

Appendix 2

Nature of start codons associated with putative strong $\sigma 70$ promoters - trends over 32 prokaryotic genomes -

Focusing on potentially strong promoters, we study the distribution of start codons between classes AUG, GUG, UUG and Other. The class frequencies vary through species but AUG frequency is constantly in the lead. We check that start codon is AUG for 80.36% of the genes (average over the 32 genomes); the average percentage of AUG start codons is 81.45% when calculated over the 26 largest genomes. As expected, start codon GUG ranks second (12.18% and 11.75% on average, over all 32 genomes and over the 26 largest genomes respectively). Then start codon UUG's average frequency is 7.19% when computed over all genomes and 6.47% when the 26 largest genomes are considered. The percentage of putative strong promoters associated with start codons different from the three previous ones is 0.27% over all 32 genomes and 0.33% over the 26 largest genomes. Considering the 8 large *Firmicutes* genomes, we observe that the average percentage of AUG start codons is equal to 82.94% (minimum 77.63%: *Bacillus subtilis*; maximum 90.53%: *Streptococcus pneumoniae*). The average is slightly higher for *Proteobacteria*, (86.47%), whereas the minimum and the maximum respectively amount to 78.68% (*Yersinia pestis*) and 94.59% (*Helicobacter pylori*).

In particular, for *Escherichia coli* model, the distribution of genes into classes AUG, GUG, UUG and Other is 85.88%, 10.59%, 3.14% and 0.39%. All previous trends were expected. However, 5.79% of the putative strong promoters of *Mycobacterium leprae* are associated with start codons other than AUG, GUG and UUG: GUA (twice), AGC (twice), AUA, CCU and ACC. The two other *Actinobacteria* studied do not show such deviations: *Mycobacterium tuberculosis* has only one AUA start codon over 289 genes with putative strong promoters; *Streptomyces coelicolor* has an AUC start codon over 320 genes. Otherwise, six species exhibit other start codons with very low frequencies (below 5×10^{-3}): these are *Bacillus subtilis*, *Clostridium perfringens*, *Listeria innocua*, *Listeria monocytogenes*, *Escherichia coli* and *Rickettsia prowazekii*. The start codons encountered are: AUU (7 times), CUG (7 times), AUA (3 times), AUC (once).

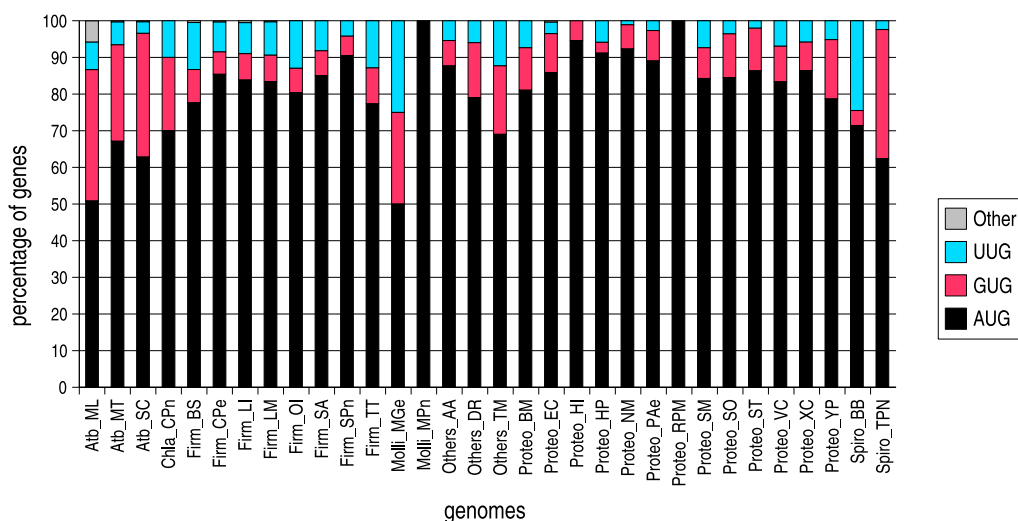


Figure 2.1 Nature of start codons for the genes harbouring potentially strong $\sigma 70$ promoters in 32 prokaryotic genomes. See Figure 3 for species nomenclature.