Appendix 1

Frequencies of putative strong $\sigma 70$ promoters obtained for 32 prokaryotic genomes

For the 32 genomes studied, the genome size and n, the total number of genes encoding proteins in this genome, are proven to be correlated (linear correlation coefficient: 0.93). With p denoting the number of genes harbouring a putative strong promoter, we compute the frequency $f = \frac{p}{n}$ for each genome, to escape the size bias when comparing genomes. We retain here that the percentage of genes harbouring a putative strong promoter varies between 0.75% (*Rickettsia prowazekii*) and 39.22% (*Thermotoga maritima*). The average over the 8 large *Firmicutes* genomes amounts to 28.25%, whereas the average over the 12 large *Proteobacteria* genomes is 5.46%. Besides, the percentage of genes harbouring putative strong promoters with UP elements varies in the range [0%, 32.98%]; the minimum and maximum are observed for *Deinococcus radiodurans* and *Clostridium perfringens* respectively. The percentage of putative strong promoters harbouring an UP element amounts to 81.86% on average for *Firmicutes*, whereas it is 8.53% for *Proteobacteria*. All *Firmicutes*'genomes except *Streptococcus pneumoniae*'s and *Mollicutes*' are characterized by a percentage over 50% (*Mollicutes* are bacteria with small genomes also belonging to the *Firmicutes* phylum). The percentages obtained for the *Firmicutes Bacillus subtilis* and *Thermoanaerobacter* are 54.8% and 56.2%. Otherwise, all percentages are 100% for *Firmicutes*.

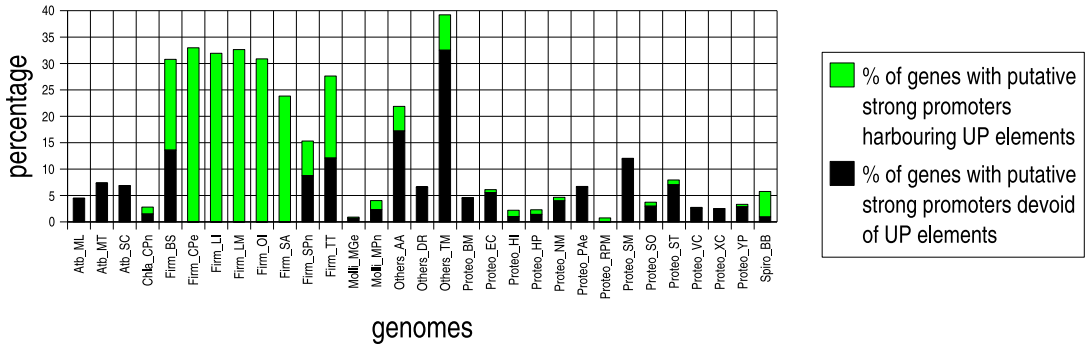

Figure 1.1 Comparison of percentages of genes harbouring putative strong $\sigma 70$ promoters over all genes coding for mRNAs, in 32 prokaryotic genomes. In particular, the *Firmicutes Listeria innocua*, *Listeria monocytogenes*, *Streptococcus pneumoniae* and *Thermoanaerobacter tengcongensis* show higher percentages than the similarly AT-rich *Proteobacteria Haemophilus influenza* and *Helicobacter pylori*. All six genomes are characterized with average 5'UTR AT-richness in the range $[60.2\%, 62.4\%]$. See Figure 3 for species nomenclature.