

Appendix 1.

Frequencies of putative strong promoters observed over 32 prokaryotic genomes

<i>n</i>	<i>a</i>	<i>s</i> (Mbp)	<i>at</i>	<i>g</i>	<i>CI</i>		<i>CII</i>	
					<i>sp_{CI}</i>	<i>p1_{CI}</i> (%)	<i>sp_{CII}</i>	<i>p1_{CII}</i> (%)
<i>Aquifex aeolicus</i> <i>vf5</i>	Others_AA	1.55	57.62	1522	188	12.4	333	21.9
<i>Bacillus subtilis</i> <i>168</i>	Firm_BS	4.21	59.27	3979	767	19.3	1225	30.8
<i>Borrelia burgdorferi</i> <i>b31</i>	Spiro_BB	0.91	73.07	850	43	5.1	49	5.8
<i>Brucella melitensis</i> <i>16m chr1</i>	Proteo_BM	3.28	45.56	2059	23	1.1	95	4.6
<i>Chlamydomonas reinhardtii</i> <i>at3</i>	Chla_CPn	1.22	62.55	1069	23	2.2	30	2.8
<i>Clostridium perfringens</i> <i>str13</i>	Firm_CPe	3.03	74.93	2532	790	31.2	835	33.0
<i>Deinococcus radiodurans</i> <i>r1 chr1</i>	Others_DR	3.05	34.01	2521	14	0.6	167	6.6
<i>Escherichia coli</i> <i>k12</i>	Proteo_EC	4.63	52.34	4173	97	2.3	255	6.1
<i>Haemophilus influenzae</i> <i>rd kw20</i>	Proteo_HI	1.83	64.22	1673	21	1.3	37	2.2
<i>Helicobacter pylori</i> <i>j99</i>	Proteo_HP	1.64	62.95	1478	24	1.6	34	2.3
<i>Listeria innocua</i>	Firm_LI	3.01	64.25	2962	662	22.3	946	31.9
<i>Listeria monocytogenes</i> <i>strain EGD</i>	Firm_LM	2.94	63.97	2837	662	23.3	926	32.6
<i>Mycobacterium leprae</i> <i>tn</i>	Atb_ML	3.26	49.55	2670	3	0.1	12	0.4
<i>Mycobacterium tuberculosis</i> <i>h37rv</i>	Atb_MT	4.41	35.47	3909	17	0.4	289	7.4
<i>Mycoplasma genitalium</i> <i>G37</i>	Molli_MGe	0.58	69.54	441	3	0.7	4	0.9
<i>Mycoplasma pneumoniae</i> <i>M129</i>	Molli_MPn	0.81	61.74	644	7	1.1	26	4.0
<i>Neisseria meningitidis</i> <i>mc58</i>	Proteo_NM	2.27	51.90	1954	41	2.1	91	4.7
<i>Oceanobacillus iheyensis</i> <i>hte831</i>	Firm_OI	3.63	66.41	3398	852	25.1	1049	30.9
<i>Pseudomonas aeruginosa</i> <i>pa01</i>	Proteo_PAE	6.26	35.28	5565	51	0.9	374	6.7
<i>Rickettsia prowazekii</i> <i>madrid e</i>	Proteo_RPM	1.11	73.16	796	6	0.8	6	0.8
<i>Salmonella typhimurium</i> <i>lt2</i>	Proteo_ST	4.85	51.22	4334	118	2.7	344	7.9
<i>Shewanella oneidensis</i> <i>mr1</i>	Proteo_SO	4.96	56.69	4501	73	1.6	168	3.7
<i>Sinorhizobium meliloti</i> <i>1021</i>	Proteo_SM	3.65	39.50	3272	58	1.8	394	12.0
<i>Staphylococcus aureus</i> <i>mw2</i>	Firm_SA	2.82	69.77	2610	546	20.9	622	23.8
<i>Streptococcus pneumoniae</i> <i>r6</i>	Firm_SPn	2.03	62.39	1861	204	11.0	285	15.3
<i>Streptomyces coelicolor</i> <i>a3 (2)</i>	Atb_SC	8.66	29.15	4665	15	0.3	320	6.9
<i>Thermoanaerobacter tengcongensis</i>	Firm_TT	2.68	63.83	2588	557	21.5	715	27.6
<i>Thermotoga maritima</i>	Others_TM	1.86	54.59	1790	294	16.4	702	39.2
<i>Treponema pallidum</i> <i>nichols</i>	Spiro_TPN	1.13	47.00	980	41	4.2	125	12.8
<i>Vibrio cholerae</i> <i>n16961 chr1</i>	Proteo_VC	4.03	54.80	2618	20	0.8	72	2.8
<i>Xanthomonas campestris</i> <i>atcc 33913</i>	Proteo_XC	5.07	35.55	4120	4	0.1	103	2.5
<i>Yersinia pestis</i>	Proteo_YP	4.6	55.36	4090	51	1.2	136	3.3

Table 1.1. Frequencies of genes harbouring a potentially strong $\sigma 70$ promoter, under two constraint sets, in 32 prokaryotic genomes. *n*: micro-organism name; *a*: abbreviation for micro-organism name (Atb: *Actinobacteria*, Chla: *Chlamydia*, Firm: *Firmicutes* (among which Molli.: *Mollicutes*), "Others" group, Proteo: *Proteobacteria*, Spiro: *Spirochaetales*); *s*: genome size (bp millions); *at*: average percentage of nucleotides T and A over the 350 bp-long region upstream of start codon, computed over the total number *g* of genes encoding proteins in the genome; *sp_{CI}*: number of genes harbouring a putative Strong Promoter identified under constraint set *CI*, *sp_{CII}*: *idem*, under the more relaxed constraint set *CII* (see text, Subsection "Genome analysis upon request" for the definition of *CI* and *CII* constraints); $p1_{CI} = 100 \times sp_{CI}/g$, $p1_{CII} = 100 \times sp_{CII}/g$.

<i>n</i>	<i>a</i>	<i>at</i>	<i>g</i>	<i>CI</i>		<i>CII</i>		$p1_{CI} \times p2_{CI}$ (%)	$p1_{CII} \times p2_{CII}$ (%)
				<i>upsp_{CI}</i>	<i>p2_{CI}</i> (%)	<i>upsp_{CII}</i>	<i>p2_{CII}</i> (%)		
Aquifex aeolicus vf5	Others_AA	57.62	1522	18	9.6	71	21.3	1.2	4.7
Bacillus subtilis 168	Firm_BS	59.27	3979	98	12.8	684	54.8	2.4	16.4
Borrelia burgdorferi b31	Spiro_BB	73.07	850	23	53.5	41	83.7	2.7	4.9
Brucella melitensis 16m	Proteo_BM	45.56	2059	0	0	4	4.2		0.2
Chlamydomonas reinhardtii	Chla_CPn	62.55	1069	4	17.4	14	46.7	0.4	1.3
Clostridium perfringens str13	Firm_CPe	74.93	2532	451	57.1	835	100	17.0	31.4
Deinococcus radiodurans r1	Others_DR	34.01	2521	0	0	0	0		
Escherichia coli k12	Proteo_EC	52.34	4173	3	3.1	23	9.1	0.1	0.5
Haemophilus influenzae rd kw20	Proteo_HI	64.22	1673	8	38.1	20	54.1	0.5	1.2
Helicobacter pylori j99	Proteo_HP	62.95	1478	3	12.5	13	38.2	0.2	0.9
Listeria innocua	Firm_LI	64.25	2962	131	19.8	946	100	4.4	31.7
Listeria monocytogenes strain EGD	Firm_LM	63.97	2837	128	19.3	926	100	4.5	32.4
Mycobacterium leprae tn	Atb_ML	49.55	2670	0	0	0	0		
Mycobacterium tuberculosis h37rv	Atb_MT	35.47	3909	1	5.9	3	1.0		0.1
Mycoplasma genitalium G37	Molli_MGe	69.54	441	0	0	1	25.0		0.2
Mycoplasma pneumoniae M129	Molli_MPn	61.74	644	0	0	11	42.3		1.6
Neisseria meningitidis mc58	Proteo_NM	51.90	1954	2	4.9	13	14.3	0.1	0.6
Oceanobacillus ihayensis hte831	Firm_OI	66.41	3398	174	20.4	1049	100	5.0	30.1
Pseudomonas aeruginosa pa01	Proteo_PAe	35.28	5565	0	0	6	1.6		0.1
Rickettsia prowazekii madrid e	Proteo_RPM	73.16	796	1	16.7	6	100	0.1	0.7
Salmonella typhimurium lt2	Proteo_ST	51.22	4334	11	9.3	39	11.3	0.3	0.9
Shewanella oneidensis mr1	Proteo_SO	56.69	4501	4	5.5	34	20.2	0.1	0.7
Sinorhizobium meliloti 1021	Proteo_SM	39.50	3272	0	0	5	1.3		0.2
Staphylococcus aureus mw2	Firm_SA	69.77	2610	183	33.5	622	100	6.9	23.6
Streptococcus pneumoniae r6	Firm_SPn	62.39	1861	53	26.0	122	42.8	2.7	6.3
Streptomyces coelicolor a3 (2)	Atb_SC	29.15	4665	0	0	1	0.3		
Thermoanaerobacter tengcongensis	Firm_TT	63.83	2588	123	22.1	402	56.2	4.8	15.5
Thermotoga maritima	Others_TM	54.59	1790	22	7.5	120	17.1	1.2	6.5
Treponema pallidum nichols	Spiro_TPN	47.00	980	1	2.4	8	6.4	0.1	0.8
Vibrio cholerae n16961	Proteo_VC	54.80	2618	1	5.0	4	5.6		0.1
Xanthomonas campestris atcc 33913	Proteo_XC	35.55	4120	0	0	0	0		
Yersinia pestis	Proteo_YP	55.36	4090	1	2.0	19	14.0		0.5

Table 1.2. Frequencies of genes with a putative $\sigma 70$ promoter harbouring an UP element, under two constraint sets, in 32 prokaryotic genomes. *n*: micro-organism name; *a*: abbreviation for micro-organism name (see Table 1.1.); *at*: average percentage of nucleotides T and A over the 350 bp-long region upstream of start codon, computed over the total number *g* of genes in the genome; *upsp_{CI}*: number of genes with an UP element harboured in the Strong Promoter region, identified under constraint set *CI*; *upsp_{CII}*: *idem*, under the more relaxed constraint set *CII* (see text, Section "Genome analysis upon request" for the definition of *CI* and *CII* constraints); $p2_{CI} = 100 \times upsp_{CI}/sp_{CI}$, $p2_{CII} = 100 \times upsp_{CII}/sp_{CII}$, with *sp_{CI}* (resp. *sp_{CII}*) the number of genes harbouring a Strong Promoter under constraints *CI* (resp. *CII*). The percentages indicated in the two rightmost columns are respectively $p1_{CI} \times p2_{CI} = upsp_{CI}/g$ and $p1_{CII} \times p2_{CII} = upsp_{CII}/g$, that is the ratio of the number of genes identified with an UP element in their putative strong promoter to the number *g* of genes encoding proteins in the genome considered; no mention means that this percentage is lower than 0.01.