# Strand asymmetries in DNA evolution

M. PILAR FRANCINO (pilar@ho-lab.biology.rochester.edu)

HOWARD OCHMAN (ochman@ho-lab.biology.rochester.edu)

***The complementary strands of DNA differ with respect to replication and transcription. Both of these processes are asymmetric and can bias the occurrence of mutations between the strands: during replication, the discontinuous lagging strand undergoes certain errors at higher rates, and transcription overexposes the nontranscribed strand to DNA damage while targeting repair enzymes to the transcribed strand. While biases introduced during replication apparently have little impact on sequence evolution, the effects of transcription are observed in the asymmetric patterns of substitution in bacterial genes and might be influencing genome-wide patterns of base composition.***

Molecular evolutionary geneticists often consider mutation as a completely random process; however, the different types of base substitutions are not equally common and the likelihood of a given mutation varies with location in the genome. Evolutionary rates of nucleotide substitution are determined mainly by the mutation rate and the effects of natural selection. Differential selective constraints govern much of the existing variation in substitution rates, such as the large differences observed between genes and intergenic regions, or between synonymous and nonsynonymous sites within a gene. At sites under little or no selective constraints, such as intergenic regions and degenerate codon positions, all mutations are neutral, or nearly so, and have an equal probability of being fixed in the population. Thus, substitutions at these sites are expected to reflect the underlying rates and patterns of mutation.

Sequence comparisons reveal that mutation rates, in addition to selective constraints, are not uniform throughout the genome. For example, synonymous substitution rates increase with distance from the origin of replication in bacteria[1] and differ among regions of mammalian chromosomes[2]. Moreover, biochemical evidence and comparative sequence analysis have identified particular DNA sequences – such as palindromes or tandem repeats – that are prone to mutation[3,4], and have demonstrated that the adjacent nucleotides promote specific mutations at a given site, for example, the rate of transition from T.A is enhanced by the presence of a 3′A.T or 3′G.C pair[5,6].

The likelihood of a particular mutation can also depend on the DNA strand in which the nucleotide is located. Replication is an asymmetric process: on the leading strand it proceeds continuously, whereas on the lagging strand it proceeds discontinuously by the synthesis and joining of short Okazaki fragments[7]. In bacteria, both DNA strands are synthesized coordinately by the same apparatus, the DNA polymerase III (pol III) holoenzyme. For lagging-strand replication, at least one additional enzyme, DNA primase, is needed to form the short RNA primers that will be elongated by pol III (Ref. 8), and it remains unclear whether pol III or pol I fills the gaps after these primers are removed[9]. In mammalian cells, four nuclear DNA polymerases have been described, but the exact role of each during leading- and lagging-strand replication has yet to be resolved. It seems clear that pol α with its associated primase activity is at least partially involved in lagging-strand synthesis, whereas pol δ or ϵ could be responsible for synthesis of the leading strand[10,11]. Because the various polymerases have substantially different error rates[11], the use of different enzymes during leading- and lagging-strand synthesis allows for variation in error rates between the strands.

In addition to the different enzymatic requirements for synthesis of the leading and lagging strands, the replication fork is structurally asymmetrical: leading-strand replication proceeds by unwinding very short templates before replication, while lagging-strand synthesis requires the exposure of long single-stranded regions. This might facilitate template–primer misalignments[11,12] and increase damage to the template strand.

Transcription is also asymmetric and can introduce biases in the patterns of mutation on the two strands. In the transcription elongation complex *in vivo*, the DNA helix is unwound over 17–18 bp, and the DNA template and the nascent RNA form a hybrid of 10–12 bp, while the nontranscribed DNA strand remains transiently single stranded[13]. Thus, transcription might bias mutation patterns between the transcribed and nontranscribed strands by exposing the nontranscribed strand to DNA damage.

## Evolutionary consequences of strand asymmetries

Strand asymmetries in the frequency of mutations would influence the evolution and composition of DNA sequences. With no strand bias, patterns of substitution are expected to be symmetrical, with complementary changes (e.g. C→T and G→A, or T→A and A→T) occurring at equal frequencies. Strand biases disrupt this symmetry by creating patterns of change where certain substitutions are more common than their complements, thereby generating inequalities between the frequencies of complementary bases (Fig. 1). In addition, if the nucleotide composition of a strand is influenced by the way it is replicated, strand asymmetries could contribute to the compositional structure of entire chromosomes.

On the basis of computer simulations, it has also been suggested that asymmetric error rates between the DNA strands might be adaptive at the population level[14,15]. In these simulations, strand-symmetric and -asymmetric models of evolution were compared, and populations under the asymmetric regime were found to evolve faster, to attain higher fitness and to avoid extinction. Presumably, asymmetry confers an advantage by increasing the variance in number of mutations accumulated per individual (because individuals descending from leading-strand replication have fewer than the average number of errors, whereas those descending from lagging-strand replication have more), thereby maintaining individuals with few or no mutations as a baseline, from which the population can only evolve towards higher fitness.

In this review, we examine the evidence relating to asymmetrical mutation rates between the DNA strands, and address some of the controversies about their impact on sequence evolution.




PII: S0168-9525(97)01118-9

## Experimental evidence of strand asymmetries

*Leading- versus lagging-strand asymmetry*

Strand bias in rates of deletion has been monitored in ColE1-derived plasmids, whose unidirectional replication is performed by *Escherichia coli* proteins. To detect deletions on each strand, two versions of the plasmid are constructed, each containing the gene for which deletions are being scored in the opposite orientation (thereby reversing the direction in which the transcribed strand is replicated). In two systems – one involving deletions of a palindrome[12] and the other monitoring single- and two-nucleotide deletions caused by *N*-2-acetylaminofluorene[16] – deletions were nearly 20-fold higher on the lagging strand.

Using an *in vitro* system of replication in human cell extracts, Kunkel and coworkers have scored mutations on two M13 vectors having the SV40 origin of replication on opposite sides of the mutational target. The results of these studies vary depending on the types of mutations scored and on the method employed to induce them. Replication errors induced by an excess of dTTP are much more prevalent on the lagging strand[17]; however, there is little difference in error rates between the strands with excess dCTP, dGTP or the base analog 8-O-dGTP, which pairs with adenine to cause A.T→C.G transversions[18–20]. Experiments involving DNA damaging agents have detected similar rates, but sometimes different patterns, of base substitutions on each strand and higher frequencies of deletions on the lagging strand[21–23]. Thus, the relative fidelity of leading- and lagging-strand replication varies with the type of mutation and the site considered.

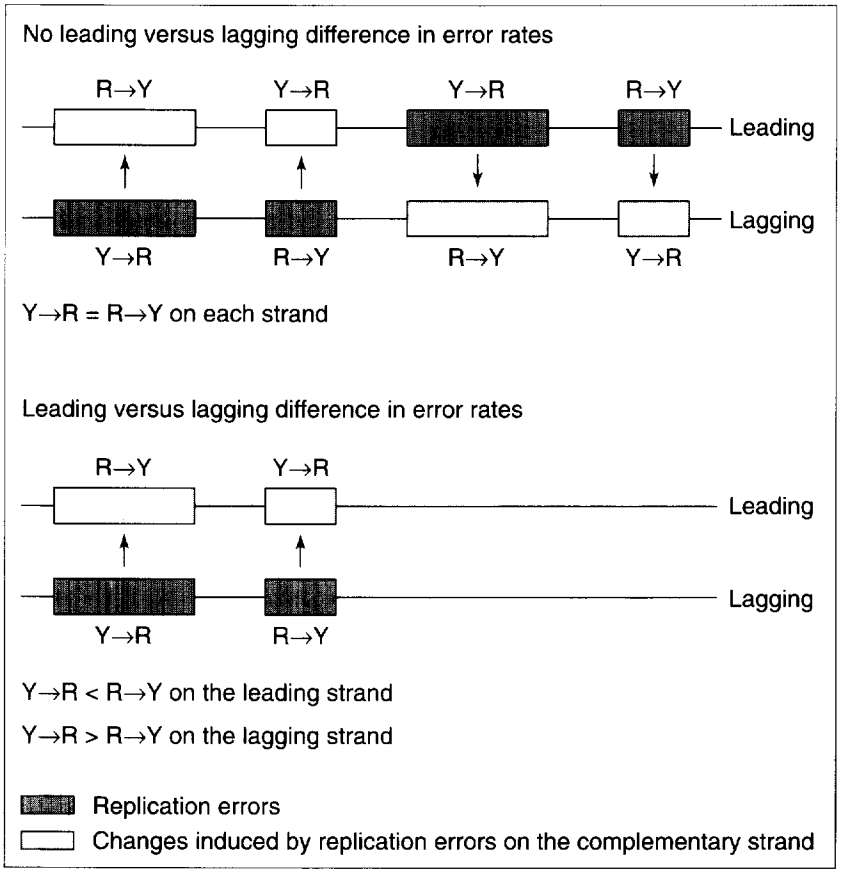



**FIGURE 1.** Unequal error rates for the leading and lagging strands as a cause of asymmetry between complementary substitution rates. Most transversions arise when a purine is introduced opposite a template purine[40–44], causing a pyrimidine to purine (Y→R) error on the synthesized strand that will induce a purine to pyrimidine (R→Y) replacement on the template strand at the next round of replication if not corrected. If error rates were the same on both strands, the total frequencies of Y→R and R→Y changes on a strand (those incurred during synthesis of that strand plus those induced by complementary errors on the opposite strand) would be equal. On the contrary, if there was a bias between the strands, errors occurring on the strand with a higher error rate would not be compensated by complementary errors on the other and, therefore, the frequencies of Y→R and R→Y changes would differ on each strand. On the strand with a higher error rate, Y→R changes would be more frequent and an accumulation of purines would ensue. The bottom half of the figure illustrates a hypothetical case where all replication errors occur on the lagging strand.

*Transcribed- versus nontranscribed-strand asymmetries*

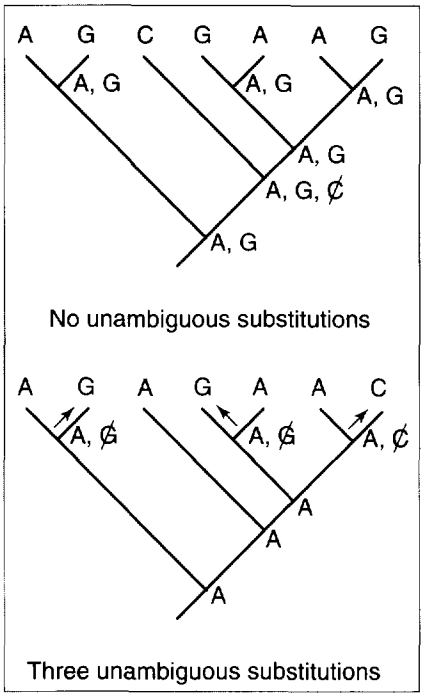
There are two transcription-dependent processes that result in different mutation rates on the transcribed and nontranscribed strands of a gene: transcription-coupled repair and deamination. Transcription-coupled repair is a process that corrects lesions in the transcribed strand of expressed genes and is known to occur in bacteria, yeast and mammalian cells. Bulky lesions, such as UV-induced cyclobutane pyrimidine dimers, block transcription when present on the template strand by causing the RNA polymerase to stall. These stalled polymerases are recognized by a transcription–repair coupling factor, which promotes the activity of nucleotide-excision-repair enzymes at the site. After completion of repair, transcription can resume or recommence[24]. As predicted by the transcription-coupled repair process, C→T mutations induced by pyrimidine dimers are recovered at a much higher frequency on the nontranscribed strand of active genes[25].

While RNA is being synthesized on the transcribed strand of DNA, a portion of the nontranscribed strand remains single stranded, which makes it prone to deamination. C deaminates to U, or T if methylated, over 100 times faster in single- than in double-stranded DNA (Ref. 26), and it has been shown recently that transcription causes an approximately fourfold increase in the frequency of C→U, or 5-methylcytosine→T, deaminations in the nontranscribed strand[27].

## Evidence from sequence analyses

Inequalities between the strands in replication error rates, propensity to damage or efficiency of repair can also be detected through sequence analysis and phylogenetic reconstruction of substitutions (Box 1).

**Box 1. Detecting asymmetries in substitutional patterns: phylogenetic reconstruction of substitutions**




On an evolutionary timescale, strand biases in the introduction of mutations could strongly affect the patterns of change in DNA sequences, and asymmetries would be evident when reconstructing substitutions among homologous regions. To study strand asymmetries, homologous sequences are aligned and their phylogenetic relationships established. Maximum parsimony methods can then be employed to infer the ancestral states at the nodes of the phylogeny and the directionality of the substitutions.

These methods use the topology of the phylogenetic tree and the states observed in the terminal taxa to reconstruct the states at the ancestral nodes that require the fewest evolutionary steps. Algorithms have been developed that include information from the whole tree when reconstructing ancestral states at any given node. During an initial pass down the tree – from the terminal taxa to the root – preliminary sets of states are sequentially determined for each node by combining the sets already calculated for the two nodes immediately above: a node is either assigned the states that are shared by both these nodes or, when no shared states exist, is assigned all of the states in these two nodes. In a subsequent pass up the tree, the sets of states from the ancestor and sister node are combined in the same manner with the set established during the downpass to yield the most parsimonious states at that node[47]. When more than one most parsimonious state is possible at a given node, the directionality of substitutions in the descendants of that node cannot be unambiguously determined, as shown in the tree at the top.

Once the substitutions have been reconstructed on the phylogeny, the frequencies of complementary changes are compared. Inequalities between these frequencies are indicative of different mutation rates between the DNA strands. In particular, different frequencies of pyrimidine to purine (Y→R) and purine to pyrimidine (R→Y) transversions are expected if error rates are sufficiently different for the leading and lagging strands (Fig. 1).

*Phylogenetic reconstruction of substitutions: asymmetries between the leading and lagging strands*

Wu and Maeda[28] tested for asymmetries between complementary substitution rates in a 3.1 kb intergenic region of the β-globin complex from six species of primates. Because this region has no known function and is not constrained by selection, the accumulation of substitutions should reflect the underlying rate and pattern of mutations. On the nontranscribed strand, pyrimidine to purine (Y→R) substitutions were significantly less common than the complementary change, suggesting that error rates differed on the two strands.

Bulmer[29] re-analyzed the mutational patterns of this intergenic region and, in contrast to Wu and Maeda[28], detected no significant difference between complementary substitution rates. Several factors probably contributed to this discrepancy: first, the data sets analyzed were not completely identical (Wu and Maeda[28] used two human sequences and discarded hypervariable regions) and second, Wu and Maeda[28] included substitutions that were indirectly inferred rather than being unambiguously reconstructed on the phylogeny by parsimony. Furthermore, in the adjacent 8 kb fragment from the β-globin complex, Bulmer[29] did not detect any strand asymmetries in the pattern of substitutions.

The asymmetry originally reported by Wu and Maeda[28] is opposite of that expected based on the position of a nearby origin of replication[30]. The region that they analyzed is downstream of this replication origin and probably synthesized as the lagging strand. Therefore, if lagging-strand synthesis is error-prone, as suggested by some of the experimental evidence, this region should have an excess of Y→R over R→Y transversions (Y, pyrimidine; R, purine; Fig. 1). However, the direction of synthesis of this region might change if secondary replication origins are used, or if the locations of replication origins have not been conserved among these species[31], thereby confounding the analysis of strand-specific substitution patterns.

In contrast to the numerous origins of replication in eukaryotic chromosomes, bacteria have a unique origin, so that the mode of replication of a given region is likely to be conserved during evolution. Moreover, whenever the map position and orientation of a bacterial sequence are known, it can be unequivocally established whether it is replicated as a leading or as a lagging strand. Analysis of the substitutional patterns in genes of natural strains of the enteric bacteria *Escherichia coli* and *Salmonella enterica* revealed no asymmetries between complementary transversions Y→R and R→Y, suggesting that error rates are similar for the leading and lagging strands during replication of the bacterial chromosome[32].

*Phylogenetic reconstruction of substitutions: asymmetries between the transcribed and nontranscribed strands*

Although bacterial genes have no asymmetry in substitution rates on the leading and lagging strands, there is an asymmetry between the frequencies of certain complementary substitutions: C→T transitions are more frequent than G→A when substitutions are scored on the nontranscribed strand, independently of how this strand is replicated (Fig. 2). Therefore, this asymmetry is based on a process that distinguishes between the transcribed and the nontranscribed strands of a gene, not on the mode of replication[32].

Although natural selection can favor certain substitutions over their complementary changes, it is probably not the main source of asymmetries in enterobacterial

genes. Because most substitutions occur at third codon positions (where transitional changes almost invariably result in synonymous substitutions), only selection acting on codon choice could potentially create the detected asymmetry. Analyses restricted to positions where transitions do not change the adaptiveness of the codon still detect the C→T versus G→A asymmetry, showing that codon preferences cannot explain this asymmetry, although they can certainly contribute to it (M.P. Francino, unpublished).

Both transcription-coupled repair and deamination have been shown experimentally to produce an excess of C→T mutations on the nontranscribed strand in *E. coli*[25,27] and could account for the observed difference between C→T and G→A substitution frequencies in genes from natural isolates of enteric bacteria. According to the distribution of C→T transitions in the genes analyzed, both processes are probably contributing to the asymmetry and, because both involve DNA damage, a substantial fraction of naturally-occurring mutations might be attributable to unrepaired damage to the DNA molecule.

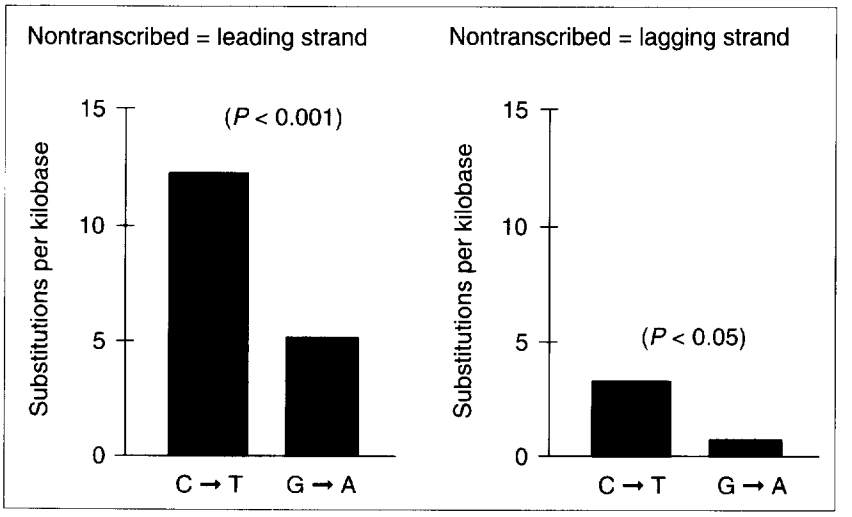



**FIGURE 2.** Complementary transition frequencies of genes coded on the leading and lagging strands of the *Escherichia coli* chromosome. Sequences from a minimum of 11 *E. coli* natural strains were analyzed for each of the following genes: *phoA*, *putP*, *gnd* and *mdh* for the leading strand, and *crr* and *gutB* for the lagging. For each gene, substitutions were scored on the nontranscribed strand along the most parsimonious phylogenetic tree relating the analyzed sequences. Total substitution frequencies for each class of genes (those encoded on the leading strand and those encoded on the lagging strand) were obtained by summing the numbers of substitutions for each gene within each class and dividing by the total number of the original nucleotide (C or G) across genes in that class (×1000). (Reproduced with permission[32].)

*Analysis of intraspecific sequence polymorphisms*

Tanaka and Ozawa[33] compared mitochondrial DNA (mtDNA) sequences from 43 humans and detected a strong bias in the occurrence of nucleotide substitutions at fourfold degenerate sites: C→T and A→G transitions were nine and 1.8 times more frequent on the heavy (H) strand than on the light (L) strand. This asymmetric substitution pattern accounts for the high A+C content at third codon positions on the L strand of human mtDNA (Ref. 34); and intergenic regions have an equally biased base composition[35], indicating that the same strand asymmetry in the pattern of mutations is affecting all of the mitochondrial genome. Although the same polymerase is thought to be responsible for the synthesis of both strands of mtDNA, replication is highly asymmetric: H strand synthesis starts at OriH long before synthesis of the L strand is initiated at OriL, 11 kb away[36]. Given that replication of mtDNA is very slow (an entire cycle is completed in two hours, which represents a polymerization rate about 200-fold slower than that of *E. coli*[36]), the parental H strand remains single stranded and exposed to DNA damage for an extended period of time. This could produce a much higher rate of cytosine deamination on the H strand, which might explain why this strand has an increased occurrence of C→T transitions in mitochondria. In contrast to the situation in bacteria, transcription cannot account for the C→T asymmetry, because the transcribed strand of most mitochondrial genes is the H strand. Apparently, the prolonged exposure of the H strand during mitochondrial replication induces a much higher frequency of cytosine deaminations than does the transient single-strandedness of the nontranscribed L strand during transcription. Moreover, transcription-coupled repair does not contribute to strand asymmetry in mitochondria, because these organelles lack nucleotide excision repair.

*Detecting asymmetries based on nucleotide composition*

If there were no biases between the two DNA strands for either mutation or selection, the equilibrium frequencies of the four bases in each strand would be [A]=[T] and [C]=[G]. Therefore, significant deviations from intrastrand parity indicate different substitutional patterns in each of the strands[28,37–39].

Because the introduction of a pyrimidine opposite a template pyrimidine is a very rare event relative to other mismatches[40–44], an error-prone lagging strand would accumulate an excess of purines, and the leading strand would accumulate pyrimidines (Fig. 1). This would result in eukaryotic chromosomes consisting of alternating tracts with opposite biases, with a pyrimidine excess on one side of the origin and a purine excess on the other. In the 73 kb β-globin region in humans, the frequency of pyrimidines in 1 kb segments is more variable than expected for a random sequence, switching frequently from low to high. This small-scale patterning might still be consistent with a higher error rate on the lagging strand, but only if germ-line cells recruited many secondary origins for rapid replication, thereby restricting replicon size to a few kilobases[29].

In bacteria, highly asymmetric patterns of base composition have been detected in the genomes of *E. coli*, *Bacillus subtilis*, *Haemophilus influenzae* and *Mycoplasma genitalium*[39,45]. Analyses of nucleotide sequences spanning extensive regions of the chromosomes of

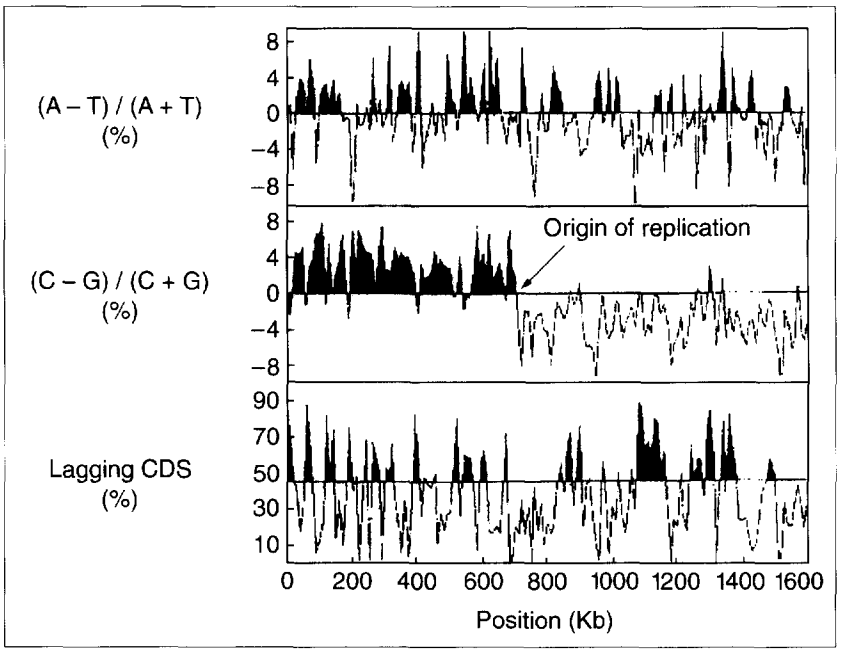



**FIGURE 3.** Base frequency deviations and abundance of coding sequences (CDS) in the lagging strand along a fragment of the *Escherichia coli* chromosome. Two indices of base frequency are plotted: (A–T)/(A+T), which measures deviations from [A]=[T], and (C–G)/(C+G), which measures deviations from [C]=[G]. A nonoverlapping moving window of 10 kb was used in the calculations. The horizontal line for (A–T) and (C–G) deviations is drawn at 0, corresponding to what is expected under conditions of no strand bias. There is a small, but statistically significant, deviation from [A]=[T] and a very large deviation from [C]=[G], and both change sign at the origin of replication. The bottom panel shows the frequency of codon positions present in the lagging strand of this chromosomal segment; the horizontal line is drawn at the mean value of 46%. The (A–T) and (C–G) deviations are dominated by the deviation resulting from codon positions in the leading strand. When there is a local abundance of lagging coding sequences, the general trend is reversed, yielding reverse asymmetry spikes. (Reproduced with permission[39].)

these species have revealed that the frequencies of A and T, as well as those of C and G, deviate from equality, thus indicating strand-bias in the patterns of substitution (Fig. 3). Moreover, the (A–T) and (C–G) deviations are more pronounced in intergenic regions and at third codon positions, suggesting that these asymmetries are generated by a mutational bias[39].

The (A–T) and (C–G) deviations detected in bacteria switch direction at the replication origin. This is due to differences in base composition between genes on the leading and lagging strands, which suggests that the two strands undergo unequal patterns of replication errors[39]. However, genes are not randomly distributed between the two strands and, in fact, for genes within the 1.6 Mb region of the *E. coli* chromosome in which these deviations were detected, the majority of highly expressed genes are coded on the leading strand. Analysis of the codon adaptation index (CAI), a measure of codon bias, which is strongly associated with levels of gene expression[46], for genes within this region reveals that the leading strand contains 55% of all genes, but 74% of those with CAI values ≥0.5 and 84% of those with CAI values ≥0.6, and has an average CAI significantly above that of the lagging strand (CAI = 0.37 versus CAI = 0.34, $P = 0.0001$; M.P. Francino, unpublished).

Differences in the overall levels of gene expression can contribute to the observed disparities in base composition between genes on the leading and lagging strands in two ways. (1) Highly expressed genes utilize a very limited set of codons, and the third codon positions for these genes will differ from those with less codon bias. (2) Frequent expression will increase the effects of single-strand deamination and propitiate transcription-coupled repair. Therefore, deviations from [A]=[T] and [C]=[G] introduced by the transcription-induced C→T versus G→A asymmetry should be more pronounced on the leading strand of *E. coli*, which contains the majority of highly expressed genes. Indeed, genes on the leading strand have larger excesses of Gs and Ts than do those coded on the lagging strand, indicating that the effects of transcription on mutagenesis influence the compositional structure of the *E. coli* genome.

## Conclusions

Substantial experimental evidence has established that the two strands of DNA are not equivalent with respect to their rates and patterns of mutation. Clearly, the transcribed status of a DNA strand will affect the levels of damage it incurs and the likelihood that damage will be repaired. In addition, the lagging strand is more susceptible to certain errors, such as some types of mismatches during *in vitro* replication in human cell extracts and deletions during plasmid replication in *E. coli*, but it is not known whether such strand biases also arise during *in vivo* chromosomal replication.

How are these differences manifested over the long-term evolution of DNA sequences? The large-scale nucleotide composition of bacterial chromosomes might be attributable to differences between the patterns of mutation in the leading and lagging strands, but it might also reflect the organization of genes in these genomes. Phylogenetic reconstructions of substitutions in eukaryotes and in bacteria have detected no unequivocal asymmetries caused by differences between the leading and lagging strands. Therefore, mitochondrial genomes provide the only sound evidence of sequence evolution being affected by different rates of mutation on the two strands of replication. However, the H and L strands of mtDNA are not equivalent to the leading and lagging strands of bacterial and eukaryotic chromosomes, for example, neither strand of mtDNA is replicated discontinuously, and the asymmetries observed in mtDNA clearly result from the particularities of mitochondrial replication.

By contrast, the highly asymmetric effects of transcription on mutagenesis, through unequal exposure of the strands to DNA damage and differential opportunity for repair, are strongly reflected in the substitution patterns of bacterial genes and might be involved in shap-

ing genome-wide patterns of base composition. Future work will evaluate whether transcription is influencing the evolution of genes in eukaryotes and how the tissue-specific expression of genes affects their rates and patterns of evolution.

## References


1 Sharp, P.M., Shields, D.C., Wolfe, K.H. and Li, W-H. (1989) *Science* 246, 808–810
2 Wolfe, K.H., Sharp, P.M. and Li, W-H. (1989) *Nature* 337, 283–285
3 Benzer, S. (1961) *Proc. Natl. Acad. Sci. U. S. A.* 47, 403–415
4 Gojobori, T., Li, W-H. and Graur, D. (1982) *J. Mol. Evol.* 18, 360–369
5 Blake, R.D., Hess, S.T. and Nicholson-Tuell, J. (1992) *J. Mol. Evol.* 34, 189–200
6 Bloom, L.B., Otto, M.R., Beechem, J.M. and Goodman, M.F. (1993) *Biochemistry* 32, 11247–11258
7 Kornberg, A. and Baker, T. (1992) in *DNA Replication*, pp. 471–510, Freeman
8 Stillman, B. (1994) *Cell* 78, 725–728
9 Stukenberg, P.T., Turner, J. and O'Donnell, M. (1994) *Cell* 78, 877–887
10 Waga, S. and Stillman, B. (1994) *Nature* 369, 207–212
11 Kunkel, T.A. (1992) *Bioessays* 14, 303–308
12 Trinh, T.Q. and Sinden, R.R. (1991) *Nature* 352, 544–547
13 Kainz, M. and Roberts, J. (1992) *Science* 255, 838–841
14 Furusawa, M. and Doi, H. (1992) *J. Theor. Biol.* 157, 127–133
15 Wada, K-N. *et al.* (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 11934–11938
16 Veaute, X. and Fuchs, R.P.P. (1993) *Science* 261, 598–600
17 Roberts, J.D., Izuta, S., Thomas, D.C. and Kunkel, T.A. (1994) *J. Biol. Chem.* 269, 1711–1717
18 Roberts, J.D., Thomas, D.C. and Kunkel, T.A. (1991) *Proc. Natl. Acad. Sci. U. S. A.* 88, 3465–3469
19 Minnick, D.T., Pavlov, Y.I. and Kunkel, T.A. (1994) *Nucleic Acids Res.* 22, 5658–5664
20 Izuta, S., Roberts, J.D. and Kunkel, T.A. (1995) *J. Biol. Chem.* 270, 2595–2600
21 Basic-Zaninovic, T., Palombo, F., Bignami, M. and Dogliotti, E. (1992) *Nucleic Acids Res.* 20, 6543–6548
22 Thomas, D.C., Nguyen, D.C., Piegorsch, W.W. and Kunkel, T.A. (1993) *Biochemistry* 32, 11476–11482
23 Thomas, D.C., Svoboda, D.L., Vos, J.M. and Kunkel, T.A. (1996) *Mol. Cell. Biol.* 16, 2537–2544
24 Hanawalt, P. (1995) *Mutat. Res.* 336, 101–113
25 Oller, A.R., Fijalkowski, I.J., Dunn, R.L. and Schaaper, R.M. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 88, 11036–11040
26 Lindahl, T. and Nyberg, B. (1974) *Biochemistry* 13, 3405–3410
27 Beletskii, A. and Bhagwat, A.S. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 13919–13924
28 Wu, C-I. and Maeda, N. (1987) *Nature* 327, 169–170
29 Bulmer, M. (1991) *J. Mol. Evol.* 33, 305–310
30 Kitsberg, D., Selig, S., Keshet, I. and Cedar, H. (1993) *Nature* 366, 588–590
31 Wu, C-I. (1991) *Nature* 352, 114
32 Francino, M.P., Chao, L., Riley, M.A. and Ochman, H. (1996) *Science* 272, 107–109
33 Tanaka, M. and Ozawa, T. (1994) *Genomics* 22, 327–335
34 Asakawa, S. *et al.* (1991) *J. Mol. Evol.* 32, 511–520
35 Jermiin, L.S., Graur, D. and Crozier, R.H. (1995) *Mol. Biol. Evol.* 12, 558–563
36 Clayton, D.A. (1982) *Cell* 28, 693–705
37 Sueoka, N. (1995) *J. Mol. Evol.* 40, 318–325
38 Lobry, J.R. (1995) *J. Mol. Evol.* 40, 326–330
39 Lobry, J.R. (1996) *Mol. Biol. Evol.* 13, 660–665
40 Hibner, U. and Alberts, B.M. (1980) *Nature* 285, 300–305
41 Fersht, A.R. and Knill-Jones, J.W. (1981) *Proc. Natl. Acad. Sci. U. S. A.* 78, 4251–4255
42 Sinha, N.K. and Haimes, M.D. (1981) *J. Biol. Chem.* 256, 10671–10683
43 Kunkel, T.A. and Alexander, P.S. (1986) *J. Biol. Chem.* 261, 160–166
44 Echols, H. and Goodman, M.F. (1991) *Annu. Rev. Biochem.* 60, 477–511
45 Lobry, J.R. (1996) *Science* 272, 745–746
46 Sharp, P.M. and Li, W-H. (1987) *Mol. Biol. Evol.* 4, 222–230
47 Maddison, W.P. and Maddison, D.R. (1992) in *MacClade v3.0. Analysis of Phylogeny and Character Evolution*, pp. 86–92, Sinauer


***M.P. Francino** and **H. Ochman** are in the Department of Biology, University of Rochester, Rochester, NY 14627, USA.*