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 CG% of coding sequences (CDSs) should be positively correlated. This prediction is generally supported empirically by prokaryotic genomes. However, the correlation is weak 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 whipplei* 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} 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 CG-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 $R_{L,\mathrm{GC}}$ 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 $R_{L,\mathrm{GC}}$ (or a genome with a small A_{C} should have a large $R_{L,\mathrm{GC}}$). In short, A_{C} and $R_{L,\mathrm{GC}}$ should be negatively correlated across genomes. This can explain why the 4 genomes listed in the previous paragraph do not have a positive $R_{L,\mathrm{GC}}$—they may have a large A_{C}.

The second factor that may contribute to a negative $R_{L,\mathrm{GC}}$ 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

Key words: selection against cytosine, differential nucleotide availability, genomic evolution, CDS length, CG content.

E-mail: xxia@uottawa.ca.

Mol. Biol. Evol. 23(7):1450–1454. 2006
doi:10.1093/molbev/msl012
Advance Access publication May 10, 2006

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:

$$P = \prod_{i=1}^{N} p_i, \tag{1}$$

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 (μ), and the probability that the mutation is disabling (p_d). The average number of such hits for a gene of length L is $L\mu p_d$, which can be used as the λ parameter in the Poisson distribution. So the probability of the gene not being hit by any disabling mutation is

$$p_i = \mathrm{e}^{-L_i \mu p_d}. \tag{2}$$

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

$$p_i = \mathrm{e}^{-(L_i p_{\mathrm{C}\cdot i}\mu_{\mathrm{C}} + L_i p_{D\cdot i}\mu_D)p_d} \approx \mathrm{e}^{-L_i p_{\mathrm{C}\cdot i}\mu_{\mathrm{C}} p_d}, \tag{3}$$

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

$$\mathrm{e}^{-L_i p_{\mathrm{C}\cdot i}\mu_{\mathrm{C}} p_d} \approx \mathrm{cor}$$

$$L_i = \frac{-\ln(\mathrm{c})}{\mu_{\mathrm{C}} p_d p_{\mathrm{C}\cdot i}} = \frac{\mathrm{C}'}{p_{\mathrm{C}.i}}, \quad \text{where } \mathrm{C}' = \frac{-\ln(\mathrm{c})}{\mu_C p_d}, \tag{4}$$

where c and C′ denote a constant. Equation (4) predicts a negative correlation between L and p_{C} of CDSs. The negative $R_{L,\mathrm{GC}}$ 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 $R_{L,\mathrm{GC}}$. 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 $R_{L,\mathrm{GC}}$. 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):

$$A_{\mathrm{C}} = \frac{N_{\mathrm{G}} - N_{\mathrm{C}}}{N_{\mathrm{G}} + N_{\mathrm{C}}}, \tag{5}$$

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 $R_{L,\mathrm{GC}}$. In other words, a genome with a small A_{C} should have a large and positive $R_{L,\mathrm{GC}}$, but a genome with a large A_{C} should have a small $R_{L,\mathrm{GC}}$. The prediction gained a hint of truth when our preliminary study showed that the 4 genomes mentioned above with a negative $R_{L,\mathrm{GC}}$ 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 Novemeber 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 $R_{L,\mathrm{GC}}$ were computed for each genome to test the prediction that genomes with a small A_{C} should have a large $R_{L,\mathrm{GC}}$ and vice versa.

We examine the relationship between $R_{L,\mathrm{GC}}$ and A_{C} in 2 ways. First, all 281 pairs of ($R_{L,\mathrm{GC}}$, 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

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 ($R_{\mathrm{L,GC}}$), Leading to a Consistently Negative Correlation between the 2 (third column)

Taxon	N	$r(R_{\mathrm{L,GC}}, \mathrm{Ac})$	P[a]
Archaea	24	−0.3501	0.0468
Actinobacteria	20	−0.7182	0.0002
Alphaproteobacteria	29	−0.4424	0.0082
Betaproteobacteria	17	−0.0212	0.4678
Chlamydiae/Verrucomicrobia	10	−0.9288	0.0001
Cyanobacteria	15	−0.5440	0.0180
FirmicutesA[b]	14	−0.1994	0.2472
FirmicutesB[c]	53	−0.4171	0.0010
Gammaproteobacteria	63	−0.2259	0.0376
Delta/Epsilonproteobacteria	14	−0.6359	0.0073
Mixed[d]	22	−0.2050	0.1801

[a] One-tailed test.

[b] Wall-less Mollicutes.

[c] All Firmicutes excluding wall-less Mollicutes.

[d] 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.

model (Tamura and Nei 1993) between the 2 *T. thermophilus* strains are 0.523 and 0.921, respectively, whereas the corresponding distances between *T. thermophilus* HB27 and the phylogenetic more remote *Thermotoga maritime* are only 0.178 and 0.205, respectively. This is not due to sequencing error. The 2 *T. thermophilus* 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 A_{C}, GC%, or CDS length. Thus, these 3 variables may be considered independent of their phylogenetic affiliation at this particular taxonomic level. We can study the correlation between $R_{L,\mathrm{GC}}$ and A_{C} within each taxonomic group. A negative correlation between $R_{L,\mathrm{GC}}$ and A_{C} in one taxonomic group contributes one point in favor of the C-minimization hypothesis, and a positive correlation between the 2 from a taxonomic group contribute one data point against the C-minimization hypothesis. When $R_{L,\mathrm{GC}}$ and A_{C} are not related (which is the null hypothesis), then the number of positive and negative correlations between the 2 from those major taxonomic groups should be equal to each other. Our alternative hypothesis is that there should be significantly more negative correlations than positive correlations.

Results and Discussion

The Pearson correlation between A_{C} and $R_{L,\mathrm{GC}}$ is −0.21143 for all 281 genomes ($P = 0.0004$). This is consistent with our prediction that genomes with a large A_{C} should have a small $R_{L,\mathrm{GC}}$, 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 A_{C} and $R_{L,\mathrm{GC}}$ 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 A_{C} and $R_{L,\mathrm{GC}}$ 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 epsilonproteobacteria 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 A_{C} and $R_{L,\mathrm{GC}}$, 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 A_{C} and $R_{L,\mathrm{GC}}$ being consistently negative.

As a more conservative test, we take one negative correlation between $R_{L,\mathrm{GC}}$ and A_{C} in one taxonomic group as only one data point in favor of the C-minimization hypothesis and one positive correlation between $R_{L,\mathrm{GC}}$ and A_{C} in one taxonomic group as one data point against the C-minimization hypothesis. The null hypothesis of r between $R_{L,\mathrm{GC}}$ and A_{C} 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

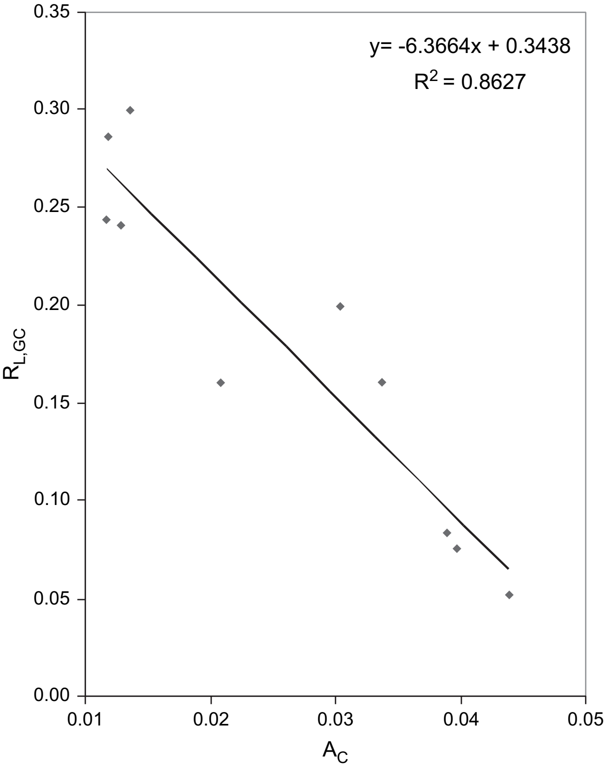

FIG. 1.—Negative relationship between A_C (selection against C usage) and $R_{L,GC}$ (correlation between the length and GC% of CDSs) in 10 Chlamydiae/Verrucomicrobia genomes.

$r(R_{L,GC}, A_C)$ values above or below zero, that is, the probability of having an $r(R_{L,GC}, A_C)$ value above zero is 0.5. This expectation can also be conclusively rejected because $P = 0.5^{11} = 0.00049$.

It may also be worth noting that, if we separate the last mixed group (table 1) into 1) the Bacteroidetes with 7 genomes, 2) Spirochaetes with 6 genomes, and 3) a mixed group with 3 Deinococcus–Thermus genomes, 2 Chloroflexi genomes, one Thermotogae genome, one Aquificae genome, one Planctomycetes genome and one Fusobacteria genome, then the correlation between A_C and $R_{\mathrm{L,GC}}$ is negative in each of the 3 groups, therefore, providing 3 data points in favor of the C-minimization hypothesis.

A more critical test of the C-minimization can be carried out with reference to the 4 bacterial genomes with a negative $R_{L,GC}$: *T. pallidum* (NC_000919), *T. whipplei* (NC_004551), *T. whipplei* (NC_004572), and *U. parvum* (NC_002162). If the C-minimization hypothesis is correct and powerful, then we should expect these species to have relatively high A_C values. *Treponema pallidum* has its $R_{L,GC} = -0.0976$ and $A_C = 0.0898$. Its sister species, *Treponema denticola* (NC_002967) has its $R_{L,GC} = 0.2350$ and $A_C = 0.0766$. This is consistent with the C-minimization hypothesis. *Tropheryma whipplei* (NC_004551) and *T. whipplei* (NC_004572) are the only 2 genomes within Actinobacteria that exhibit a negative $R_{L,GC}$ (equal to -0.0927 and -0.0745, respectively). They also have the largest A_C values within Actinobacteria (equal to 0.0740 and 0.0781, respectively). The mean A_C value within Actinobacteria is only 0.0029. *Ureaplasma parvum* belongs to the wall-less Mollicutes (labeled as FirmicutesA in table 1) and is the only species within the group that has a negative $R_{L,GC}$ ($= -0.0355$), and it also has the largest A_C value ($=0.1027$) within the group. Such comparisons with local phylogenetic information nicely illustrate the excellent predictive power of the C-minimization hypothesis.

In short, our comparative genomic analysis substantiates the C-minimization hypothesis that explains why some bacterial genomes do not show a positive correlation between the length and GC% of CDSs. We have provided 2 plausible factors that would contribute to selection against C usage in long CDSs, that is, the limited availability of C in cells and the high C → U mutations mediated by spontaneous deamination. In general, when there are such factors contributing to selection against C usage in long CDSs, the originally predicted positive correlation between the length and CG% of CDSs (Oliver and Marin 1996; Xia et al. 2003) is reduced and may even become negative.

Acknowledgments

This study is supported by grants from the Natural Sciences and Engineering Research Council's Discovery, Research Tools and Instruments, and Strategic Grants to X.X. We thank S. Aris-Brousou, E. Betran, P. Higgs, A. Morin, and E. Rocha for their insightful comments and suggestions.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–10.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–402.

Carels N, Bernardi G. 2000. Two classes of genes in plants. Genetics 154:1819–25.

Colby C, Edlin G. 1970. Nucleotide pool levels in growing, inhibited, and transformed chick fibroblast cells. Biochemistry (Mosc) 9:917.

Duret L, Mouchiroud D, Gautier C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol 40:308–17.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat 125:1–15.

Felsenstein J. 1988. Phylogenies and quantitative methods. Annu Rev Ecol Syst 19:445–71.

Frederico LA, Kunkel TA, Shaw BR. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry (Mosc) 29:2532–7.

Frederico LA, Kunkel TA, Shaw BR. 1993. Cytosine deamination in mismatched base pairs. Biochemistry (Mosc) 32:6523–30.

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC. 2005. Whole-genome prokaryotic phylogeny. Bioinformatics 21:2329–35.

Hofer A, Steverding D, Chabes A, Brun R, Thelander L. 2001. *Trypanosoma brucei* CTP synthetase: a target for the treatment of African sleeping sickness. Proc Natl Acad Sci USA 98:6412–6.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709–15.

Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–5.

Maynard Smith J. 1998. Evolutionary genetics. Oxford: Oxford University Press.

Oliver JL, Marin A. 1996. A relationship between GC content and coding-sequence length. J Mol Evol 43:216–23.

Rocha EP, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet 18:291–4.

Sancar A, Sancar GB. 1988. DNA repair enzymes. Annu Rev Biochem 57:29–67.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–26.

Tanaka M, Ozawa T. 1994. Strand asymmetry in human mitochondrial DNA mutations. Genomics 22:327–35.

Wang HC, Singer GA, Hickey DA. 2004. Mutational bias affects protein evolution in flowering plants. Mol Biol Evol 21:90–6.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol 1:8.

Xia X. 1995. Body temperature, rate of biosynthesis and evolution of genome size. Mol Biol Evol 12:834–42.

Xia X. 2001. Data analysis in molecular biology and evolution. Boston: Kluwer Academic Publishers.

Xia X. 2005. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene 345:13–20.

Xia X, Hafner MS, Sudman PD. 1996. On transition bias in mitochondrial genes of pocket gophers. J Mol Evol 43:32–40.

Xia X, Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92:371–3.

Xia X, Xie Z, Li WH. 2003. Effects of GC content and mutational pressure on the lengths of exons and coding sequences. J Mol Evol 56:362–70.

Takashi Gojobori, Associate Editor

Accepted May 8, 2006