate this asymmetry, the lagging strand polymerase needs to recycle on the template (Marians, 1992). It has been shown, however, that the core enzymes are not functionally distinct (Yuzhakov et al., 1996). They both have the properties required for lagging strand synthesis, the asymmetry being implied by the helicase (Yuzhakov et al., 1996).

Therefore, neither base incorporation nor proofreading by the polymerase should differ between strands. It could, however, be that one of the strands is synthesised faster in terms of stepwise progression, as pointed out by Radman (1998). For example, it is possible that the lagging strand polymerase synthesises faster to compensate for the time of its recycling, and therefore commits more errors in the base insertion step.

Another model for differential replication was proposed by Fijalkowska et al. (1998). According to this model, it is the difference in processivity between lagging and leading strand complexes that causes different mutation rates. Even though core enzymes are functionally symmetric, their participation in replication is highly asymmetric, and demands a difference in processivity (tendency to remain on a single template) between the polymerase complexes (Marians, 1992). The leading strand complex needs to be highly processive to stay on the template throughout replication, while its counterpart on the lagging strand is considerably less processive to allow rapid recycling. In *E. coli*, the difference in processivity is in the order of 1000-fold (Marians, 1992). The model is built on results from previous studies by the same authors (Fijalkowska and Schaaper, 1996) showing that, in addition to exonucleolytic proofreading, dissociation of the DNA polymerase from the terminal mismatch is an alternative mode of error removal since it leaves the mismatched 3′ terminus free for excision by some cellular exonuclease. The differences in fidelity of diverse DNA polymerases are in fact manifested much more in their capacity to elongate after a mistake than in their tendency to make the misinsertion (Echols and Goodman, 1991), and only those nucleotide misinsertions that are followed by elongation of the DNA can become mutations. Being less processive, the lagging strand polymerase dissociates more easily from the template and leaves a mismatch free for excision. Therefore, according to this model, the lagging strand would be less error-prone than the leading one.

Mismatch repair would not act differently on the leading and lagging strand since the involved enzymes distinguish between template and newly synthesised strand rather than between leading and lagging strand. However, Radman (1998) proposes that since mismatch repair requires nicks in DNA, error correction by mismatch repair could be more efficient on the lagging strand. The discontinuous replication of the lagging strand provides such nicks, at least in vitro, which is why the lagging strand could replicate more accurately than the leading strand. Moreover, if DNA ligase is highly discriminative, sealing only correctly base-paired termini, ligases could confer a degree of error checking (Housby and Southern, 1998). This would occur only in the lagging strand, where Okazaki fragments are ligated after the removal of the RNA primer. However, Okazaki fragments are 1000–2000 nucleotides long, and such error screening would not contribute considerably to fidelity.

### 3.2. Experimental evidence for asymmetric replication

The distinct modes of replication in the two DNA strands have triggered hypotheses that suggest differential replication fidelity between the leading and the lagging strand. Several groups have tested for fidelity difference between strands. The general belief seems to ascribe a higher error rate to the lagging strand, but results of experimental studies on the matter are contradictory. Since it is difficult to test error rates in structural genes in natural systems, most tests involve special cases that unfortunately tell us little about a possible connection with the asymmetric base composition observed in bacterial genomes.

Kunkel and coworkers have scored mutations in eucaryotic systems (Simian 40-dependent replication in human cell extracts) (Izuta et al., 1995; Roberts et al., 1994; Thomas et al., 1993). The results are, however, difficult to interpret because of the likely operation of more than one DNA polymerase at the replication fork in eucaryotes. They report some strand biases but these seem to depend on the mutagenic site.

A number of tests have been performed in ColE1-derived plasmids (Rosche et al., 1995; Trinh and Sinden, 1991; Veaute and Fuchs, 1993). In these, a unidirectional replication carried out by *E. coli* proteins is used. To measure the number of deletions in each strand, two plasmids were constructed. The choice of the lagging or leading strand as the transcription template depends on the orientation of the gene with respect to ori. Therefore, in each plasmid a gene was inserted in the opposite direction, so that the direction of replication was reversed. Using a palindromic DNA constitution to measure deletions (Rosche et al., 1995; Trinh and Sinden, 1991) and a carcinogen-adducted gene to measure frame-shifts (Veaute and Fuchs, 1993), strand bias in deletion rate was shown to be about 20-fold higher in the lagging strand than in the leading strand. However, these systems do not address the real difference between the two strands. The former probably involves secondary structure formation of the lagging strand template, and the latter is a special case, scoring mutation at a lesion.

The experimental design of the study by Iwaki et al. (1996) seems to reflect natural conditions more closely. Still, in ColE1 plasmids, reversion frequencies of inactivated drug resistance in a reporter gene were measured. To detect error rates generated during lagging and leading strand replication before the proofreading step, a mutator strain dnaQ49 was used as it is deficient in 3′–5′ exonuclease activity. Results showed that frequencies of three frame-shift and one point mutation were at least 10–100-fold higher in the lagging strand than in the leading. The authors consider it unlikely that mismatch repair mechanisms contribute to the bias.

However, ColE1 plasmids do not replicate in the same manner as the *E. coli* chromosome (Marians, 1992), requiring for example an extensive synthesis by DNA polymerase I. A study by Fijalkowska et al. (1998) was performed in an *E. coli* chromosome and involves the measurement of *lac* reversion frequency by base substitution for the two orientations. Mismatch and proofreading deficient strains were used to detect intrinsic error rates between strands. In contrast to the study of Iwaki et al. (1996), the results propose that the lagging strand is more accurate.

Thus both the studies of Fijalkowska et al. and Iwaki et al. measure intrinsic error rates between leading and lagging strand using strains deficient in proofreading (and mismatch repair in the former study). They both propose that these rates differ between strands, but the results are in contrast concerning which strand is more error-prone (Fijalkowska's model was discussed in Section 3.1). The question is therefore which (if any) result could explain the observed strand asymmetries. Differences in mutation rates could yield asymmetries in base composition if error rates differ between complementary mismatches. For transitions, the T · G and G · T mispairs dominate over complementary A · C and C · A mispairs (Mendelman, 1990). Therefore, the more error-prone strand would be relatively richer in GT. For transversions, the data are more heterogeneous, but it seems as if pu · pu mismatches are more common at the insertion step (Fersht and Knill-Jones, 1981; Topal and Fresco, 1991). The strand committing more errors would then accumulate purines. The leading strand in bacterial genomes is enriched in GT (Perrière et al., 1996), and it has also been proved purine-rich in some genomes (Freeman et al., 1998) (although this is probably not a result of biased mutation, see Section 5). This would be more in agreement with Fijalkowska's model, in which the leading strand is the more error prone. Moreover, only base substitutions would be relevant for the explanation of asymmetric substitution patterns and the study of Iwaki et al. involves three frame-shifts and only one point mutation. However, as both of the studies use strains deficient in proofreading or mismatch repair, one cannot dismiss the possibility that the detected fidelity difference can be compensated for in natural systems, so that no asymmetric substitution patterns would be generated by these fidelity differences.

### 3.3. *Cytosine deamination theory*

The asymmetric structure of the replication fork introduces a difference between strands in the amount of time spent single-stranded. This is important because single-stranded DNA is more exposed to damage. Concerning base substitution mutagenesis, not only replication or repair errors but also spontaneous chemical modifications such as oxidation, deamination and alkylation may be frequent sources of mutations (Lindahl, 1993). DNA base residues are susceptible to hydrolytic deamination, and the main targets for this reaction are cytosine and its homologue 5-methylcytosine (Kreutzer and Essigmann, 1998; Lindahl, 1993). The high frequency of C→T deaminations may explain why G · C→A · T transitions dominate the spectra of mutations in *E. coli* (Echols and Goodman, 1991). Deamination of cytosine leads to the formation of uracil, which pairs with adenine during replication causing a C to T mutation. In normal circumstances, because of the Watson–Crick base pairing, nucleotides are effectively protected against hydrolytic deamination. C deaminates 140 times faster when present in single-stranded DNA than in double-stranded DNA (Frederico et al., 1990). According to the present-day replication model (Baker and Wickner, 1992; Kelman and O'Donnel, 1995; Marians, 1992, 1996), stretches of the template for lagging strand synthesis are temporarily in a single-stranded state. The length of such a stretch should be at least equal to the size of the most nascent Okazaki fragment. The model involves the looping of the lagging strand that enables the simultaneous replication of both strands, and its discontinuous synthesis in Okazaki fragments. The replication fork must advance enough to allow for the looping and the synthesis of the next Okazaki fragment, leaving the lagging strand template temporarily single-stranded. Single-stranded binding proteins should not protect considerably against deamination. The study where single-stranded deamination was found to be more frequent (Frederico et al., 1990) was carried out on nuclear DNAs, which should be protein-coated. Thus structural asymmetry could introduce a mutational bias between the two strands, generating compositional asymmetry. The theory of asymmetric deamination is compatible with the observed GT-richness of the leading strand, since C→T deamination in one strand would increase G% and T% in that strand, and increase C% and A% content in the complementary strand. In the same way, the less common deamination of A to hypoxantine which base-pairs with C rather than T (Lindahl, 1993) will result in an increase in G% and T% in the exposed strand. However, if asymmetric deamination during replication were a universal phenomenon, asymmetric substitution should be seen in all bacteria, and no skew was detected in *Synechocystis* (McLean et al., 1998).

Studies on mitochondria (Jermiin et al., 1995; Perna and Kocher, 1995; Reyes et al., 1998; Tanaka and Ozawa, 1994) and viruses (Grigoriev, 1998, Mrázek and Karlin, 1998) confirm the existence of an asymmetric substitution matrix in these genomes and support the deamination theory. A strong compositional asymmetry is observed in mitochondrial genomes, and the skew is clearly higher at synonymous codon positions, suggesting the existence of an asymmetric directional mutation pressure. The replication of mitochondrial DNA is highly asymmetric: the daughter H strand displaces the parental H strand so that the parental H strand remains in a single-stranded state until paired with the newly synthesised L strand. The parental H strand is thus exposed to damage during mitochondrial replication like the template for the lagging strand during chromosomal replication. A study on human adenovirus type 40 (Grigoriev, 1998) shows similar results, probably because the genome replicates leaving one of the strands single-stranded while the other is being duplicated. For both mitochondrial (Reyes et al., 1998) and adenoviral (Grigoriev, 1998) genomes a correlation has been suggested between skews and time spent single-stranded. Therefore, asymmetric deamination is likely to be the reason for compositional asymmetries in mitochondrial and viral genomes (Grigoriev, 1998; Reyes et al., 1998) (see Note added in proof).

The results from mitochondrial (and possibly viral) genomes thus provide us with evidence that differences in the substitution matrix could be due to differences in the damage spectra of single- and double-stranded DNA. It is, however, possible that this model for asymmetric substitution is specific to mitochondria, and does not explain the skews in bacteria. For instance, the template for the lagging strand should spend considerably less time single-stranded than does the parental H strand in mitochondria. The mitochondrial replication can take up to 2 h, and the H strand will only be partially covered with ss binding protein (Reyes et al., 1998). Prokaryotic replication, on the other hand, proceeds with a rate of 1–2 kb/s (Kelman and O'Donnel, 1995), and the region of single-stranded DNA is considerably shorter, so the time spent single-stranded is only transient.

Both bacterial and mitochondrial genomes show lower absolute values of AT than GC skews (McLean et al., 1998; Reyes et al., 1998), which raises questions as to the soundness of the deamination hypothesis. An excess of C→T deaminations in one strand would yield equal decrease in C and increase in T in that strand which is not consistent with the greater skews often observed for GC compared to AT. However, absolute values of GC and AT skews should depend on GC content, since a GC content that differs from 50% would create a relative difference between GC and AT deviation values. Reyes et al. (1998) propose that differences between GC and AT absolute values are due to different mutation rates in $\alpha$ (AT) and $\gamma$ (GC) bases.

## 4. Combination of selection and mutation

There is at least one possible mechanism that involves both selection and mutation. Francino et al. (1996) and Francino and Ochman (1997) suggested that processes which distinguish between *transcribed/non-transcribed* strand can account for DNA asymmetry. Transcription alone would not distinguish between leading and lagging strand, but in combination with biased gene orientation (discussed in Section 2.2.2), transcription-induced mutations could generate the compositional asymmetry between leading and lagging strand that has been observed in bacterial genomes. Therefore such a theory must be classified as being based on both mutation and selection.

### *4.1. Differential repair*

Using phylogenetic reconstructions, Francino et al. (1996) scored frequencies of complementary substitutions in enteric bacteria. They detected asymmetries between the complementary transitions C→T and G→A, when comparing the sense and antisense strand.

The authors explain this bias by asymmetric transcription-coupled repair on the antisense strand. Transcription-coupled repair is highly strand-specific in *E. coli*, the antisense strand being preferentially repaired. The strand-specificity is further amplified in transcriptionally active genes (Hanawalt, 1991; Mellon and Hanawalt, 1989), which are probably also the ones with the highest asymmetry of distribution (Francino and Ochman, 1997; McLean et al., 1998). Transcription-coupled repair is known to act preferentially on pyrimidine dimers (Hanawalt, 1991; Mellon and Hanawalt, 1989). C:G to T:A mutations by insertion of A opposite C, as well as by C→T deaminations, are common in pyrimidine dimers (Hutchinson, 1996), which could explain the observed asymmetries. It could also be that preferential repair of pyrimidine dimers in the antisense strand is evolutionary favoured because it increases the content of the less vulnerable purines in the sense strand, preventing deleterious mutations during transcription. A pyrimidine-rich antisense strand has been reported for several bacteria, bacteriophages and higher organisms (Szybalski et al., 1996).

### *4.2. Asymmetry of the transcription bubble*

An alternative explanation for the observed C→T versus G→A asymmetry between sense and antisense strand is given by Beletskii and Bhagwat (1996, 1998). Their study shows that C to T deaminations are induced by transcription in *E. coli*, because transcription of a gene promotes deamination of cytosine only when it is present in the sense strand. This is because during transcription, like during replication (see Section 3.3), the DNA strands will be unequally exposed to damage. The antisense strand is temporarily associated with mRNA and the transcription complex, whereas the sense strand may be considered to be single-stranded. Therefore, C→T deaminations would dominate in the sense strand, as observed by Francino et al. (1996).

AT and CG deviation switch at origin of replication could be explained by uneven distribution of genes between leading and lagging strands (see Section 2.2.2). Francino and Ochman (1997) consider this likely in *E. coli* because highly expressed genes tend to be coded on the leading strand. If transcription is responsible for creating strand asymmetries, we would expect higher mutation rates in highly expressed genes. There is some evidence that transcription increases mutation in yeast (Datta and Jinks-Robertson, 1995). However, no studies have reported the correlation between GC(AT) deviations and gene expressivity that would be expected in the case of transcription-induced mutation. On the contrary, Sharp and Li (1989) and Sharp et al. (1989) showed that highly expressed genes have a lower degree of divergence, which is not consistent with a transcription-based theory.

Asymmetric deamination could occur during transcription, replication or both. The time spent single-stranded during transcription should be very transient since the transcription bubble only contains a 12 nucleotide single-stranded region. During replication, the single-stranded stretch should be considerably longer [at least one Okazaki fragment (Marians, 1996)] which would compensate for the higher rate of DNA polymerase relative to RNA polymerase. Furthermore, transcription-induced deamination only causes a premutagenic lesion that will have to wait for the next round of replication to become fixed. During this time the resulting uracil can be removed by uracil–DNA glycosylase and the origin sequence can be restored. When deamination occurs in the lagging strand template, the resulting U will almost immediately base-pair with an incoming A in the synthesis of the lagging strand. This generates a fixed mutation since a T instead of a C will replace the U excised by uracil–DNA glycosylase.

Mrázek and Karlin (1998) remark that the *E. coli* genome has C>T at codon site 1 and C∼T at position 3 (Nakamura et al., 1999), and that this is contrary to what would be expected from C→T deaminations. However, first codon position is not free from selective pressure and it is logical that mutational bias has weaker impact on its base composition. Equal C and T values at codon position 3 should not be seen if asymmetric deamination during transcription were an important source for base composition skews, but does not contradict the theory of asymmetric deamination during replication.

## 5. Discussion

Compositional studies of bacterial, mitochondrial and viral genomes has established the existence of deviations from the frequencies A=T and G=C expected under *no-strand-bias* conditions. Skew values differ depending on what part of the genome is studied and different genomes conform differently to the predicted models. Therefore, compositional asymmetry could be a result of superposition of different mechanisms that influence base composition to different extents, and act differently in different organisms. There are, however, some common traits. Base composition skews measured at intergenic regions and third codon position (where the selective pressure is greatly decreased) are pronounced, which highlights the occurrence of a mutational bias. Skews at the genome level could be affected by amino acid constraints and codon usage when genes are organised to avoid collisions between transcription and replication polymerases.

Mutational bias could result from genuine replication or repair errors or from DNA decay generated during the process of replication or transcription. There is some experimental evidence for fidelity difference between the leading and lagging strand of replication, but experimental conditions are rather special and results are not consistent regarding which strand is the more accurate, which makes it difficult to draw any general conclusion. This does not, however, exclude the possibility that asymmetric replication errors cause asymmetric substitution patterns.

Codon position 3 of leading and lagging coding sequence in bacteria generally show different skew values (Lobry, 1996a; McLean et al., 1998), reflecting a mutational bias that could be replicational, transcriptional or both. The GT-rich leading strand could be explained by replication- and/or transcription-induced deamination events since asymmetric C→T deamination would increase GT content in one strand. Both replication and transcription have been suggested to cause asymmetric deamination; replication-related deamination was proposed to cause asymmetric directional mutation pressure in mitochondria (Reyes et al., 1998), and deamination was shown to occur more frequently in the sense strand of *E. coli* (Beletskii and Bhagwat, 1996) than in the antisense strand.

Deaminations should occur both in the lagging strand template during replication and the coding strand during transcription, but their relative influence on base composition asymmetry should depend on organism-specific factors such as chromosome organisation, growth rate, reparation enzymes and frequency of replication cycles. Since the degree of asymmetry varies among genomes and can be as low as 54% in *E. coli*, it is difficult to determine whether the phenomenon of biased gene distribution has a global impact or not, although it probably has an important influence on the base compositions of some genomes. Because highly expressed genes seem to show a higher degree of asymmetric orientation (Brewer, 1988; McLean et al., 1998), and since transcription-induced mutation events should occur more often in these, it can be argued that transcription-induced mutations could still be important, even if asymmetric orientation is low. However, this hardly explains the strong base composition in organisms such as *E. coli* where the gene distribution is almost random. Moreover, Rocha et al. detected that the strong compositional asymmetries between leading and lagging strand genes cause asymmetry in codon usage between the strands (Rocha et al., 1999). The codon asymmetry was distinct from the usual codon bias resulting from gene expression levels, which is in conflict with the possibility that the bias is caused by the over-representation of highly expressed genes on the leading strand.

Total genome skews are sometimes opposite to the ones observed for third codon position. This suggests an impact of amino acid constraints on codon positions 1 and 2 on the global skew, but should only be seen in genomes where distribution of genes is strongly uneven (*B. subtilis*, *M. genitalium*, *M. pneumoniae*) (McLean et al., 1998). This supports the idea that base composition skews could be influenced by differences between coding and non-coding strand because of biased gene orientation. Mrázek and Karlin (1998) suggest that biased gene orientation is the main cause of strand compositional asymmetries in *Mycoplasma* species and in *B. subtilis*. In the *E. coli* genome, on the other hand, the G-rich tendency of the leading strand is still seen when analysing all the positions of the genome (Mrázek and Karlin, 1998), probably because gene distribution is not asymmetric enough to let amino acid constraints and codon usage contribute to the skew. Freeman et al. (1998) detected a purine excess correlated with coding excess in several bacteria, most strikingly in *M. jannaschii* and *M. thermoautotrophicum*. The authors suggested a link with replication, but a purine-rich coding strand should rather reflect amino acid constraints, as pointed out by McLean et al. (1998).

Among the genomes studied by McLean et al. (1998), the largest skews were detected in *B. burgdorferi* and *T. pallidum*. This probably indicates a greater tendency in these genomes to be influenced by asymmetry-causing mechanisms; asymmetry between strands in *B. burgdorferi* has even been shown to be strong enough to cause separate codon usage between strands (McInerey, 1998). As mentioned above, the *Mycoplasma* species have a different GC skew at third codon position, the leading strand being enriched in CT (McLean et al., 1998). The deamination theory could still hold for these genomes. When the discriminant power of synonymous codons was compared (in *Mycoplasma* species as well as in *E. coli* and *H. influenzae*), leading coding sequences were found to be enriched in codons containing G and T (Perrière et al., 1996). The CT-richness in the leading strand of *Mycoplasma* species contrasting with the GT-richness in other genomes might have a simple explanation. *Mycoplasma* species globally contain more C than G at codon position 3 by about 6%. Since the gene orientation is strongly biased (more than 70% of the genes are encoded in the leading strand), counting all codon position 3 on the leading strand could generate an excess of C over G.

There is no skew present in archaeal genomes of *M. jannaschii* and *A. fulgidus* (Mrázek and Karlin, 1998; McLean et al., 1998), and only a weak skew was observed in *M. thermoautotrophicum* (McLean et al., 1998). It has been speculated that archaeal, like eucaryotic genomes, possess multiple origins of replication (Olsen and Woese, 1997), which could explain the absence of skew. Several eucaryotes show no distinct strand asymmetry (Karlin et al., 1998). However, although GC and AT skews are weak or do not exist in archaebacteria, studies that show a skew of one or several oligomers around a single pair of points provide evidence for a single origin in *M. thermoautotrophichum*

(Salzberg et al., 1998; Lopez et al., 1999), *P. horikoshii* and *P. furiosus* (Lopez et al., 1999). Moreover, Lopez et al. (1999) found AT-rich elements and inverted repeats near the putative origins in these genomes, which further supports the existence of a unique origin of replication in archaebacteria. If asymmetries in eubacteria are produced by the replication system, it could be that the reduced skew in archae is due to differences in the replication system compared to eubacteria. In fact, archaeal replication proteins look more like eucaryotic than prokaryotic replication proteins (Edgell and Doolittle, 1997).

In conclusion, different mechanisms seem to contribute to the base compositional skews, generating results that depend in part on the chromosome organisation of the organism studied. To make a long story short, at the present time we have strong evidence for asymmetrical directional mutation pressure in mitochondria and some viruses, evidence in some eubacteria, little evidence in archaebacteria and no evidence in eucaryotes. Evidence in chloroplasts appeared during revision of this article (see Note added in proof). The cytosine deamination theory seems to be the best explanation for the origin of asymmetrical directional mutation pressure. If gene distribution is biased, skews produced by replicational asymmetry could be counteracted or pronounced by transcriptional bias, codon usage and amino acid constraints.

The biological significance of asymmetrical directional mutation pressure is close to zero for regular biologists as it does not fit into the structure/adaptation scheme: there is no meaning in terms of fitness for the chirochore structure in bacteria or for the skew gradient in mitochondria. There is no doubt that if the chirochore structure were not able to predict, as a by-product, the location of terminus and origin of replication, biologists would not care about it. From an evolutionary point of view, however, existence of asymmetric substitution patterns could be less meaningless. It has been suggested (Furusawa and Hirofumi, 1998) that an asymmetric global mutation rate between the two DNA strands could be advantageous by allowing high mutation rates while still preserving optimal genotypes at a population level. If the asymmetric deamination of C→T is the major source of spontaneous mutation, global mutation rate between the two strands should be different.

## 6. Note added in proof

During the revision of this paper, some additional articles of great interest were published. A study by Grigoriev (1999, Virus Res. 60, 1–19) reveals compositional asymmetry between leading and lagging strand in 22 complete sequences of ds DNA viruses. Possible contributions of transcription and replication (and their associated repair mechanisms) are discussed along with other potential sources of strand bias. Similarly, in the chloroplast genome of *Eugena gracilis* (Morton, 1999, Proc. Natl. Acad. Sci. 96, 5123–5128), the two strands are asymmetric with regard to nucleotide composition. Li (Computer and Chemistry Special Issue, 23, 283–301) has examined all completely genome sequences for strand asymmetry, analysing the base composition asymmetry at the three codon positions separately. Furthermore, Cerbrat et al. (1999, Physica A 265, 78–84) have confirmed asymmetric base composition of third codon positions in *E. coli* and proposed that asymmetric replication is the source for this. The relation between replication-associated mutational pressure and amino-acid composition of proteins has been investigated by the same authors (Mackiewicz et al., 1999, Genome Res. 9, 409–416), who have also analysed the sources of asymmetry in nucleotide composition in prokaryotic chromosomes (Mackiewicz et al., 1999, J. Appl. Genet. 40, 1–14). Finally, Lafay et al. (1999, Nucleic Acids Res. 27, 1642–1649) have detected asymmetric codon usage and amino-acid composition between leading and lagging strand genes in the spirochaetes *B. burgdorferi* and *T. pallidum*. They suggest that translational selection is absent or ineffective in these spirochaetes, and consider it unlikely that the different compositions of genes and proteins between the two strands is the result of natural selection.

## References

Andersson, S., et al., 1981. Sequence and organization of the human mitochondrial genome. Nature 290, 457–465.

Andersson, S.G., et al., 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396, 133–143.

Baker, T.A., Wickner, S.H., 1992. Genetics and enzymology of DNA replication in *Escherichia coli*. Annu. Rev. Genet. 26, 447–477.

Beletskii, A., Bhagwat, A.S., 1996. Transcription-induced mutations: increase in C to T mutations in the non-transcribed strand during transcription in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 93, 13919–13924.

Beletskii, A., Bhagwat, A.S., 1998. Correlation between transcription and C to T mutations in the non-transcribed DNA strand. Biol. Chem. 379, 549–551.

Bernardi, G., et al., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–957.

Bernardi, G., 1989. The isochore organization of the human genome. Annu. Rev. Genet. 23, 637–661.

Blattner, F.R., et al., 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1474.

Boore, J.L., Collins, T.M., Stanton, D., Daehler, L.L., Brown, W.M., 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376, 163–165.

Brewer, B.J., 1988. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. Cell 53, 679–686.

Bulmer, M., 1991. Strand symmetry of mutation rates in the β-globin region. J. Mol. Evol. 33, 305–310.

Chargaff, E., 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia 6, 201–240.

Cole, S.T., et al., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544.

Cornet, F., Louarn, J., Patte, J., Louarn, J.M., 1996. Restriction of the activity of the recombination site *dif* to a small zone of the *Escherichia coli* chromosome. Genes Dev. 10, 1152–1161.

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.

Datta, A., Jinks-Robertson, S., 1995. Association of increased spontaneous mutation rates with high levels of transcription in yeast. Science 268, 1616–1619.

Dunn, J.J., Studier, F.W., 1983. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J. Mol. Biol. 166, 477–535.

Echols, H., Goodman, M.F., 1991. Fidelity mechanisms in DNA replication. Annu. Rev. Biochem. 60, 477–511.

Edgell, D.R., Doolittle, F.W., 1997. Archaea and the origin(s) of DNA replication proteins. Cell 89, 995–998.

Fersht, A.R., Knill-Jones, J.W., 1981. DNA polymerase accuracy and spontaneous mutation rates: frequencies of purine · purine, purine · pyrimidine and pyrimidine · pyrimidine mismatches during DNA replication. Proc. Natl. Acad. Sci. USA 78, 4251–4255.

Fijalkowska, I.J., Schaaper, R.M., 1996. Mutants in the Exo I motif of *Esherichia coli* dnaQ: defective proofreading and inviability due to error catastrophe. Proc. Natl. Acad. Sci. USA 93, 2856–2861.

Fijalkowska, I.J., Jonczyk, P., Tkaczyk, M.M., Bialoskorska, M., Schaaper, R.M., 1998. Unequal fidelity of leading and lagging strand DNA replication on the *Escherichia coli* chromosome. Proc. Natl. Acad. Sci. USA 95, 10020–10025.

Filipski, J., 1990. Evolution of DNA sequence, contributions of mutational bias and selection to the origin of chromosomal compartments. In: Ole, G. (Ed.), Advances in Mutagenesis Research 2. Springer, Berlin, pp. 1–54.

Fleischmann, R.D., et al., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

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, G.K., Makeev, V.J., 1997. G and T nucleotide content show specie invariant negative correlation for all three codon positions. J. Biomol. Struct. Dynam. 14, 629–639.

Fraser, C.M., et al., 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397–403.

Fraser, C.M., et al., 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390, 580–586.

Fraser, C.M., 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 rate 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 organization in bacteria. Science 279, 1827.

Furusawa, M., Hirofumi, D., 1998. Asymmetrical DNA replication promotes evolution: disparity theory of evolution. Genetica 102/103, 333–347.

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.

Hanawalt, P.C., 1991. Heterogenity 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.

Housby, J.N., Southern, E.M., 1998. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. Nucleic Acids Res. 26, 4259–4266.

Hutchinson, F., 1996. Mutagenesis. In: Neidhardt, F.C. (Ed.), *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology, 2nd edn., ASM, Washington, DC, pp. 749–763.

Ikemura, T., 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. 146, 1–21.

Iwaki, T., et al., 1996. Preferential replication-dependent mutagenesis in the lagging DNA strand in *Escherichia coli*. Mol. Gen. Genet. 251, 657–664.

Izuta, S., Roberts, J.D., Kunkel, T.A., 1995. Replication errors rates for T · dGTP and A · dGTP mispairs and evidence for differential proofreading by leading and lagging strand DNA replication in human cells. J. Biol. Chem. 270, 2595–2600.

Jermiin, L.J., Graur, D., Crozier, R.H., 1995. Evidence from analyses of intergenic regions for strand-specific directional mutation pressure in metazoan mitochondrial DNA. Mol. Biol. Evol. 12, 558–563.

Karlin, S., Mrázek, J., Campbell, A.M., 1996. Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. Nucleic Acids Res. 24, 4263–4272.

Karlin, S., Campbell, A.M., Mrázek, J.M., 1998. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. 32, 185–225.

Kelman, Z., O'Donnel, M., 1995. DNA polymerase III holoenzyme: structure and function of a chromosomal replicating machine. Annu. Rev. Biochem. 64, 171–200.

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.

Kunkel, T.A., 1992a. Biological asymmetries and the fidelity of eucaryotic DNA replication. BioEssays 14, 303–308.

Kunkel, T.A., 1992b. DNA replication fidelity. J. Biol. Chem. 267, 18251–18254.

Kunst, F., et al., 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390, 249–256.

Kuzminov, A., 1995. Collapse and repair of replication forks in *Escherichia coli*. Mol. Microbiol. 16, 373–384.

Lagunez-Otero, J., Trifonov, E.N., 1992. mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. J. Biomol. Struct. Dynam. 10, 455–464.

Lin, H.J., Chargaff, E., 1967. On the denaturation of deoxyribonucleic acid. II. Effects of concentration. Biochim. Biophys. Acta 145, 398–409.

Lindahl, T., 1993. Instability and decay of the primary structure of DNA. Nature 362, 709–715.

Liu, B., Alberts, B.M., 1995. Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. Science 267, 1131–1137.

Lobry, J.R., Gautier, C., 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. 22, 3174–3180.

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. Origin of replication of *Mycoplasma genitalium*. Science 272, 745–746.

Lobry, J.R., 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205, 309–316.

Lobry, J.R., Lobry, C., 1999. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. Mol. Biol. Evol. 16, 719–723.

Lopez, P., Philippe, H., Myllykallio, H., Forterre, P., 1999. Identification of putative chromosomal origins of replication in Archaea. Mol. Microbiol. 32, 883–891.

Marians, K.J., 1992. Prokaryotic DNA replication. Annu. Rev. Biochem. 61, 673–719.

Marians, K.J., 1996. Replication fork propagation. In: Neidhardt, F.C. (Ed.), *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology, 2nd edn., ASM, Washington, DC, pp. 749–763.

McInerney, J.O., 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. USA 95, 10698–10703.

McLean, J.M., 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.

Mendelman, L.V., 1990. Base mispair extension kinetics. Comparison of DNA polymerase alpha and reverse transcriptase. J. Biol. Chem. 265, 2338–2346.

Mrázek, J., Karlin, S., 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA 95, 3720–3725.

Nakamura, Y., Gojobori, T., Ikemura, T., 1999. Codon usage tabulated from the international DNA sequence databases, its status 1999. Nucleic Acids Res. 27, 292.

Olsen, G.J., Woese, C.R., 1997. Archaeal genomics: an overview. Cell 89, 991–994.

Perna, N.T., Kocher, T.D., 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J. Mol. Evol. 41, 353–358.

Perrière, G., Lobry, J.R., Thioulouse, J., 1996. Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid 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 center of the *Borrelia burgdorferi* linear chromosome. Mol. Microbiol. 32, 437–445.

Radman, M., 1998. DNA replication: one strand may be more equal. Proc. Natl. Acad. Sci. USA 95, 9718–9719.

Reyes, A., Gissi, C., Pesole, G., Saccone, 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. Mispair, site-, and strand-specific error rates during simian virus 40 origin-dependent replication in vitro with excess deoxythymine triphoshate. J. Biol. Chem. 269, 1711–1717.

Rocha, E.P.C., Viari, A., Danchin, A., 1996. Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. Nucleic Acids Res. 26, 2971–2980.

Rocha, E.P.C., Danchin, A., Viari, A., 1998. Universal replication biases in bacteria. Mol. Microbiol. 32, 11–16.

Rosche, W.A., Trinh, T.Q., Sinden, R.R., 1995. Differential DNA secondary structure-mediated deletion mutation in the leading and lagging strands. J. Bacteriol. 177, 4385–4391.

Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., Tomb, J.F., 1998. Skewed oligomers and origins of replication. Gene 217, 57–67.

Sanger, F., et al., 1977. Nucliotide sequence of bacteriophage phi X174 DNA. Nature 265, 687–695.

Schaaper, R.M., 1993. Base selection, proofreading and mismatch repair during DNA replication in *Escherichia coli*. J. Biol. Chem. 268, 23762–23765.

Sharp, P.M., Li, W.H., 1989. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4, 222–230.

Sharp, P.M., Shields, D.C., Wolfe, K.H., Li, W.H., 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. Science 246, 808–810.

Sharp, P.M., Matassi, G., 1994. Codon usage and genome evolution. Curr. Opin. Genet. Dev. 6, 851–860.

Sueoka, N., 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc. Natl. Acad. Sci. USA 47, 1141–1149.

Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 48, 2653–2657.

Sueoka, N., 1992. Directional mutation pressure, selective constraints and genetic equilibria. J. Mol. Evol. 34, 95–114.

Sueoka, N., 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40, 318–325. Erratum 42, 323.

Szybalski, W., Kubinski, H., Sheldrick, P., 1966. Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis, Cold Spring Harbor Symp. on Quantum Biology Vol. 31, 123–127.

Tanaka, M., Ozawa, T., 1994. Strand asymmetry in human mitochondrial DNA mutations. Genomics 22, 327–335.

Thomas, D.C., Nguyen, D.C., Piegorsch, W.W., Kunkel, T.A., 1993. Relative probability of mutagenic translesion synthesis on the leading and lagging strands during replication of UV-irradiated DNA in human cell extract. Biochemistry 32, 11476–11482.

Topal, M.D., Fresco, J.R., 1991. Complementary base pairing and the origin of substitution mutations. Nature 263, 285–289.

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.

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, T.V., Smith, K.C., 1989. Discontinuous DNA replication in a lig-7 strain of *Escherichia coli* is not the result of mismatch repair, nucleotide-excision repair, or the base-excision repair of DNA uracil. Biochem. Biophys. Res. Comm. 165, 685–688.

Watson, J.D., Crick, F.C.H., 1953. A structure for deoxyribose nucleic acid. Nature 171, 737–738.

Wu, C.I., Maeda, N., 1987. Inequality in mutation rates of the two strands of DNA. Nature 327, 169–170.

Yoda, K., Okazaki, T., 1991. Specificity of recognition sequence for *Escherichia coli* primase. Mol. Gen. Genet. 227, 1–8.

Yuzhakov, A., Turner, J., O'Donnel, M., 1996. Replisome assembly reveals the basis for asymmetric function in leading and lagging strand replication. Cell 86, 877–886.

Zeigler, D.R., Dean, D.H., 1990. Orientation of genes in the *Bacillus subtilis* chromosome. Genetics 125, 703–708.