Correlation of GC content with replication timing and repair mechanisms in weakly expressed \textit{E.coli} genes

Patrick Deschavanne and Jan Filipski*

Laboratoire de Mutagenèse, Institut Jacques Monod, 2 place Jussieu, Tour 43, 75251 Paris, France

Received January 13, 1995; Revised and Accepted March 10, 1995

ABSTRACT

Regional variations of DNA GC content are observed in species as different as \textit{S.cerevisiae} and humans. In vertebrates and yeast they are correlated with replication timing; late replicating chromosomal regions are more AT-rich than early replicating regions. We show here that gene composition in \textit{E.coli} also has long range variations which are similarly correlated with replication timing. We suggest that the enrichment in AT base pairs in late replicating DNA reflects differences in DNA repair modes. These sequences, which are in single copy for a greater part of the cell cycle than origin-linked genes, have less opportunity to engage in repair via homologous recombination and therefore may resort more often to translesion synthesis involving the misincorporation of adenine opposite modified nucleotides.

INTRODUCTION

There is no generally accepted explanation for the long range fluctuations of DNA composition found in the genomes of most species studied. In vertebrates these fluctuations are correlated with chromosomal banding and replication timing (1,2). The GC-rich chromosomal bands, pale on Giemsa staining, replicate earlier in the cell cycle than more AT-rich dark bands. Recently the megabase sequencing of yeast chromosomes has unexpectedly demonstrated that in this species as well, DNA composition shows long range variability exhibiting sine wave-like fluctuations of GC content (3–5). The GC-rich peaks are spaced ~100 kb apart. Sequences located in a GC-rich peak in chromosome III replicated early in S phase while the AT-rich telomeric regions replicated late (6,7). This led us to speculate that the AT enrichment of late replicating DNA may be quite a general phenomenon.

It has been suggested that the long range variability of DNA composition along the eukaryotic chromosome is due to the evolutionary accumulation of mutations arising during DNA repair processes biased differently for early and late replicating sequences (8). However, our ignorance of DNA repair mechanisms in mammalian germline cells makes it difficult to test this hypothesis.

The situation is different in prokaryotes, especially in \textit{E.coli} where various repair systems, including damage-inducible, error prone SOS repair, are quite well characterised (9). Bacterial chromosome location-dependent variations in repair have already been suggested to be responsible for variability of silent site divergence between \textit{E.coli} and \textit{S.typhimurium} (10). Sharp et al. have found that the genes located near the origin of replication, (oriC) has diverged less than genes located on the other side of the chromosome. An explanation was proposed that the sequences near oriC might be repaired by mechanisms which are more efficient when sequences are present in more than one copy (10). It was interesting to see whether gene GC content also changes along the \textit{E.coli} chromosome, as one would expect, if these repair mechanisms have different mutational bias (11). The main problem of this type of study is that genes coding for highly expressed proteins use primarily codons recognised by the most abundant tRNA species (12) which is reflected, for example, by the Codon Adaptation Index (CAI,13). The evolutionary codon adaptation affects DNA composition and may also contribute to variations of base composition along the chromosome. To minimise this effect, we considered only weakly expressed genes (class 1 genes as defined in ref. 14). We expected that when the pressure of selection related to codon usage is weak, the contribution of mutational bias of various repair systems could be demonstrated. Indeed we have found that the genes belonging to class 1 and located near the terminus of replication are enriched in AT if compared with genes belonging to this class located close to oriC.

MATERIALS AND METHODS

The 587 weakly expressed genes (belonging to class 1 as defined in ref. 14) from the nucleic acid sequences data bank GenBank were analysed using a retrieval package installed on a central computer in Paris (Centre Interuniversitaire de Traitement de I’Information). Statistical significance of the results obtained were checked by t-test and anova test as indicated in the text.

RESULTS AND DISCUSSION

When the GC content of the first, second and third codon positions of 587 class 1 genes is plotted against the total gene GC content (Fig. 1), the slope is highest for the third position and lowest for the second position. (We want to stress that this presentation, traditionally used in the literature is illustrative and not diagnostic because the X and Y values are not independent. In order to make sure that the correlations exist, the GC values for

* To whom correspondence should be addressed
different codon positions should be checked against each other. In the set of genes analysed here the GC content in the third position is correlated with the GC content of the second position at the confidence level 0.001 and with the first codon position at the confidence level 0.009.) A similar relationship has already been observed in the case of heterologous set of genes (15) or in the set of recA genes (16) from various bacterial species plotted against total bacterial GC content. This finding confirms the relative functional importance of the three positions and suggests that even in a single bacterial genome, different genes are under different mutation pressure. The third codon positions in genes under AT pressure can be more heavily loaded with As and Ts with negligible consequences for the phenotype than the first and especially the second positions.

The overall GC content of genes and the GC content of each codon position varies along the chromosome (Fig. 2). These variations are statistically significant as demonstrated by an anova test taking 5 min wide segments on the chromosome map as categories. The probabilities that the first, second and third codon position do not vary along the chromosome are : $P_1 = 0.0046$, $P_2 = 0.0013$ and $P_3 = 0.0001$, respectively. Regional variations are highest for the third codon position, for which the normalised variations from the mean GC content are twice as high as those for the first or the second position. All curves, in particular that for the third codon position, show a minimum near the replication terminus (ter). Genes close to the replication origin (oriC) show the highest GC content. The difference in GC content in all three codon positions for genes located in these two regions is significant at the 95% confidence level ($t$-test on a 20 min segment centred on the ter and oriC regions, at 30 and 84 min on the map, respectively. Probabilities that there is no difference in GC content in the first, second and third codon positions between these two regions are: $P_1 = 0.0027$, $P_2 = 0.04$ and $P_3 < 0.0001$, respectively).

We find it likely that the enrichment of the late replicating region of the chromosome in AT is a consequence of different contributions of differently biased repair mechanisms to the genome evolution. E.coli possesses an error-prone repair system which inserts adenine opposite modified non-coding bases. The pattern of compositional variability demonstrated above could be explained by different contributions of this ‘translesion’ DNA synthesis in the overall repair of damaged DNA sequences depending on their localisation with respect to oriC. It is known, for example, that UV light-induced DNA lesions in E.coli are repaired in the dark by three repair processes (17). 85% are removed from the genome by UvrA, B, C dependent excision repair and 13% are repaired by a recombination-dependent process; only if these attempts fail will the bacteria activate UmuC, D dependent mutagenic translesion synthesis, leading to the preferential incorporation of adenine (the A-rule, 18). Similar reasoning applies to DNA lesions caused by a number of other mutagenic agents and there are hints that it could also be true for other organisms (19).
Figure 3. Repair of mutagenic lesions in DNA. (A) Most DNA lesions are repaired by the excision repair system, which removes a damaged single-stranded DNA segment and fills the gap by a high-fidelity DNA synthesis complex (light-hatched oval). (B) Lesions remaining unrepaired by this system, (such as inter-strand cross-links and other lesions causing dissociation of the high fidelity complex) are repaired by recombination in the region of the chromosome already replicated. (C) The single-copy regions are repaired by mutagenic translesion necessitating induction of an adenine specific, low-fidelity DNA synthesis system (dark-hatched oval).

We propose that such a cascade of repair systems acting on a genome over evolutionary time resulted in systematic variations of composition along the chromosome (see Fig. 3). The further the damaged sequence is from oriC, the lower are the chances that the sister DNA molecule has already been synthesised, allowing the recombinational repair pathway to be effective. Another possible obstacle for recombination in this region might be the presence of ter sequences which, when complexed with tus protein, block the movement of DNA helicases along the DNA (20). The sequences located close to the terminus of replication are thus more likely to be repaired by translesion DNA synthesis, causing different, more AT biased, mutation pressure. This higher contribution of mutagenic translesion synthesis may also contribute to higher mutability of sequences near ter as revealed by the study of divergence between genes of E.coli and Salmonella (21) or by the study of mutations in closely related laboratory strains of E.coli (16). The mechanism proposed above explains only one of the features revealed by the DNA sequence analysis. Another striking phenomenon is the variation of the third codon positions (Fig. 2) close to the origin of replication. We do not have any obvious explanation for its evolutionary origin.

A cascade of repair mechanisms operates also in higher organisms influencing mutation rates and direction, although, as more DNA sequence data accumulate, it becomes clear that there is no one simple explanation for DNA divergence patterns there. Recent calculations (22), show that although synonymous substitution rates and silent site base composition are correlated in mammals, this correlation is not linear (as indicated by earlier analysis, 11) and is less strong than previously suggested (23). In mammals there is no simple linear correlation between gene expression rate and the frequency of 'optimal' codon usage either. Some other effects, like the direction of gene transcription with respect to the closest origin of replication may influence the gene composition. In E.coli this factor seems to have only secondary importance as opposed to the λ phage chromosome (24) but it may contribute to local variations of GC content in other genomes.

The analysis of nucleotide sequence in yeast shows that the autonomously replicating sequences (ARS, considered to represent the origins of replication), are located in both GC-rich and GC-poor regions. One particular class of these ARS, called ‘functional ARS’, is preferentially localised in the AT-rich regions of the chromosome II (25). This may suggest that these regions replicate early in the S phase of the cell cycle. On the other hand, determination of replication timing performed for several chromosomal segments has shown that activation of specific origins of replication occurs at different periods in the cell cycle. Some of them are 'early' whilst the others are 'late' (5,6,25). The AT-rich peritelomeric regions dozens of kilo bases long were systematically late replicating at least in the yeast chromosomes II and V (6,25) whilst the GC-rich region on the short arm of the chromosome III has been replicating early in the S phase (5).

The bias towards AT of E.coli DNA sequences close to the ter is only one of various compositional biases caused by mutations occurring during replication and repair (26). The finding that it could be observed in the late replicating sequences in mammals, E.coli and yeast might suggest that a high contribution of mutagenic translesion synthesis to the evolution is quite a general phenomenon.
ACKNOWLEDGEMENTS

We express our thanks to R. D'Ari, G. Holmquist, T. H. Jukes and M. Radman for interest and for helpful discussions. This project was supported by the grant from GREG 520634.

REFERENCES