**ASYMMETRICAL CODING SEQUENCE REPARTITION AND CODON ADAPTATION INDEX VALUES BETWEEN LEADING AND LAGGING STRANDS IN SEVEN BACTERIAL SPECIES.**

PERRIERE G., LOBRY J.R.+

Laboratoire BGBP – UMR CNRS n° 5558, Universite Claude Bernard – Lyon 1, 43 bd. du 11 novembre 1918, 69622 Villeurbanne Cedex, France;
e-mail: perriere@biomserv.univ-lyon1.fr; lobry@biomserv.univ-lyon1.fr

+Corresponding author

Keywords: bacterial species, asymmetrical coding sequences, DNA replication, transcription, selective advantage, asymmetrical repartition

**1. Introduction**

Because DNA replication and gene transcription are working simultaneously in *Escherichia coli*, it is believed (1-4) that there is a selective pressure against head-on collisions between a DNA replication apparatus and RNA polymerase transcription complex, yielding a selective advantage to genomes whose genes are in the leading strand to avoid such collisions. As a matter of consequence, an asymmetrical repartition of genes is expected between the leading and the lagging strands and this asymmetry is expected to be stronger for genes with high expressivity. We have explored this hypothesis further by analysing seven bacterial genomes.

**2. Material and methods**

The genes from the seven complete genomes used were taken from the NRBact (Non-Redundant Bacterial) database. This system can accessed at URL <http://pbil.univ-lyon1.fr/search/query.html>. NRBact contains the genomes of all completely sequenced bacteria and the yeast genome. Genomes used were those from *Borrelia burgdorferi*, *Bacillus subtilis*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium* and *Mycoplasma pneumoniae* (Table 1).

These genomes were selected because the location of their replication origin and terminus is known from experimental evidence (5) or could be predicted from base compositional asymmetries (6). This allowed us to split coding sequences into two groups: those which are on the the leading strand and those which are on the lagging strand.

Gene expressivity was estimated using the Codon Adaptation Index (CAI) values (7). To establish CAI tables we have used samples of putatively highly expressed genes. These samples were obtained by correspondence analyses computed on codon usage variability in the seven species. Corresponding CAI tables are available at
URL:  <http://pbil.univ-lyon1.fr/datasets/bgrs98.html>.

**3. Results and Discussion**

At a critical level of 5% experimental data are alway in contradiction with the null hypothesis of an even repartition of coding sequences between leading and lagging strands: there is always an excess of coding sequences on the leading strand.

**Table 1.** Proportion of genes in the leading strand

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | #lead. | #lag. | #total | %lead. |
| *B.burgdorferi* | 543 | 278 | 821 | 66 |
| *B.subtilis* | 2999 | 1053 | 4052 | 74 |
| *E.coli* | 2337 | 1917 | 4254 | 55 |
| *H.influenzae* | 884 | 763 | 1647 | 54 |
| *H.pylori* | 882 | 631 | 1513 | 58 |
| *M.genitalium* | 371 | 95 | 466 | 80 |
| *M.pneumoniae* | 531 | 143 | 674 | 79 |

Hence, at least in the bacterial species under study, the selective pressure against head-on collisions is a general phenomenon, but with different intensities since the proportion of coding sequences in the leading strand ranges from 54% to 80%.

The within-species dispersion of CAI values varies greatly with a very narrow distribution for *B.burgdorferi*, *H.pylori*, *M.genitalium* and *M.pneumoniae*; a narrow distribution for *B.subtilis* and a broad distribution for *E.coli* and *H.influenzae*. Hence, codon usage variability linked to gene expressivity is different between species.

CAI values were found to be very highly significantly different between leading and lagging coding sequences in three species: *B.burgdorferi, E.coli*and *H.influenzae*. As expected, for these three species CAI values are higher on the leading strand. For the remaining species there is no significant differences between the two groups (Table 2.).

**Table 2.** CAI mean values comparison between leading and lagging coding sequences

|  |  |  |  |
| --- | --- | --- | --- |
| Species | CAI lead. | CAI lag. | p |
| *B.burgdorferi* | 0.708 ± 0.033 | 0.649 ± 0.033 | < 10-4 |
| *B.subtilis* | 0.454 ± 0.069 | 0.456 ± 0.064 | 0.38 |
| *E.coli* | 0.309 ± 0.111 | 0.297 ± 0.093 | 2.10-4 |
| *H.influenzae* | 0.472 ± 0.094 | 0.454 ± 0.085 | < 10-4 |
| *H.pylori* | 0.689 ± 0.033 | 0.689 ± 0.036 | 0.67 |
| *M.genitalium* | 0.720 ± 0.029 | 0.724 ± 0.031 | 0.26 |
| *M.pneumoniae* | 0.687 ± 0.039 | 0.681 ± 0.046 | 0.13 |

This last result is more surprising because it is apparently unconsistent with the excess of gene in the leading strand in theses species: the selective pressure against head-on collisions is apparently working but no correlation is detected with gene expressivity.

**References**

1. B.J. Brewer. "When polymerase collide: replication and the transcriptional organization of the
*E. coli* Chromosome" Cell 53, 679 (1988)
2. S. French. "Consequence of replication fork movement through transcription units *in vivo*" Science **258**, 1362 (1992)
3. B. Liu and B.M. Alberts "Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex" Science **267**, 1131 (1995)
4. M.P. Francino and H. Ochman "Strand asymmetries in DNA evolution" Trends in Genetics **13**, 240 (1997)
5. J.M. Freeman, T.M. Plasterer, T.F. Smith and S.C. Mohr "Patterns of genome organization in bacteri" Science **279**, 1827 (1998)
6. J.R. Lobry "Origin of replication of *Mycoplasma genitalium*" Science **272**, 745 (1996)
7. P.M. Sharp and W.-H. Li "The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications" Nucleic Acids Res., **15**, 1281 (1987)