J Mol Evol (1996) 43:161–163

JOURNAL OF MOLECULAR EVOLUTION

Letter to the Editor

A Possible Relationship Between VSP Mismatch Repair and Gene Expression Level

Gabriel Gutiérrez,[1] Josep Casadesús,[1] José L. Oliver,[2] Antonio Marín[1]

[1] Departamento de Genética, Universidad de Sevilla, Apartado 1095, E-41080 Sevilla, Spain
[2] Departamento de Genética, Universidad de Granada, Campus de Fuentenueva, E-18071 Granada, Spain

Received: 14 August 1995 / Accepted: 16 February 1996

In a Letter to the Editor, Eyre-Walker (1995) disputes a previous interpretation relating G+C variation in the *Escherichia coli* genome to differential activity of Very Short Patch (VSP) repair and, in turn, to gene expression level (Gutiérrez et al. 1994).

VSP repair corrects T:G mismatches to C:G when they are embedded in special sequence contexts (see Glässner et al. 1995 for a list of substrates and preferences). If the T:G mismatch lies at a site where T is the correct base, a T-to-C transition takes place. Statistical analysis of oligonucleotide frequencies unveiled the mutagenic effect of VSP repair on the *E. coli* genome; as a consequence of VSP activity the genome has been partly depleted of VSP substrates and enriched in VSP outputs (Bhagwat and McClelland 1992; Merkl et al. 1992). The degree of underrepresentation of the different potential VSP repair substrates correlates to their reactivities (Glässner et al. 1995).

While studying two samples of *E. coli* nucleotide sequences selected according to the frequency of the CTAG tetranucleotide, we found that the codon usage of the genes harbored in the CTAG (>0.07%)-containing sample was more similar to that of lowly expressed genes than to that of highly expressed genes. It should be noted that the tetranucleotide CTAG used to divide the samples is the most susceptible to VSP processing, and, indeed, it is the rarest *E. coli* tetranucleotide. Since VSP activity depletes the genome of CTAG, the regions where the observed frequency of CTAG is high can be expected to have been less affected by VSP repair than the regions where CTAG is absent. These observations prompted us to consider a tentative link of VSP repair activity to gene expression level. The latter was established by means of the connection between codon bias and gene expressivity (Gouy and Gautier 1982; Ikemura 1985).

The purpose of the letter by Eyre-Walker appears to be, first, to deny the possibility of variation in the efficiency of VSP repair across the *E. coli* genome, and second to rule out that it may do so in relation to gene expression level. To support his view, he has repeated our analysis on a copious database, reporting that when he compared the average CAI (an indirect measure of gene expressivity, Sharp and Li 1987) of a sample of genes with no CTAG motif against that of a sample of genes with a gene-wide frequency of CTAG higher than 0.07%, he actually found significant differences as we did. Next, he failed to detect significant CAI differences using a narrower CTAG-rich gene sample which contained only genes lacking CTAG in the 1231 codon frame but with the frequency of the CTAG motif higher than 0.1% in the 2312 frame.

The reason for moving from a gene-wide stratification to the restrictive one in the 2312 frame was said to be the presence of the rare CTA leucine codon in the 1231 frame; on this ground, performing the analysis in the

Correspondence to: A. Marín

Table 1. Average CAI values in gene samples stratified according to the frequency of the tetranucleotide CTAG

Sample[a]	Gene number	Mean CAI[b]	*P*
0	1294	0.315 ± 0.003	
1	50	0.268 ± 0.020	<0.0085
2	43	0.260 ± 0.022	<0.0044
3	27	0.226 ± 0.024	<0.0003
4	20	0.246 ± 0.029	<0.0141

[a] Sample 0 contains genes lacking CTAG in all frames. Samples 1, 2, and 3 contain genes with a CTAG frequency in the 2312 frame higher than 0.1%, 0.2%, and 0.3%, respectively, disregarding the possible occurrence of CTAG in the other two frames. Sample 4 contains genes having CTAG only in the 2312 frame with a frequency higher than 0.3%.

[b] CAI values were computed by the method of Sharp and Li (1987) using as reference the data given by Sharp et al. (1988). *P* is the probability of the Student's *t*-tests of comparisons between sample 0 and the others

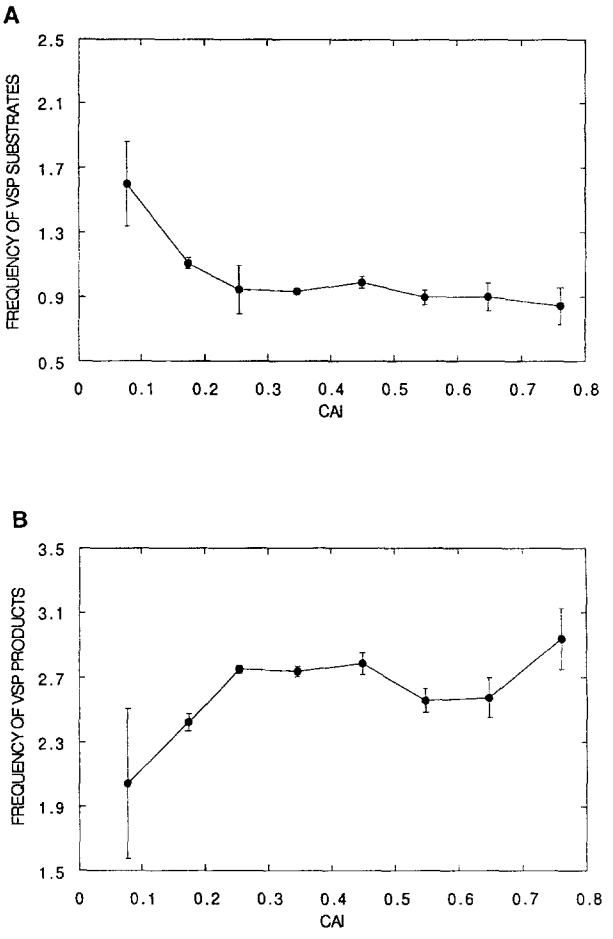

Fig. 1. Fluctuation of VSP substrates **(A)** and products **(B)** along CAI values. Each point corresponds to the mean value of VSP substrates (CTAG, CTTG, CCTA, CCAA, TTGG, TAGG, and CAAG) and products (CCTG, CCAG, CTGG, and CAGG) within a 0.1 CAI interval. *Error bars* indicate the standard error of the mean in each interval.

2312 frame would be similarly questionable as then the rare AGN serine and arginine codons would be present.

We think that using CTAG to stratify the data is justified because CTAG is the tetranucleotide most affected by VSP repair; its high reactivity to Vsr mismatch endonuclease has been biochemically quantitated by measuring differences in the cleavage rates of different Vsr substrates (Glässner et al. 1995); obviously, this fact is not a consequence of the translational fitness of CTA or AGN codons. Translational selection may (in highly expressed genes) or may not (in weakly expressed genes) deal with substitutions to avoid CTA or AGN codons. However, the presence of CTAG tetranucleotides (as well as the higher and lower frequencies of the other VSP substrates and products, respectively) in weakly expressed genes should primarily reflect low VSP repair activity rather than positive selection for rare codons. The possibility that minor proteins are preferentially encoded by rare codons to keep their expression levels low was refuted by Sharp and Li (1986).

There is another shortcoming in the analysis performed by Eyre-Walker: By considering CTAG solely in the 2312 frame, some genes (119, Eyre-Walker personal communication) which have the motif in the 1231 frame are included in the non-CTAG sample and excluded from the CTAG-containing one.

Perhaps a proper stratification should contemplate solely the CTAG frequency in the 2312 frame, disregarding the occurrence of CTAG in the 1231 frame. We have now performed such a comparison by analyzing a purged sample of *E. coli* genes from ECD (Wahl et al. 1994), and we have found (Table 1) that the average CAI of the genes with no CTAG motif is significantly higher than that of samples of genes having a frequency in the 2312 frame greater than 0.1%, 0.2%, and 0.3%. It can be also seen that the average CAI decreases as the CTAG frequency increases. We have also found a significant difference between the mean CAI of the no-CTAG sample and that of a sample with a frequency greater than 0.3% in the 2312 frame (but lacking CTAG in the other two frames). We acknowledge that threshold setting is somewhat arbitrary; this is particularly true because of the uneven distribution of CTAG by frames: In a recent study of CTAG frequencies across a 1.6-Mb sequence of *E. coli* (Burland et al. 1995) the frequencies found in the 1231, 2312, and 3123 frames were 0.03%, 0.01%, and 0.002%, respectively.

It is difficult to ascertain whether the CTAG-containing genes have the CTAG motif because they are poorly expressed or because they are weakly processed by VSP repair. However, the latter view is supported by the observation that the frequencies of other VSP substrates are significantly higher in the CTAG-containing sample than in the CTAG-lacking one. Further, since poorly expressed genes are under weak selection constraints and their compositional biases are better explained by contextual mutations (Sharp and Li 1986; Bulmer 1988), if VSP repair would be equally active on all genes, then weakly expressed genes should show a depletion of VSP substrates. The hindrance of VSP re-

pair in weakly expressed genes provides a reasonable explanation for the phenomenon.

The second point in the letter by Eyre-Walker concerns the possible correlations between the ratios of observed-to-expected frequencies of the tetranucleotides involved in VSP repair. If one considers only the 2312 frame, the variation coefficients of both the observed and expected tetranucleotide frequencies cast doubts on the value of the correlations computed by using their quotients. We have reconsidered the nature of the variation between the CAI value and the observed frequency of VSP substrates and products (Fig. 1). It can be seen that at low CAI values (<0.3) an increase in CAI corresponds to a decrease in VSP substrates and to an increase in VSP outputs. Thus, in our judgment, the contention that VSP repair seems to be hindered in poorly expressed genes still holds.

Finally, we would like to remark that it may actually be the case that VSP repair and gene expression level are unlinked; in fact, we declared in our paper that the notion was speculative and not proved. However, it seems to us that the arguments provided by Eyre-Walker (1995) do not suffice to declare this idea unfounded.

Acknowledgments. We thank Dr. A. Eyre-Walker for information about the gene samples he used and Dr. L.M. Corrochano for helpful comments.

References

Bhagwat AS, McClelland M (1992) DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. Nucleic Acids Res 20:1663–1668

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by selection-mutation balance? J Evol Biol 1:15–26

Burland V, Plunkett G III, Sofia HJ, Daniels DL, Blattner FR (1995) Analysis of the *Escherichia coli* genome VI: DNA sequence of the region from 92.8 through 100 minutes. Nucleic Acids Res 23:2105–2119

Eyre-Walker A (1995) Does Very Short Patch (VSP) repair efficiency vary in relation to gene expression levels? J Mol Evol 40:705–706

Gläsner W, Merkl R, Schellenberger V, Fritz HJ (1995) Substrate preferences of Vsr DNA mismatch endonuclease and their consequences for the evolution of the *Escherichia coli* K-12 genome. J Mol Biol 245:1–7

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

Gutiérrez G, Casadesús J, Oliver JL, Marín A (1994) Compositional heterogeneity of the *Escherichia coli* genome: a role for VSP repair? J Mol Evol 39:340–346

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Merkl R, Kröger M, Rice P, Fritz HJ (1992) Statistical evaluation and biological interpretation of non-random abundance in the *E. coli* K-12 genome of tetra- and pentanucleotide sequences related to VSP mismatch repair. Nucleic Acids Res 20:1657–1662

Sharp PM, Li WH (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. Nucleic Acids Res 14:7737–7749

Sharp PM, Li W-H (1987) The codon adaption index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F (1988) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens;* a review of the considerable within-species diversity. Nucleic Acids Res 17:8207–8211

Wahl R, Rice P, Rice CM, Kröger M (1994) ECD—a totally integrated database of *Escherichia coli* K12. Nucleic Acids Res 22:3450–3455