Appendix 4.
Detailed comments relative to Sections
Are potentially strong $\sigma 70$ promoter-like sequences frequent?
and
Are potentially strong $\sigma 70$ promoter-like sequences harbouring an UP-like element frequent?

Are potentially strong $\sigma 70$ promoter-like sequences frequent?

We recall that $p1_{CI}$ and $p1_{CII}$ respectively denote the percentages obtained under constraints CI and CII. *E. coli* model ($g = 4173$) is characterized by $p1_{CI} = 2.3\%$ and $p1_{CII} = 6.1\%$. With a number of genes not quite so different ($g = 3979$), *B. subtilis* model is described by $p1_{CI} = 20.0\%$ and $p1_{CII} = 30.8\%$. The highest $p1$ percentages are observed for *C. perfringens* ($g = 2532$; $p1_{CI} = 31.5\%$; $p1_{CII} = 33.0\%$) and *T. maritima* ($g = 1790$; $p1_{CI} = 17,5\%$; $p1_{CII} = 39.5\%$) whereas low percentages are observed for the genome with the highest number of genes (*P. aeruginosa*; $g = 5565$; $p1_{CI} = 0.8\%$; $p1_{CII} = 6.3\%$).

Are potentially strong $\sigma 70$ promoter-like sequences harbouring an UP-like element frequent?

We show that the differentiation between *Firmicutes* and other genomes holds, but it is more subdued for $p2$ percentage than for $p1$ percentage. In addition to *Firmicutes* and together with *A. aeolicus*, *T. maritima* and *B. burgdorferi*, already pointed out since having the highest $p1$ percentages (over 5%), eight more species other than *Firmicutes*, *H. pylori*, *R. prowazekii*, *C. pneumoniae*, *N. meningitidis*, *S. oneidensis*, *S. typhimurium*, *Y. pestis* and *H. influenza*, show the highest $p2$ percentages (over 10%).

No UP element can be identified under any constraint for *S. coelicolor* ($sp_{CI} = 22$; $sp_{CII} = 347$) and *X. campestris* ($sp_{CI} = 6$; $sp_{CII} = 97$) (see Figure 1 (c)). *M. genitalium*, the genome having the lowest number of putative strong promoters ($sp_{CI} = sp_{CII} = 4$), shows a single UP element under both constraints CI and CII.

More interestingly, another result is that some genomes having relatively few strong promoters show in contrast a high ($p2$) percentage of them harbouring an UP element, whatever the constraint: *H. influenza* ($sp_{CI} = 31$; $sp_{CII} = 37$; $upsp_{CI} = 8$; $upsp_{CII} = 20$; $p2_{CI} = 25.8\%$, $p2_{CII} = 54.1\%$), *B. burgdorferi* ($sp_{CI} = 43$; $sp_{CII} = 49$; $upsp_{CI} = 25$; $upsp_{CII} = 40$; $p2_{CI} = 58.1\%$, $p2_{CII} = 81.6\%$). Other such AT-rich genomes with few strong promoters show a high proportion of them harbouring an UP element, only under CII constraints (*C. pneumoniae*, *H. pylori* and *M. pneumoniae*). As an extreme trend, we observe that the very few promoters identified for *R. prowazekii* under CII relaxed constraint are all associated with an UP element ($sp_{CII} = upsp_{CII} = 6$; $p2_{CII} = 100\%$).

Nature of the genes harbouring putative strong promoters in their regulatory region

It is attractive to consider the following question: what are the genes found in association with the putative strong promoters coding for? Since various gene nomenclatures are used in GenBank files, pairwise genome comparison is not straightforward. One way to circumvent this difficulty is to perform pairwise alignments between the coding sequences of the genes, for each possible pair of genomes among the 32 genomes. The program BL2SEQ, which is part of BLAST software, is devoted to the comparison of two sequences. In this framework, the stand-alone executable BL2SEQ publicly available from NCBI ftp site ftp://ftp.ncbi.nlm.nih.gov/blast/executables/ was run to compare protein sequences. Two filtering steps were implemented: (i) a gene g_1 is compared to a gene g_2 if their lengths do not differ by more than a given percentage threshold; (ii) two genes are considered to be homologs if the identity percentage of the alignment amounts to a minimal percentage threshold. The threshold relative to the maximal difference allowed between gene lengths is meant to reduce pairwise comparison time. The two previous threshold values were tuned to 5% and 30% respectively. For every pair of genomes, this method points out all pairs of homologous genes both associated with strong promotors in each genome. Examining the limited set of homologous genes obtained (unpublished data), we could not reveal the existence of a pattern common to the different organisms. The nature of these housekeeping genes is totally different in the various organisms.

Stringency selectivity; confirmed specificity of *Firmicute* genomes

It has been observed that Firmicutes need longer SD sequences for translation initiation (31-32). As a consequence, we observe that the average percentage of genes encoding proteins associated with an optimal SD sequence calculated over the 18 large non Firmicute genomes is 9.17% whereas the average computed over the 8 large Firmicute genomes is nearly three times as much as the former (28.25%). Therefore, in Table 4.1, it is essential to compare the average ρ ratio calculated over the 18 large non Firmicute genomes to the average ρ ratio computed over the 8 large Firmicute genomes. Regarding the non Firmicute genomes, not only is the UP element presence a selective factor ($\rho < 0.15$); *CI* condition alone still only selects around a third of the genes associated with an optimal SD. Regarding Firmicute averages, when the UP element is required, the request is certainly not under-constrained; nonetheless, relatively high ratios are observed (0.27 and 0.55 respectively under *CI* and *CII* conditions). This previous remark, added to the fact that ρ is a normalized ratio suppressing the bias due to longer SDs in Firmicute genomes, inclines us to think that the ratio of 0.80 is meaningful under *CI* condition alone, for Firmicutes. On the contrary, the lack of difference between Firmicutes' and non Firmicutes' (high) percentages shows that *CII* conditions alone are too relaxed to describe intrinsically strong promoters.

ρ (average)	UP element optional		UP element required	
	CI	*CII*	*CI*	*CII*
32 genomes	0.55	0.94	0.13	0.31
26 large genomes	0.50	0.93	0.10	0.26
18 large non Firmicute genomes	0.30	0.90	0.04	0.13
8 large Firmicute genomes	0.80	0.98	0.27	0.55

Table 4.1 Comparison of the ratio ρ of the number of $\sigma 70$ promoter-like sequences associated with an optimal Shine-Dalgarno sequence, to the total number of genes associated with an optimal Shine-Dalgarno sequence, for 32 bacterial genomes and under four constraint sets.