Cytosine Usage Modulates the Correlation between CDS Length and CG Content in Prokaryotic Genomes

Xuhua Xia,*† Huaichun Wang,‡ Zheng Xie,* Malisa Carullo,* Huang Huang,* and Donal Hickey§

*Department of Biology, University of Ottawa, Ottawa, Ontario, Canada; †Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, Ontario, Canada; ‡Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia; and §Biology Department, Concordia University, Montreal, Quebec, Canada

Previous studies have argued that, given the AT-rich nature of stop codons, the length and GC% of coding sequences (CDSs) should be positively correlated. This prediction is generally supported empirically by prokaryotic genomes. However, the correlation is week for a number of species, with 4 species showing a negative correlation. Here we formulate a more general hypothesis incorporating selection against cytosine (C) usage to explain the lack of strong positive correlation between the length and GC% of CDSs. Two factors contribute to the selection against C usage in long CDSs. First, C is the least abundant nucleotide in the cell, and a long CDS should have fewer Cs to increase transcription efficiency. Second, C is prone to mutation to U/T and selection for increased reliability should reduce C usage in long CDSs. Empirical data from prokaryotic genomes lend strong support for this new hypothesis.

Introduction

The length and GC% of coding sequences (CDSs) are in most cases positively correlated in prokaryotic genomes (Oliver and Marin 1996; Xia et al. 2003). This positive correlation is attributed to the effect of mutation on the distribution and frequency of stop codons that are relatively AT-rich and should be encountered less frequently in GC-rich genomes. However, there are at least 4 bacterial genomes, Treponema pallidum subsp. pallidum str. Nichols (NC_000919), Tropheryma whippyli TW08/27 (NC_004551), T. whipplei str. Twist (NC_004572), and Ureaplasma parvum serovar 3 str. ATCC 700970 (NC_002162) that, instead of exhibiting the predicted positive correlation between the length and GC% of CDSs, have a negative correlation between the two (Xia et al. 2003). There are also many other species exhibiting a positive correlation that is not significantly different from zero. In addition, the positive correlation is mostly absent in eukaryotic genomes (Duret et al. 1995; Carels and Bernardi 2000; Xia et al. 2003; Wang et al. 2004). In short, although the effect of mutation on the distribution and frequency of stop codons can account for a significant proportion of variation in prokaryotic CDS length, the proportion is small and other explanatory factors need to be sought. In this paper, we propose a more general hypothesis on factors contributing to the variation of CDS lengths and provide empirical substantiation of the hypothesis.

Two factors may modulate the correlation between the length and the GC% of the CDSs, both invoking selection against cytosine (C) usage. The first is based on the observation of generally low intracellular C availability in both eukaryotes and prokaryotes. For example, in the exponentially proliferating chick embryo fibroblasts in culture, the concentrations of adenosine triphosphate, cytidine triphosphate (CTP), guanosine triphosphate, and uridine triphosphate, in the unit of (moles \times 10^{-12} \text{ per 10^6 cells}) are 1890, 53, 190, and 130, respectively, in 2-hour culture and 2390, 73, 220, and 180, respectively, in 12-hour culture (Colby and Edlin 1970). The protozoan parasite, Trypanosoma brucei, exemplifies C-limitation in mammalian blood. The parasite maintains its de novo synthesis pathway for CTP and inhibiting its CTP synthetase effectively eradicates the parasite population in the host (Hofer et al. 2001). This suggests that little CTP can be salvaged from the host. In contrast, the parasite does not have de novo synthesis pathways for purines, suggesting that the parasite can obtain the purines by its salvage pathway. C-limitation appears to be a general feature in bacterial species, and a biochemical explanation has been offered to explain the general C-limitation in bacterial species (Rocha and Danchin 2002). Thus, a long CDS with many Cs may take inordinately long to be transcribed and should therefore be selected against. The selection against C usage (and consequently against GC%) in long CDSs naturally will result in long CDSs to become less C-rich (and consequently less GC-rich), leading to a decrease in the correlation between the length and GC% of CDSs.

Designate $A_C$ as an index measuring the intensity of selection against C usage in a bacterial genome (where the letter “A” confer the meaning of “against”) and $R_{L,GC}$ as the correlation between the length and GC% of CDSs within a genome. The reasoning in the previous paragraph involving selection against C usage would lead to the prediction that a genome with a large $A_C$ should have a small $R_{L,GC}$ (or a genome with a small $A_C$ should have a large $R_{L,GC}$). In short, $A_C$ and $R_{L,GC}$ should be negatively correlated across genomes. This can explain why the 4 genomes listed in the previous paragraph do not have a positive $R_{L,GC}$—they may have a large $A_C$.

The second factor that may contribute to a negative $R_{L,GC}$ is derived from the observation of a much higher spontaneous deamination rate of C to U or methylated C to T than other chemical processes leading to the decay of DNA and RNA (Sancar A and Sancar GB 1988; Frederico et al. 1990; Frederico et al. 1993; Lindahl 1993). The single-stranded mRNA is particularly prone to mutations caused by the spontaneous deamination because the deamination rate is about 100 times higher in single-stranded nucleotide sequences than in double-stranded ones (Frederico et al. 1990). We here present a simple model to link the high deamination rate to the correlation between the length and GC% of the CDSs, by following the same logic as the late
Maynard Smith (1998) in linking the mutation rate and the genome length.

Assume that the probability of an organism to survive and reproduce \( (P) \) depends on the probability of all essential genes functioning normally and can be described approximately as follows:

\[
P = \prod_{i=1}^{N} p_i,
\]

where \( p_i \) is the probability of gene \( i \) functioning normally and \( N \) is the number of essential genes in the genome. Note that, given the potentially large number of essential genes (i.e., a large \( N \)) in an organism, \( p_i \) values should be roughly a constant close to 1, otherwise \( P \) would be too small. The value of \( p_i \) should depend on the number of sites (i.e., the length of gene \( i \), designated by \( L_i \)), the mutation rate per site \( (\mu) \), and the probability that the mutation is disabling \( (p_d) \). The average number of such hits for a gene of length \( L \) is \( L_ip_d \), which can be used as the \( \lambda \) parameter in the Poisson distribution. So the probability of the gene not being hit by any disabling mutation is

\[
p_i = e^{-L_ip_d}.
\]

The nucleotide C has a relatively high rate of mutating to T (when C is methylated) or U through spontaneous deamination, and these C \( \rightarrow \) T and C \( \rightarrow \) U mutations dominate the mutation spectrum (Sancar A and Sancar GB 1988; Frederico et al. 1990; Lindahl 1993; Tanaka and Ozawa 1994). Designating \( A \) as a non-C nucleotide according to the International Union of Biochemistry convention and \( p_{C,i} \) and \( p_{A,i} \) as the proportion of \( C \) and \( A \) in the CDS, we rewrite equation (2) as follows:

\[
p_i = e^{-(L_ip_{C,i} + L_ip_{A,i} + p_d)\mu_i} \approx e^{-L_ip_{C,i}p_d},
\]

where the approximation assumes that \( p_{C,i} \gg p_{A,i} \). We have noted before, given equation (1), that \( p_i \) should be roughly a constant smaller than (but close to) 1. This implies that

\[
e^{-L_ip_{C,i}p_d} \approx \text{cor}
\]

\[
L_i = \frac{-\ln(c)}{p_{C,i}p_d} = C', \quad \text{where} \quad C' = \frac{-\ln(c)}{p_{C,i}}, \tag{4}
\]

where \( c \) and \( C' \) denote a constant. Equation (4) predicts a negative correlation between \( L \) and \( p_{C} \) of CDSs. The negative \( R_{L,GC} \) in the 4 bacterial genomes mentioned above may then be viewed as a consequence of the negative correlation between \( L \) and \( p_{C} \) of CDSs because \( p_{C} \) and GC\% are almost always highly positively correlated. We should mention here that, although equation (4) suggests a quantitative relationship between \( L \) and \( p_{C} \) of CDSs, the relationship should be interpreted only qualitatively because of the oversimplicity of the model.

The 2 factors above, that is, the limited availability of C and the high deamination rate involving C may both contribute to selection against C usage in long CDSs and consequently decrease \( R_{L,GC} \). We may therefore hypothesize that, if we can devise an index, say \( A_{C} \), measuring the selection against C usage, we expect this index to be negatively correlated with \( R_{L,GC} \). This hypothesis will henceforth be referred to as the hypothesis of C-minimization in long CDSs, or C-minimization hypothesis for short.

One simple index to measure the selection against C usage in a genome is the GC skew (Lobry 1996):

\[
A_{C} = \frac{N_G - N_C}{N_G + N_C},
\]

where \( N_C \) and \( N_G \) designate the number of nucleotides C and G, respectively, and are calculated from all CDSs in the genome. The expected \( A_{C} \) is 0 when there is no selection against C usage, greater than 0 when C usage is selected against, and smaller than 0 when C usage is favored. Note that \( A_{C} \) is a genomic property because it is calculated from all CDSs in the genome. Also note that \( A_{C} \) is not our invention but just a shortened name for the GC skew (Lobry 1996).

We can now predict a negative correlation between \( A_{C} \) (i.e., selection against C usage) and \( R_{L,GC} \). In other words, a genome with a small \( A_{C} \) should have a large and positive \( R_{L,GC} \), but a genome with a large \( A_{C} \) should have a small \( R_{L,GC} \). The prediction gained a hint of truth when our preliminary study showed that the 4 genomes mentioned above with a negative \( R_{L,GC} \) happen to all have positive \( A_{C} \) values. It is interesting to examine the generality of the prediction by studying all complete prokaryotic genomic sequences.

Materials and Methods

All 281 prokaryotic genomes as of 16 November 2005 were retrieved from http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. CDSs were extracted by using DAMBE (Xia 2001; Xia and Xie 2001). \( A_{C} \) and \( R_{L,GC} \) were computed for each genome to test the prediction that genomes with a small \( A_{C} \) should have a large \( R_{L,GC} \) and vice versa.

We examine the relationship between \( R_{L,GC} \) and \( A_{C} \) in 2 ways. First, all 281 pairs of \((R_{L,GC}, A_{C})\) values from the 281 complete prokaryotic genomes were studied for a negative correlation without reference to their phylogenetic relationships, that is, there were treated as independent data points. This is not a statistically rigorous approach because of the shared ancestry among prokaryotic lineages that violates the assumption of data independence (Felsenstein 1985, 1988). A conceptually more appropriate approach is to construct a phylogenetic tree of all these genomes and perform a phylogeny-based comparative analysis (Felsenstein 1985, 1988) that we have used in analyzing the relationship between genome size and temperature (Xia 1995), in tracing transitions and transversions along branches in a phylogenetic tree (Xia et al. 1996), and in relating the length and GC% of CDSs in a small set of species (Xia et al. 2003). However, there are practical difficulties in using the method for analyzing the diverse array of prokaryotic species in this study. The rRNA sequences do poorly as universal evolutionary yardsticks, often with sequence differences between strains of the same species much greater than that between different species. Take the 16S rRNA sequences from the 2 Thermus thermophilus strains (HB27 and HB8) for example. The p-distance and the genetic distance based on the TN93
model (Tamura and Nei 1993) between the 2 T. thermophiles strains are 0.523 and 0.921, respectively, whereas the corresponding distances between T. thermophilus HB27 and the phylogenetic more remote Thermotoga maritima are only 0.178 and 0.205, respectively. This is not due to sequencing error. The 2 T. thermophiles strains each contain 2 copies of 16S rRNA, and the 2 copies are identical within each strain. No phylogenetic methods could possibly group the 2 Thermus thermophilus strains together as a monophyletic group. There are other less extreme but also troubling cases, making the phylogeny-based comparative method difficult to apply. An alternative is to use universally shared protein-coding sequences such as RNA polymerase or concatenated ribosomal proteins (Wolf et al. 2001), but similar problems exist. A third alternative is to Blast (Altschul et al. 1990, 1997) genes in each genome against all other genomes and use the gene sharing as an index to build a tree (Wolf et al. 2001; Henz et al. 2005). However, gene sharing changes substantially with the Blast cutoff E value, for example, changing the E value from 0.01 to 0.001 leads to substantially different phylogenies. In addition, this approach would incorporate the information from genes that must have undergone extensive horizontal gene transfer, for example, only one of the 7 genes in the urease gene cluster in Helicobacter pylori is shared with its close relatives such as Campylobacter jejuni, but all 7 are shared with several remotely related gram-positive bacterial species. Such results weaken our confidence in taking this Blast approach in building genomic trees. For these reasons, we limit phylogeny-based comparisons within species known to be closely related.

The large number of genomes available, however, does allow a statistically sound alternative method. The 281 genomes span many major taxonomic groups (table 1) whose phylogenetic relationships are poor predictors of 

<table>
<thead>
<tr>
<th>Taxon</th>
<th>N</th>
<th>r(RL,G,Ac)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaε</td>
<td>24</td>
<td>-0.3501</td>
<td>0.0468</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>20</td>
<td>-0.7182</td>
<td>0.0002</td>
</tr>
<tr>
<td>Alphaproteobacteria</td>
<td>29</td>
<td>-0.4424</td>
<td>0.0082</td>
</tr>
<tr>
<td>Betaproteobacteria</td>
<td>17</td>
<td>-0.0212</td>
<td>0.4678</td>
</tr>
<tr>
<td>Chlamydiae/Verrucomicrobia</td>
<td>10</td>
<td>-0.9288</td>
<td>0.0001</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>15</td>
<td>-0.5440</td>
<td>0.0180</td>
</tr>
<tr>
<td>FirmicutesA</td>
<td>14</td>
<td>-0.1994</td>
<td>0.2472</td>
</tr>
<tr>
<td>FirmicutesB</td>
<td>53</td>
<td>-0.4171</td>
<td>0.0010</td>
</tr>
<tr>
<td>Gammaproteobacteria</td>
<td>63</td>
<td>-0.2259</td>
<td>0.0376</td>
</tr>
<tr>
<td>Delta/Epsilonproteobacteria</td>
<td>14</td>
<td>-0.6359</td>
<td>0.0073</td>
</tr>
<tr>
<td>Mixed</td>
<td>22</td>
<td>-0.2050</td>
<td>0.1801</td>
</tr>
</tbody>
</table>

* One-tailed test.
* Wall-less Mollicutes.
* All Firmicutes excluding wall-less Mollicutes.
* Including Bacteroidetes (7), Spirochaetes (6), Thermotogae (1), Aquificae (1), Chloroflexi (2), Planctomycetes (1), Deinococcus–Thermus (3), and Fusobacteria (1), where the number within parenthesis indicates the number of genomes.

Table 1

Genomes with Strong Selection against C Usage (a large AC) Are Associated with a Low Correlation between the Length and GC% of CDSs (RL,G), Leading to a Consistently Negative Correlation between the 2 (third column)

Negative Correlation between the 2 (third column) may be considered independent of their phylogenetic affil-

Results and Discussion

The Pearson correlation between Ac and RL,G is −0.21143 for all 281 genomes (P = 0.0004). This is consistent with our prediction that genomes with a large Ac should have a small RL,G, that is, a negative correlation between the 2. However, as we have mentioned earlier in the Materials and Methods section, such a correlation, computed from all 281 pairs of Ac and RL,G values, suffers from the problem of nonindependence, and the P value consequently cannot be interpreted in a strict probabilistic sense.

Given the difficulty in performing a phylogeny-based comparative analysis (see the Materials and Methods section for details), we have adopted a conservative alternative by computing the correlation between Ac and RL,G separately for each of the taxonomic groups in table 1. The grouping follows that of National Center for Biotechnology Information (NCBI)’s microbial genome page at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, except for the following 2 treatments. First, the wall-less Mollicutes within Firmicutes are analyzed separately from the rest of Firmicutes. Second, the deltaproteobacteria and epsilon-proteobacteria each contain only 7 genomes and are merged because they belong to the same delta/epsilon division according to NCBI’s Taxonomic browser at http://www.ncbi.nlm.nih.gov/Taxonomy. All other taxonomic groups with 7 or fewer genomes were all lumped into the “mixed” group that is taxonomically heterogeneous (table 1).

The 10 genomes in the Chlamydiae/Verrucomicrobia group exhibit the highest negative correlation between Ac and RL,G, with the Pearson r being −0.9288 (P = 0.0001, fig. 1 and table 1). Other taxonomic groups (table 1) also provide consistent support for the C-minimization hypothesis, with the correlation between Ac and RL,G being consistently negative.

As a more conservative test, we take one negative correlation between RL,G and Ac in one taxonomic group as only one data point in favor of the C-minimization hypothesis and one positive correlation between RL,G and Ac in one taxonomic group as one data point against the C-minimization hypothesis. The null hypothesis of r between RL,G and Ac equal to 0 can be conclusively rejected by the 11 consistently negative r values (mean r = −0.4262, standard error = 0.08014, t = 5.3183, df = 10, P = 0.00034, 2-tailed test). One can also express the expectation of the null hypothesis of r equal to 0 as having equal number of
Cytosine Usage Modulates the Correlation between Length and GC% of CDSs

We observed that the correlation between length and GC% of CDSs is negative for some bacterial genomes. This is consistent with the C-minimization hypothesis, which suggests that cytosine deamination (C → U) is favored in long coding sequences (CDSs) due to the limited availability of C in these sequences. The C-minimization hypothesis is supported by the negative relationship between the length and GC% of CDSs, as shown in Fig. 1. The equation for the line of best fit is

\[ y = -0.3664x + 0.3438 \]

with a coefficient of determination \( R^2 = 0.8627 \). This indicates a strong negative correlation.

The C-minimization hypothesis can be further tested by comparing the correlation coefficient \( r(R_{L,GC}, A_C) \) for each genome. If the hypothesis is correct, we would expect a negative correlation between these variables. We calculated the correlation coefficients for each genome and found that the correlation coefficient is negative for all genomes in the Bacteroidetes group, as well as for two genomes in the Actinobacteria group.

In summary, the C-minimization hypothesis provides a coherent explanation for the observed negative correlation between the length and GC% of CDSs in some bacterial genomes. Further research is needed to confirm this hypothesis and understand its implications for the evolutionary biology of these organisms.

Takashi Gojobori, Associate Editor

Accepted May 8, 2006