Appendix 6.
Supplement to the discussion about the *Firmicutes* case

Due to mismatches and a maximal distance of 200 specified for the gap between SD and -10 box, we could not compute the theoretical expected probabilities to encounter $\sigma 70$ promoter-like sequences in randomly generated genomes. We were therefore compelled to implement simulations. In the case of the occurrence of an optimal Shine-Dalgarno sequence located between 2 and 10 bp upstream of the start codon, the calculation of the *exact* probability, denoted p_{rand}, is tractable. In the sequel, we will consider the language $\mathcal{L}$ of words of length 15 (maximal bp distance added to SD sequence length), constructed on alphabet $\mathcal{A} = \{A,\ C,\ T,\ G,\ n\}$, n being the usual IUPAC character indifferently coding for A, C, T or G. O_i $(1 \leq i \leq 9)$ will denote the event of an optimal SD sequence occurring at bp distance $i+1$, upstream of the start codon. Such 9 events simply correspond to the enumeration $nnnnnnnnGGAGGnn$, $nnnnnnnGGAGGnnn$, $\cdots$,$nGGAGGnnnnnnnnn$, $GGAGGnnnnnnnnnn$. Second, $O_{i_1} \cap O_{i_2} \cdots \cap O_{i_k}$ $(i_1 < i_2 < \cdots i_k)$ will represent the event corresponding to k overlappings. Finally, the probability of a word w belonging to $\mathcal{L}$ is denoted $p(w)$ and is merely computed as the product of its character probabilities (depending on the bacterial genome considered).

To take account of possible overlappings between occurrences, probability p_{rand} is successively refined following the Poincaré formula:

$$p_{rand}(\bigcup_{i=1}^{9} O_i) = \sum_{i=1}^{9} p(O_i) - \sum_{1 \leq i < j \leq 9} p(O_i \cap O_j) + \sum_{1 \leq i < j < k \leq 9} p(O_i \cap O_j \cap O_k) - \ \cdots\cdots\ + (-1)^8 p(O_1 \cap \ldots \cap O_9).$$

Namely, at level 1, the approximate probability p_{rand} amounts to the sum S_1 of the probabilities of all 9 occurrences $nnnnnnnnGGAGGnn$, $nnnnnnnGGAGGnnn \cdots nGGAGGnnnnnnnnn$, $GGAGGnnnnnnnnnn$. At level 2, the sum S_2 of the probabilities for pairwise intersections $nnnnn\overline{GGAGGAGG}nn$, $nnnn\overline{GGAGGGAGG}nn$, $\cdots$ $\overline{GGAGGGAGG}nnnnnn$, $\overline{GGAGGAGG}nnnnnnn$ is substracted from S_1. The process is iterated successively adding to or substrating from current p_{rand} decreasing terms.

Finally, given the probability p_{rand} and the number g of genes encoding proteins in the genome considered, we calculate the mean and standard deviation for the expected number of genes associated with SD optimal sequences as the parameters of a normal law: $M_{rand} = p_{rand} \times g$, $\sigma_{rand} = \sqrt{g \times p_{rand} \times (1 - p_{rand})}$. Then we compute the corresponding Z-score as $Z\text{-}score = \frac{|obs - M_{rand}|}{\sigma_{rand}}$, where *obs* is the number of genes associated with optimal SD sequences observed in the bacterial genome considered.

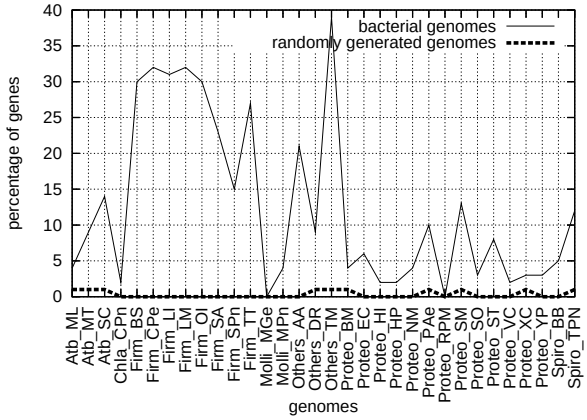

Figure 6.1 Percentage of genes encoding proteins associated with the optimal Shine-Dalgarno sequence GGAGG - comparison between the ratios observed in 32 bacterial genomes (p_{bact}) and the ratios expected from similarly-AT rich genomes generated at random (p_{rand}); p_{rand} is expressed as a percentage.

Table 6.1 describes the Z-scores observed over the 32 genomes analysed.

genome name	abbreviation	p_{bact} (%)	p_{rand} (%)	Z-score
Mycobacterium leprae tn	Atb_ML	4.87	1.08	18.95
Mycobacterium tuberculosis h37rv	Atb_MT	9.24	1.88	35.05
Streptomyces coelicolor a3 (2)	Atb_SC	14.26	1.98	61.4
Chlamydophila pneumoniae ar 39	Chla_CPn	2.81	0.37	13.56
Bacillus subtilis 168	Firm_BS	30.79	0.64	**251.25**
Clostridium perfringens str13	Firm_CPe	32.98	0.23	**363.89**
Listeria innocua	Firm_LI	31.94	0.48	**262.17**
Listeria monocytogenes strain EGD	Firm_LM	32.64	0.48	**247.38**
Oceanobacillus iheyensis hte831	Firm_OI	30.87	0.40	**304.7**
Staphylococcus aureus mw2	Firm_SA	23.83	0.27	**235.6**
Streptococcus pneumoniae r6	Firm_SPn	15.31	0.55	**86.82**
Thermoanaerobacter tengcongensis	Firm_TT	27.63	0.72	**168.19**
Mycoplasma genitalium G37	Molli_MGe	0.91	0.21	3.18
Mycoplasma pneumoniae M129	Molli_MPn	4.04	0.37	15.96
Aquifex aeolicus vf5	Others_AA	21.88	0.64	106.20
Deinococcus radiodurans r1 chr1	Others_DR	9.40	1.78	29.31
Thermotoga maritima	Others_TM	39.83	1.03	168.7
Brucella melitensis 16m chr1	Proteo_BM	4.86	1.19	15.96
Escherichia coli k12	Proteo_EC	6.21	0.90	37.93
Haemophilus influenza rd kw20	Proteo_HI	2.21	0.38	12.20
Helicobacter pylori j99	Proteo_HP	2.30	0.47	10.76
Neisseria meningitidis mc58	Proteo_NM	4.71	0.79	19.60
Pseudomonas aeruginosa pa01	Proteo_PAe	10.37	1.66	51.24
Rickettsia prowazekii madrid e	Proteo_RPM	0.75	0.11	5.33
Sinorhizobium meliloti 1021	Proteo_SM	13.39	1.43	59.8
Shewanella oneidensis mr1	Proteo_SO	3.73	0.60	28.45
Salmonella typhimurium lt2	Proteo_ST	8.01	0.90	50.79
Vibrio cholerae n16961 chr1	Proteo_VC	2.75	0,69	12.88
Xanthomonas campestris atcc 33913	Proteo_XC	3.74	1.66	10.95
Yersinia pestis	Proteo_YP	3.35	0.67	22.33
Borrelia burgdorferi b31	Spiro_BB	5.76	0.16	40.00
Treponema pallidum nichols	Spiro_TPN	12.96	1.43	30.34

Table 6.1 Percentage of genes encoding proteins associated with the optimal Shine-Dalgarno sequence GGAGG - comparison between the Z-scores observed for 32 genomes. The Z-scores are computed from the percentages observed on the bacterial genomes (p_{bact}) and the mean (p_{rand}) and standard deviation calculated from similarly-AT rich genomes generated at random.

If an organism like *B. subtilis* exhibits such a high frequency of putative strong promoters, one would then expect to encounter a higher concentration of mRNAs in these cells as compared to *E. coli*. Facing a similar question in view of the high densities of putative *functional* promoters identified in bacterial genomes, Huerta and co-authors suggest that the majority of putative *functional* promoters could simply not proceed further than the formation of the closed complex with RNA polymerase (Huerta *et al.*, 2006). Amongst such sequences, the ones that could be activated through single point mutation are postulated to be sequences inherited from the ancestral genome. Selection would maintain them to circumvent deleterious mutations of the main promoter(s), or to adjust gene expression depending on environmental changes. Secondly, according to these authors, within regulatory regions, some of the numerous promoter-like sequences detected might actually be functional, but only be active under restricted conditions. Though our study does not deal with densities in regulatory regions but frequencies over genomes, it is attractive to transpose such explanations to our case. Some genomes would be favoured by evolution as harbouring more potentially strong promoters than other genomes. However, the conditions under which these cryptic promoters would contribute to high gene expression are unknown. So far, as we will see in a further section, for cost reasons, only few experimentations relative to putative strong promoters identified by BacTrans^2 have been performed.