Supplementary material

Table S1. Nucleotide base composition and average free energy (AFE) in different regions of bacterial promoters. Genome sequences corresponding to 491, 282 and 40 TSSs, which are at least 500 nucleotides apart, are considered from *E. coli*, *B. subtilis* and *M .tuberculosis* respectively. Sequences are aligned with respect to the TSS. Standard deviation from the respective mean values is given in parenthesis. The second row in each case corresponds to AFE values for the same sequence after being shuffled 10 times. AFE values are given in kcal/mol.

Location of sequence analyzed with respect to TSS (Length of the region)		***E. coli***		***B. subtilis***		***M. tuberculosis***	
		AFE	**G+C**	**AFE**	**G+C**	**AFE**	**G+C**
Whole genome		-20.1 (2.4)	0.51	-18.9 (2.3)	0.44	-22.5 (2.1)	0.66
Upstream region (401 nt)	-500 to -100	-19.9 (1.0)	0.49 (0.06)	-18.8 (0.8)	0.43 (0.05)	-22.4 (0.7)	0.65 (0.03)
	-500 to -100 shuffled sequence	-19.6 (1.0)		-18.6 (0.8)		-22.1 (0.6)	
Downstream region (401 nt)	100 to 500	-20.1 (0.7)	0.49 (0.04)	-19.0 (0.7)	0.44 (0.04)	-22.5 (0.5)	0.66 (0.03)
	100 to 500 shuffled sequence	-19.9 (0.7)		-18.7 (0.7)		-22.3 (0.5)	
Promoter region (101 nt)	-80 to +20	-18.6 (1.3)	0.42 (0.08)	-17.1 (1.0)	0.33 (0.06)	-21.4 (1.0)	0.61 (0.05)
	-80 to +20 shuffled sequence	-18.5 (1.2)		-17.0 (0.9)		-21.4 (0.9)	

Longer region (1001 nt)	-500 to +500	-19.8 (0.7)	0.49 (0.04)	-18.6 (0.5)	0.42 (0.03)	-22.3 (0.4)	0.65 (0.02)
	-500 to +500 shuffled sequence	-19.5 (0.6)		-18.4 (0.5)		-22.1 (0.33)	

Table S2.1: Number of promoter sequences for protein coding genes when grouped according to their %GC content. The categorization is done based on the %GC over 1001nt long promoter sequences ranging from -500 to +500 w.r.t TSS) in the three bacterial systems *E. coli, B. subtilis* and *M. tuberculosis*.

	No of sequences in each group*			
%GC	***E. coli***	***B. sub***	***M. tb***	**Combined**
30 – 35	-	6	-	6
35 – 40	16	61	-	77
40 – 45	47	168	-	215
45 – 50	183	47	-	230
50 – 55	193	-	-	193
55 – 60	18	-	-	18
60 – 65	-	-	25	25
65 – 70	-	-	15	15
Total	457	282	40	779

* Only protein coding TSSs which are 500nt apart are considered.

Table. S2.2: Average free energy variation for different regions of promoter sequences, grouped according to their %GC content**.** Standard deviation from the respective mean values is given in parenthesis. AFE values are given in kcal/mol.

%GC	**Average free energy over -40 to +20 region**			**Average free energy over -80 to +20 region**			**Average free energy over -500 to +500 region**		
	E. coli	***B. sub***	***M. tb***	***E. coli***	***B. sub***	***M. tb***	***E. coli***	***B. sub***	***M. tb***
30 – 35	-	-16.5 (1.0)	-	-	-16.5 (0.8)	-	-	-17.1 (0.3)	-
35 – 40	-17.2 (1.1)	-16.8 (1.2)	-	-17.2 (0.7)	-16.7 (0.9)	-	-17.9 (0.2)	-18.0 (0.3)	-
40 – 45	-17.6 (1.5)	-17.0 (1.1)	-	-17.6 (1.1)	-17.2 (0.9)	-	-18.8 (0.3)	-18.7 (0.3)	-
45 – 50	-18.3 (1.2)	-17.4 (1.0)	-	-18.3 (1.0)	-17.7 (0.9)	-	-19.6 (0.2)	-19.4 (0.2)	-
50 – 55	-18.9 (1.3)	-	-	-19.1 (1.1)	-	-	-20.3 (0.2)	-	-
55 – 60	-19.9 (1.5)	-	-	-20.2 (1.3)	-	-	-21.0 (0.2)	-	-
60 – 65	-	-	-20.7 (1.1)	-	-	-21.2 (1.0)	-	-	-22.1 (0.2)
65 – 70	-	-	-21.2 (1.5)	-	-	-21. 9 (0. 8)	-	-	-22.6 (0.2)

Table S3: Whole genome annotation for ten *E. coli* strains. The *E. coli* K12 sub str. MG1655 (NCBI Accession: NC_000913) is the reference sequence analyzed in detail.

NCBI Accession	**NC_000913**	**AC_000091**	**NC_002655**	**NC_002795**	**NC_004431**	**NC_007946**	**NC_008253**	**NC_008563**	**NC_009800**	**NC_009801**
***E. coli* Strain**	K12 substr. MG1655	K12 substr. W3110	O157:H7 EDL933	O157:H7 sakai	CFT073	UTI189	536	APEC01	HS	E24377A
Genome size (Mbps)	4.6	4.65	5.59	5.6	5.2	5.21	4.9	5.51	4.6	5.27
%GC	50.8	50.8	50.3	50.5	50.5	50.6	50.8	50.3	50.8	50.7
Protein genes	4274	4226	5312	5230	5339	5021	4620	4428	4378	4749
RNA genes	182	157	128	141	116	110	103	114	158	114
Gene density [a]	929.13	908.82	950.27	933.93	1026.73	963.72	942.86	803.63	951.74	901.14
Number of promoter predictions (PP)	9136	9130	11254	10800	10330	9888	9826	9932	9175	9855
Number of TP w.r.t TLS of all genes [b]	3157	3088	3728	3614	3198	3231	3301	2936	3032	3238
Protein genes for which TP is identified [c]	2483 (58%)	2454 (58%)	2969 (56%)	2911 (56%)	2634 (49%)	2632 (52%)	2654 (57%)	2363 (53%)	2424 (55%)	2610 (55%)
RNA genes for which TP	127 (70%)	109 (69%)	92 (72%)	98 (70%)	68 (57%)	78 (71%)	78 (76%)	82 (72%)	70 (55%)	86 (75%)

is identified [c]										
Number of promoters identified within coding sequence (FP w.r.t TLS of all genes [b] **)**	2063 (23% of all PP)	2043 (22% Of all PP)	2568 (23% of all PP)	2394 (22% of all PP)	2985 (29% of all PP)	2443 (25% of all PP)	2233 (23% of all PP)	2337 (24% of all PP)	2089 (23% of all PP)	2228 (23% of all PP)

[a] Gene density was calculated as the number of genes per Mbps.

[b] True Positives (TP) and False Positives (FP) were calculated with respect to the TLS of each gene as described in methods

[c] A gene is counted only once even if there is more than one predicted region that satisfies the TP criteria.

Diverse training and test data set used to arrive at optimum E and D cut-off values

The initial data set has a total of 779 TSSs which are 500nt apart for protein coding genes (457, 282 and 40 TSSs from *E. coli, B. subtilis* and *M. tuberculosis* respectively are considered). Based upon the GC content of the 1001nt long promoter sequence (-500 to +500 with respect to TSS), the initial data set has been categorized at 5% GC interval. Five fold analysis has been carried out over the %GC groups (35-40, 40-45, 45-50 and 50-55 %GC) with more than 50 TSSs in each data set. Three fold analyses have been carried out for remaining %GC groups (30-35, 55-60, 60-65 and 65-70 %GC) (data not shown). Cutoff values obtained from each training set have been validated by applying to the respective test set as shown Table S4.

Table S4. Cutoff values calculated in five fold analysis with different training sets and their validation over test sets.

Training set :1 Total No. of promoter sequences : 599					
%GC		**35-40**	**40-45**	**45-50**	**50-55**
No of sequences		64	174	199	162
E		-16.81 (0.9)	-17.28 (1.0)	-18.19 (1.0)	-19.16 (1.1)
RE_{av}		-18.25 (0.5)	-18.82 (0.6)	-19.69 (0.6)	-20.19 (0.5)
D = RE_{av} -E		1.44	1.54	1.50	1.03
Cut-off	**E**	-16.8	-17.3	-18.2	-19.2
	D	1.4	1.5	1.5	1.0

Test set : 1 Total No. of promoter sequences : 150			
%GC	**35-40**	**No of sequences in each %GC category**	**16**
	40-45		**43**
	45-50		**51**
	50-55		**40**
TP		118	
FP		84	
FN	**After I cycle**	23	
	After II cycle	0	
Sensitivity		1	
Precision		0.58	

Table S4 (continued)

Training set :2 Total No. of promoter sequences : 599					
%GC		**35-40**	**40-45**	**45-50**	**50-55**
No of sequences		64	173	200	162
E		-16.85 (0.9)	-17.26 (1.0)	-18.26 (1.0)	-19.17 (1.2)
RE_{av}		-18.22 (0.5)	-18.82 (0.6)	-19.67 (0.5)	-20.18 (0.5)
D = RE_{av} -E		1.37	1.56	1.41	1.01
Cut-off	**E**	-16.8	-17.3	-18.3	-19.2
	D	1.4	1.6	1.4	1.0

Test set : 2 Total No. of promoter sequences : 150			
%GC	**35-40**	**No of sequences in each %GC category**	**16**
	40-45		**44**
	45-50		**50**
	50-55		**40**
TP		120	
FP		91	
FN	**After I cycle**	19	
	After II cycle	2	
Sensitivity		0.98	
Precision		0.57	

Table S4 (continued)

Training set :3 Total No. of promoter sequences : 599					
%GC		**35-40**	**40-45**	**45-50**	**50-55**
No of sequences		64	173	201	161
E		-16.92 (0.9)	-17.31 (1.0)	-18.29 (1.0)	-19.16 (1.1)
RE_{av}		-18.18 (0.5)	-18.85 (0.6)	-19.68 (0.6)	-20.20 (0.4)
D = RE_{av} -E		1.26	1.54	1.39	1.04
Cut-off	**E**	-16.9	-17.3	-18.3	-19.2
	D	1.3	1.5	1.4	1.0

Test set : 3 Total No. of promoter sequences : 150			
%GC	**35-40**	**No of sequences in each %GC category**	**16**
	40-45		**44**
	45-50		**49**
	50-55		**41**
TP		119	
FP		91	
FN	**After I cycle**	20	
	After II cycle	2	
Sensitivity		0.98	
Precision		0.57	

Table S4 (continued)

Training set :4 Total No. of promoter sequences : 599					
%GC		**35-40**	**40-45**	**45-50**	**50-55**
No of sequences		64	174	200	161
E		-16.87 (1.0)	-17.33 (1.0)	-18.26 (1.0)	-19.17 (1.1)
RE_{av}		-18.18 (0.5)	-18.85 (0.6)	-19.69 (0.6)	-20.17 (0.5)
D = RE_{av} -E		1.31	1.62	1.43	1.00
Cut-off	**E**	-16.9	-17.2	-18.3	-19.2
	D	1.3	1.6	1.4	1.0

Test set : 4 Total No. of promoter sequences : 150			
%GC	**35-40**	**No of sequences in each %GC category**	**16**
	40-45		**43**
	45-50		**50**
	50-55		**41**
TP		122	
FP		85	
FN	**After I cycle**	24	
	After II cycle	1	
Sensitivity		0.99	
Precision		0.59	

Table S4 (continued)

Training set :5 Total No. of promoter sequences : 600					
%GC		**35-40**	**40-45**	**45-50**	**50-55**
No of sequences		64	174	200	162
E		-16.77 (0.9)	-17.25 (1.0)	-18.26 (1.1)	-19.10 (1.1)
RE_{av}		-18.20 (0.5)	-18.81 (0.6)	-19.68 (0.6)	-20.18 (0.5)
D = RE_{av} -E		1.43	1.56	1.4	1.08
Cut-off	**E**	-16.8	-17.3	-18.3	-19.1
	D	1.4	1.6	1.4	1.1

Test set : 5 Total No. of promoter sequences : 149			
%GC	**35-40**	**No of sequences in each %GC category**	**16**
	40-45		**43**
	45-50		**50**
	50-55		**40**
TP		117	
FP		78	
FN	**After I cycle**	23	
	After II cycle	3	
Sensitivity		0.98	
Precision		0.60	

Fig. S1. Sequence logos for different classes of *E. coli* RNA gene promoter sequences (spanning a region from -100 to +50nt and aligned with respect to TLS at '0' position). (A) 86 tRNA gene promoter sequences, (B) 22 rRNA gene promoter sequences, (C) 74 other RNA gene promoter sequences. 5' end of RNA genes are referred as TLS of RNA genes in the present study and are indicated by black arrows.

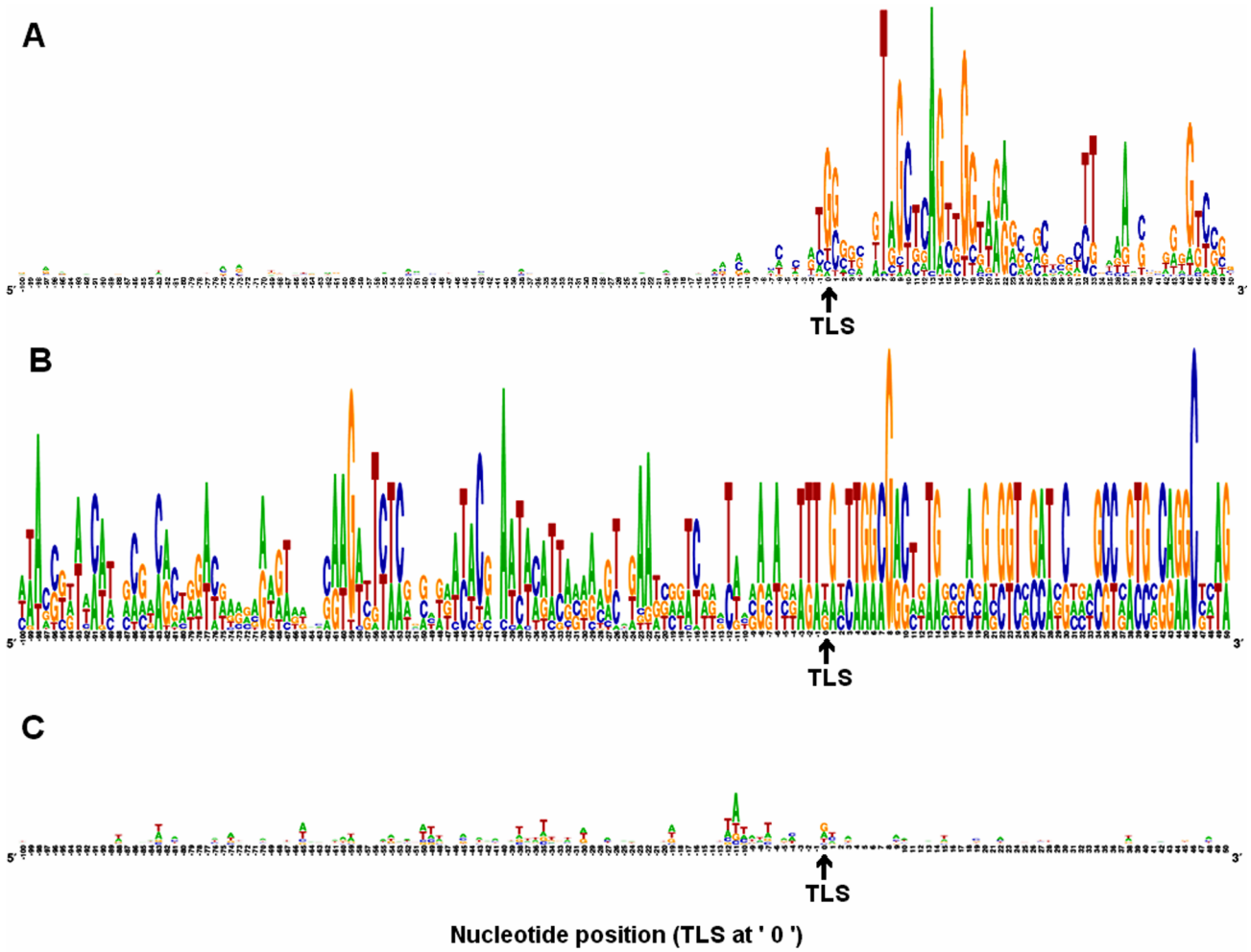

Fig. S2. Distribution of gene clusters with high false positives in PromPredict results for *E. coli* genome. The neighboring genes which have high number of predicted promoter regions within their coding regions are grouped into one cluster.

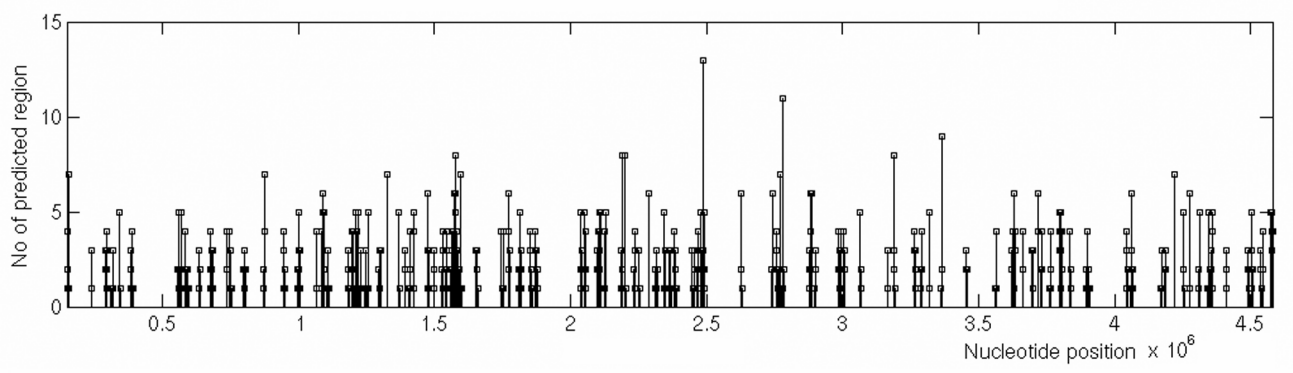

Fig. S3. Functional classification of genes with high number of predicted sites within coding region (FP) count in PromPredict results for *E. coli* genome.

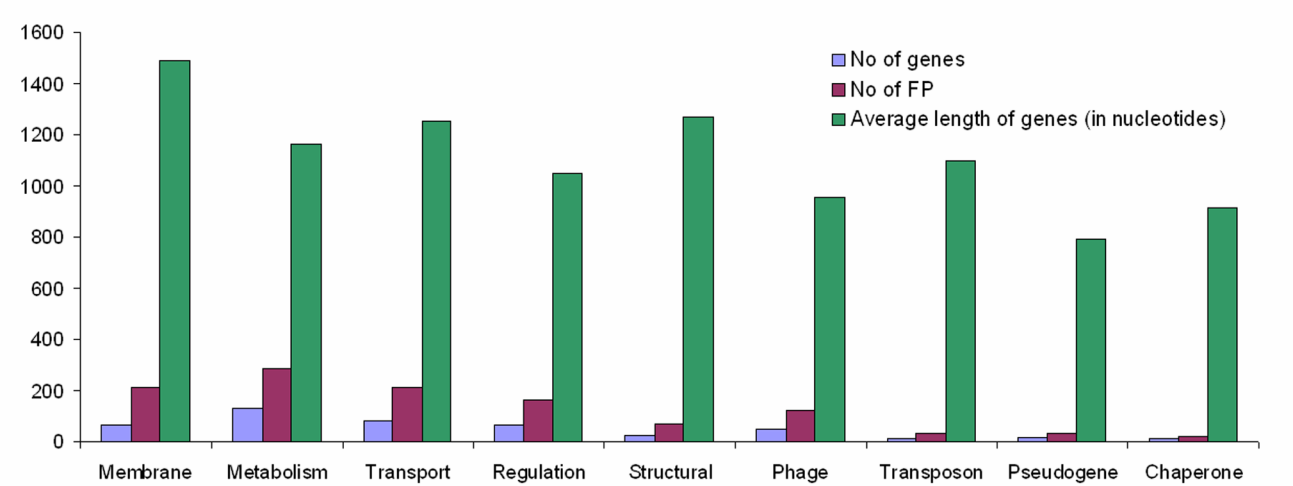

Fig. S4. Histogram showing the length distribution of predicted promoter regions in whole genome annotation of *E. coli.* All 9136 predicted regions in forward and reverse strand are considered. The length of the predicted promoter regions is indicated on X-axis.

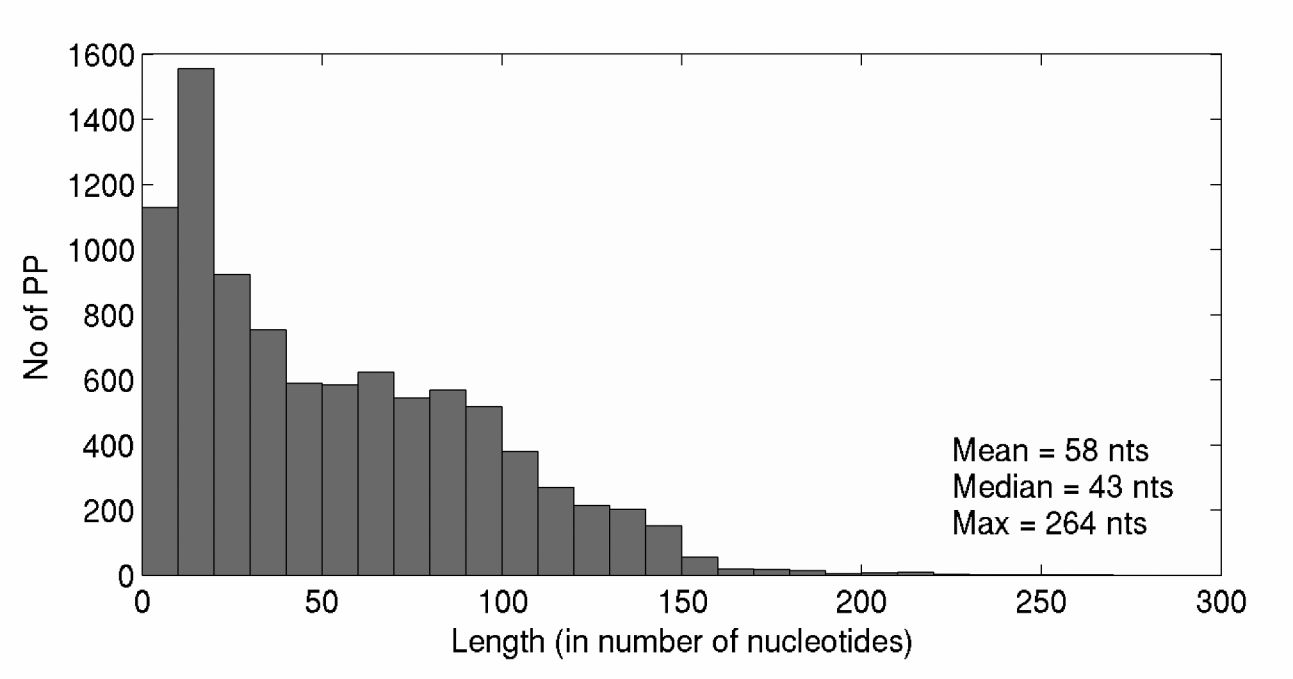

Fig. S5. Histogram showing the distance of 2026 transcription factor binding site (TFBS) relative centre position from its respective TSS (taken from Regulon DB version6.2 last updated 10th July 2008). The distance varies from -887 nts to +1848 nts with 95% of TFBS occurring within -300 to +100 region with respect to the TSSs.

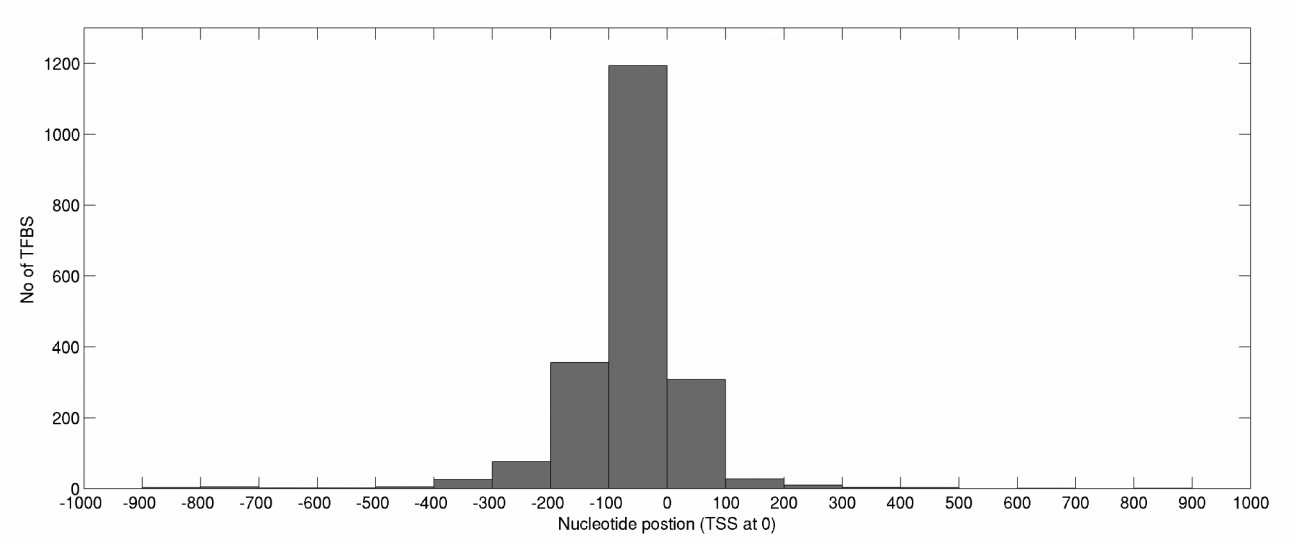

Fig. S6. Histogram showing the distance of the annotated TSS from its respective gene TLS for (A) 1145 *E. coli* TSSs (B) 615 *B. subtilis* TSSs. If a TSS is associated with more than one gene, then the gene nearest to TSS is considered. Minimum distance '0 nt' denotes that the TSS and the respective TLS coincide with each other. A negative value for distance indicates that the TSS lies within the coding region of its own gene.