Title
Comparing the fine structure of promoter regions across bacterial species

Authors
Huerta, Araceli M.
Collado-Vides, Julio
Francino, M. Pilar

Publication Date
2005-06-14

Comparing the fine structure of promoter regions across bacterial species.

Araceli M. Huerta[1*], Julio Collado-Vides[2] and M. Pilar Francino[1]

[1]Genomics Division, Lawrence Berkeley National Lab and Evolutionary Genomics Department and DOE-Joint Genome Institute.
[2]Program of Computational Genomics, CIFN-UNAM, Mexico.
*correspondence to: amhuerta@jgi.doe.gov

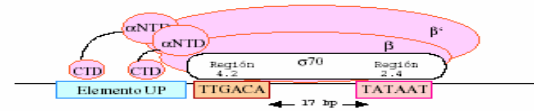
MBE JUNE 2005

ABSTRACT

The selective mechanisms operating in regulatory regions of bacterial genomes are poorly understood. In *Escherichia coli*, we have found that regulatory regions contain high densities of overlapping and probably competing promoter-like signals, in contrast to coding regions and regions located between convergently-transcribed genes. Functional promoter sites identified experimentally are often found in the subregions of highest density of signals, even when individual sites with higher binding affinity for RNA polymerase exist elsewhere within the region [Huerta & Collado-Vides 2003]. In order to explore whether this trend discovered in *E. coli* promoter regions is common in other bacterial species, we conducted similar analyses for a representative set containing 40 additional genomes belonging to different genera across all major bacterial phyla. This comparison is validated by the fact that RNAPs are evolutionarily conserved across all bacteria.

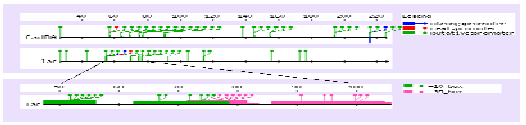
Canonical Model of Sigma 70 Promoters.

The Escherichia. coli RNAP is composed of a core complex of alpha, beta, and beta‘ subunits and one of a variety of sigma factors, the principal one being sigma70, which is capable of binding to the -10 (TATAAT) and -35 (TTGACA) promoter sequences and is essential for general transcription in exponentially growing cells.



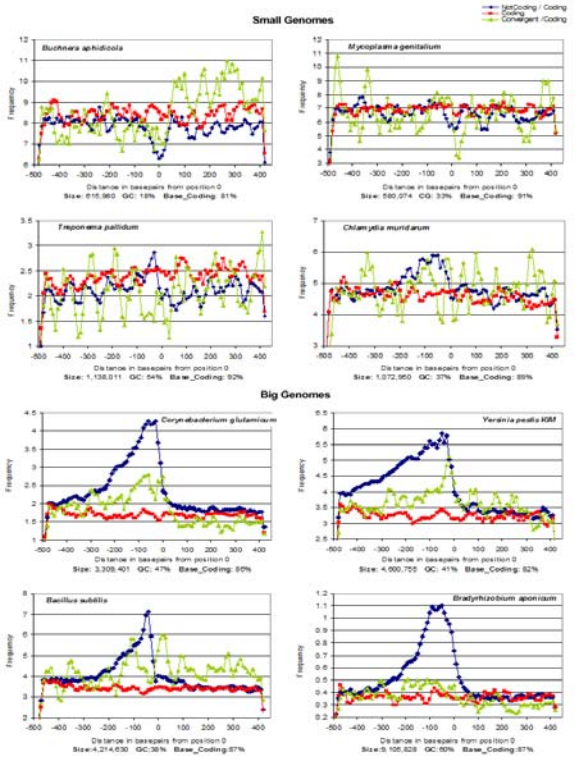
High density and clustering of promoter-like signals in *E. coli*.

The cases of *fadBA* and *lac* promoter regions. 60% of promoter regions in *E. coli* have a promoter-like signal with higher score than the mapped functional promoter (as in *fadBA*).



Density of promoter-like signals in other genomes.

The essential primary sigma factors, which are present in all known eubacteria are closely related to sigma70 of *E. coli.* Recent data suggest that there is only one primary sigma factor present in any given eubacterial species [Wösten 1998]. Of 42 analyzed representative bacterial genomes, we have found that regulatory regions contain an excess of promoter-like signals in at least 24 different cases. Only genomes with small size do not present this abundance of signals for housekeeping promoters.



These observations strongly suggest that the high density of promoter-like signals in regulatory regions of large bacterial genomes is maintained by natural selection. First, the fact that this pattern is observed in phylogenetically-distant genomes argues against mutational biases as its main source, since mutational biases are known to vary among genomes, particularly when there are large differences in GC content. Second, signal density is highly dependent on large genome size, and is completely absent from most of the small genomes of animal parasites and symbionts with an intracellular or predominantly host-restricted lifestyle. We propose that the loss of promoter signal densities in the regulatory regions of reduced genomes represents an instance of genome degradation due to high mutation rate and diminished response to selection, analogous to the accumulation of other types of moderately deleterious changes in these genomes. In agreement with this hypothesis, there is recent evidence that small genomes have in fact impoverished regulatory mechanisms [Wilcox et al. 2003]. Conversely, this implies that the pattern of high signal density in regulatory regions detected in large, free-living bacterial genomes is maintained by natural selection.

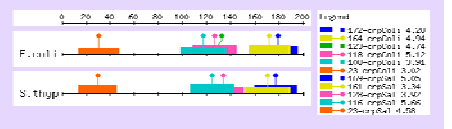
Promoter-like Signal Density in Eubacterial Genomes.

Species	Genome Size(Mb)	%GC	%Overall Similarity to E.coli	RpoD% Identity	OverRepresentation of Signals	%Genes with Clusters	Signals by Cluster	%Signals in Clusters
M_genitalium	0.58	33	28	--	no	-	-	-
B_aphidicola	0.62	18	57	74	no	-	-	-
B_aphidicola_Sg	0.64	20	61	80	no	-	-	-
Buchneria_sp	0.64	19	61	82	no	-	-	-
W_brevipalpis	0.70	15	55	73	no	-	-	-
U_urealyticum	0.75	23	28	49	no	-	-	-
M_pneumoniae	0.82	34	27	--	no	-	-	-
B_burgdorferi	0.91	23	29	31	no	-	-	-
C_muridarum	1.07	37	28	38	no	-	-	-
T_pallidum	1.14	55	27	36	no	-	-	-
C_pneumoniae_AR39	1.23	35	28	40	no	-	-	-
R_conorii	1.27	31	32	45	no	-	-	-
A_aeolicus	1.55	41	30	38	no	-	-	-
C_jejuni	1.64	25	31	38	no	-	-	-
H_pylori_26695	1.67	32	30	37	yes	100	6.38	88
H_influenzae	1.83	34	57	67	no	-	-	-
T_maritima	1.86	42	28	57	yes	88	3.63	75
S_pneumoniae_R6	2.04	34	30	65	yes	98	5.08	83
B_melitensis	2.12	50	34	49	yes	80	3.05	73
F_nucleatum	2.17	25	31	60	no	-	-	-
P_multocida	2.26	35	56	71	yes	99	5.30	85
N_meningitidis_MC58	2.27	46	42	52	yes	90	3.80	78
L_lactis	2.37	30	31	61	yes	99	5.28	85
X_fastidiosa	2.68	47	40	59	yes	83	3.48	75
A_tumefaciens_C58_Cereon	2.84	53	34	47	yes	79	2.79	67
S_aureus_Mu50	2.88	29	31	61	no	-	-	-
V_cholerae	2.96	43	54	76	yes	95	4.29	84
L_innocua	3.01	34	32	61	yes	98	5.07	83
C_perfringens	3.03	24	31	68	no	-	-	-
M_leprae	3.27	54	29	48	no	-	-	-
C_glutamicum	3.31	48	29	54	yes	89	3.89	81
Synechocystis_PCC6803	3.57	42	29	59	yes	96	4.58	81
S_meliloti	3.65	58	34	48	yes	63	2.10	65
R_solanacearum	3.72	62	40	57	yes	63	1.97	70
C_acetobutylicum	3.94	27	31	63	no	-	-	-
C_crescentus	4.02	62	32	49	yes	43	1.24	55
B_subtilis	4.21	38	32	68	yes	90	3.78	76
Y_pestis_KIM	4.60	42	70	91	yes	97	4.99	84
S_flexneri_2a	4.61	44	97	100	yes	96	4.61	82
E_coli_K12	4.64	43	100	100	yes	92	4.23	80
S_typhi	4.81	44	87	98	yes	97	4.67	84
P_aeruginosa	6.26	61	45	66	yes	60	2.07	69
Nostoc_sp	6.41	36	28	60	yes	98	5.28	84
M_loti	7.04	58	33	48	yes	55	1.74	60
B_japonicum	9.11	60	33	46	yes	58	1.89	64

Log-likelihood test was used to test the over/under representation of densities of promoter-like signals. Two main alternative profiles were obtained, as illustrated in the previous figure. Regulatory regions contain an excess of promoter-like signals in 24 genomes (P<0.001), including genera belonging to phyla distantly related to *E. coli*, such as the *Firmicutes*, the *Actinobacteria*, the *Cyanobacteria* and the *Thermotogae*. Clearly, presence of the pattern is highly dependent on genome size [Huerta et al. in preparation].

Phylogenetic Footprinting *in silico* in Enterobacteriaceae genomes.

Functional sequences. i.e. regulatory modules, are more conserved evolutionarily than nonfunctional regions. Using the ***Cover*** program [Huerta & Collado-Vides 2003] in the regulatory region of the *crp* gene of *E. coli*, we found a cluster of promoter like-signals. The orthologous sequence in *S. typhimurium* showed a similar cluster.



We have initiated a comparative study of regions containing high densities of promoter-like signals in different enteric species. We are analyzing conservation in these regions, in terms of sequence and information content, both globally and separately for subregions of different signal density, as well as for individual sites within subregions. Analysis of these patterns of conservation will shed light on the molecular and selective mechanisms that maintain this level of signal redundancy within regulatory regions.

Bibliography.

1. Wosten MM. (1998) Eubacterial sigma-factors. *FEMS Microbiol Rev.* 22(3):127-50. Review.
2. Huerta A.M. and Collado-Vides J. (2003) Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. *J.Mol.Biol.* 333, 261-278.
3. Wilcox JL, Dunbar HE, Wolfinger RD, Moran NA. Consequences of reductive evolution for gene expression in an obligate endosymbiont. Mol Microbiol. 2003 Jun;48(6):1491-500.
4. Huerta AM, Francino MP, Morett E, Collado-Vides J. Natural selection for high density of s70 promoter-like signals in the regulatory regions of large bacterial genomes. (in preparation 2005).

Acknowledgments.
We thank Dr. Enrique Morett and Jeff Froula for fruitful discussion and support.

*This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36. LBNL-57728