J Mol Evol (1999) 48:712–716

JOURNAL OF MOLECULAR EVOLUTION

Compositional Correlation Between Open Reading Frames with Opposite Transcriptional Orientations in *Escherichia coli*

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

[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: 2 April 1998 / Accepted: 18 June 1998

Abstract. This paper analyzes correlations in base composition between pairs of neighboring genes in *Escherichia coli.* The G + C contents of nearby, but convergently or divergently transcribed, genes show weak but significant correlations, and this is attributed to compositional variation among genomic regions. The finding that the base composition varies among intergenic regions, depending upon whether the adjacent genes are transcribed convergently, divergently, or in the same orientation, seems to indicate that transcription affects the patterns of mutation and, therefore, the overall base composition of the region.

Key words: *Escherichia coli* — G + C content — Intergenic DNA — Transcription-induced mutation

Introduction

The complete genome sequence of *Escherichia coli* K-12 reported by Blattner and co-workers (1997) has opened the way for a detailed description of the genomic characteristics of this model organism. Early work suggested that bacterial genomes do not show differences in base composition among different pieces in the same genome (Sueoka 1959; Rolfe and Meselson 1959). However, it is becoming evident that some regional compositional variation occurs along the chromosome of bacteria. Although the mechanisms responsible for bacterial intragenomic G + C variation are not fully understood, a number of causes have been invoked as a source of regional compositional homogeneities, including local differences in selective constraints, directional mutational pressure, DNA horizontal transfer, and DNA repair (Nomura et al. 1987; Médigue et al. 1991; D'Onofrio and Bernardi 1992; Sueoka 1992; Gutiérrez et al. 1994; Deschavanne and Filipski 1995; Lawrence and Ochman 1997).

Data and Methods

In this work we have used the *E. coli* sequence and annotation, version M52 (September 5, 1997), downloaded from the *E. coli* Genome Center's Home Page (http://www.genetics.wisc.edu:80/), which contains 4405 actual and proposed protein coding genes (ORFs). From this sequence we selected all the ORF pairs oriented in opposite directions with respect to transcription and separated by a spacer length in the range of 100–500 bp. The spacer lower length limit was set to avoid random fluctuations in base composition associated with shorter sequences and to minimize the effect of biased composition of possible regulatory signals, and the upper limit to avoid unidentified ORFs. The sample analyzed consists of 196 pairs of convergently oriented ORFs and 457 pairs of divergently oriented ORFs. Since convergent spacer lengths are shorter than divergent spacer ones, the latter ordinations are overrepresented when the spacer length range is set to 100–500 bp. In this work, we define spacers as noncoding sequences separating ORFs. According to the orientation of the flanking ORFs, we use the term convergent spacer to denote the noncoding sequence between two neighboring ORFs convengently transcribed, divergent spacer to denote the noncoding sequence between divergently transcribed ORFs,

Correspondence to: A. Marín; e-mail: anmarin@cica.es

Table 1. Mean and standard error of length and G + C content of ORFs and spacers in ORF pairs with opposite orientations with respect to transcription[a]

	Convergent orientations (n = 196)			Divergent orientations (n = 457)		
	w-ORF	Spacer	c-ORF	c-ORF	Spacer	w-ORF
	5′ |-------->·····································3′			5′·····································|-------->3′		
	3′·····································<--------| 5′			3′<--------|·····································5′		
Length						
Mean	1012.4	212.8	1008.6	955.7	240.5	962.0
SE	53.8	7.3	48.2	28.8	4.4	28.1
G + C						
Mean	0.502	0.477	0.507	0.509	0.403	0.507
SE	0.003	0.004	0.003	0.002	0.002	0.002
GC3						
Mean	0.538		0.542	0.543		0.543
SE	0.006		0.006	0.003		0.003

[a] Only pairs separated by spacer lengths in the range 100–500 bp are considered. w-ORF is in the strand stored in the database, and c-ORF in the complementary strand.

and tandem spacer to denote the noncoding sequence between ORFs transcribed in the same direction.

Results and Discussion

The search for compositional (G + C) variation along the *E. coli* genome was made by correlation analyses between the G + C content of ORF pairs oriented in opposite directions with respect to transcription. The reason to choose transcriptionally opposite ORFs is because a possible source of compositional variation depends upon codon choices for translational efficiency (Ikemura 1981; Gouy and Gautier 1982; Lobry and Gautier 1994; Gutiérrez et al. 1996, *inter alia*), and it is not expected that such translational constraints covary between convergently or divergently oriented genes as could be the case for tandemly oriented ones, which often constitute operons where the same constraints (codon choice) are expected. Further insight into the existence of regional compositional homogeneities was sought through correlations between the G + C content of the spacer (where no translational constraints occur) and the G + C content of the flanking ORFs.

The statistics of ORF length, spacer length, ORF G + C content (total and at the third codon positions, GC3), and spacer G + C content in the selected sample are given in Table 1. We note that neither ORF length nor ORF G + C contents differ significantly (Mann–Whitney test) between the ORFs flanking the convergent and those flanking the divergent orientations.

In the first step we computed Pearson and Spearman correlation coefficients between the GC3 values of ORFs in convergent and divergent pairs and also between the GC3 value in each flanking ORF and the G + C content of the spacer separating them. Prior to correlation computations, G + C proportions were arc sine root transformed. Our results (Table 2) show significant ($p <$ 0.001) positive correlations between the GC3 values of convergently oriented ORFs and also between the GC3 values of divergently oriented ORFs. Likewise, it can be seen that in both kinds of ordinations, the GC3 in each of the two flanking ORFs is positively correlated with the G + C content of the corresponding spacer. Although the correlation coefficients are statistically significant, they are certainly low, hence the considerable scatter of points in the plots (Fig. 1). Correlation coefficients of a similar magnitude have been reported between the GC3 and the G + C content of introns in *Drosophila melanogaster* (Kliman and Hey 1994) and between the CAI index [codon adaptation index, a measurement thought to be related to gene expression level (Sharp and Li 1987)] and the intergenic distance in *E. coli* (Eyre-Walker 1995).

Table 2. Correlation coefficients (Pearson and Spearman) between GC3 in ORFs and G + C content in spacers[a]

	Convergent orientations		Divergent orientations	
Correlation	Pearson	Spearman	Pearson	Spearman
w-ORF vs. c-ORF	0.251	0.208	0.386	0.364
w-ORF vs. spacer	0.451	0.357	0.248	0.190
c-ORF vs. spacer	0.389	0.319	0.253	0.194

[a] w-ORF and c-ORF as in Table 1. All correlation coefficients are statistically significant ($p < 0.001$).

From the preceding results it seems that some kind of regional compositional variation does occur along the *E. coli* genome. It is worth mentioning that topological compositional variation in *E. coli* has been reported by Deschavanne and Filipski (1995), who found compositional differences between genes close to the replication origin and genes close to the replication terminus (see also Sharp et al. 1989). These differences have been interpreted as the consequence of mutational bias in-

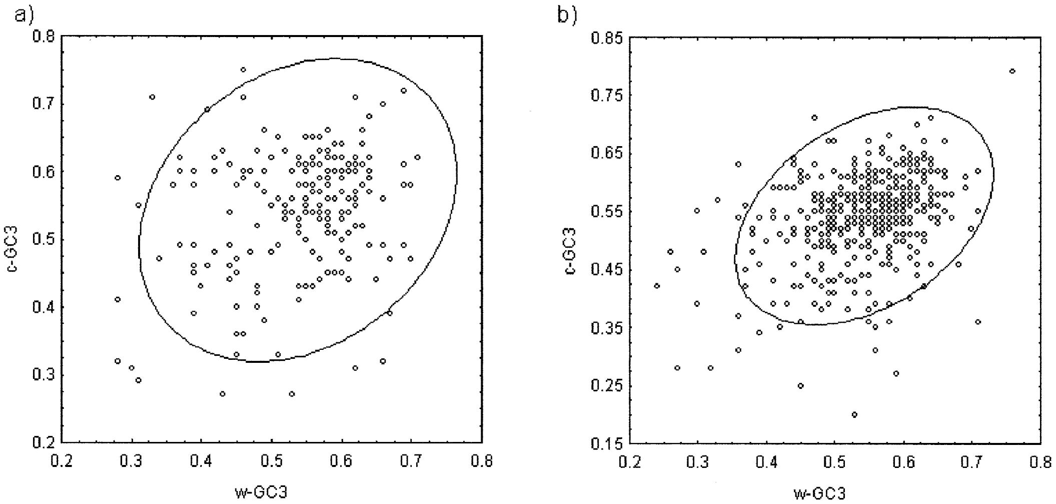

Fig. 1. c-GC3 vs. w-GC3 scatter plot. **a** Convergent arrangements. **b** Divergent arrangements. The confidence ellipse ($p = 0.95$) is centered on the sample means of the *x* and *y* variables, and the unbiased sample standard deviations of *x* and *y* determine its major axes.

duced by differential DNA repair in relation to replication timing. Other compositional changes concerning strand asymmetry have been noted recently to correlate with the direction of DNA replication (Lobry 1996; Blattner et al. 1997; Mrázek and Karlin 1998; Freeman et al. 1998).

We think that regional variation in mutation pressure, whatever its cause, may explain the correlations found in this work, which provide further support and substantiation for the ubiquity of genome compositional heterogeneities anticipated by Bernardi et al. (1985). We assume that there is no covariation of expression level in transcriptionally opposite genes. This assumption has been put to the test by computing the correlation between the CAI values of the opposite ORFs pairs. No correlation was found between the CAI values of convergently transcribed ORF pairs ($r = 0.087$, $p = 0.222$), and a weak, but significant, correlation ($r = 0.135$, $p = 0.004$) was found between divergently transcribed ORF pairs. The same results (not shown) were obtained when using only ORF pairs longer than 300 codons, which are almost certainly genuine coding sequences and do maintain the aforementioned compositional correlations. Regarding the CAI correlation, we note that since CAI and GC3 are in turn correlated, it might be that a part of the weak correlation between the CAI values of divergent ORFs is a reflection of the GC content effect on the CAI estimates, rather than true covariation in gene expression.

In the second step we analyzed the G + C content variation found in spacer DNA. That the G + C content of spacer DNA is lower than that of coding DNA seems a rather general phenomenon (see, e.g., Guigó and Fickett 1996; Clay et al. 1996, and references therein). In *E. coli,* the average G + C content of coding sequences is 0.510 and that of spacer DNA is 0.444. In the genome sample analyzed here, the average spacer G + C content is significantly higher in convergent (0.477) than in divergent (0.403) orientations, and both, in turn, are significantly different from that of tandem orientations, which has an intermediate value (0.429 ± 0.002, measured over 907 tandem spacers in the length range 100–500 bp). The Mann–Whitney test for all pairwise comparisons was always significant ($p < 0.001$).

It is thought that base composition is under fewer selective constraints in spacer DNA than in coding DNA, and should reflect the result of mutational biases, but selective constraints acting on important regulatory signals cannot be discounted. Thus, some bias in intergenic base composition could be associated with gene regulation due to the occurrence of transcription termination signals between convergently transcribed genes, promoters and modifiers between divergently transcribed genes, and probably both kind of signals between genes transcribed in tandem. However, as an alternative or concomitant explanation, the compositional difference between convergent and divergent spacers seem to indicate that transcription affects the patterns of mutation and, therefore, the overall base composition of these sequences. In the following, we speculate on the possibility of transcriptionally induced mutations as an explanation for (at least a part of) the compositional variation found among spacers.

Although the matter is unsettled, some variation in the spontaneous mutation rate of genes with transcription level has been suggested; such variation would rely on differential damage sensitivity of transcribing DNA and/or transcription-coupled repair (Fix and Glickman 1987; Davis 1989; Fox et al. 1994; Holmquist and Filipski 1994; Datta and Jinks-Robertson 1994; Francino et al. 1996; Beletskii and Bhagwat 1996; Francino and Ochman 1997). Intergenic DNA composition could be affected by the transcription-induced mutational bias since transcription in *E. coli* may proceed beyond the stop codon until the transcriptional apparatus meets a terminator and even continue past the terminator sequence (the so-called read-through phenomenon).

A different intensity of transcription, and, therefore, of the transcription-induced mutation/repair effect, on the three kinds of spacers is expected. In convergent spacers, two opposite transcriptional waves (one in each strand) may proceed past the stop codons. In tandem spacers, the transcriptional wave affects only one strand. In divergent spacers, it is not expected that transcription occurs between the two opposite initiation transcription points. These differential expectations are in accordance with the gradation of G + C content in the three kinds of spacers.

Indirect support for a role of transcription-induced mutation in shaping spacer base composition can be obtained regarding the relationship between spacer length and G + C content. In convergent and tandem orientations it is expected that the longer the spacer, the lower the proportion of transcribed nucleotides and the lower the transcription-induced mutational bias. We have indeed found that the relationship between convergent spacer length and G + C content is negative (Pearson's $r = -0.285$, Spearman's $r = -0.307$; both p's < 0.001), as is the relationship between tandem spacer length and G + C content (Pearson's $r = -0.215$, Spearman's $r = -0.231$; both p's < 0.001). On the other hand, in divergent spacers, where transcription should be less extensive, no such correlation is found (Pearson's $r = 0.011$, $p = 0.81$; Spearman's $r = -0.013$, $p = 0.78$).

In conclusion, our analyses provide results suggesting the existence of compositional variation along the *E. coli* chromosome, although the intragenomic variability of G + C content in this species is extremely low in comparison with that found in Mycoplasma (McInerney 1997; Kerr et al. 1997). A gradual decrease in the G + C content ORFs > convergent > tandem > divergent spacers has been shown; that, given its parallelism with decreasing transcription likelihoods, led us to speculate about a role of transcription in shaping DNA composition.

Acknowledgements. We thank J. Casadesús, L.M. Corrochano, F. González, J.R. Lobry, and one anonymous referee for helpful comments on the manuscript. We also thank Dr. E. Zuckerkandl for valuable criticism. This work was supported by Grant PB96-144-C02 from the Ministerio de Educación y Cultura.

References

Beletskii A, Bhagwat AS (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

Bernardi G, Oloffson B, Filipski J, et al. (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958.

Blattner FR, Plunkett G III, Bloch CA, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462

Clay O, Cacciò S, Zoubak S, Mouchiroud D, Bernardi G (1996) Human coding and noncoding DNA: Compositional correlations. Mol Phylogenet Evol 5:2–12

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

Davis BD (1989) Transcriptional bias: A non-Lamarckian mechanism for substrate-induced mutations. Proc Natl Acad Sci USA 86:5005–5009

Deschavanne P, Filipski J (1995) Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. Nucleic Acids Res 23:1350–1353

D'Onofrio G, Bernardi G (1992) A universal compositional correlation among codon positions. Gene 110:81–88

Eyre-Walker A (1995) The distance between *Escherichia coli* genes is related to gene expression levels. J Bacteriol 177:5368–5369

Fix DF, Glickman BW (1987) Asymmetric cytosine deamination revealed by spontaneous mutational specificity in an Ung$^-$ strain of *Escherichia coli.* Mol Gen Genet 209:78–82

Fox MS, Radicella JP, Yamamoto K (1994) Some features of base pair mismatch repair and its role in the formation of genetic recombinants. Experientia 50:253–260

Francino MP, Ochman H (1997) Strand asymmetries and DNA evolution. TIG 13:240–245

Francino MP, Chao L, Riley MA, Ochman H (1996) Asymmetries generated by transcription-coupled repair in Enterobacterial genes. Science 272:107–109

Freeman JM, Plasterer TN, Smith TF, Mohr SC (1998) Patterns of genome organization in bacteria. Science 279:1827

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

Guigó R, Fickett JW (1995) Distinctive features in protein coding, genic non-coding, and intergenic human DNA. J Mol Biol 253:51–60

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

Gutiérrez G, Márquez L, Marín A (1996) Preference for guanosine at firts codon position in highly expressed *Escherichia coli* genes. Nucleic Acids Res 24:2525–2527

Holmquist GP, Filipski J (1994) Organization of mutations along the genome: a prime determinant of genome evolution. TREE 9:65–69

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

Kerr ARW, Peden JF, Sharp PM (1997) Systematic base composition variation around the genome of *Mycoplasma genitalium,* but not *Mycoplasma pneumoniae.* Mol Microbiol 25:1177–1184

Kliman RM, Hey J (1994) The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics 137:1049–1056

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: Rates of change and exchange. J Mol Evol 44:383–397

Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–665

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

McInerney JO (1997) Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb Comp Genom 2:1–10

Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 222:851–856

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

Nomura M, Sor F, Yamagishi M, Lawson M (1987) Heterogeneity of GC content within a single bacterial genome and its implication for evolution. Cold Spring Harbor Symp Quant Biol 52:658–663

Rolfe R, Meselson M (1959) The relative homogeneity of microbial DNA. Proc Natl Acad Sci USA 45:1039–1042

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

Sharp PM, Shields DC, Wolfe KH, Li W-H (1989) Chromosomal location and evolutionary rate variation in enterobacterial genes. Science 246:808–810

Sueoka N (1959) A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc Natl Acad Sci USA 45:1480–1490

Sueoka N (1992) Directional mutation pressure, selection constraints, and genetic equilibria. J Mol Evol 34:95–114