J Mol Evol (1995) 40:705–706

JOURNAL OF MOLECULAR EVOLUTION

Letter to the Editor

Does Very Short Patch (VSP) Repair Efficiency Vary in Relation to Gene Expression Levels?

Received: 21 October 1994 / Accepted: 17 November 1994

The very short patch (VSP) repair system in *Escherichia coli* repairs G:T mismatches in specific contexts to G:C; in doing it is expected to reduce the frequency of the CTAG, CTTG, CCTA, CCAA, TTGG, TAGG, and CAAG tetranucleotides (henceforth referred to as group I tetramers) and increase the frequency of the CCTG, CCAG, CTGG, and CAGG (group II) tetranucleotides. Gutierrez et al. (1994) have recently suggested that the efficiency of VSP repair varies across the *E. coil* genome in relation to gene expression level. They showed that genes which have low frequencies of CTAG also have low frequencies of the other group I tetramers, and high frequencies of the group II tetramers compared to genes which have a high frequency of CTAG. Furthermore, genes which have low CTAG frequencies have relatively high levels of codon bias. Since codon bias and gene expression level are correlated (Gouy and Gautier 1982) these two observations taken together suggest that the frequency of VSP repair varies between genes in relation to gene expression level.

However, synonymous codon bias complicates the issue. *E. coli* preferentially uses those synonymous codons which bind the most common tRNAs with normal base pairing (Ikemura 1985), particularly in highly expressed genes. This means that there are correlated changes in di- and tri- and tetranucleotide frequencies across genes in relationship to gene expression level. Since all group II tetramers (in the 3123 and 1231 frames) encode at least one codon that is preferentially used in highly expressed genes, whereas the group I tetramers do not, it is perhaps not surprising that the frequencies of group I and group II tetramers are correlated to other group I and group II tetramer frequencies (respectively) and to the level of gene expression. For instance, CTAG can only appear in two reading frames in protein coding sequences (because TAG is a stop codon). In one of these (1231) the motif forms a CTA leucine codon; this is one of the rarest codons in high codon bias *E. coli* genes (Sharp 1989). In the other reading frame (2312) the AG codes for the twofold degenerate codons of serine and arginine, both of which are also very rare in high-codon-bias genes. One therefore expects genes with relatively high frequencies of CTAG to have lower levels of codon bias; the motif CTAG can only ever be found in low-bias genes because those are the only genes which have CTA and AGN codons. To demonstrate the importance of these biases I have calculated the CAI value, a measure of synonymous codon bias, by the method of Sharp and Li (1987) (with minor modifications suggested by Bulmer [1988]) for 1289 *E. coli* genes from Ecoseq6 (Rudd 1992, Rudd pers. comm.). As did Gutierrez et al. (1994), I found the mean CAI value of those genes with no CTAG tetranucleotides (0.392 ± 0.004) was significantly higher than that of those with greater than 0.07% CTAG (0.328 ± 0.011) (t-test, $t = 5.231$, $P < 0.0001$). However, if the data are divided according to the frequency of CTAG in the 2312 frame the difference is reduced; the mean CAI value of those genes with no CTAG (0.388 ± 0.003) is not significantly different from

Table 1. Spearman's rank correlation coefficients between CAI values and tetramer ratios

Tetramer	All genes	Genes >300 codons
Group I		
CTAG	–0.035	–0.019
CTTG	0.001	0.029
CCTA	0.139**	0.159**
CCAA	0.120**	0.105
TTGG	0.012	–0.010
TAGG	–0.063*	–0.049
CAAG	–0.001	–0.023
Group II		
CCTG	–0.008	–0.019
CCAG	–0.030	–0.004
CTGG	0.037	–0.002
CAGG	–0.018	–0.001

* Significant at 5%

** Significant at 1%

Significance values not corrected for multiple simultaneous tests

Table 2. Spearman's rank correlation coefficients between tetramer ratios[a]

	Group I							Group II			
	CTAG	CTTG	CCTA	CCAA	TTGG	TAGG	CAAG	CCTG	CCAG	CTGG	CAGG
CTAG	—	0.003	0.041	0.000	0.015	0.003	0.022	–0.012	0.026	0.015	0.006
CTTG	–0.051	—	0.048	0.029	0.035	0.078**	0.020	–0.020	0.033	–0.034	0.042
CCTA	–0.000	0.016	—	–0.011	0.052	0.076	0.018	0.039	0.032	0.070*	0.055
CCAA	–0.012	–0.012	–0.093*	—	0.027	0.069*	–0.014	–0.011	0.010	0.032	0.028
TTGG	0.023	0.021	0.002	0.023	—	0.016	0.024	0.037	0.061*	0.010	0.026
TAGG	–0.005	0.046	–0.002	0.037	–0.058	—	0.025	0.000	0.050	–0.025	0.030
CAAG	0.012	–0.034	0.027	–0.093*	0.012	0.016	—	0.064*	–0.013	0.017	–0.008
CCTG	0.008	–0.083	–0.027	–0.037	0.044	0.017	0.015	—	–0.014	0.031	0.007
CCAG	0.043	–0.009	–0.017	–0.042	0.017	–0.020	–0.076	–0.063	—	0.043	0.006
CTGG	–0.015	–0.063	0.036	–0.025	–0.010	–0.041	–0.034	0.012	0.062	—	–0.010
CAGG	0.030	0.046	0.036	–0.012	–0.036	–0.046	–0.045	–0.054	–0.007	–0.015	—

[a] Figures are Spearman's rank correlation coefficients for all 1,289 *E. coli* genes in sample (above diagonal) and those genes with more than 300 codons (666 genes) (below diagonal). Significance levels are not corrected for multiple simultaneous tests

* Significant at 5%

** Significant at 1%

that of those with a frequency greater than 0.1% (0.358 ± 0.021) (t-test, $t = 1.471$, $P < 0.15$). (Note that since CTAG can only generally be present in the 1231 and 2312 frames, a gene-wide frequency of 0.07% is equivalent to a frequency of 0.1% in each frame.)

To overcome the problem of correlations between the level of codon bias and di- and trinucleotide frequencies, I considered the frequencies of the type I and type II tetramers in the 2312 frame, thus eliminating trinucleotide effects. To remove the effect of the dinucleotide correlations the observed frequency of a tetranucleotide was divided by the expected frequency calculated from the dinucleotide frequencies. For instance, in the *accD* gene, the frequency of CC is 0.0492 in the last two codon positions and the frequency of TG is 0.0295 in the first two codon positions; thus we expect ~0.0015 CCTG tetramers. The ratio of observed to expected in this case is 2.266. This ratio gives an indication of whether the tetranucleotide is over- or underrepresented in the gene, controlling for the dinucleotide frequencies.

The correlations between these ratios and the CAI values are shown in Tables 1 and 2. Under the hypothesis of Gutierrez et al. the frequencies of the group I ratios should be positively correlated to each other and negatively correlated to the group II ratios and CAI values, whereas group II ratios should be positively correlated to each other and the CAI value but negatively correlated to the group I ratios. None of these predictions is consistently met, and any significant correlations are as likely to go in the wrong direction as in the right direction. There therefore seems little reason to believe that the efficiency of VSP repair varies across the *E. coli* genome, and that it does so in relation to gene expression level. This is not necessarily unexpected. Although it is known that excision repair is linked to transcription, the mechanism by which this is achieved is probably unique to the bulky lesions, like pyrimidine dimers, which excision repair deals with. These cause the RNA polymerase to stall at the affected lesion, hence signaling to the repair machinary that a lesion is present (Friedberg et al. 1994). It seems unlikely that a similar process would apply to the base mismatches repaired by the VSP repair system.

References

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

Friedberg EC, Bardwell AJ, Bardwell L, Wang Z, Dianov G (1994) Transcription and nucleotide excision repair—reflections, considerations and recent biochemical insights. Mutat Res 307:5–14

Gouy G, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acid Res 10:7055–7073

Gutierrez G, Casadeus J, Oliver JL, Marin 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

Rudd KE (1992) Alignment of *E. coli* DNA sequences to a revised integrated genomic restriction map. In: Miller J (ed) A short course in bacterial genetics: a laboratory manual and handbook for *Escherichia coli* and related bacteria. Cold Spring Harbor Press, Cold Spring Harbor

Sharp PM (1989) Evolution at silent-sites in DNA. In: Hill WG, McKay TFC (eds) Evolution and animal breeding: reviews of molecular and quantitative approaches in honour of Alan Robertson. CAB International, Wallingford, Britain, p 23

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

Adam Eyre-Walker
Department of Biological Sciences
Rutgers University
Piscataway, NJ 08855-1059, USA