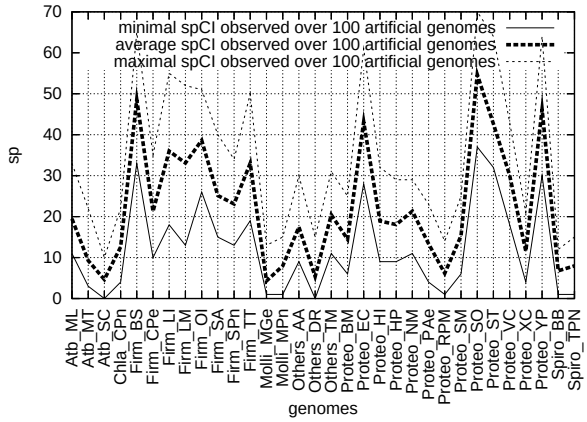
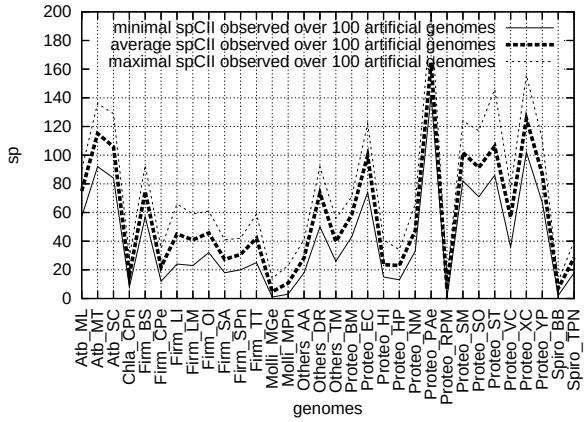
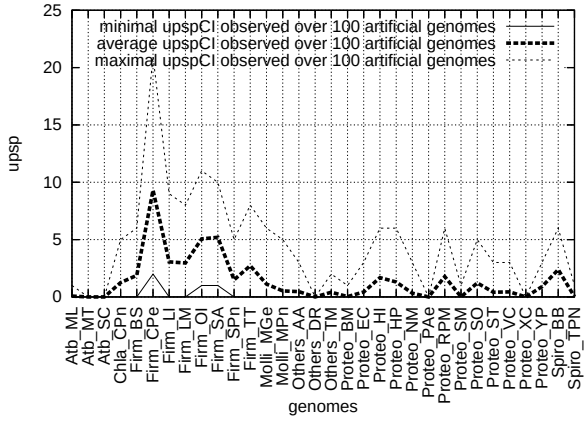
Appendix 5.
Comparing observations in bacterial genomes with expectations in randomly generated genomes



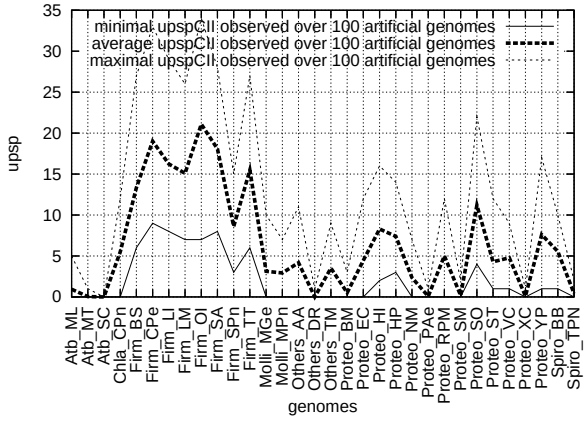
(a) CI, UP element optional



(b) CII, UP element optional



(c) CI, UP element required



(d) CII, UP element required

Figure 5.1. 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 text, 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.

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							
		5	2	5	2	5	2	5	2
Mycobacterium leprae tn	Atb_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
Deinococcus radiodurans r1	Others_DR		DR	DR	DR	-	-		
Thermotoga maritima	Others_TM	TM	TM	TM	TM	TM	TM	TM	TM
Brucella melitensis 16m	Proteo_BM		BM	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
Helicobacter pylori j99	Proteo_HP				HP				HP
Neisseria meningitidis mc58	Proteo_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
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		XC		XC	-	-		
Yersinia pestis	Proteo_YP			YP	YP				YP

Table 5.1. Evaluation of the 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 Subsection "Genome analysis upon request". The significance is evaluated through the Z-score value (see text, Subsection "Empirical approach"). Two thresholds are considered (5 and 2). − means that the Z-score is not calculable. In a given column, the mention of a species points out statistical significance.

		Z-scores				
	UP element presence	*E. coli*	min	max	average	standard deviation
CI	optional	**7.33**	0.51	165.32	38.42	50.16
CII	"	**15.38**	2.02	172.62	51.12	56.87
CI	required	**3.95**	0.11	128.4	33.66	38.33
CII	"	**8.43**	0.11	239.71	59.94	82.06

Table 5.2. Evaluation of the significance of $\sigma 70$ promoter frequencies: comparison of *E. coli* with respect to 26 species with large genomes. For details about Z-scores, see Subsection "Comparing observations in bacterial genomes with expectations in genomes randomly generated".