The Consequences of Base Pair Composition Biases for Regulatory Network Organization in Prokaryotes

Otto X. Cordero and Paulien Hogeweg*
*Theoretical Biology and Bioinformatics, University of Utrecht,Utrecht, The Netherlands

Given the dramatic variation in guanine-cytosine (GC) content observed in prokaryotes, from ∼20% to ∼75% GC, one wonders if these extreme biases in base pair composition affect the evolution of transcription factor–binding sites (BS). This letter shows that, along the wide range of GC content variation in bacteria, bacterial BS keep a high frequency of AT bases, roughly independently of the background (BG) base pair composition of intergenic regions. As a result, the equilibrium base pair frequencies of BS depart the most from those of BS DNA in GC-rich genomes. This not only implies a higher specificity but also a higher coding barrier for BS in GC-rich genomes. In accordance, we observe that the average percentage of divergently transcribed regions increases with the GC content of the genome, suggesting the use of a more efficient coding strategy.

One of the basic tenets in the paradigm of gene regulatory networks is that transcription factors (TF) are able to interact with specific sites on the DNA, by recognizing small sequence patterns in the DNA out of millions of other possible sites (Gerland et al. 2002). Given the huge biases in nucleotide composition observed across prokaryotes, from ∼20% guanine-cytosine (GC) in intracellular parasites to ∼75% GC in large species with complex metabolisms such as *Streptomyces coelicolor*, one wonders how the organization of the regulatory network copes with the potential problems that may be posed by changes in the information content of the DNA sequence. Intuitively, as nucleotide frequencies deviate from the maximum entropy configuration, in which all nucleotides are equally frequent, some motifs will appear more often and some will become more rare (Schneider et al. 1986). A naive expectation would be that evolution exploits GC biases by selecting TF recognition sites with an opposite GC content than the average of intergenic regions. However, as noticed by Mitchison (2005), intrinsic requirement for higher DNA flexibility at places of TF binding (Protozanova et al. 2004) may push binding sites (BS) to be AT richer than the intergenic background (BG). One question we ask here is whether this constraint holds across bacteria with large differences in GC content. Moreover, if there is a general bias for BS to be AT rich, we would like to know if it is relative to the base pair composition of the BG, or if there is a BG independent constraint on the AT content of BS. Depending on the answers to these questions, regulatory networks in GC-rich and GC-poor genomes may evolve under different evolutionary pressures. In this letter we will clarify these issues, discuss the evolutionary implications of our findings, and show evidence indicating that, as a consequence of BS base pair composition biases, GC-rich and AT-rich genomes use different strategies to code their regulatory interactions.

We use the RegTransBase database v4 (Kazakov et al. 2007), containing 141 alignments and position weight matrices (PWM) and including more than 1000 different binding sequences, to study the relationship between BG and BS base pair biases on a large phylogenetic scale. The first step is to quantify the relationship between the base pair frequencies of the genome and of the binding motifs. Figure 1*A* shows that, over wide GC content range, the nucleotide frequencies of motifs remain in average constrained to a high percentage of AT bases. Moreover, because of the of relatively flat trend between the AT content of the intergenic BG and of BS, genomes with extreme base pair frequency biases have different regimes of "specificity": whereas in AT-rich genomes BS sequences are "typical" with respect to the BG (and in many cases even have AT contents which are lower than for the average intergenic region), in GC-rich genomes BS base pair frequencies depart the most from BG DNA.

To account for possible human biases in the curation of BS, we used the unsupervised BS predictions contained in the SwissRegulon database (Pachkov et al. 2007). These predictions were produced (for a small set the species) by a model of sequence evolution in orthologous intergenic regions, which detects conservation and assigns a posterior probability of being a binding site to sequences in these regions. Figure 1*B* shows that the results obtained with this predictions match the flat trend obtained with curated BS in fig. 1*A*, confirming the idea that binding sequences depart the most from the BG in GC-rich genomes.

The trend observed between BG and BS base pair frequencies implies that AT-rich genomes should have a higher frequency of partial BS matches than GC-rich genomes. This is not so trivial as higher order correlations between base pairs (Djordjevic et al. 2003) or different weights per position match/mismatch ultimately determine the chance of finding a given sequence pattern in a genome. To address this question, we use PWM to calculate the density of PWM hits per DNA molecule (see Supplementary Material online).

With a cutoff of 60% of the average score of the corresponding BS found in the genome, we counted PWM hits on both strands of the DNA molecules where the corresponding BS are found. Figure 1*C* shows the distribution of the density of 60% PWM hits for GC-rich (GC $>$ 50%) and AT-rich (GC $<$ 50%) genomes. In accordance with the trends in Figure 1*A* and *B*, we observe that AT-rich genomes have a higher density of sequences that partially match the consensus recognition motif.

Although it has been shown that short eukaryotic BS (5–9 bp) can quickly evolve by point mutations (Stone and Wray 2001), this is not necessarily the case for

Key words: evolution, gene regulation, GC content, binding sites.
E-mail: ottoxcordero@gmail.com.

Mol. Biol. Evol. 26(10):2171–2173. 2009
doi:10.1093/molbev/msp132
Advance Access publication June 30, 2009

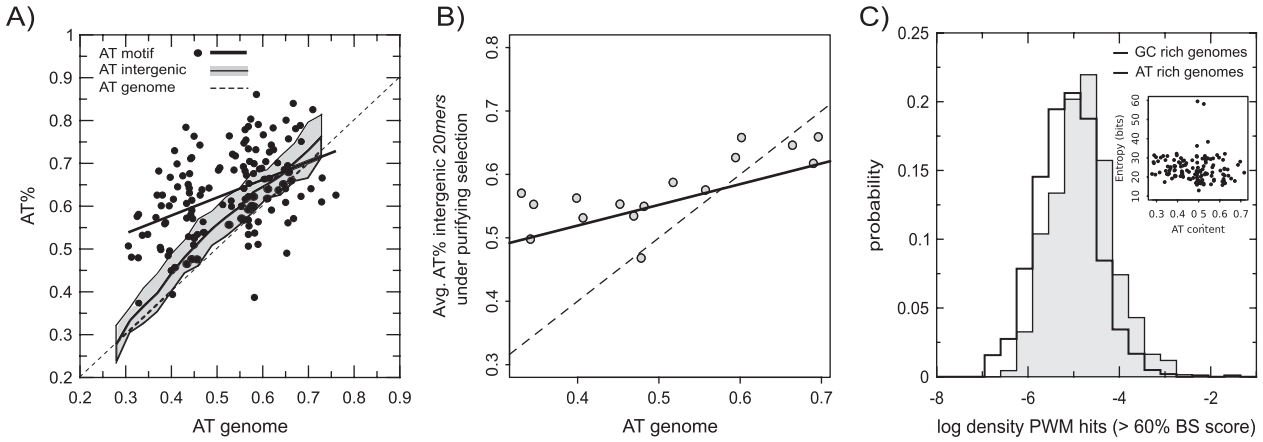

FIG. 1.—The relationship between AT content and specificity. (*A*) Average AT content of motifs vs. the average AT content of the genomes where they occur. The Pearson correlation is 0.45 ($P < 10^{-3}$). The average AT content of motifs is calculated by sampling 10 000 sequences from the nucleotide frequency matrices. The area on the BG corresponds to the average AT content of intergenic regions ±2 standard deviations and the dashed line is AT of the whole genome. (*B*) Results obtained with unsupervised prediction of binding sequences. The results are consistent with those of fig. A. (*C*) Distribution of the density of PWM hits in GC-rich (>50%) and AT-rich genomes. The inset shows that the shift in the distribution of PWM hits is not caused by higher alignment entropies in AT rich genomes ($r = -0.1$, $P = 0.18$).

longer (15–20 bp) prokaryotic BS. Our results show that in GC-rich genomes the distribution of hamming distances between BG and target BS is shifted toward larger values (see Supplementary Material), so we can expect that the evolution of new BS requires a larger number of mutations in these genomes. Moreover, because the BS fitness landscape is a nonlinear function of the number of mismatches (Gerland and Hwa 2002; Berg et al. 2004), most mutations will not have a fitness effect unless the hamming distance to the target BS is small enough, which can lead to long neutral waiting times (Berg et al. 2004). For the case of BS already present in the genome, if there is an intrinsic pressure keeping a high frequency of GC bases, these sites will require a specific and opposite selection pressure to remain AT rich. Altogether, this implies that BS evolution faces a higher coding barrier in GC-rich genomes.

If high BG GC content is indeed an obstacle for the evolution of BS, we can imagine evolution would find ways to "optimize" the coding of BS in genomes with high GC frequencies. One way to do this would be to pack more BS per intergenic region, by coding adjacent genes in divergent orientations. This strategy would allow to reuse regions that may already have a high tolerance for AT sequences, or even to code overlapping BS. Figure 2 shows that according to this expectation, GC-rich genomes tend to use a larger percentage of divergent intergenic regions than AT richer genomes. It is important to notice that, in spite of the correlation between GC content and genome size (Musto et al. 2006), there is no strong correlation between genome size and density of divergent regions (% of divergent regions vs. size of DNA molecule in base pairs, $r = 8.5 \times 10^{-4}$), which suggests that the trend in percentage of divergent regions is indeed explained by the AT content of the genome. This suggests that base pair composition biases can pose limitations to the coding of regulatory interactions, affecting the organization of the regulatory network. The high percentage of divergently transcribed regions in GC-rich genomes may reflect the need for a more efficient coding strategy.

One aspect that has not been discussed in this paper is whether the apparent convergence between BG and BS sequences in AT-rich genomes has consequences for the functional specificity of the network. One possibility is that TF in AT-rich genomes spend more time bound to spurious

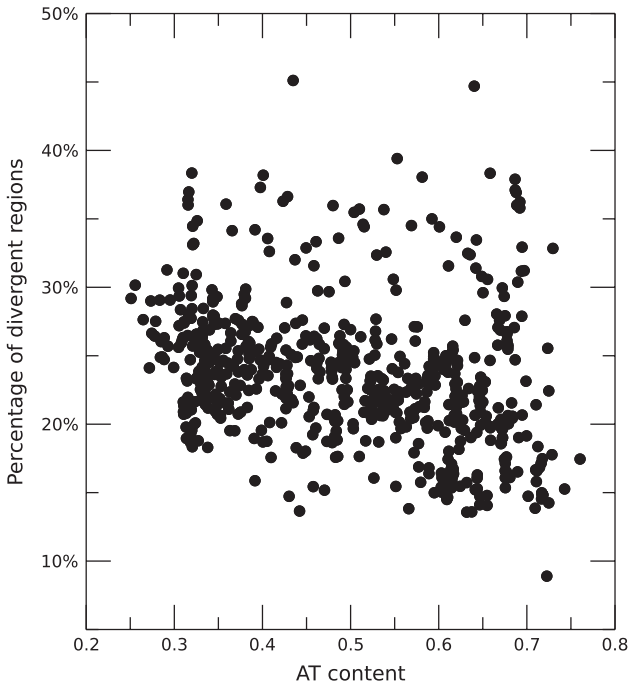

FIG. 2.—Density of divergently transcribed regions as a function of AT content. In accordance to our predictions, there is an increasing trend in density of divergent regions in GC-rich genomes. The Pearson correlations between AT content and density of divergent regions is −0.24 ($P < 10^{-3}$). Leaving out points above 30% density of divergent regions, the correlation is −0.45 ($P < 10^{-3}$). Those outliers with extremely high density of divergent regions are mostly *archaeas*, *thermatogas* and some *cyanobacteria*. To avoid inclusion of intraoperonic pairs, the figure shows regions with intergenic length >50bp.

sites than in their GC rich counterparts, slowing down the kinetics of transcription (Kolesov et al. 2007). If this is the case, such scenario could be permissible in obligate parasites, which have lower demand for regulation (Borenstein et al. 2008). However, in large and facultative AT-rich gram positives such as *Firmicutes*, it is unclear to what extent the convergence between BS and BG base pair frequencies may represent a relevant physiological limitation.

Our results show that in GC-rich genomes BS evolve under very different constraints with respect to the BG, whereas in AT- rich genomes BS are more "typical." The general pattern of high AT frequencies in BS gives us strong reason to believe that there are species independent biophysical constrains, possibly related to DNA flexibility. Our results show that these constraints may have led to the evolution of different coding strategies in GC-rich and AT-rich genomes. In particular, we have seen that GC-rich genomes tend to use a large percentage of divergently transcribed regions, which could potentially allow for a more efficient coding of regulatory interactions. As we have recently suggested (Cordero and Hogeweg 2009), these different strategies may underlie the lineage-specific trends between genome size and proportion of regulatory proteins.

Supplementary Material

Supplementary Material are available online at Molecular Biology and Evolution (http://www.mbe.oxfordjournals.org/).

Acknowledgement

The authors wish to express their gratitude Dr. Juan Poyatos for ideas and suggestions which helped shape this research.

Literature Cited

Berg J, Willmann S, Lassig M. 2004. Adaptive evolution of transcription factor binding sites. BMC Evol Biol. 4:42.

Borenstein E, Kupiec M, Feldman MW, Ruppin E. 2008. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci. 105:14482–14487.

Cordero OX, Hogeweg P. 2009. Regulome size in prokaryotes: universality and lineage specific variations. Trends Genet. PMID: 19540614.

Djordjevic M, Sengupta AM, Shraiman BI. 2003. A biophysical approach to transcription factor binding site discovery. Genome Res. 13:2381–2390.

Gerland U, Hwa T. 2002. On the selection and evolution of regulatory DNA motifs. J Mol Evol. 55:386–400.

Gerland U, Moroz JD, Hwa T. 2002. Physical constraints and functional characteristics of transcription factor-DNA interaction. Proc Natl Acad Sci U S A. 99:12015–12020.

Kazakov AE, Cipriano MJ, Novichkov PS, Minovitsky S, Vinogradov DV, Arkin A, Mironov AA, Gelfand MS, Dubchak I. 2007. Regtransbase – a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. Nucleic Acids Res. 35:D407–D412.

Kolesov G, et al. 2007. How gene order is influenced by the biophysics of transcription regulation. Proc Natl Acad Sci U S A. 104:13948–13953.

Mitchison G. 2005. The regional rule for bacterial base composition. Trends Genet. 21:440–443.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. 2006. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun. 347:1–3.

Pachkov M, Erb I, Molina N, Van Nimwegen E. 2007. Swissregulon: a database of genome-wide annotations of regulatory sites. Nucleic Acids Res. 35:D127–D131.

Protozanova E, Yakovchuk P, Frank-Kamenetskii MD. 2004. Stacked-unstacked equilibrium at the nick site of DNA. J Mol Biol. 342:775–785.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. J Mol Biol. 188:415–431.

Stone JR, Wray GA. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol. 18: 1764–1770.

Michele Vendruscolo, Associate Editor

Accepted 23 June 2009