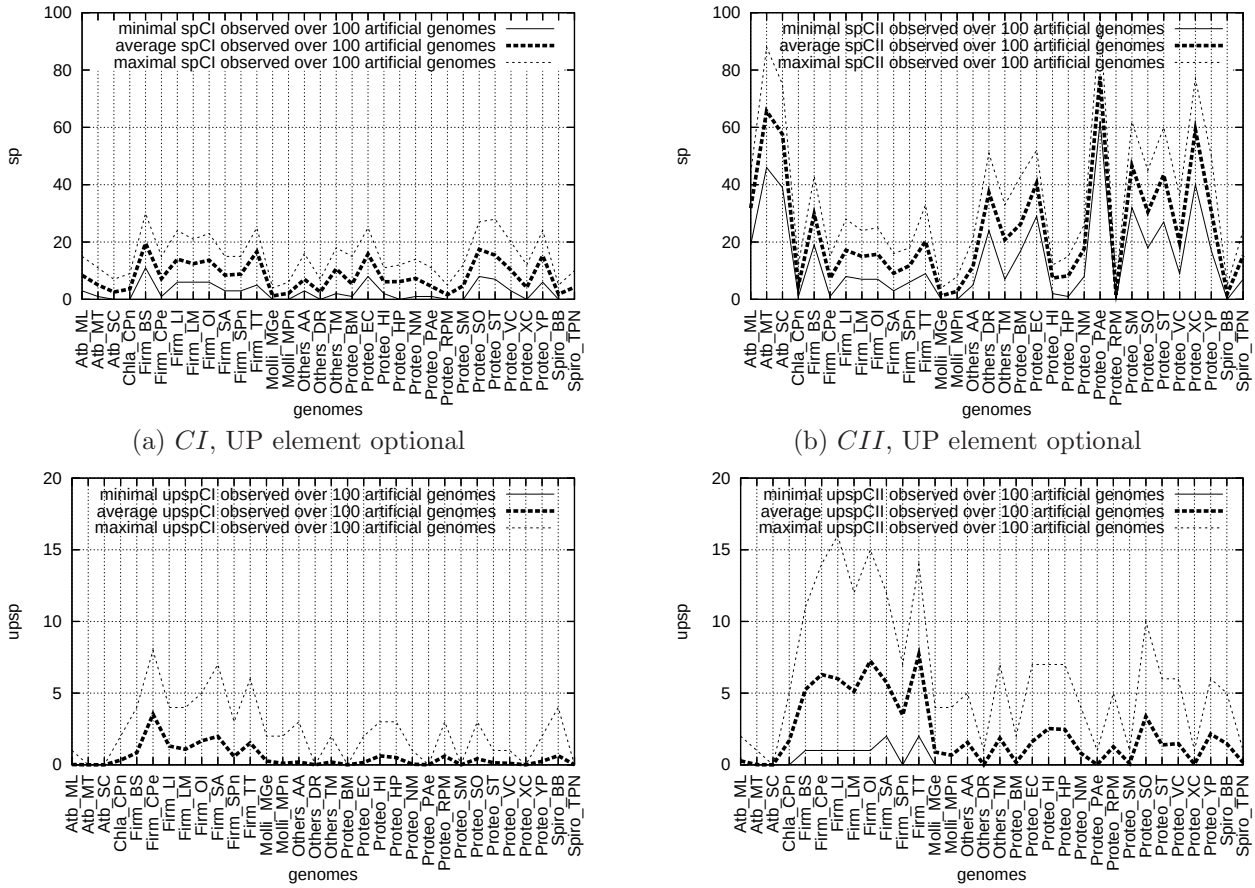
Appendix 6.
Comparing observations in bacterial genomes with expectations in randomly generated genomes

Magnification of the results obtained for randomly generated genomes



(a) *CI*, UP element optional

(b) *CII*, UP element optional

(c) *CI*, UP element required

(d) *CII*, UP element required

Figure 6.1 Minimal, average and maximal values observed over 100 randomly generated genomes, for *sp* and *upsp* respectively; 32 bacterial genomes are considered (see Figure 1 for genome nomenclature). For each such bacterial genome, 100 artificial genomes are generated at random, which have same proportions of A, C, T, G nucleotides and same total number of genes encoding proteins as the bacterial genome. sp_{CI} denotes the number of genes with a putative **S**trong **P**romoter identified under constraint set *CI*. sp_{CII} is defined similarly for constraint set *CII* (see Section "Systems and methods", Subsection "Genome analysis upon request" for the definition of *CI* and *CII* constraints). $upsp_{CI}$ denotes the number of genes with an **UP** element in their putative **S**trong **P**romoter, and identified under constraint set *CI*. $upsp_{CII}$ is defined similarly.

Description of statistical significance through Z-scores

In Section "Are potentially strong σ70 promoter-like sequences harbouring an UP-like element frequent?", *CII* condition has been shown to be under-constrained when not combined with the presence of the UP element. Therefore, in the sequel, the data relative to *CII* are only provided for the record.

Table 6.1 compares the Z-scores obtained under all four conditions. When Z-scores could not be calculated, we noticed that the *sp* value (respectively the *upsp* value) observed for the bacterial genome and the corresponding value expected for the average artificial genome are both equal to 0 (exceptionally, such previous values are equal to 1 and 0 respectively).

	CI UP element optional				*CII* UP element optional				*CI* UP element required				*CII* UP element required			
	min	max	av	std	min	max	av	std	min	max	av	std	min	max	av	std
8 *Firmicutes*	81.3	308.5	193.0	66.1	92.9	324.6	216.9	66.8	74.9	291.7	155.9	64.3	69.4	311.9	188.3	66.9
13 *Proteobacteria*	1.0	32.4	16.0	9.5	3.4	53.6	23.5	15.2	2.0 (1)	15.6	11.0	7.9	4.4	66.6	16.8	17.0
26 species with large genomes	1.0	308.5	74.9	89.9	4.9	324.6	91.4	97.0	0.15 (2)	291.7	76.9	81.6	0.2 (3)	311.9	75.9	88.8

Table 6.1 Evaluation of the statistical significance of $\sigma 70$ promoter frequencies: comparison between different bacterial groups. For details about Z-scores, see text, Subsection "Comparing observations in bacterial genomes with expectations in randomly generated genomes"; av: average, std: standard deviation. (1)(2)(3): Z-scores were calculable respectively for 10, 19 and 25 genomes.

genome name	abbreviation	*CI* UP element optional			*CII* UP element optional			*CI* UP element required			*CII* UP element required		
		genomes for which Z-score value is above threshold											
		7	**15**	**80**	**7**	**15**	**80**	**7**	**15**	**80**	**7**	**15**	**80**
Mycobacterium leprae tn	Atb_ML	ML	ML		ML	ML							
Mycobacterium tuberculosis h37rv	Atb_MT	MT			MT	MT		-	-	-	MT	MT	
Streptomyces coelicolor a3 (2)	Atb_SC	SC			SC	SC		-	-	-	-	-	-
Aquifex aeolicus vf5	Others_AA	AA	AA	AA	AA	AA	AA	AA	AA		AA	AA	
Deinococcus radiodurans r1	Others_DR	DR			DR	DR		-	-	-	DR		
Thermotoga maritima	Others_TM	TM	TM	TM	TM	TM	TM	TM	TM		TM	TM	TM
Brucella melitensis 16m	Proteo_BM	BM			BM			-	-	-	BM		
Escherichia coli k12	Proteo_EC	EC	EC		EC	EC		EC			EC	EC	
Haemophilus influenza rd kw20	Proteo_HI	HI			HI			HI			HI		
Helicobacter pylori j99	Proteo_HP	HP			HP			HP			HP		
Neisseria meningitidis mc58	Proteo_NM	NM	NM		NM	NM		NM	NM		NM		
Pseudomonas aeruginosa pa01	Proteo_PAe	PAe	PAe		PAe	PAe		-	-	-	PAe	PAe	
Sinorhizobium meliloti 1021	Proteo_SM	SM	SM		SM	SM		-	-	-	SM	SM	
Shewanella oneidensis mr1	Proteo_SO	SO	SO		SO	SO		SO			SO	SO	
Salmonella typhimurium lt2	Proteo_ST	ST	ST		ST	ST		ST	ST		ST	ST	
Vibrio cholerae n16961	Proteo_VC				VC								
Xanthomonas campestris atcc 33913	Proteo_XC							-	-	-			
Yersinia pestis	Proteo_YP	YP			YP	YP					YP		

Table 6.2 Evaluation of the statistical significance of $\sigma 70$ promoter frequencies for 18 bacterial species selected as non *Firmicutes* species with large genomes. For definition of CI and CII constraints, see Section "Systems and methods", Subsection "Genome analysis upon request". Three significance thresholds are considered (7, 15 and 80). − means that the Z-score is not calculable. In a given column, the mention of a species points out statistical significance for the corresponding threshold.

		Z-scores				
	UP element presence	*E. coli*	min	max	average	standard deviation
CI	optional	**21.7**	1.0	308.5	74.9	89.9
CII	"	**38.2**	4.9	324.6	91.4	97.0
CI	required	**7.3**	0.15	291.7	76.9	81.6
CII	"	**15.3**	0.2	311.9	75.9	88.8

Table 6.3 Evaluation of the statistical significance of $\sigma 70$ promoter frequencies: comparison of *E. coli* with respect to 26 species with large genomes.

		Z-score threshold 7	15	80	number of genomes with calculable Z-scores
	UP element presence				
CI	optional	24	15	10	26
CII	"	25	21	10	26
CI	required	16	12	7	19
CII	"	22	16	8	25

Table 6.4 Numbers of large genomes (among 26) for which the total number of putative strong promoters identified is shown to be significantly different from that of the corresponding "average" randomly generated genome.

Description of statistical significance through relative differences

In complement to Figure 2, Table 6.5 describes the relative differences between bacterial and randomly generated genomes. With an overwhelming majority, this relative difference is always higher than 0.70. However, we would insist that an asserted difference must not be mistaken for biological significance, whose proof can only be brought through experimental validation. Besides, we recall that results relative to *CII* constraint are only provided for the record.

	UP element optional		UP element required	
	CI	*CII*	*CI*	*CII*
Aquifex aeolicus vf5	0.97	0.97	0.99	0.95
Bacillus subtilis 168	0.98	0.98	0.99	0.95
Borrelia burgdorferi b31	0.96	0.96	0.97	0.99
Brucella melitensis 16m chr1	0.79	0.72		0.95
Chlamydophila pneumoniae ar39	0.84	0.84	0.95	0.97
Clostridium perfringens str13	0.99	0.99	0.99	1.00
Deinococcus radiodurans r1 chr1	0.88	0.82		0.98
Escherichia coli k12	0.84	0.84	0.95	0.85
Haemophilus influenza rd kw20	0.80	0.80	0.92	0.95
Helicobacter pylori j99	0.80	0.76	0.93	0.87
Listeria innocua	0.98	0.98	0.99	0.99
Listeria monocytogenes strain EGD	0.98	0.98	0.99	0.99
Mycobacterium leprae tn	0.73	0.74		
Mycobacterium tuberculosis h37rv	0.84	0.77		0.99
Mycoplasma genitalium G37	0.70	0.68	0.73	0.85
Mycoplasma pneumoniae M129	0.88	0.89	0.95	0.90
Neisseria meningitidis mc58	0.85	0.81	0.99	0.92
Oceanobacillus iheyensis hte831	0.99	0.99	0.99	0.98
Pseudomonas aeruginosa pa01	0.90	0.78		1.00
Rickettsia prowazekii madrid e	0.74	0.73	0.70	0.97
Salmonella typhimurium lt2	0.87	0.87	0.99	0.97
Shewanella oneidensis mr1	0.85	0.82	0.96	0.88
Sinorhizobium meliloti 1021	0.91	0.88		0.94
Staphylococcus aureus mw2	0.99	0.99	0.99	0.99
Streptococcus pneumoniae r6	0.96	0.96	0.99	0.98
Streptomyces coelicolor a3 (2)	0.88	0.83		
Thermoanaerobacter tengcongensis	0.97	0.97	0.99	0.97
Thermotoga maritima	0.97	0.97	0.99	0.99
Treponema pallidum nichols	0.90	0.88	1.00	0.97
Vibrio cholerae n16961 chr1	0.66	0.73	0.95	0.82
Xanthomonas campestris atcc 33913	0.31	0.39		
Yersinia pestis kim	0.71	0.77	0.92	0.89
average (global)	0.86	0.85	0.95	0.95
average (large genomes)	0.86	0.85	0.97	0.95
average (large non Firmicute genomes)	0.81	0.79	0.96	0.93
average (large Firmicute genomes)	0.96	0.96	0.99	0.98

Table 6.5 Another description of the difference between bacterial and randomly generated genomes: the ratio displayed is computed as $\frac{f_{bact}-f_{rand}}{f_{bact}}$, where f_{bact} and f_{rand} respectively denote the frequency of genes encoding proteins harbouring a putative strong promoter, in a bacterial genome, and the corresponding frequency in the similarly-AT rich genome generated at random.

Is there a genome size bias? Comparison of *Firmicutes* genomes with similarly AT-rich *Proteobacteria* genomes

Hereafter, we complete our analysis, checking that the specificity observed for *Firmicutes* is not due to genome size bias. For this purpose, we compare two *Proteobacteria* genomes (*H. influenza*, *H. pylori*) and four *Firmicutes* genomes (*L. innocua*, *L. monocytogenes*, *S. pneumoniae*, *T. tengcongensis*). All six (high) AT-richnesses range in the narrow interval [60.8%, 62.6%]. Successively considering the number of promoters observed in each *Proteobacteria* genome as a reference, we calculate a corrected frequency for each *Firmicute* genome, applying a correction based on proportionality relative to genome sizes. Then we compare observed values *versus* corrected values under each of the four conditions studied. Not only do we implement such corrections for bacterial genomes, we also process the six average random genomes in a similar way (see Tables 6.5 and 6.6). When considering the average genomes generated at random corresponding to *Firmicutes*, we show that the order of magnitude is identical for expected values and corrected expected values. In contrast, when dealing with bacterial genomes, the size bias correction does not smooth out the difference between *Firmicutes* and *Proteobacteria*.

	Proteobacteria		*Firmicutes*							
	HI	HP	LI		LM		SPn		TT	
AT-content	61.9%	60.8%	62.6%		62.0%		60.28%		62.4%	
g	**1673**	**1478**	**2962**	*HI - HP*	**2837**	*HI - HP*	**1861**	*HI - HP*	**2588**	*HI - HP*
	obs	obs	obs	corr	obs	corr	obs	corr	obs	corr
sp_{CI}	31	31	713	*54-62*(*)	707	*52-59*	213	*34-39*	581	*47-54*
sp_{CII}	37	34	946	*65-68*	926	*62-65*	285	*41-42*	715	*57-59*
$upsp_{CI}$	8	7	145	*14-14*	147	*13-13*	59	*8-8*	150	*12-12*
$upsp_{CII}$	20	17	501	*35-34*	488	*33-32*	120	*22-21*	407	*30-29*

Table 6.6 Comparison of the observed numbers of putative strong promoters between two *Proteobacteria* and four *Firmicutes* genomes characterized by close (high) AT-contents (range $[60.2\%, 62.4\%]$), under conditions CI and CII, and with or without UP element required. g: total number of genes encoding proteins in the genome considered; HI: *Haemophilus influenza*, HP: *Helicobacter pylori*; LI: *Listeria innocua*, LM: *Listeria monocytogenes*, SPn: *Streptococcus pneumoniae*, TT: *Thermoanaerobacter tengcongensis*; obs: observed values, corr: corrected values. (*) With HI then HP taken as a reference, these columns in italics yield the corrected values sp_{CI}, ..., $upsp_{CII}$ on the basis of proportionality to total gene number; the two corrected values in italics have to be compared with the value on their left. Example: the reference being HI, corrected $sp_{CI,corr}$ for LI is $sp_{CI,corr}(LI) = \frac{sp_{CI,obs}(HI)\times 2962}{1673} = 54$, with $sp_{CI,obs}(HI) = 31$; $sp_{CI,corr}(LI)$ has to be compared with the value of 713 observed for LI.

	Proteobacteria		*Firmicutes*							
	HI	HP	LI		LM		SPn		TT	
AT-content	61.9%	60.8%	62.6%		62.0%		60.28%		62.4%	
g	**1673**	**1478**	**2962**	*HI - HP*	**2837**	*HI - HP*	**1861**	*HI - HP*	**2588**	*HI - HP*
	exp	exp	exp	corr	exp	corr	exp	corr	exp	corr
sp_{CI}	6	6	14	*10-12*(*)	12	*10-11*	8	*6-7*	16	*9-10*
sp_{CII}	7	8	17	*12-16*	14	*11-15*	11	*7-10*	20	*10-12*
$upsp_{CI}$	0	0	1	*0-0*	1	*0-0*	0	*0-0*	1	*0-0*
$upsp_{CII}$	2	2	6	*3-4*	5	*3-3*	3	*2-2*	7	*3-3*

Table 6.7 Comparison of the expected numbers of putative strong promoters between six average genomes generated at random and characterized by the same AT-contents as two *Proteobacteria* and four *Firmicutes* genomes (range $[60.2\%, 62.4\%]$), under conditions CI and CII, and with or without UP element required. See Table 6.5 caption for explanations. Example: the reference being HI, corrected $sp_{CI,corr}$ for LI is $sp_{CI,corr}(LI) = \frac{sp_{CI,exp}(HI)\times 2962}{1673} = 10$, with $sp_{CI,exp}(HI) = 6$; $sp_{CI,corr}(LI)$ has to be compared with the value of 14 expected for the average artificial genome having the same AT-richness as the LI genome.