Comparing the Fine Structure of Promoter Regions across Enteric Bacteria.

Araceli M. Huerta[1,2], 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, Center of Genomics Sciences, UNAM, México.

WALNUT CREEK MARCH 2006

UNAM-MEXICO

ABSTRACT

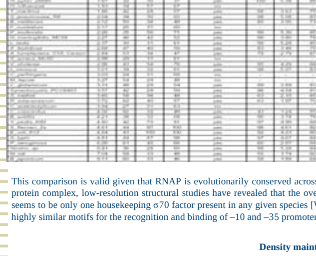
In *Escherichia coli*, we have found that regulatory regions contain high densities of overlapping and probably competing promoter-like signals, in contrast to coding regions and regions located between convergently-transcribed genes [Huerta & Collado-Vides 2003]. It has been shown that the differential densities of promoter-like signals between regulatory and nonregulatory regions are detectable in large eubacterial genomes, in contrast to genomes that have experienced severe reductive evolution, which do not show the differential pattern [Huerta et al. submitted]. Presence of this pattern mirrors that of genes and other genomic features that require weak selection to be effective in order to persist. This implies that the differential distribution of promoter-like signals between regulatory and nonregulatory regions confers a significant fitness advantage. Here we present a more detailed comparative analyses of promoter regions among enteric γ-proteobacteria. This analysis detects significant positional conservation of promoter-like signal clusters across enterics, some times in spite of strong primary sequence divergence. This suggests that the conservation of the nature and exact position of specific nucleotides is not necessarily the priority of selection for maintaining the transcriptional function in these bacteria.

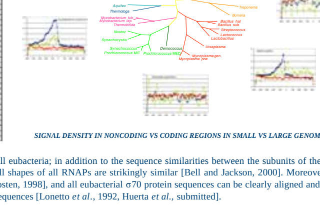
Introduction.

The *Escherichia coli* RNAP is composed of a core complex of α, β, and β' subunits and one of a variety of σ factors, the principal one being σ70, which is capable of binding to the -10 (TATAAT) and -35 (TTGACA) promoter sequences and is essential for general transcription in exponentially growing cells. We have shown that regulatory regions in *E. coli* contain high densities of σ70 promoter-like signals, in contrast to coding regions and regions located between convergently transcribed genes. Moreover, functional promoter sites identified experimentally are often found within the subregions of highest density of overlapping signals, even when individual sites with higher binding affinity for RNAP exist elsewhere within the region [Huerta and Collado-Vides, 2003].

Densities of Promoter-like Signals in Eubacteria.

The essential primary sigma factors, which are present in all known eubacteria, are closely related to σ70 of *E. coli*. Recent data suggest that there is only one primary sigma factor present in any given eubacterial species. We have found that regulatory regions contain an excess of promoter-like signals in at least 43 different representative bacterial genomes. Only genomes with small size do not present this abundance of signals for housekeeping promoters, which is in agreement with evidence that small genomes have impoverished regulatory mechanisms.





SIGNAL DENSITY IN NONCODING VS CODING REGIONS IN SMALL VS LARGE GENOMES.

This comparison is valid given that RNAP is evolutionarily conserved across all eubacteria; in addition to the sequence similarities between the subunits of the RNAP protein complex, low-resolution structural studies have revealed that the overall shapes of all RNAPs are strikingly similar [Bell and Jackson, 2000]. Moreover, there seems to be only one housekeeping σ70 factor present in any given species [Wosten, 1998], and all eubacterial σ70 protein sequences can be clearly aligned and contain highly similar motifs for the recognition and binding of –10 and –35 promoter sequences [Lonetto *et al.*, 1992, Huerta *et al.*, submitted].

Density maintained by natural selection.

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 (*Mycoplasma, Ureaplasma, Treponema, Borrelia, Campylobacter, Rickettsia, Buchnera, Wigglesworthia*).

We suggest that such density might be generated by some form of natural selection, and that the difference between regulatory and nonregulatory regions is due to the joint effects of selection for promoter-like signals in regulatory regions and selection against them in the other areas of the genome. However, it is important to note that such scenario does not imply that all promoter-like signals in regulatory regions are being maintained by selection because a specific function is required from each one of them. Several different types of selective regimes could act on regulatory regions and influence the abundance and distribution of promoter-like signals.

Models implying turnover of functional promoter sequences.

Compensatory Selection

Compensatory evolution occurs when a pair of mutations at different sites that would be singly deleterious produces normal fitness in combination and can therefore easily becomes fixed in a population [Kimura 1985].

• Promoter turnover should be facilitated by the fact that the level of sequence and structural degeneracy tolerated by functional promoter sites.

• Compensatory evolution is especially likely when the population size is large [Carter and Wagner 2002]. Given that bacteria have large populations, short generation times and high mutation rates, the neutral turnover of redundant sites should be even faster than in animal genomes.

• Potentially, the target of purifying selection for promoter function could shift often enough to preserve a large number of promoter-like sites at a given point in time. The preservation of multiple redundant promoter-like sites could be promoted by selection for genetic robustness, the existence of multiple potential promoter sites capable of initiating transcription at relatively similar rates would minimize the deleterious effect of genetic mutations on gene expression.

Stabilizing selection

With basis on their chimaeric enhancer experiments, Ludwig et al. [2000] have proposed an alternative model of regulatory evolution in eukaryotes.

• Because promoter seem to contain multiple binding sites, and can accommodate sequence and structural variability, many independent mutations could actually contribute to variation in gene expression [Ludwig et al. 2000; Ludwig 2002].

•Stabilizing selection would then have to act regulatory regions to maintain stable levels of gene expression within boundaries not affecting fitness.

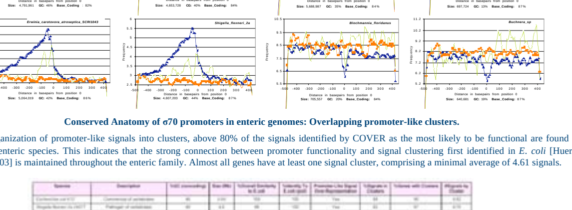
•When the number of sites affecting a quantitative trait is large, the average selection coefficient per mutant site under stabilizing selection is small, and the rate of substitution can be high [Kimura 1981].

•Therefore, stabilizing selection on promoters could maintain phenotypic constancy while allowing mutational turnover of functionally important sites. The stabilizing selection could preserve a substantial level of hidden genetic variation capable of affecting gene expression in bacterial populations. Such variation would permit rapid evolution of subtle changes in gene expression and could contribute substantially to the impressive ability of bacteria to adapt swiftly and precisely to environmental change.

Evolutionary Genomics Department of the Joint Genome Institute. Poster #
This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawe nce Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36. And under auspicies of C.O.N.A.C.y.T. (Mexico) under contracts 37023-N & 0028, and UNAM Mexico.

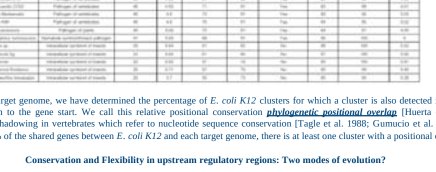
Densities of Promoter-like Signals in Enteric Bacteria.

In order to investigate the factors that maintain a high density of promoter-like signals in bacterial regulatory regions, we are undertaking detailed structural analyses of homologous regulatory regions across the Enterobacteria [Huerta et al. 2006]. The availability of numerous complete genomes for enteric species and strains with different degrees of relatedness enables the use of comparative analyses to asses the conservation of the regulatory features defined in *E. coli*, and to reveal the mode of evolution of regulatory sequences within this group. Most enteric species are inhabitants of the mammalian intestine, but the genomes analyzed here encompass a wide variety of bacterial lifestyles. The species analyzed include commensals or pathogens of vertebrate animals, one pathogen of plants, intracellular symbionts of insects and *Photorhabdus luminiscens*, which has a complex lifestyle involving pathogenic and symbiotic relationships with two different invertebrate hosts [Joyce, Watson and Clarke 2005]. Overall, the enterics represent an organismal radiation that started at least 200 MYA [Clark, Moran and Baumann 1999].



Conserved Anatomy of σ70 promoters in enteric genomes: Overlapping promoter-like clusters.

Regarding the organization of promoter-like signals into clusters, above 80% of the signals identified by COVER as the most likely to be functional are found within clusters in every enteric species. This indicates that the strong connection between promoter functionality and signal clustering first identified in *E. coli* [Huerta and Collado-Vides, 2003] is maintained throughout the enteric family. Almost all genes have at least one signal cluster, comprising a minimal average of 4.61 signals.

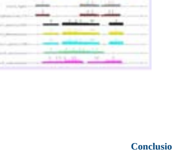


In comparisons against every target genome, we have determined the percentage of *E. coli K12* clusters for which a cluster is also detected in the second genome at an overlapping position in relation to the gene start. We call this relative positional conservation ***phylogenetic positional overlap*** [Huerta et al. 2006], in contrast to phylogenetic footprinting and shadowing in vertebrates which refer to nucleotide sequence conservation [Tagle et al. 1988; Gumucio et al. 1993; Aparicio et al. 1995; Sumiyama et al. 2001]. For 90% of the shared genes between *E. coli K12* and each target genome, there is at least one cluster with a positional overlap.

Conservation and Flexibility in upstream regulatory regions: Two modes of evolution?

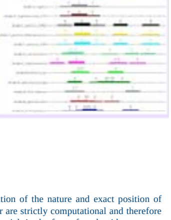
The CRP protein is a transcription factor that provides versatility in mediating the effects of environmental and metabolic signals on transcriptional regulation. The best-characterized promoter of crp, crp1, is located 167 bp away from the gene start.Experimental evidence indicates that an additional promoter, crp2, is located downstream of crp1 and can promote transcription initiation at three closely spaced sites, located 73, 79 and 80 bp downstream from the crp1initiation site.

The experimentally mapped locations of –10 and –35 boxes for crp1 in *E. coli K12* do not correspond exactly with the motif positions of the predicted promoters but the general position of crp1 overlaps that of two promoter-like signals. This positional overlap occurs in spite of strong sequence divergence between the *E.coli* clade and the other species, including numerous indels which obstruct nucleotide sequence alignment. In terms of rate and mode of evolution, the crp region appears to be more variable among enteric species, both in terms of the overall feature landscape and in the conservation of the experimentally mapped promoters.



The dnaQ gene encodes the epsilon subunit of DNA polymerase III, which performs 3' to 5' exonucleolytic proofreading activity. This proofreading function is essential and dnaQ is present across eubacteria. dnaQ has been experimentally shown to be regulated by two promoters, with +1 sites positioned at –50 (P1) and –133 (P2) base pairs from the gene start. At low RNAP concentrations, the downstream promoter (P1) is utilized preferentially, but the upstream promoter (P2) is utilized as well when RNAP concentration increases.

Inspection of the sequence alignments for this region indicates that the bases overlapping E.coli K12 P1 in the endosymbionts are poorly conserved, implying that promoter-signal positional overlap has been maintained in face of high sequence divergence. This suggests that the identification
of phylogenetic positional overlap can be employed to detect probable functional conservation among regions undergoing high rates of evolution.



Conclusions.

Our preliminary analyses of structural and organizational evolution in bacterial regulatory regions suggest that the conservation of the nature and exact position of specific nucleotides is not necessarily the priority of selection for maintaining regulatory function. Although our analyses so far are strictly computational and therefore can't establish functional conservation, the positional conservation of similar promoter-like signal clusters with regulatory potential, in the face of nucleotide sequence divergence, suggests that function could be selectively maintained without requiring the presence of specific nucleotide sequences. This implies that phylogenetic positional overlap should be a good technique for detecting probable functional conservation among regions undergoing high rates of evolution, and even to identify homoplastic functional similarity among regulatory regions having independent evolutionary origins.

Bibliography.

- Aparicio, S., A., Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf, and S. Brenner. 1995. Detecting Conserved Regulatory Elements with the model genome of the Japanese Puffer Fish, *Fugu rubripes*. PNAS. USA 92:1684-1688.
- Bell, S. D., and Jackson, S. P. (2000). Charting a course through RNA polymerase. Nat Struct Biol 7.
- Carter, A. J., and G. P. Wagner. 2002. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc Biol Sci. 269:953-960
- Clark, M. A., N. A. Moran, and P. Baumann. 1999. Sequence evolution in bacterial endosymbionts having extreme base compositions. Mol. Biol. Evol. 16:1586-1598.
- Gumucio, D. L., D. A. Shelton, W. J. Bailey, J. L. Slightom, and M. Goodman. 1993. Phylogenetic footprinting reveals unexpected complexity in trans-factor binding upstream from the epsilon globin gene. PNAS USA 90:6018-6022.
- 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.
- Huerta A.M, Collado-Vides J. and Francino M.P. (2006) Positional Conservation of Clusters of Overlapping Promoter-Like Sequences in Enterobacterial Genomes. Mol. Biol. Evol. E (accepted for publication)
- Huerta A.M, Francino M.P., Morett E., and Collado-Vides J. Selection for Unequal Densities of Promoter-like Signals in Different Regions of Large Bacterial Genomes. (submitted)
- Joyce, S.A., R. J. Watson, and D. H. Clarke. (2005) The regulation of pathogenicity and mutualism in Photorhabdus. Curr. Opin. Microbiol. 9:1-6.
- Kimura, M. 1981. Possibility of extensive neutral evolution under stabilizing selection with special reference to non-random usage of synonymous codons. Proc. Natl. Acad. Sci. USA 78:5773-5777.
- Kimura, M. 1985. The role of compensatory neutral mutations in molecular evolution. J. Genet. 64:7-19.
- Lonetto, M., Gribskov, M., and Gross, C. A. (1992). The sigma 70 family: sequence conservation and evolutionary relationships. *J Bacteriol* **174**.
- Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564-567.
- Ludwig, M.Z., 2002 Functional evolution of non-coding DNA. Curr. Opin. Genet. Dev. 12:634-639.
- Sumiyama, K., C. B. Kim, and F. Ruddle. 2001. An efficient cis-element discovery method using multiple sequence comparisons based on evolutionary relationships. Genomics 71:260-262.
- Tagle, D.A., B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago craussicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. 203:439-455.
- Wosten, M. M. (1998). Eubacterial sigma-factors. FEMS Microbiol Rev 22.

Acknowledgments.
We thank Jeff L. Froula, Victor Del Moral, and Cesar Bonavides for fruitful discussion and computational support.