Letter to the Editor

Deamination as the Basis of Strand-Asymmetric Evolution in Transcribed *Escherichia coli* Sequences

M. Pilar Francino and Howard Ochman*†*

*Department of Biology, University of Rochester; †Department of Ecology and Evolutionary Biology, University of Arizona

Analyses of sequence evolution in *Escherichia coli* and *Salmonella enterica* have revealed that the pattern of nucleotide substitutions in enterobacterial genes is asymmetric. The incidence of C→T transitions is strongly biased toward the nontranscribed strand of DNA, which accumulates such changes at a two- to threefold higher rate than the complementary transcribed strand. We previously proposed that the asymmetric distribution of C→T substitutions was caused by strand-specific biases in the occurrence and repair of DNA damage during transcription (Francino et al. 1996; Francino and Ochman 1997). Two processes render mutations less likely to originate on the transcribed template strand than on its complement: (1) transcription-coupled repair is induced by RNA polymerases stalled at lesions on the template strand (Hanawalt 1995), and (2) cytosine deamination is less frequent on the template strand, which is shielded by the RNA polymerase and the nascent mRNA, than on the more exposed nontranscribed strand (Beletskii and Bhagwat 1996). However, the pattern of nucleotide substitutions in coding regions may reflect not only the underlying mutational process, but also the action of natural selection. Given that the majority of C→T substitutions in bacterial sequences occur at synonymous sites, selection on codon usage could potentially contribute to the generation of the observed asymmetry.

To determine whether transcription alone, without the intervention of natural selection on codon usage, produces substitutional asymmetry in bacterial sequences, we analyzed patterns of substitution in two different noncoding regions: a transcribed but untranslated region, and the adjacent nontranscribed sequence. The detection of substitutional bias in the transcribed but untranslated region and the absence of bias in the nontranscribed sequence would confirm that transcription is necessary and sufficient to generate asymmetry, without a requirement for selection on codon usage. Furthermore, an increase in C→T substitutions with transcription would implicate deamination events in the nontranscribed strand as the principal cause of asymmetry, whereas a decrease in G→A substitutions with transcription would implicate transcription-coupled repair of pyrimidine dimers on the complementary template strand.

Because bacterial chromosomes are so tightly packed with genes, noncoding sequences of appropriate length for analysis of rates and patterns of substitution are scarce. Intergenic regions in *E. coli* are typically very short—averaging only 118 bp in length—but the complete genomic sequence of *E. coli* MG1655 has revealed some intergenic regions of much larger size (Blattner et al. 1997). However, when we investigated sequence variation in several of the longest untranslated regions and nontranscribed sequences among natural strains of *E. coli*, we found that surprisingly low levels of divergence in most of these regions precluded the analysis of substitutional patterns (data not shown). Therefore, we restricted our analysis to a region which contained sufficient nucleotide sequence diversity to investigate the effect of transcription on the pattern of nucleotide substitutions. This region, the *cysB*-*acnA* region at 28.7 min on the *E. coli* chromosome, is well suited for this type of analysis because the transcription start points of *acnA* have been experimentally established (Cunningham, Gruer, and Guest 1997) and a very likely rho-independent terminator has been located near the end of *cysB* (Prodromou, Artymiuk, and Guest 1992). These transcription signals clearly delineate the transcribed and nontranscribed sequences between the *cysB* and *acnA* genes (fig. 1*A*).

The noncoding regions were amplified using polymerase chain reaction (PCR) primers located within the surrounding coding sequences in *E. coli* MG1655. PCR amplification was performed in 12 strains encompassing each of the major phylogenetic subdivisions of the ECOR collection (Ochman and Selander 1984; Herzer et al. 1990): strains ECOR 1 and ECOR 6 from group A; strains ECOR 28, ECOR 58, ECOR 69, and ECOR 71 from group B1; strains ECOR 51, ECOR 61, and ECOR 66 from group B2; strains ECOR 35 and ECOR 50 from group D; and strain ECOR 37 from group E. Purified PCR products (QIAquick PCR Purification or Gel Extraction Kits, QIAGEN) were used as templates for automated sequencing (ABI Prism BigDye Terminator Cycle Sequencing Kit, Applied Biosystems). Internal sequencing primers were designed as needed from accumulated sequence data. Sequences were assembled, edited, and aligned with SEQUENCHER 3.1RC12 (GeneCodes, Ann Arbor, Mich.).

The pattern of nucleotide substitutions was reconstructed on the phylogeny relating the *cysB*-*acnA* sequences as estimated by the neighbor-joining method. The sequence alignment and tree topology were analyzed with MACCLADE, version 3.0 (Maddison and Maddison 1992) to reconstruct the most parsimonious ancestral states and the directionality of the nucleotide

Present address: Laboratoire de Génétique Moléculaire Evolutive et Médicale, E9916 INSERM Faculté de Médecine-''Necker-Enfants malades'', Université René Descartes-Paris V, Paris, France.

Key words: cytosine deamination, substitution patterns, strand asymmetry, transcription-induced mutations, noncoding DNA, *Escherichia coli*.

Address for correspondence and reprints: Howard Ochman, Department of Ecology and Evolutionary Biology, 233 Life Sciences South, University of Arizona, Tucson, Arizona 85721. E-mail: hochman@email.arizona.edu.

Mol. Biol. Evol. 18(6):1147–1150. 2001

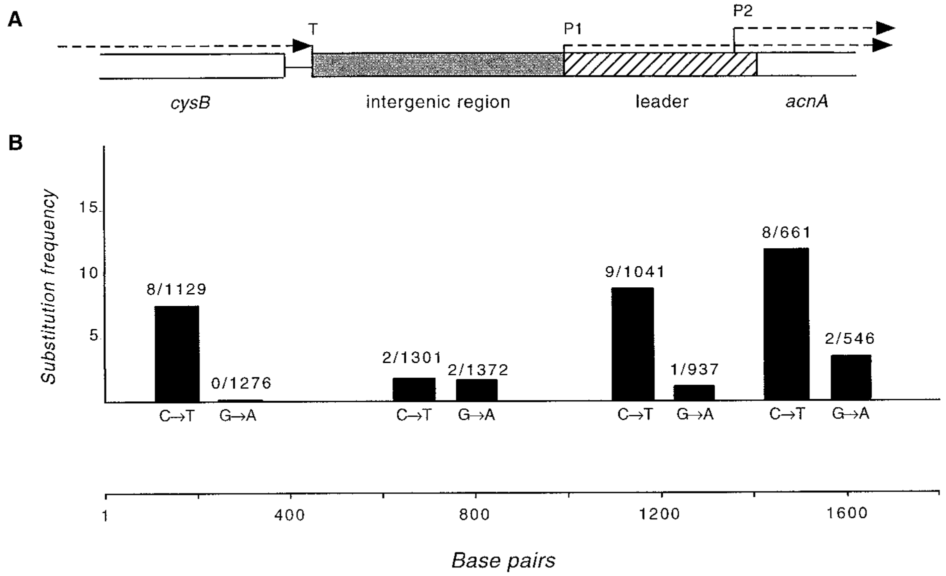

FIG. 1.—*A,* Structure of the *cysB-acnA* region at 28.7 min on the leading strand of the *Escherichia coli* MG1655 chromosome. Dashed arrows indicate the direction of transcription; T = terminator (Prodromou, Artymiuk, and Guest 1992), P1 and P2 = promoters 1 and 2 (Cunningham, Gruer, and Guest 1997). *B,* Complementary transition frequencies in the different portions of the *cysB-acnA* region, expressed as the minimal number of C→T (or G→A) transitions divided by the total number of C's (or G's) over all strains, × 1,000. The ratio of the minimal number of C→T (or G→A) transitions over the total number of C's (or G's) for each region is given above the corresponding histogram bar. Because of the small number of changes, the statistical significance of the asymmetry in each transcribed region was determined by calculating the binomial probabilities of obtaining the observed or larger excesses of C→T over G→A given an equal rate of C→T and G→A substitutions and taking into account the numbers of C's and G's in each region. *P* values are 0.017 (leader), 0.002 (*cysB*), and 0.100 (*acnA*). An analogous procedure was applied to test the statistical significance of the excesses of C→T substitutions in each of the transcribed regions over the C→T substitutions in the nontranscribed region. *P* values for the comparisons between the nontranscribed region and each of the transcribed regions are 0.032 (vs. *cysB*), 0.012 (vs. untranslated region), and 0.004 (vs. *acnA*).

substitutions. Minimal numbers of substitutions were used in subsequent analyses, i.e., substitutions that occurred with certainty but whose localization within the tree topology may or may not be determined. Substitution frequencies were obtained by dividing the observed occurrences of a given substitution by the total number of nucleotides of the type undergoing the substitution among all sequences (×1,000).

The nontranscribed intergenic region and the adjacent mRNA leader underwent different patterns of nucleotide substitution. The transcribed but untranslated leader showed a significant excess of C→T over G→A transitions, similar to that in the sequenced portions of the surrounding coding sequences, *cysB* and *acnA.* In contrast, C→T-over-G→A asymmetry was not apparent in the nontranscribed intergenic sequence. For each region, C→T and G→A substitution frequencies are graphed in fig. 1*B*, which also specifies the absolute numbers of C→T and G→A substitutions, over the numbers of C's and G's. Although the number of changes in the nontranscribed region was too small to affirm that no asymmetry existed, the fact that this region underwent a significantly lower frequency of C→T changes than transcribed regions containing fewer C's strongly suggests a lack of asymmetry in the region. Therefore, the situation in the *cysB-acnA* region supports the hypothesis that the C→T-versus-G→A asymmetry is generated during transcription by indicating (1) that the asymmetry is not apparent in regions that are not transcribed, and (2) that translation and selection on amino acid or codon usage are not necessary for asymmetry.

During transcription, two processes have been shown experimentally to affect the generation of mutations in an asymmetric manner between the two DNA strands: transcription-coupled repair (TCR) and cytosine deamination. TCR corrects bulky lesions, particularly UV-induced pyrimidine dimers, which block transcription when present on the template strand by causing the RNA polymerase to stall. Given that C→T transitions are the primary mutations induced by pyrimidine dimers, they are recovered at a much higher frequency on the nontranscribed strand of active genes (Oller et al. 1992). C→T transitions due to deamination are also more frequent on the nontranscribed strand, presumably because cytosines on this strand remain unpaired longer as the transcription bubble proceeds (Beletskii and Bhagwat 1996; Beletskii et al. 2000). However, TCR and cytosine deamination have different effects on the rates of C→T and G→A transitions generated during transcription: deamination causes an increase in C→T transitions (and no change in G→A), whereas transcription-coupled repair causes a decrease in G→A transitions (and no change in C→T) when the nontranscribed strand of regions that undergo transcription is compared with regions that are not transcribed (fig. 2). Hence, comparison of C→T (and G→A) frequencies between transcribed and nontranscribed regions can reveal which of the two strand-asymmetric processes is the main gen-

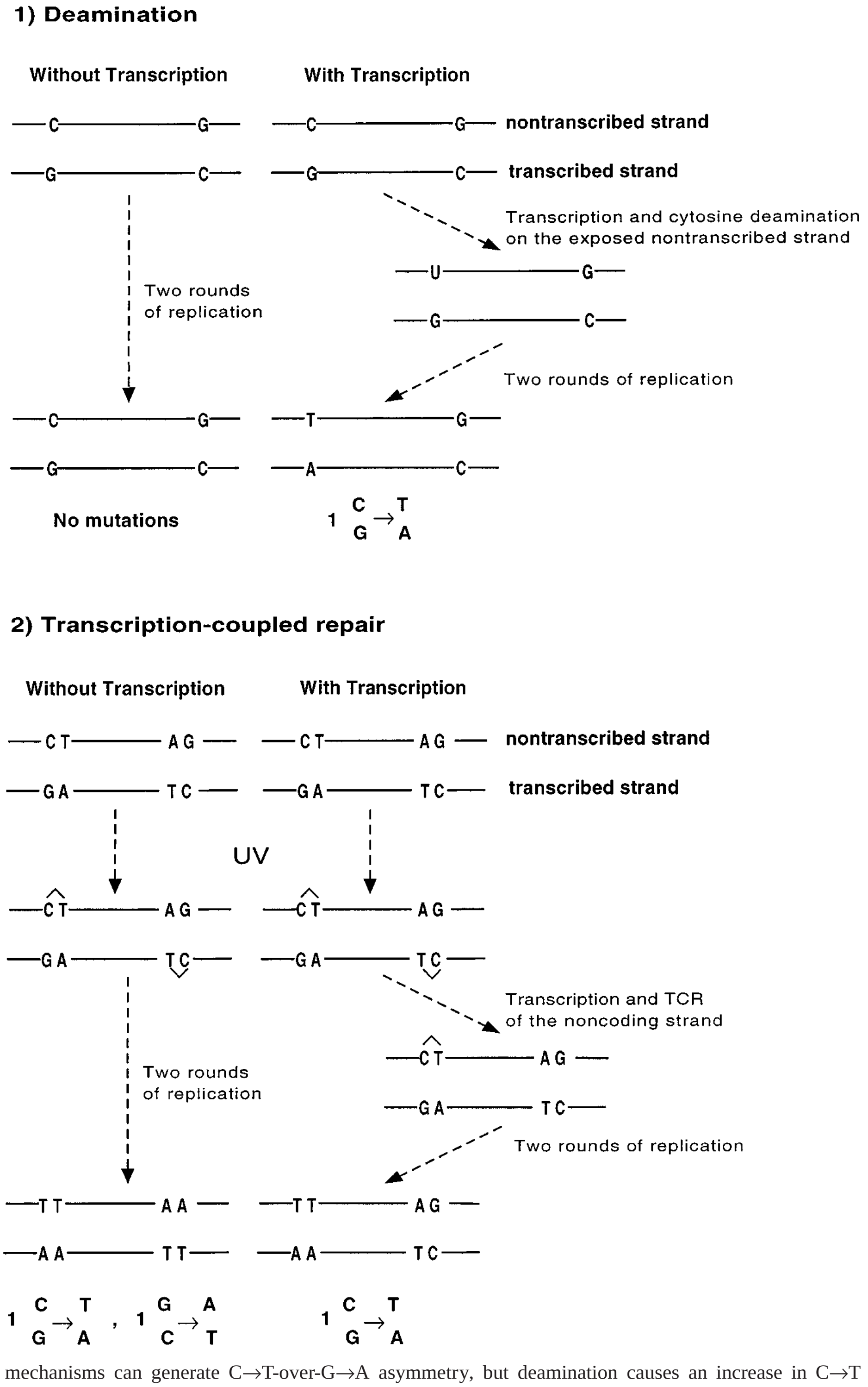
1) Deamination
Without Transcription
With Transcription
nontranscribed strand
transcribed strand
C
G
Two rounds of replication
Transcription and cytosine deamination on the exposed nontranscribed strand
U
Two rounds of replication
T
A
No mutations
1 C/G → T/A
2) Transcription-coupled repair
Without Transcription
With Transcription
CT
AG
GA
TC
nontranscribed strand
transcribed strand
UV
Two rounds of replication
Transcription and TCR of the noncoding strand
Two rounds of replication
TT
AA
1 C/G → T/A , 1 G/C → A/T
1 C/G → T/A

FIG. 2.—Two mechanisms can generate C→T-over-G→A asymmetry, but deamination causes an increase in C→T transitions (and no change in G→A), whereas transcription-coupled repair causes a decrease in G→A transitions (and no change in C→T) when the nontranscribed strand of regions that undergo transcription is compared with regions that are not transcribed.

erator of the asymmetric substitutional pattern observed in transcribed sequences. The frequencies of C→T substitutions in each of the sequences that undergo transcription were all significantly higher than the frequency of C→T substitutions in the nontranscribed intergenic sequence (fig. 1*B*). Numbers of G→A substitutions in both the transcribed and the nontranscribed sequences were similarly low, although there were too few changes for statistical comparison. Nevertheless, the steep increase in C→T substitutions that accompanies transcription suggests that the C→T-versus-G→A asymmetry is generated by excessive deamination of cytosines on the nontranscribed strand rather than by TCR of lesions on the complementary strand.

Mutational asymmetries incurred during transcription will affect base composition, as has been recently shown for bacteriophage T7, for which genes have accumulated excesses of T proportional to their expression levels, presumably because of transcription-induced deamination events on the coding strand (Beletskii et al. 2000). Furthermore, in most bacteria, genes are unevenly distributed between the complementary DNA strands, and therefore transcription-induced asymmetries will contribute to compositional strand biases on a genomewide scale (Francino and Ochman 1997, 1999; Freeman et al. 1998; McLean, Wolfe, and Devine 1998; Mrázek and Karlin 1998; Frank and Lobry 1999; Tillier and Collins 2000). The availability of complete genomic sequences from widely divergent taxa has revealed that base composition is indeed strand-biased in most prokaryotic chromosomes, with the leading strand of replication generally containing an excess of G over C and of T over A (Rocha, Danchin, and Viari 1999). Both a difference in the rates and patterns of mutations on the leading and lagging strands and the overrepresentation of coding sequences on the leading strand have been shown to independently produce base-compositional strand biases in bacterial and viral genomes (Grigoriev 1999; Mackiewicz et al. 1999; Tillier and Collins 2000). By extending the C→T-over-G→A substitutional asymmetry to a transcribed but untranslated region of *E. coli,* we have shown that transcription can promote an excess of G and T within genes and also within the noncoding regions of prokaryotic chromosomes. Furthermore, comparison of the substitutional patterns in transcribed and nontranscribed sequences pinpoints cytosine deamination as the mechanism behind C→T-over-G→A asymmetry. Cytosine deamination has also been implicated in the generation of biased base composition in mitochondria (Reyes et al. 1998) and in the origin of mammalian isochores (Fryxell and Zuckerkandl 2000) and is emerging as a principal agent of large-scale base-compositional structuring in the genomes of viruses and organelles, bacteria, and eukaryotes.

LITERATURE CITED

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

Beletskii, A., A. Grigoriev, S. Joyce, and A. S. Bhagwat. 2000. Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. J. Mol. Biol. **300**:1057–1065.

Blattner, F. R., G. Plunkett III, C. A. Bloch et al. (14 coauthors). 1997. The complete nucleotide sequence of *Escherichia coli* K-12. Science **277**:1453–1462.

Cunningham, L., M. J. Gruer, and J. R. Guest. 1997. Transcriptional regulation of the aconitase genes (*acnA* and *acnB*) of *Escherichia coli.* Microbiology **143**:3795–3805.

Francino, M. P., L. Chao, M. A. Riley, and H. Ochman. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science **272**:107–109.

Francino, M. P., and H. Ochman. 1997. Strand asymmetries in DNA evolution. Trends Genet. **13**:240–245.

———. 1999. A comparative genomics approach to DNA asymmetry. Ann. N.Y. Acad. Sci. **870**:428–431.

Frank, A. C., and J. R. Lobry. 1999. A symmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238**:65–77.

Freeman, J. M., T. N. Plasterer, T. F. Smith, and S. C. Mohr. 1998. Patterns of genome organization in bacteria. Science **279**:1827.

Fryxell, K. J., and E. Zuckerkandl. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. **17**:1371–1383.

Grigoriev, A. 1999. Strand-specific compositional asymmetries in double-stranded DNA viruses. Virus Res. **60**:1–19.

Hanawalt, P. C. 1995. DNA repair comes of age. Mutat. Res. **336**:101–113.

Herzer, P. J., S. Inouye, M. Inouye, and T. S. Whittam. 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli.* J. Bacteriol. **172**:6175–6181.

Mackiewicz, P., A. Gierlik, M. Kowalczuk, M. R. Dudek, and S. Cebrat. 1999. How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. **9**:409–416.

McLean, M., K. H. Wolfe, and K. M. Devine. 1998. Base composition skews, replication orientation and gene orientation in 12 prokaryotic genomes. J. Mol. Evol. **47**:691–696.

Maddison, W. P., and D. R. Maddison. 1992. MacClade v3.0. Analysis of phylogeny and character evolution. Sinauer, Sunderland, Mass.

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

Ochman, H., and R. K. Selander. 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157**:690–693.

Oller, A. R., I. J. Fijalkowska, R. L. Dunn, and R. M. Schaaper. 1992. Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in *Escherichia coli.* Proc. Natl. Acad. Sci. USA **88**:11036–11040.

Prodromou, C., P. J. Artymiuk, and J. R. Guest. 1992. The aconitase of *Escherichia coli.* Eur. J. Biochem. **204**:599–609.

Reyes, A., C. Gissi, G. Pesole, and C. Saccone. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. **15**:957–966.

Rocha, E. P., A. Danchin, and A. Viari. 1999. Universal replication biases in bacteria. Mol. Microbiol. **32**:11–16.

Tillier, E. R., and R. A. Collins. 2000. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. **50**:249–257.

Dan Graur, reviewing editor

Accepted February 13, 2001