Appendix 7.
Putative strong promoters *versus* experimentally verified promoters in *Escherichia coli* genome

					CI					
gene	functional promoter identifier (*)		score (**)	strand	BacTrans^2 TSS	RegulonDB TSS	Δ_{BR}	PromEC TSS	Δ_{BP}	Δ_{RP}
infC	P2		1.0	reverse	1898842	1898842	**0**	see RegulonDB	see RegulonDB	0
ahpC			1.0	forward	638105	638144	39	see RegulonDB	see RegulonDB	0
manX			1.0	forward	1899938	1899957	**19**	see RegulonDB	see RegulonDB	0
ansB	P2		1.4	reverse	3098822	3098773	49	3098770	52	3
phnC			1.4	reverse	4322774	4323229	455	4322784	**10**	445
gltB			1.4	forward	3352071	3352531	460	3352144	73	387
argI			1.4	reverse	4475921	4476367	446	4475907	**14**	460
frdA			6.4	reverse	4379994	4380435	441	4379990	**4**	445
carA	P1		1.4	forward	29502	29551	49	-	-	-
flgB			11.4	forward	1130088	1130216	128	-	-	-
clpB			0.6	reverse	2732222	-	-	2732225	**3**	-
pfkB			1.4	forward	1804299	-	-	1804376	77	-
					CII					
gene	functional promoter identifier (*)	(***)	score (**)	strand	BacTrans^2 TSS	RegulonDB TSS	Δ_{BR}	PromEC TSS	Δ_{BP}	Δ_{RP}
infC	P2	x					**0**			
ahpC		x					**39**			
manX		x					**19**			
argI		x					446		**14**	
ansB	P2	x					49		52	
phnC		x					455		**10**	
gltB		x					460		73	
clpA	P1		1.6	forward	922302	922303	**1**	see RegulonDB	see RegulonDB	0
oxyR			1.6	forward	4156036	4156480	444	4156036	**0**	444
otsB			1.8	reverse	1980464	1980489	**25**	see RegulonDB	see RegulonDB	0
acnA	P1		1.8	forward	1333810	1333448	362	see RegulonDB	see RegulonDB	0
	P2		”	”	”	1333805	**5**	see RegulonDB	see RegulonDB	0
frdA		d	1.8	reverse	4379980	4380435	455	4379990	**10**	445
carA	P1	x					49	-	-	-
rhaS			1.6	forward	4095229	4095734	963	-	-	-
cpxR			1.6	reverse	4103443	4103709	266	-	-	-
flgB		d	1.8	forward	1130088	1130216	128	-	-	-
mgtA	P1		2.0	forward	4465137	4465386	249	-	-	-
	P2		”	”	”	4465303	166	-	-	-
zntA			2.0	forward	3604017	3604445	428	-	-	-
clpB		x				-	-		**3**	-
pfkB		x				-	-		77	-
					CII, UP element required					
gene	functional promoter identifier (*)	(***)	score (**)	strand	BacTrans^2 TSS	RegulonDB TSS	Δ_{BR}	PromEC TSS	Δ_{BP}	Δ_{RP}
carA	P1		15.4	forward	29552	29551	**1**	-	-	-
pfkB			6.0	forward	1804332	-	-	1804376	44	-

Table 7.1 Comparison of the transcription start sites (TSSs) relative to putative strong σ70 promoters identified by BacTrans^2 in *E. coli* genome, with respect to the TSS(s) of the functional σ70 promoter(s) harboured in the same gene. The functional promoters of *E. coli* are annotated in RegulonDB and PromEC databases. (*) when several functional promoters are known, they are denoted $P1$, $P2$ and so on; (**) BacTrans^2 score (see Subsection "Scoring function used"); (***) x indicates whether the putative strong promoters identified under CI and CII constraints are identical; d indicates a difference between them; Δ_{BR}: bp distance between BacTrans^2 TSS and RegulonDB TSS; Δ_{BP}: bp distance between BacTrans^2 TSS and PromEC TSS; Δ_{RP}: bp distance between RegulonDB TSS and PromEC TSS; - indicates that the gene is not referred to in the corresponding database; low bp distances (≤ 25) are highlighted in boldface characters.

In the sequel, we will refer to the two repositories for functional σ70 promoters known in *E. coli*: RegulonDB, (http://regulondb.ccg.unam.mx/data/Promoter Set.txt) and PromEC (http://margalit.huji.ac.il/). *E. coli* genome contains 4173 genes (coding for m-RNAs). In 5.7 RegulonDB release (june 2007), we listed 1632 genes, among which 601 (36%) are associated with a σ70 annotation. The field showing evidence of experimental validation contains at least one of the following items (or possibly sub-items): transcription initiation mapping, RNA polymerase footprinting, inferred from mutant, inferred from direct assay, inferred from genetic interaction, inferred from experiment, inferred from physical interaction. Some promoters among the earlier

listed in RegulonDB had not been assigned an evidence field. Nevertheless, they are indeed experimentally verified promoters and are in the process of being fully annotated (Salgado, personal communication).

Regarding CI constraints, 12 out of the 96 genes harbouring potentially strong $\sigma 70$ promoters are present in RegulonDB or PromEC databases. Gene $infC$ is the only one having the TSS relative to the putative promoter exactly coinciding with a functional promoter referred to in RegulonDB or PromEC. Besides, we note that genes $infC$, $ahpC$ and $manX$ each have identical TSSs for the functional promoter in RegulonDB and the functional promoter in PromEc database. On the contrary, $ansB$, $phnC$, $gltB$, $frdA$ and $argI$ RegulonDB TSSs differ from those reported in PromEC database. Two genes ($carA$, $flgB$) are referred to in RegulonDB only. Two genes ($clpB$, $pfkB$) are only mentioned in PromEC database. For some genes, the TSSs relative to the putative strong promoter and the functional promoter are close (0 ($infC$), 3 ($clpB$), 4 ($frdA$)) or relatively close (10 ($phnC$), 14 ($argI$), 19 ($manX$)). All three remaining genes show a bp distance between the putative strong promoter and the functional promoter over 39: 39 ($ahpC$), 49 ($ansBP2$), 49 ($carAP1$), $gltBB$ (73), $pfkB$ (77), $flgB$ (128).

When constraints are relaxed (CII), 20 out of the 254 genes identified with putative strong promoters are present in RegulonDB or PromEC databases. Among the 12 genes already identified under CI constraints and mentioned in at least one of the RegulonDB or PromEC databases, we check that the BacTrans^2 score is improved for genes $flgB$ and $frdA$. Again, $infC$ is the single gene whose functional promoter mentioned in RegulonDB or PromEC is also an intrinsically strong promoter. Eight more genes identified by our software are encountered in both repositories. Three of them have the TSS of their putative strong promoter in the close vicinity of a functional TSS ($oxyR$ (0), $clpA$ (1), $acnAP2$ (5)), or at a rather small bp distance ($otsB$ (25)). On the other hand, the bp distances between the putative strong promoter and the functional promoter identified for genes $rhaS$, $cpxR$, $mgtA$ and $zntA$ are all over 100.

To recapitulate, under CI condition, coincidence between the putative strong promoter's TSS and the functional promoter's is verified for one gene, close proximity ($\leq$ 5 bp) is verified for two genes, relative proximity ($\leq$ 25 bp) is reported for three genes. Under CII condition, coincidence is observed for two genes; close proximities are reported for four genes, as well as relative proximity. We note that Regulon DB alone contributes for 6 genes not referred to in PromEC. PromEC alone contributes for 3 genes. Moreover, when a gene is referred to by both databases, the two functional promoters' TSSs may be located at remote positions. Comparing the distances between putative strong promoters and their nearest functional promoters, we observe that 6 distances out of 12 are below 25, under CI conditions. Ten distances out of 20 are below 25 under CII conditions, including a superposition of the TSSs in two cases.

Finally, under the most stringent constraint (CI, UP element required), none of the 3 genes identified by BacTrans^2 is referred to in the two repositories devoted to functional promoters. In contrast, under the more relaxed similarity constraint CII, we identify two genes, $carA$ and $pfkB$, also mentioned in RegulonD or PromEC. The bp distance between the putative strong promoter's TSS and the functional promoter's is rather high for $pfkB$ (44). However, this distance is outstandingly small (1), for gene $carAP1$.

We now discuss the reasons likely to explain the results observed. New σ 70 models were recently compiled from 684 functional promoters listed in RegulonDB and PromEC databases (Shultzaberger *et al.*, 2007) under the form of sequence logos, corresponding to Position-Specific Scoring Matrices. One such model is provided for each possible length of the gap located between -35 and -10 boxes (in range [15-20]) (see latter reference, Figure 2). Such models are consistent with our specification of the $\sigma 70$ strong promoter in terms of bp distance constraints. Nonetheless, the specificities of these models are rather low. The sequence logos of the -35 boxes indicate that the two first nucleotides of the consensus TTGAA are more likely to be encountered simultaneously in the functional promoters than any other pair of nucleotides in the consensus; this description is compatible with constraint CII but not with constraint CI. However, in view of the sequence logos relative to the -10 box, it is looking unlikely that more than 3 nucleotides are simultaneously conserved with respect to consensus TATAAT; the least drastic condition, CII, requires that no more than 2 nucleotides differ with respect to the -10 consensus. Hence, we did constrain our strong $\sigma 70$ promoter model in a way consistent with biological reality, that is with -35 box less specific than -10 box; anyway, we constrained it a degree higher with regard to known functional promoters, which is the least expected for intrinsically strong promoters.