Title

Selection for Unequal Densities of Sigma70 Promoter-like Signals in Different Regions of Large Bacterial Genomes

Permalink

https://escholarship.org/uc/item/3mv7j8qq

Authors

Huerta, Araceli M.
Francino, M. Pilar
Morett, Enrique
et al.

Publication Date

2006-03-01

Selection for Unequal Densities of Sigma70 Promoter-like Signals in Different Regions of Large Bacterial Genomes.

Running head: unequal promoter signal densities

Araceli M. Huerta[1*], M. Pilar Francino[1], Enrique Morett[2], and Julio Collado-Vides[3].

[1]Evolutionary Genomics Department, DOE Joint Genome Institute
and Genomics Division, Lawrence Berkeley National Laboratory.
Walnut Creek, CA 94598; USA.

[2]Department of Cell Engineering and Biocatalysis
Institute of Biotechnology, UNAM
Cuernavaca, Morelos 62210; Mexico

[3]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México
Cuernavaca, Morelos 62100; México

*Correspondence:
Araceli M. Huerta
Evolutionary Genomics Department, DOE Joint Genome Institute
2800 Mitchell Dr
Walnut Creek, CA 94598, USA.
E-mail: amhuerta@ccg.unam.mx, ahuertamoreno@lbl.gov Telephone: (925) 296 5872,
Fax: (925) 296 5666.

Abstract

The evolutionary processes operating in the DNA regions that participate in the regulation of gene expression are poorly understood. In *Escherichia coli,* we have established a sequence pattern that distinguishes regulatory from nonregulatory regions. The density of promoter-like sequences, that are recognizable by RNA polymerase and may function as potential promoters, is high within regulatory regions, in contrast to coding regions and regions located between convergently-transcribed genes. Moreover, functional promoter sites identified experimentally are often found in the subregions of highest density of promoter-like signals, even when individual sites with higher binding affinity for RNA polymerase exist elsewhere within the regulatory region. In order to investigate the generality of this pattern, we have used position weight matrices describing the –35 and –10 promoter boxes of *E. coli* to search for these motifs in 43 additional genomes belonging to most established bacterial phyla, after specific calibration of the matrices according to the base composition of the noncoding regions of each genome. We have found that all bacterial species analyzed contain similar promoter-like motifs, and that, in most cases, these motifs follow the same genomic distribution observed in *E. coli*. Differential densities between regulatory and nonregulatory regions are detectable in most bacterial genomes, with the exception of those that have experienced evolutionary extreme genome reduction. Thus, the phylogenetic distribution of this pattern mirrors that of genes and other genomic features that require weak selection to be effective in order to persist. On this basis, we suggest that the loss of differential densities in the reduced genomes of host-restricted pathogens and symbionts is the outcome of a process of genome degradation resulting from the decreased efficiency of purifying selection in highly structured small populations. This implies that the differential distribution of promoter-like signals between regulatory and nonregulatory regions detected in large bacterial genomes confers a significant, although small, fitness advantage. This study paves the way for further identification of the specific types of selective constraints that affect the organization of regulatory regions and the overall distribution of promoter-like signals through more detailed comparative analyses among closely-related bacterial genomes.

Introduction

For both prokaryotes and eukaryotes, understanding of the organizational structure and mode of evolution of regulatory DNA sequences is incomplete. Although in the Bacteria gene regulation involves fewer proteins and cis-regulatory DNA sites, and less complex interactions among them, recent findings suggest that the classical description of bacterial promoter regions may have been significantly oversimplified [1]. In *Escherichia coli,* RNA polymerase (RNAP) is composed of a core complex of α, β, β' and ω subunits and one of a variety of σ factors, the primary one being σ^{70}, which is essential for general transcription in exponentially growing cells. The canonical model of the σ^{70} promoter is defined as a simple pair of hexamers, positioned at –35 and –10 bp from the transcription start (+1), with respective consensus sequences TTGACA and TATAAT, and separated by a spacer of 15 to 21 bp [2]. σ^{70} can recognize and bind –35 and –10 motifs that differ substantially from their consensus sequences, although mutations that bring these motifs closer to the consensi generally increase promoter strength [3]. On average, *E. coli* promoters preserve only 8 of the 12 canonical bases of the –35 and –10 hexamers [4,5]. It has been recently shown that most of the regulatory regions in *E. coli* do not contain a single promoter sequence, but rather display high densities of potential σ^{70} binding sites, forming clusters of overlapping promoter-like signals. In contrast, such signal densities are not detected in coding regions and regions located between convergently-transcribed genes [1]. 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 [1].

Even though the degeneracy of the σ^{70}–binding promoter motifs ensures that new sites can evolve (i.e., appear and become fixed in populations) via local point mutation on short timescales [6], random fixation of mutations by neutral drift could not explain the different promoter-like signal densities between regulatory and nonregulatory regions of *E. coli* [1]. Moreover, it has been shown that natural selection acts to remove spurious occurrences of the two consensus words of the σ^{70} promoter (TTGACA & TATAAT) from both coding and noncoding regions in several eubacterial genomes, implying that it is disadvantageous to maintain misplaced sites which can strongly bind σ^{70} and interfere with proper gene expression [7]. This suggests that the observed excess of promoter-like signals in regulatory regions is likely to be the result of natural selection for some past or present function.

Here we report that differential densities of promoter-like signals between regulatory and nonregulatory regions are detectable in most bacterial genomes, with the exception of those that have experienced severe size reduction. We argue that the phylogenetic distribution of this differential density pattern implies that this genomic feature is maintained by weak natural selection, and we discuss possible functional roles for the high redundancy of promoter-like signals in the regulatory regions of large bacterial genomes.

Results and Discussion

In order to explore whether the differential promoter signal density pattern discovered in *E. coli* is common to other bacterial species, we conducted similar analyses for a representative set containing 43 additional genomes belonging to different genera across all major bacterial phyla. This comparison is valid given that RNAP is evolutionarily conserved across Bacteria. Moreover, there seems to be only one housekeeping σ^{70} factor present in any given species [8, 9], and all eubacterial σ^{70} protein sequences can be clearly aligned and contain highly similar motifs for the recognition and binding of –10 and –35 promoter sequences [10] (Figure 1). This implies that the DNA sequences of promoter motifs in these organisms must also be similar to those found in *E. coli*. Therefore, we used in our searches position weight matrices (PWM) describing the –10 and –35 sequences determined in *E. coli* [1] and specifically calibrated to the base composition of strictly noncoding regions of each analyzed genome (Figure 2).

Table 1 reports for every genome the consensus and average score of the detected –10 and –35 motifs. The consensi were obtained after applying the COVER Function [1] on the set of promoter-like signals found by the PWM search in the functional regulatory regions of each genome (see methods section). From table 1, it can be seen that all the additional 43 species analyzed have consensus sequences for both motifs highly similar to those of *E. coli*, as well as average scores of comparable magnitudes. In fact, large GC-rich genomes (*Pseudomonas*, *Ralstonia* and most of the alpha-proteobacteria) display motif scores above those of *E. coli*, probably due to the greater compositional difference between the AT-rich motifs and the background genomic sequence. In contrast, small GC-poor genomes have lower motif scores, with the insect endosymbiont *Wigglesworthia* displaying the lowest.

Table 2 shows the promoter-like signal density patterns detected for all the genomes analyzed. Two main alternative profiles were obtained, as illustrated in Figure 3 (profiles for all species analyzed are available in Figure S1 at http://www.ccg.unam.mx/Computational_Genomics/PromoterTools/Supplemental06/prof.html). Regulatory and nonregulatory regions contain differential densities of promoter-like signals in 24 genomes, including genera belonging to phyla distantly related to *E. coli*, such as the Firmicutes, the Actinobacteria, the Cyanobacteria and the Thermotogae. Clearly, the presence of the pattern is highly dependent on genome size. All genomes above 4 Mb display marked differences in promoter-like signals between regulatory and nonregulatory regions, whereas none of the genomes under 1.5 Mb do so. Among the genomes of intermediate size, the pattern is detectable in 65% of the cases. There is also a notable effect of GC content. Although the differential signal densities can be seen in genomes of very different GC contents (from 30% GC *Lactococcus lactis* to 62% GC *Ralstonia solanacearum* and *Caulobacter crescentus*), overall, 75% of GC-rich genomes display the differential pattern, in contrast to 53% of those under 50% GC.

The observations here reported strongly suggest that the differential density of promoter-like signals in regulatory and nonregulatory 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 being its main source, since mutational biases are known to vary among genomes, particularly when there are large differences in GC content. Second, differential 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* and *Wigglesworthia*).

Degradation of Regulatory Functions in the Small Genomes of Host-Restricted Bacteria

Acquisition of such obligate dependence on a host is known to have many deleterious consequences, collectively known as genome degradation or genome reduction [12]. Reduced genomes have independently evolved many times within several bacterial phyla, repeatedly undergoing rapid sequence evolution and acquiring extremely low GC content, often with clearly maladaptive consequences, including accumulation of deleterious amino acid substitutions and loss of adaptive codon biases [13,14]. Typical changes also include a large increase in the frequency of mobile elements in the early phases of genome degradation, chromosomal rearrangements mediated by recombination among these elements, pseudogene formation, and deletions of varying size. There is recent evidence that genome degradation also affects gene regulation, due to losses of certain promoter sequences, specialized σ factors and regulatory proteins [15-17]. These common changes likely reflect a diminished capacity of reduced genomes to respond to natural selection, conducing to the accumulation of all types of moderately deleterious mutations that would normally be purged from the genomes of free-living bacteria. This lowered effectiveness of purifying selection is due to the subdivided population structure imposed by confinement within a host, which limits the effective size of the bacterial population by subjecting it to recurrent bottlenecks and by thwarting opportunities for recombination with close relatives [18]. In addition, different types of molecular evolutionary analyses indicate that the rate of generation of point mutations is accelerated in reduced genomes [19, 20].

We argue that a decrease in the efficiency of purifying selection in host-restricted bacteria allows promoter-like signals to rapidly accumulate all along the genomic sequence of these organisms, causing the loss of differential signal density patterns. This interpretation is based on the following evidence: (i) simulation results demonstrating that transcription factor binding sites can rapidly appear as a consequence of local point mutations on short timescales without invoking selection [6]; (ii) the observation that natural selection can act to remove spurious transcription factor binding sites from nonregulatory regions in many bacterial genomes, with a weak strength similar to that of selection on adaptive codon bias [7]; (iii) empirical and theoretical results demonstrating that the effectiveness of selection is diminished in the small, highly-structured populations of host-restricted bacteria [12-18]; (iv) the finding that promoter-like sequences in these organisms show a certain degree of degradation reflected in their decreased scores for the –10 and –35 motifs (table 2), suggesting an overall reduction of the efficiency of selection on regulatory function; and (v) the fact that the nonregulatory regions of host-restricted bacteria present a frequency of promoter like-signals similar to the highest frequency peak found inside the regulatory regions of their commensal or free-living relatives (Fig. 4). Figure 4 shows the frequency of promoter-like signals in *Buchnera aphidicola, Mycoplasma genitalium* and their relatives *Escherichia coli and Bacillus subtilis*. The increased frequency of promoter-like signals could have regulatory implications by making it more difficult for RNAP to correctly identify regulatory regions and the functional promoters within them. In that case, one could expect a slow response to changes in cellular conditions and/or a decreased level of gene expression.

To further test whether genome degradation produces loss of differential signal densities, we decided to compare signal density patterns between a genome that has recently entered a process of degradation and close relatives evolving under stronger purifying selection. Our initial analyses pinpointed *Mycobacterium leprae,* the obligate intracellular pathogen that causes leprosy, as one of the rare genomes above three Mb for which differential promoter signal densities are not detected. However, although the genome of *M. leprae* remains relatively large (3.27 Mb) and GC-rich (58% overall), it is clearly decreasing in size and GC content relative to its closest relative, *M. tuberculosis* (4.41 Mb and 65.6% overall GC). Genome comparisons between different mycobacterial species indicate that *M. leprae* has also undergone massive gene decay by pseudogenization and deletion, as well as numerous genomic rearrangements likely due to the proliferation of IS elements [21]. As we expected, analyses of promoter-like signals in two strains of *M. tuberculosis* do reveal the differential signal densities characteristic of large genomes (Figure 4). Both *M. tuberculosis* strains display higher average scores than *M. leprae* for the –35 and –10 promoter motifs (Table 3), supporting the idea that the leprosy bacillus is undergoing a general degradation of its regulatory sequences.

Potential Functions of the Promoter-like Signals.

In the course of our analyses, we identify two types of promoter-like signal densities: a global density and a local density [Figure 4 and 7 in reference 1]. The global density is obtained during the search for the -10 and -35 motifs using position weight matrices. In *E. coli,* this search produces an average of 38 signals per 250 bp regulatory region, which may be distributed evenly across the sequence or present different levels of overlap. The second type of density, the local density, results from applying the COVER function on the set of 38 signals; it produces 4.7 signals in average, most of which exist as a series of overlapping potentially competing RNA polymerase binding sites. The average size of these clusters of overlapped signals is 42 bp. 74% of the *E. coli* experimentally mapped promoters are embedded in this kind of clusters [1]. We suggest that each of these types of densities may be maintained by natural selection for very different reasons.

Global density of promoter-like signals in regulatory regions.

It is most likely that the global density is largely built by promoter-like signals that do not substitute the function of the primary promoter which has to respond to a given cellular condition; this would be the case if the majority of detected signals could bind RNAP and form a closed complex but were not able to proceed with the subsequent steps required to initiate transcription. However, a subset of these promoter-like signals could be just a single point mutation away from being able to operate as active transcription initiation sites. These could be called cryptic promoters, in analogy to cryptic genes that can be activated by single mutations. Some cryptic promoters could be relics of ancient promoters, indicating the high frequency of changes in regulatory regions. This high frequency correlates with the observed high flexibility in the evolution of transcriptional regulators in bacteria [22].

The permanence of cryptic promoters in the regulatory regions of bacteria could be facilitated by different kinds of evolutionary processes. First, they could constitute a collection of “back up” promoter sequences, maintained by selection for robustness. Bacteria having redundant signals capable of acquiring functionality with a single base change would increase in frequency in the population, because a fraction of their descendent cells would carry such activating mutations and would be able to survive in the advent of a deleterious mutation that destroyed the main promoter. In other words, the existence of multiple potential promoters would minimize the deleterious effects of genetic mutations on gene expression. In addition, for bacterial species prone to encounter a variety of environments, the existence of cryptic promoters of different strength could also allow for rapid evolution of changes in gene expression allowing adaptation to environmental changes. Taking into account that for sigma 70 transcription, the precise positioning of the regulatory proteins strongly determines their positive or negative role [23], this large availability of promoters-like sequences provides a fertile ground for quick changes in the role of regulatory proteins. Those changes have been proposed to depend on the demand of gene expression in a model where selection governs changes in gene regulation [24].

It is also possible that some of the signals detected in regulatory regions are not cryptic, but rather fully functional promoters that are utilized only in special conditions. Although, in general, the site of transcription initiation is known to be rather precise, 25% of the reported regulatory regions in *E. coli* are known to harbor multiple functional promoters, three on average [1, 25]. The simultaneous availability of alternative promoters for a given gene would provide plasticity of gene expression in response to different conditions regularly encountered by the bacterial cell.

The regulatory region of the *E. coli lac* operon exhibits signals of both types. Six promoter sequences have been experimentally detected in close proximity to the primary *lac* promoter (*lacP1*). Four signals were created via single base pair mutations in the wild sequence, and two were detected *in vivo* as weak promoters that function when the primary promoter is impaired and its activator protein is absent [26,27].

Finally, the global density of promoter-like motifs in the regulatory regions could be bringing the polymerase near to the promoter during the random DNA search prior to forming closed complexes, by attracting the enzyme to the general vicinity of the functional promoter. However, kinetic studies suggest that this might only result in a minor increase of the rate of formation of closed complexes (Jay D. Gralla personal communication).

Local density of overlapping promoter-like signals.

Overlapping promoter-like signals could play a regulatory role through functional interaction with the true promoter sequence, and their effect on regulation could be negative or positive.

Overlapping signals could negatively affect regulation by:

1) Competition. The overlapping promoter-like signals might play a negative role if their interaction with RNAP were competitive. When two or more promoters are in close proximity, the potential exists for competition between them for the binding of RNAP [28]. Regulatory proteins play an important role in helping RNAP to choose the functional promoter sequence according to specific conditions. The regulation of the *gal* gene is an example of the effect of regulatory proteins on the positioning of RNAP between two competing promoters [29].

2) Pause induction. The promoter-like signals could also induce pauses in the early steps of elongation when σ^{70} is still bound to core RNAP. The strongest evidence indicating that σ^{70} can play a functional role during elongation comes from studies of the bacteriophage λ PR' promoter and the lacUV5 promoter. Biochemical experiments have shown that σ^{70}-dependent pause occurs during early elongation in these promoters after RNAP has escaped from the promoter and synthesized a 16- or 17-nt transcript. This pause is mediated by protein-DNA interaction between σ^{70} and a DNA sequence element in the initially transcribed region that resembles a promoter -10 element [30, 31].

Positive effects of overlapping signals may include:

1) Alternate transcription. The promoter-like signals could be noncompetitive weak promoters that, in the absence of activation of the primary promoter, produce basal transcription of downstream genes. For example, transcription units that encode their own regulator would require constitutive levels of basal transcription.

2) Repositioning. The overlapping promoter-like signals might also play a positive role by collecting RNAP molecules which could then be channeled to the primary promoter sequence [32].

Conclusion.

Clearly, the distribution of the differential pattern of promoter-like signal densities among bacterial genomes mirrors that of genes and other genomic features that require weak selection to be effective in order to persist. This implies that the differential density of promoter-like signals between regulatory and nonregulatory regions confers some small but significant fitness advantage. Therefore, the outcome of gene regulation can be affected by factors much beyond the sequence of a single pair of RNAP binding sites in a given regulatory region, including the general abundance and organization of promoter-like signals in the region, as well as the presence of signals in the non-regulatory portions of the genome. Identification of the specific types of selective constraints that shape the number, position and arrangement of promoter-like signals across the different genomic regions of large bacterial genomes will require further comparative analyses among closely-related bacteria.

Materials and Methods

The promoter model we adopted for our searches was Matrix_18_15_13_2_1.5 (Figure 2), defined through a thorough evaluation of more than 200 *E. coli* matrix pairs that optimized different criteria [1]. This model contains the canonical consensus sequences for both the –10 and –35 motifs, and outperformed all others according to measures of sensitivity, specificity, precision and accuracy [1].

The strategy to find promoters-like signals in other genomes involved several steps:

(i) The base composition of the strictly noncoding regions of the genome to be analyzed was obtained and used to define the *a priori* probabilities of each base.

(ii) The frequency matrices obtained from the *E. coli* genome for the –10 and –35 boxes were calibrated with the *a priori* probabilities of the analyzed genome using the PATSER program [33, 34]:

$$W(b,l) = f(b,l)\log_2 \frac{f(b,l)}{p(b)}$$

where *f(b,l)* is the relative probability of the base *b* at the position *l* of the input *E. coli* matrix and *p(b)* is the *a priori* probability of base *b* in the analyzed genome (calculated in (*i*)).

(iii) In order to define which motifs would be considered significant promoter-like signals in the different target genomes, we determined minimal cutoff scores that would retain 98% of the original motifs from functional σ^{70} promoters that were used in generating the *E. coli* frequency matrices. With this aim, the original *E. coli* motifs were rescored according to the *a priori* probabilities by means of PATSER and the statistical distribution (mean and standard deviation) of *E. coli* motif scores in that genome was obtained. For each of the genomes, the score, I_{seq}, of a motif of size *L*, according to the corresponding *a priori* probabilities calculated in *(i)*, was obtained as:

$$I_{seq} = \sum_{l=1}^{L} f(b,l)\ \lg \frac{\mathrm{f}(b,l)\big/ n}{p(b)}$$

where *n* is the number of sequences in the input *E. coli* matrix alignment.

(iv) To conduct the density analysis, three kinds of regions were defined in each genome according to NCBI annotations: coding, convergent, and strictly noncoding (which excludes the convergent regions). Convergent regions are analyzed separately because they are not expected to contain any functional promoters. Using PATSER, we searched each genomic region for -10 and -35 motifs with the corresponding calibrated matrices, and retained the motifs that scored above the respective cutoff.

(v) An analysis of under/over-representation using a log-likelihood statistic was done to determine if the genome had an excess density of promoter-like signals in potentially regulatory regions (the strictly non-coding

regions) when compared against coding and noncoding regions between convergent genes.

(vi) For genomes with significant over-representation of promoter-like signals in regulatory vs. convergent noncoding regions ($p < 0.001$), we estimated the most likely functional promoters in the genome. To this aim, we predicted transcriptional units (single genes and operons) with the method of Moreno-Hagelsieb and Collado-Vides [11], which relies on the distribution of distances between genes in a given genome. The set of 250 bp sequences upstream of the first gene of every transcription unit is likely to constitute the smallest set of regulatory regions required for the expression of all genes in a genome, and should contain the highest proportion of true functional promoters. We applied the COVER function on the collections of PATSER motifs from this set of regulatory sequences having scores above the cutoff value. COVER has been designed to select the most likely functional promoters from a conglomerate of promoter-like signals by means of a "divide and conquer" strategy based on partial sorting [1]. The resulting COVER-predicted promoters in these regulatory regions were taken as the most likely functional promoters in that genome.

Acknowledgements

The authors thank Victor del Moral and Cesar Bonavides for computational support, and Gabriel Moreno-Hagelsieb, Jacques van Helden and Jeff L. Froula for critical discussions and advice. A.M.H. has been supported by a Ph.D. grant from CONACyT (Mexico) and postdoctoral support from the Department of Evolutionary Genomics, DOE JGI. E. M. acknowledges support from CONACYT, grant 37023-N. This work has been supported by grant 0028 from CONACyT and by grant GM62205 from NIH to J. C-V. The latest stages of this work were 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 Berkeley National Laboratory under contract No. DE-AC03-76SF00098.

References

1. Huerta AM, 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.
2. Gruber TM, Gross CA (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. Annu Rev Microbiol 57: 441-466.
3. Hawley DK, McClure WR (1983) Compilation and analysis of Escherichia coli promoter DNA sequences. Nucleic Acids Res 11: 2237-2255.
4. Lisser S, Margalit H (1993) Compilation of E. coli mRNA promoter sequences. Nucleic Acids Res 21: 1507-1516.
5. Ozoline ON, Deev AA, Arkhipova MV (1997) Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by Escherichia coli RNA polymerase. Nucleic Acids Res 25: 4703-4709.
6. Stone JR, Wray GA (2001) Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18: 1764-1770.
7. Hahn MW, Stajich JE, Wray GA (2003) The effects of selection against spurious transcription factor binding sites. Mol Biol Evol 20: 901-906.
8. Wosten MM (1998) Eubacterial sigma-factors. FEMS Microbiol Rev 22: 127-150.
9. Mittenhuber G. (2002) An inventory of genes encoding RNA polymerase sigma factors in 31 completely sequenced eubacterial genomes. J Mol Microbiol Biotechnol 4: 77-91.
10. Lonetto M, Gribskov M, Gross CA (1992) The sigma 70 family: sequence conservation and evolutionary relationships. J Bacteriol 174: 3843-3849.
11. Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18 Suppl 1: S329-336.
12. Moran NA, Plague GR (2004) Genomic changes following host restriction in bacteria. Curr Opin Genet Dev 14: 627-633.
13. Herbeck JT, Wall DP, Wernegreen JJ (2003) Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia. Microbiology 149: 2585-2596.
14. Herbeck JT, Funk DJ, Degnan PH, Wernegreen JJ (2003) A conservative test of genetic drift in the endosymbiotic bacterium Buchnera: slightly deleterious mutations in the chaperonin groEL. Genetics 165: 1651-1660.
15. Wilcox JL, Dunbar HE, Wolfinger RD, Moran NA (2003) Consequences of reductive evolution for gene expression in an obligate endosymbiont. Mol Microbiol 48: 1491-1500.
16. Madan Babu M (2003) Did the loss of sigma factors initiate pseudogene accumulation in M. leprae? Trends Microbiol 11: 59-61.
17. Moran NA, Dunbar HE, Wilcox JL (2005) Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont Buchnera aphidicola. J Bacteriol 187: 4229-4237.
18. Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93: 2873-2878.
19. Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci U S A 96: 12638-12643.

20. Itoh T, Martin W, Nei M (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc Natl Acad Sci U S A 99: 12944-12948.
21. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. (2001) Massive gene decay in the leprosy bacillus. Nature 409: 1007-1011.
22. Lozada-Chávez I., Janga, S. C. and Collado-Vides, J. (submitted) Bacterial regulatory networks are extremely flexible in evolution.
23. Collado-Vides J., Magasanik B. y Gralla J.D. (1991) Control Site Location and Transcriptional Regulation in *Escherichia coli*. Microbiol. Reviews 55:371-394.
24. Savageau MA. (1998) Demand theory of gene regulation. I. Quantitative development of the theory. Genetics 149:1665-1676.
25. Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J. (2006). RegulonDB version 5.0 RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. 34(Database issue): D394-7.
26. Reznikoff, W. (1992). The lactose operon-controlling elements: a complex paradigm. Mol. Microbiol 6: 2419–2422.
27. Czarniecki, D., Noel, R. J., Jr & Reznikoff, W. S. (1997). The 245 region of the Escherichia coli lac promoter: CAP-dependent and CAP-independent transcription. J. Bacteriol 179: 423–429.
28. Goodrich JA, McClure WR. (1991) Competing promoters in prokaryotic transcription. Trends Biochem Sci. 16: 394-397. Review.
29. Goodrich JA, McClure WR. (1992). Regulation of open complex formation at the Escherichia coli galactose operon promoters. Simultaneous interaction of RNA polymerase, gal repressor and CAP/cAMP. J Mol Biol 224: 15-29.
30. Nickels BE, Mukhopadhyay J, Garrity SJ, Ebright RH, Hochschild A. (2004) The sigma 70 subunit of RNA polymerase mediates a promoter-proximal pause at the lac promoter. Nat Struct Mol Biol 11: 544-50.
31. Brodolin K, Zenkin N, Mustaev A, Mamaeva D, Heumann H. (2004) The sigma 70 subunit of RNA polymerase induces lacUV5 promoter-proximal pausing of transcription. Nat Struct Mol Biol 11: 551-7.
32. Reznikoff WS, Bertrand, K., Donnelly, C., Krebs, M., Maquat, L.E., Peterson, M., Wray, L., Yin, J., Yu, X.M. (1987) Complex promoters. In: Reznikoff WS, Burgess, R. R., Dahlberg, J. E., Gross, C. A., Record, M. T. Jr & Wickens, M. P., editor. RNA Polymerase and the Regulation of Transcription. New York, N.Y.: Elsevier. pp. 105–113.
33. Stormo GD (1998) Information content and free energy in DNA--protein interactions. J Theor Biol 195: 135-137.
34. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563-577.
35. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

36. Moreno-Hagelsieb G, Trevino V, Perez-Rueda E, Smith TF, Collado-Vides J (2001) Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. Trends Genet 17: 175-177.

Figure Legends

Figure 1. Schematic representation of the main interactions of RNAP with promoter DNA and alignment of the σ^{70} motifs for recognition and binding of –10 (2.4 region) and –35 promoter sequences (4.2 region) for representative eubacteria. CLUSTALW was used to generate the alignment with default parameters (http://www.ebi.ac.uk/clustalw) [35].

Figure 2. Frequency matrices for the –10 and -35 motifs of σ^{70} promoters in *E. coli.* This matrix pair (Matrix_18_15_13_2_1.5) was selected for searching across bacterial genomes from a collection of optimized matrices defined for *E. coli* in [1]. Note that in order to compare these matrices with the canonical patterns (TTGACA and TATAAT), the spacers of 13 bp to 19 bp between the two boxes correspond to the 15 bp to 21 bp reported in the literature, as the TGT triplet is considered as part of the -10 box. Before searching for promoter-like signals, these matrices were calibrated using the base composition of each target genome.

Figure 3. Signal Density in regulatory vs nonregulatory regions of large and small eubacterial genomes. Regulatory regions are the ones found between divergent genes; they include 500 bases upstream and 500 bases downstream of the start of the gene (position 0). Coding regions contain genes with sizes above 1Kb; from those genes, the middle point was taken as the position 0 and 500 bases upstream and 500 bases downstream of this position were included. For Convergent regions, the end of the 3' gene was taken as position 0 and 500 bases upstream and 500 bases downstream of this position were included. The number of signals was averaged within intervals of 10 bp.

Figure 4 Signal density in regulatory vs nonregulatory regions of *M. tuberculosis* and *M. leprae.* Region definition and methodology as in Figure 3.

Tables

Table 1. -10 and -35 consensi and corresponding average scores for σ^{70} promoter-like signals in representative bacterial genomes. Promoter-signal searches were performed with *E. coli* Matrix_18_15_13_2_1.5 (Figure 2), calibrated with the base composition of the strictly noncoding regions of the corresponding genome, as described in Material and Methods. Consensi and average scores are reported for the subset of promoter-like signals most likely to contain functional promoters within each target genome.

Table 2. **Density patterns of σ^{70} promoter-like signals in eubacterial genomes.** Species are ordered by genome size to highlight the impact of this character on the presence of the differential density pattern. The overall similarity to *E. coli* is a measure based on the similarities of all orthologs between two genomes [36]. The values in the last three columns are based on the most likely functional promoters selected by COVER and were obtained only for those genomes showing an excess of signals in regulatory vs. convergent noncoding regions by a log-likelihood test ($p < 0.001$).

Table 3. -10 & -35 consensi and average scores for the degrading genome of *M. leprae* and its close relative *M. tuberculosis*. The %GC is for noncoding DNA.

#NCBI	Species	-10 box Consensus	-10 Average Score	-35 box Consensus	-35 Average Score
NC_000853	*Thermotoga_maritima*	G T T A T A A T	2.26	A A C T T G A A A	1.53
NC_000907	*Haemophilus_influenzae*	G T T A A A A T	2.15	A A A T T G A A A	1.42
NC_000908	*Mycoplasma_genitalium*	G T T A A A A T	1.86	A A A T T G A A A	1.43
NC_000911	*Synechocystis_PCC6803*	G T T A A A A T	2.62	A A A T T G A A A	1.55
NC_000912	*Mycoplasma_pneumoniae*	A T T A A A A T	2.26	C|T T A T T T A A A	1.58
NC_000913	*Escherichia_coli_K12*	G T T A T A A T	2.79	A A A T T G A A A	1.77
NC_000915	*Helicobacter_pylori_26695*	T T T A A A A T	2.44	A A A T T T A A A	1.39
NC_000918	*Aquifex_aeolicus*	T T T A A A A T	2.52	A T C T T T A A A	1.47
NC_000919	*Treponema_pallidum*	G G T A T A A T	2.52	C T C T T G A C G	1.54
NC_000964	*Bacillus_subtilis*	G T T A T A A T	2.77	A T A T T G A A A	1.66
NC_001318	*Borrelia_burgdorferi*	T T T A T A A T	1.94	A A A T T T A A A	1.33
NC_002162	*Ureaplasma_urealyticum*	T T T A A A A T	1.68	A A A T T T A A A	1.26
NC_002163	*Campylobacter_jejuni*	T T T A A A A T	2.41	A A A T T T A A A	1.34
NC_002179	*Chlamydophila_pneumoniae_AR39*	T T T A T A A T	2.34	A T A T T T A A A	1.48
NC_002488	*Xylella_fastidiosa*	G T T A T A A T	2.53	C A A T T G A A A	1.70
NC_002505	*Vibrio_cholerae*	G T T A A A A T	2.35	C A A T T G A A A	1.57
NC_002516	*Pseudomonas_aeruginosa*	G T T A T A A T	3.75	C T C T T G A A A	2.51
NC_002620	*Chlamydia_muridarum*	T T T A T A A T	2.29	A T A T T G A A A	1.44
NC_002662	*Lactococcus_lactis*	G T T A A A A T	2.28	A A A T T T A A A	1.44
NC_002663	*Pasteurella_multocida*	G T T A T A A T	2.40	A A A T T G A A A	1.50
NC_002677	*Mycobacterium_leprae*	G T T A A A A T	2.90	C A A T T G A C A	1.60
NC_002678	*Mesorhizobium_loti*	A T C A A T A|C T	2.98	A A C T T G A C A	1.95
NC_002696	*Caulobacter_crescentus*	G T T A T C A T	3.27	C G C T T G A C G	2.00
NC_002758	*Staphylococcus_aureus_Mu50*	T T T A T A A T	2.21	A A A T T T A A A	1.33
NC_003030	*Clostridium_acetobutylicum*	T T T A T A A T	2.14	A A A T T T A A A	1.24
NC_003047	*Sinorhizobium_meliloti*	G T T A T A A T	2.99	C G C T T G A A A	2.17
NC_003062	*Agrobacterium_tumefaciens_C58*	G T T A T A A T	2.67	C T A|C T T G A C A	1.96
NC_003098	*Streptococcus_pneumoniae_R6*	G T T A T A A T	2.43	A A A T T G A A A	1.55
NC_003103	*Rickettsia_conorii*	G T T A T A A T	2.06	A A A T T T A A A	1.38
NC_003112	*Neisseria_meningitidis_MC58*	G T T A A A A T	2.79	A A A T T G A A A	1.76
NC_003198	*Salmonella_typhi*	G T T A T A A T	2.99	A A A T T G A A A	1.94
NC_003212	*Listeria_innocua*	G T T A T A A T	2.36	A A A T T G A A A	1.41
NC_003272	*Nostoc_sp*	G T T A A A A T	2.34	A A A T T G A A A	1.53
NC_003295	*Ralstonia_solanacearum*	G T T A T A A T	3.75	C A A T T G A A G	2.25
NC_003317	*Brucella_melitensis*	G T T A T A A T	2.87	A A A T T G A A A	1.97
NC_003366	*Clostridium_perfringens*	T T T A T A A T	1.94	A A A T T T A A A	1.17
NC_003450	*Corynebacterium_glutamicum*	G T T A A A A T	2.77	A A A T T G A A A	1.80
NC_003454	*Fusobacterium_nucleatum*	T T T A T A A T	1.88	A A A T T T A A A	1.21
NC_004088	*Yersinia_pestis_KIM*	G T T A T A A T	2.73	C A A T T G A A A	1.66
NC_004337	*Shigella_flexneri_2ª*	G T T A T A A T	2.96	C T A T T G A A A	1.82
NC_004344	*Wigglesworthia_brevipalpis*	T T T A T A A T	1.55	A A A T T T A A A	0.67
NC_004463	*Bradyrhizobium_japonicum*	G C T A A A A T	3.32	A A C T T G A C A	2.03
NC_004545	*Buchnera_aphidicola*	T T T A A A A T	1.93	A A A T T T A A A	0.74

Table 1.

Species	Genome Size(Mb)	%GC in NonCoding DNA	%Overall Similarity to *E.coli*	%Identity to *E.coli* rpoD	Over Representation 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	-	-	-
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	3.66	88
H_influenzae	1.83	34	57	67	no	-	-	-
T_maritima	1.86	42	28	57	yes	88	2.76	75
S_pneumoniae_R6	2.04	34	30	65	yes	98	3.53	83
B_melitensis	2.12	50	34	49	yes	77	2.35	72
F_nucleatum	2.17	25	31	60	no	-	-	-
P_multocida	2.26	35	56	71	yes	99	3.64	85
N_meningitidis_MC58	2.27	46	42	52	yes	90	2.91	78
L_lactis	2.37	30	31	61	yes	99	3.59	85
X_fastidiosa	2.68	47	40	59	yes	83	2.65	75
A_tumefaciens_C58_Cereon	2.84	53	34	47	yes	70	2.13	70
S_aureus_Mu50	2.88	29	31	61	no	-	-	-
V_cholerae	2.96	43	54	76	yes	94	3.21	84
L_innocua	3.01	34	32	61	yes	98	3.52	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	3.03	81
Synechocystis_PCC6803	3.57	42	29	59	yes	96	3.31	81
S_meliloti	3.65	58	34	48	yes	62	1.82	65
R_solanacearum	3.72	62	40	57	yes	63	1.82	70
C_acetobutylicum	3.94	27	31	63	no	-	-	-
C_crescentus	4.02	62	32	49	yes	43	1.15	55
B_subtilis	4.21	38	32	68	yes	90	2.84	76
Y_pestis_KIM	4.60	42	70	91	yes	97	3.76	84
S_flexneri_2a	4.61	44	97	100	yes	95	3.38	82
E_coli_K12	4.64	43	100	100	yes	92	3.12	80
S_typhi	4.81	44	87	98	yes	97	3.32	84
P_aeruginosa	6.26	61	45	66	yes	60	1.79	69
Nostoc_sp	6.41	36	28	60	yes	98	3.63	84
M_loti	7.04	58	33	48	yes	55	1.53	60
B_japonicum	9.11	60	33	46	yes	58	1.71	64

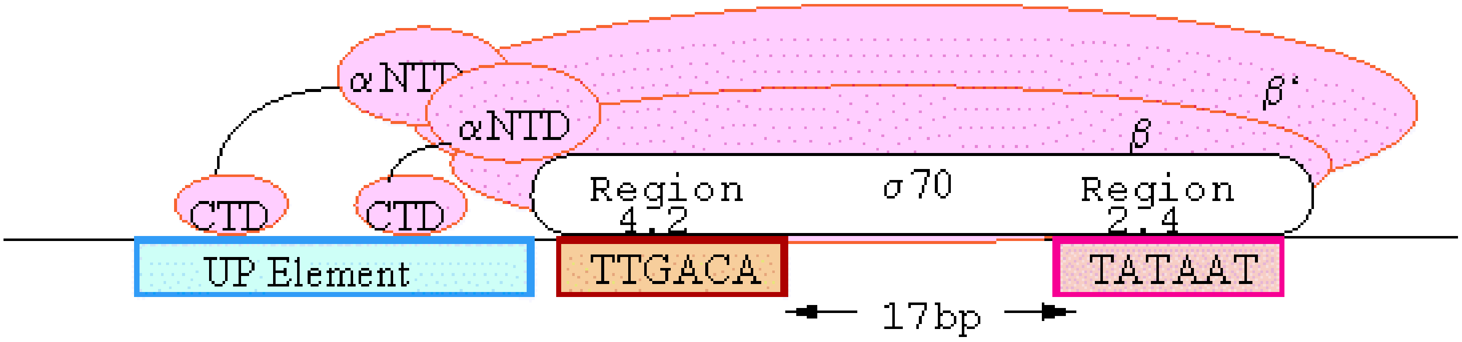
Table 2.

Species	Genome Size(Mb)	%GC	-10 box Consensus	-10 Average Score	-35 box Consensus	-35 Average Score
M. leprae	3.27	54	G T T A A A A T	2.90	C A A T T G A C A	1.60
M. tuberculosis CDC1551	4.40	64	G T T A T C A T	3.08	C A C T T G A C G	1.76
M. tuberculosis H37Rv	4.41	62	G T T A T A A T	3.17	C A C T T G A C A	1.78

Table 3.

Supporting Information

Figure S1. Signal Density in regulatory vs nonregulatory regions for all analyzed eubacterial genomes. Figure S1 can be viewed at http://www.ccg.unam.mx/Computational_Genomics/PromoterTools/Supplemental06/prof.html.

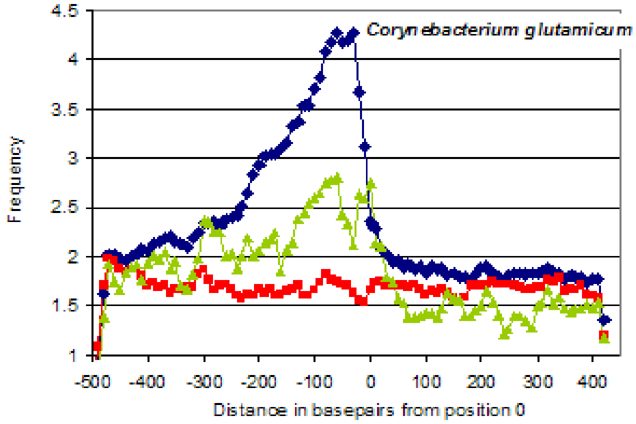


GENOME	4.2 REGION	2.4 REGION
A_tumefaciens_C58_Cereon	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
S_meliloti	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_melitensis	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
M_loti	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_japonicum	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
C_crescentus	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
R_conorii	DHTLEEVGQQFKVTRERIRQIESKALRKLQ	GYKFSTYATWWIRQAITRAIADQARTIRIP
E_coli_K12	DYTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
S_flexneri_2a	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
S_typhi	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
Y_pestis_KIM	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_aphidicola_Sg	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
Buchnera_sp	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_aphidicola	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
W_brevipalpis	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
V_cholerae	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
P_multocida	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
H_influenzae	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
P_aeruginosa	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
X_fastidiosa	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRSIADQARTIRIP
N_meningitidis_MC58	DHTLEEVGRQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
R_solanacearum	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_subtilis	TRTLEEVGKVFGVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRAIADQARTIRIP
L_innocua	TRTLEEVGRVFGVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
S_aureus_Mu50	TRTLEEVGKVFGVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
L_lactis	MHTLEDVGKQFKVTRERIRQIEAKAIKKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
S_pneumoniae_R6	MRTLEDVGKVFNVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
C_acetobutylicum	ARTLEEVGKEFNVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
C_perfringens	ARTLEEVGKEFNVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRAIADQARTIRIP
C_muridarum	PKTLEEVGSAFNVTRERIRQIEAKALRKMR	GYKFSTYATWWIRQAVTRAIADQARTIRIP
C_pneumoniae_AR39	PKTLEEVGSAFNVTRERIRQIEAKALRKMR	GYKFSTYATWWIRQAVTRAIADQARTIRIP
C_glutamicum	PRTLDEIGQVYGVTRERIRQIESKTMSKLR	GYKFSTYATWWIRQAITRAMADQARTIRIP
M_leprae	PRTLDQIGKLFGLSRERVRQIERDVMCKLR	GFKFSTYATWWIRQAITRGMADQSRTIRLP
H_pylori_26695	DRTLEEIGKELNVTRERVRQIESSAIKKLR	GFKFSTYATWWIKQAISRAIADQARTIRIP
C_jejuni	DRTLEEIGKELNVTRERVRQIESSAIKKLK	GYKFSTYATWWIRQAISRAIADQARTIRIP
Nostoc_sp	MKTLEEIGQIFNVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRAIADQSRTIRLP
Synechocystis_PCC6803	MKTLEEIGQIFNVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRAIADQSRTIRLP
A_aeolicus	EYTLEQVGKMFNVTRERIRQIENKALRKLR	GYKFSTYATWWIRQAITRAIADQARTIRIP
F_nucleatum	PKTLEEVGKIFNVTRERIRQIEVKALRKLR	GYKFSTYATWWIRQAITRAIADQGRTIRIP
T_maritima	PKTLEEVGQYFNVTRERIRQIEVKALRKLR	GYKFSTYATWWIRQAITRAIADQARTIRIP
B_burgdorferi	SLTLEEVGLMFNVTRERIRQIESKALRKLK	GFKFSTYATWWIRQAITRSISDQARTIRVP
T_pallidum	SQTLEEVGLYFDVTRERIRQIEAKALKRLR	GYKFSTYATWWIRQAITRSISDQARTIRVP
U_urealyticum	PYTLEEVGEYLGVTRERARQIESKAIRKLK	GHKFSTYATWWIRQSITRAIADQARQIRIP
-	--**:::*----::***-****-..:-:::	-.**********:*::.*.::**.*-**:*

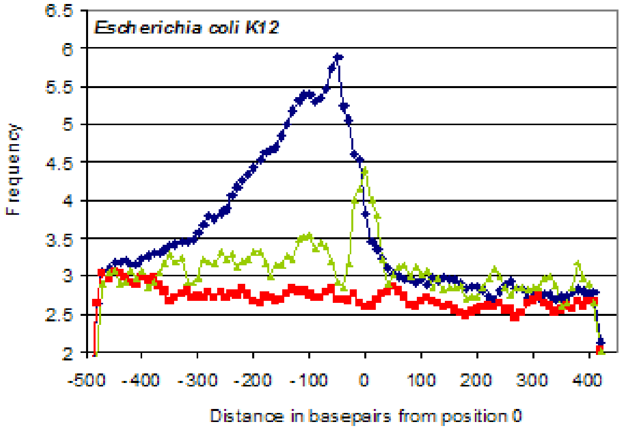
A	|	18	20	2	116	32	53	44	6	26		A	|	29	48	40	0	9	8	47	44	52
C	|	20	23	15	0	15	23	28	2	18		C	|	37	3	40	18	14	11	34	47	8
G	|	54	26	5	0	18	13	25	0	39		G	|	32	29	22	24	5	62	18	0	31
T	|	24	47	94	0	51	27	19	108	33		T	|	18	36	14	74	88	35	17	25	25
		G	T	**T**	**A**	**T**	**A**	**A**	**T**	G	[13...19]			C	A	A/C	**T**	**T**	**G**	**A**	**C**/A	**A**
						|											|					
						-10											**-35**					

◆ Regulatory / Coding
■ Coding
▲ Convergent / Coding

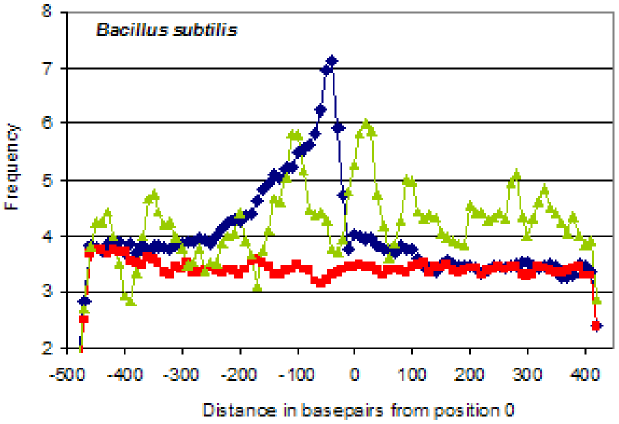
Large Genomes



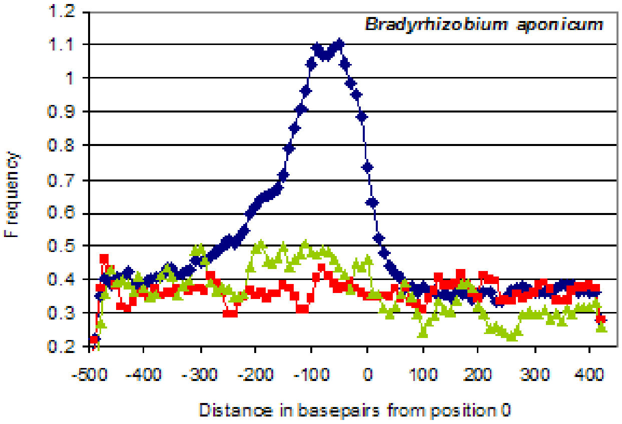
Size: 3,309,401 GC: 47% Base_Coding: 86%



Size: 4,639,221 GC: 42% Base_Coding: 88%

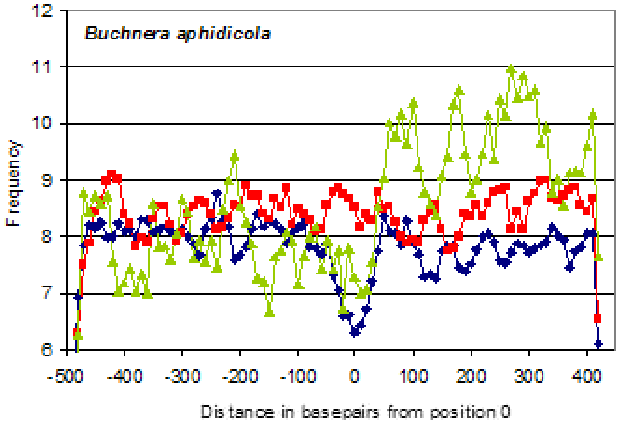


Size:4,214,630 GC:38% Base_Coding:87%

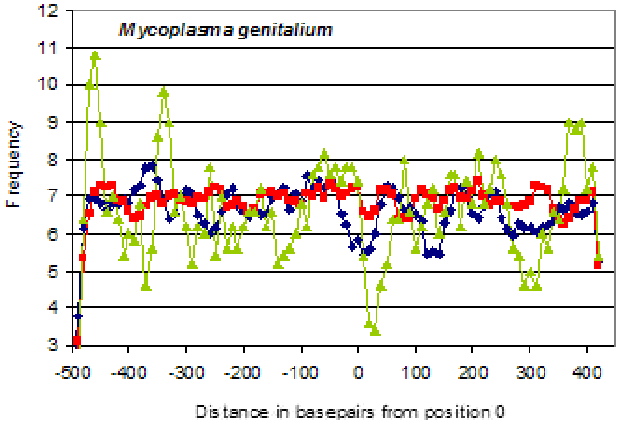


Size:9,105,828 GC:60% Base_Coding:87%

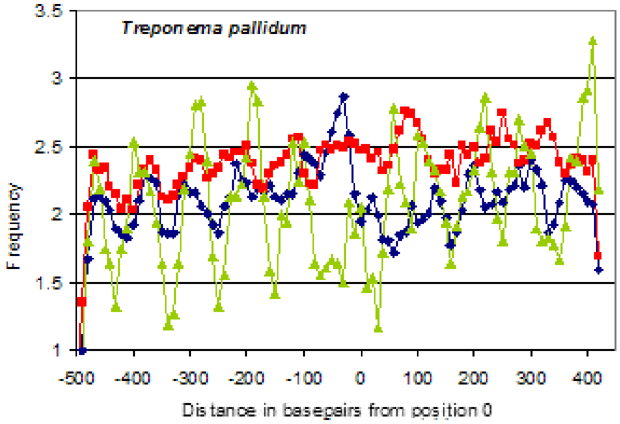
Small Genomes



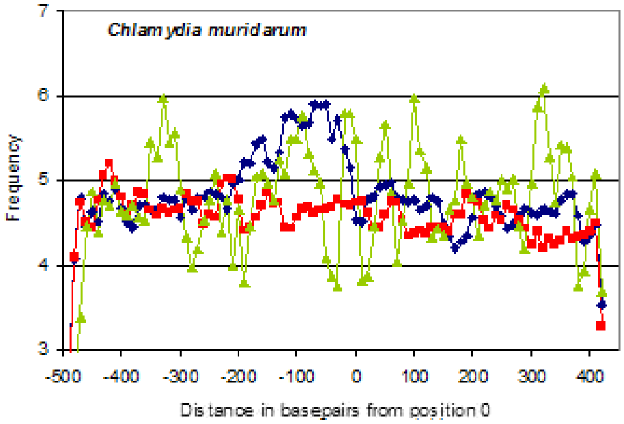
Size: 615,980 GC: 18% Base_Coding: 81%



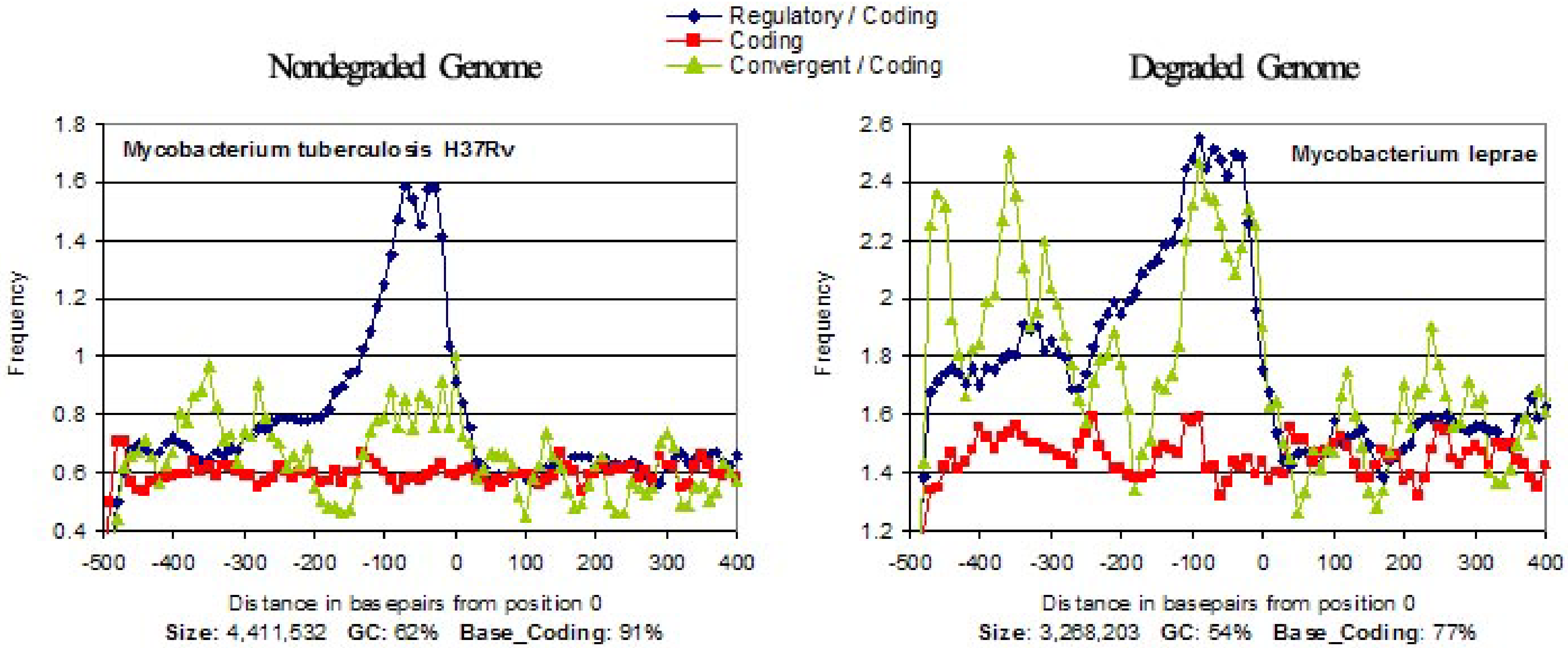
Size: 580,074 CG: 33% Base_Coding: 91%



Size: 1,138,011 GC: 54% Base_Coding: 92%



Size: 1,072,950 GC: 37% Base_Coding: 89%


Regulatory / Coding
Coding
Convergent / Coding
Nondegraded Genome
Mycobacterium tuberculosis H37Rv
Frequency
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
-500
-400
-300
-200
-100
0
100
200
300
400
Distance in basepairs from position 0
Size: 4,411,532 GC: 62% Base_Coding: 91%
Degraded Genome
Mycobacterium leprae
Frequency
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
-500
-400
-300
-200
-100
0
100
200
300
400
Distance in basepairs from position 0
Size: 3,268,203 GC: 54% Base_Coding: 77%