Genetics: Published Articles Ahead of Print, published on August 30, 2010 as 10.1534/genetics.110.121137

http://dx.doi.org/doi:10.1534/genetics.110.121137

Title:

Avoidance of long mononucleotide repeats in codon pair usage

Tingting Gu[*,§], Shengjun Tan[*,§], Xiaoxi Gou[*], Hitoshi Araki[*,**], Dacheng Tian[*]

[*] State Key Laboratory of Pharmaceutical Biotechnology, Department of Biology, Nanjing University, Nanjing, China

[**] Department of Fish Ecology and Evolution, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Center of Ecology, Evolution and Biogeochemistry, Kastanienbaum, Switzerland

§ These authors contributed equally to this work. D. T. and H. A. are equal senior authors.

Running title: Avoidance of mononucleotide repeats

Keywords: codon pair, codon usage bias, mononucleotide repeat, mutability, comparative genomics

Corresponding author:

Hitoshi Araki

Seestrasse 79, 6047 Kastanienbaum, Switzerland

Tel: +41 (0)41 349 2130

Fax: +41 (0)41 349 2168

Email: hitoshi.araki@eawag.ch

ABSTRACT

Protein is an essential component for life, and its synthesis is mediated by codons in any organisms on earth. While some codons encode the same amino acid, their usage is often highly biased. There are many factors that can cause the bias, but a potential effect of mononucleotide repeats, which are known to be highly mutable, on codon usage and codon pair preference is largely unknown. In this study we performed a genomic survey on the relationship between mononucleotide repeats and codon pair bias in 53 bacteria, 68 archaea and 13 eukaryotes. By distinguishing the codon pair bias from the codon usage bias, four general patterns were revealed: strong avoidance of five or six mononucleotide repeats in codon pairs; lower observed/expected (o/e) ratio for codon pairs with C- or G-repeats (C/G pairs) than that with A- or T-repeats (A/T pairs); a negative correlation between genomic GC contents and the o/e ratios, particularly for C/G pairs; avoidance of C/G pairs in highly conserved genes. These results support natural selection against long mononucleotide repeats, which could induce frameshift mutations in coding sequences. The fact that these patterns are found in all kingdoms of life suggests that this is a general phenomenon in living organisms. Thus, long mononucleotide repeats may play an important role in base composition and genetic stability of a gene and gene functions.

INTRODUCTION

Among the many components of life, protein is most essential because living organisms use proteins not only for body structuring but also for its functioning. After the discovery of the genetic code for protein biosynthesis (CRICK *et al.* 1961), redundancy in the genetic code attracted great attention. Highly biased use of synonymous codon is one of them, and the codon usage bias is common not only among species but also within species (GROSJEAN and FIERS 1982; AKASHI 2001). Previous studies showed that codon usage bias is linked to several factors, such as efficiency and accuracy of translation (ROBINSON *et al.* 1984; BULMER 1991; AKASHI 1994; PLOTKIN *et al.* 2004), compositional bias (MUTO and OSAWA 1987; MCLEAN *et al.* 1998), and genome size or other non-selective forces (LAWRENCE and OCHMAN 1998; DOS REIS *et al.* 2004).

One additional factor is the sequence environment. It is known that nucleotides surrounding a codon can influence the codon usage preference, called context-dependent codon bias (YARUS and FOLLEY 1985). The context-dependent codon bias affects the efficiency and accuracy of translation (TANIGUCHI and WEISSMANN 1978; IRWIN *et al.* 1995) and the suppression of both premature stop codons and missense codons (BOSSI and RUTH 1980; MURGOLA *et al.* 1984). Reflecting the context-dependent codon bias, a strong codon *pair* bias is detected in both prokaryotic and eukaryotic genomes (GUTMAN and HATFIELD 1989; BUCHAN *et al.* 2006; TATS *et al.* 2008). It has been suggested that codon pair preference is influenced by all three nucleotides of the ribosomal A-site codon and the third nucleotide of the P-site codon (BUCHAN *et al.* 2006). Therefore, tRNA geometry within the ribosome was presumed to be the key factor governing genomic codon pair patterns, as it might enhance the fidelity and/or rate of translation.

A mononucleotide repeat is a homogeneous run of the same nucleotides. Potentially deleterious effects of a mononucleotide repeat in coding sequences (CDS) have been pointed out: the mononucleotide repeats in CDS are prone to transcriptional and translational slippage, which leads to functional disruption of the corresponding gene products (WAGNER *et al.* 1990; GURVICH *et al.* 2003; BARANOV *et al.* 2005); a strong association between mononucleotide repeats and the occurrence of insertion/deletion (indel) during the DNA replication process will elevate the risk of frameshift mutations (STRAUSS 1999), which might have severe fitness consequences. The list of diseases resulting from changes of unstable repeats continues to grow (GATCHEL and ZOGHBI 2005). In addition, previous studies suggested that in long mononucleotide runs, errors during the process of DNA synthesis are easier to escape from polymerase proofreading or mismatch repair (MMR) systems (KROUTIL *et al.* 1996; TRAN *et al.* 1997). C/G mononucleotide runs are found to be more unstable than A/T runs in *Escherichia coli* (SAGHER *et al.* 1999), yeast (HARFE and JINKS-ROBERTSON 2000) and mammalian cells (BOYCHEVA *et al.* 2003). Indeed, some of the mononucleotide repeats, such as GGGGGn, are found to be among the unpreferred codon pairs in various species (TATS *et al.* 2008). Therefore, the number of mononucleotide repeats, as well as their base compositions (A/T runs or C/G runs), might affect the occurrence of indels and the genetic stability of CDS.

In this study, we conducted a systematic survey on 134 genomes in bacteria, archaea and eukaryotes to evaluate the potential influence of the mononucleotide repeats on codon pair preference. We used the observed/expected (o/e) ratio of codon pairs with mononucleotide repeats to distinguish the codon pair bias from the codon usage bias. Our results suggest a strong avoidance of long (five or six) mononucleotide runs in CDS, most likely due to natural selection against the high mutability, which may shed new light on the forces exerted on both codon and codon-pair usage.

MATERIALS AND METHODS

Databases

To cover a diverse range of species, 13 eukaryotic, 53 bacterial and 68 archaeal genomes were selected from online databases (Table S1, S2 and S3). In addition, four sequence alignments (*Saccharomyces cerevisiae* & *S. paradoxus*, *Caenorhabditis elegans* & *C. briggsae*, *Drosophila melanogaster* & *D. yakuba*, *H. sapiens & M. musculus*) were obtained from UCSC Genome Informatics website (http://hgdownload.cse.ucsc.edu). Protein coding regions were determined based on the annotations in these databases. The 13 eukaryotic genomes, including fungi, plants and animals, were randomly selected to represent a wide range of species (Table S1). The 53 bacterial genome sequences were selected based on a criterion of >4Mb to give sufficient data (Table S2), whereas this criterion was not applied to the archaeal genomes because of their small genome sizes (2.24 Mb on average; Table S3).

Studied codon pairs

We first analyzed codon pairs that have mononucleotides spanning the two codons (sense:sense pairs, Table S4). Among 4096 (= 4^6) possible codon pairs, 928 sense:sense pairs contained two to six mononucleotides in the pair junction, when excluding the pairs containing a stop codon. A/T pairs (codon pairs with As or Ts spanning two codons) or C/G pairs (codon pairs with Cs or Gs spanning two codons) were analyzed together not only because A and T or C and G are parallel in the nucleotide chain position, but also the level of bias was similar (Figure S1). Codon pairs with the same number and composition (A/T or C/G) of mononucleotide runs in the pair junction were classified as a group.

In the analysis, codon pairs containing mononucleotide repeats other than those spanning the two codons are excluded because in such codon pairs the mononucleotide run size is not affected by the adjacent codon. For example, the number of longest mononucleotide repeats in codon pair AAATCG is three, which is the same as the single codon AAA. If there are any factors contributing to the reduction in the mononucleotide repeats, the single codon (AAA) would be the actual target, independent of the adjacent codons. We refer this as codon bias, not codon pair bias.

To further confirm the effect of mononucleotide repeats on codon pair bias, we also analyzed synonymous codon pairs, which is defined as a codon pair that has a choice of nucleotide bases which alter the number of mononucleotide repeats (from two to six) without changing the encoded amino acid sequences (Table S5). The possible longest mononucleotide run in the codon pair was two, three, five or six (Table S5). For example, codon pairs encoding dipeptide lysine:lysine had four possible compositions: AAGAAG, AAGAAA, AAAAAG and AAAAAA, for which the numbers of the longest mononucleotide runs were two, three, five and six, respectively.

The two types of codon pairs above have a partial overlap especially when we consider long mononucleotide repeats. Particularly, six mononucleotide repeats (6N) are entirely shared by both types of codon pairs (Table S4 and S5). For shorter mononucleotide runs ($\leq$5N), however, these types of codon pairs generally include different sets of codon pairs. We also analyzed sense:sense codon pairs excluding synonymous codon pairs.

Normalizing codon pair frequencies

Codon pair bias could be attributed to codon usage bias. To eliminate this effect, we normalized the expectation of codon pair occurrence by the frequency of used codons (GUTMAN and HATFIELD 1989; BUCHAN *et al.* 2006). First we calculated the observed (o_{ij}) and the expected number (e_{ij}) of a codon pair (codon *i* and codon *j*), on the basis of the estimated codon frequencies in the *k* th open reading frame (ORF$_k$, BUCHAN *et al.* 2006):

$$e_{ij,k} = \frac{c_i c_j N_p}{N_{tot}^{\ \ 2}}$$

where c_i is the observed count of codon *i*, N_{tot} is the total number of codons, and $N_p = N_{tot}$-1 represents the total number of codon pairs in the ORF$_k$. The effect of dipeptide bias on codon pairing was removed by normalizing the expected values of each codon pair, to generate e_{nor} (GUTMAN and HATFIELD 1989; BUCHAN *et al.* 2006):

$$e_{nor} = e_{nor\,ij,k} = \frac{\sum_{mn}(o_{dip\,mn})}{\sum_{mn}(e_{dip\,mn})} \times e_{ij,k} = \frac{\sum_{mn}(o_{dip\,mn})}{\sum_{mn}(\frac{c_m c_n}{N_{tot}^{\ \ 2}} \times N_p)} \times e_{ij,k}$$

where $o_{dip,mn}$ and $e_{dip,mn}$ are the observed and expected codon pair counts, respectively, encoding dipeptide *mn*. Observed and expected codon pair counts were then summed up at the genomic level. The numbers of codon pairs were calculated using a Perl script.

Because the codon pairs in each type encode the same dipeptide for synonymous pairs, the sum of the observed counts (e_{nor}) is equal to the sum of expectation. For sense:sense pairs, on the other hand, the sum of the observed counts is equal to the sum of expectation only when all 4096 possible codon pairs are included (the 0-1 mononucleotides in the pair junction).

Analyzing codon pair bias

The o/e ratio for each group of codon pairs with the same number (p) of mononucleotide Q in the pair junction was calculated as follows:

$$o / e_{pQ} = \frac{\sum o_{ij}}{\sum e_{ij}}$$

The average o/e ratio for a group of genomes was calculated as the geometric mean of the respective ratio of each genome.

To measure the difference between the observed and expected values of a single codon pair, a normalized offset value defined as r was calculated (BOYCHEVA *et al.* 2003):

$$r = \frac{o_{ij} - e_{ij}}{D_{\exp}} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij} \times (1 - e_{ij} / N_{tot})}}$$

where $D_{\exp}$ is the expected random deviation. r value is considered to be significant when the absolute value is greater than 2.0 (BOYCHEVA *et al.* 2003).

Analyzing codon pair usage in conserved regions

A Perl script was written to calculate the number of nucleotide substitutions and indels throughout the following combinations of alignments: *S. cerevisiae* & *S. paradoxus*, *C. elegans* & *C. briggsae* and *D. melanogaster* & *D. yakuba*, *H. sapiens & M. musculus*. The average nucleotide divergence (D) was adjusted with the Jukes and Cantor correction (JUKES and CANTOR 1969).

First, CDS were extracted according to the annotations of *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens,* respectively. Then the CDS sequences of each of the four comparisons were classified into three groups according to *D*. Each group had an equal length of sequences, and the one with the smallest *D* was regarded as the highly conserved region, while with the largest *D* as the less conserved region. The observed and expected counts of each codon pair were analyzed in both highly and less conserved regions (see Table S6 for details).

RESULTS

Avoidance of long mononucleotide runs in codon pairs

In the sense:sense codon pairs, the o/e ratio was apparently less than one in codon pairs with long mononucleotide runs, such as 5- or 6-mononucleotide repeats for C/G pairs ($o/e_{5C/G}$ or $o/e_{6C/G}$) and 6N for A/T pairs ($o/e_{6A/T}$, Figure 1). The geometric mean of the o/e ratio for the 6N pairs was 0.528 in eukaryotes, 0.488 for bacteria and 0.596 for archaea (Table S7-S9). The ratios for codon pairs with shorter runs (<4N mononucleotides) were significantly larger than those for pairs with longer runs (>4N mononucleotides) ($P<0.05$, t-test). The consistently lower number of observations than expected values (o/e < 1.0 for the 6N pairs in 133/134 genomes, or 99.2%. Figure S2 and Tables S7-S9) suggests a universal avoidance of long mononucleotide runs for sense:sense codon pairs in these organisms. The only exception was *Geobacter uraniireducens*, which showed o/e =1.084 for the 6N pairs (Table S8).

In addition, the o/e ratios for codon pairs with long A/T mononucleotide runs ($o/e_{5A/T}$ or o/$e_{6A/T}$) were significantly higher than those for C/G pairs in both prokaryotes and eukaryotes (*P*<0.05, paired t-test, except for o/$e_{5A/T}$ *vs.* o/$e_{5C/G}$ in mammals, discussed below). For example, in eukaryotic genomes, o/$e_{6A/T}$ was 1.8 times higher than o/$e_{6C/G}$, where the o/$e_{6C/G}$ was only 0.335. This result indicates that the 6-C/G is the most unfavorable codon pairs.

On the other hand, there was a pattern unique to mammals. Unlike the other genomes, higher o/e ratios for C/G pairs than for A/T pairs were generally observed in mammals except for 6Ns (Figure 1C). In addition, the o/e ratios for A/T pairs with long mononucleotide runs were higher in prokaryotes than in eukaryotes (o/$e_{5A/T}$ = 0.952 and o/$e_{6A/T}$ = 0.696 in bacteria; 0.963 and 0.669 in archaea, versus 0.831 and 0.602 in eukaryotes, respectively. Both *P*<0.05 by t-test). For long mononucleotide C/G runs, however, no clear difference was detected between prokaryotes and eukaryotes. The only exception was o/$e_{5C/G}$ in eukaryotes, which was significantly larger than o/$e_{5C/G}$ in bacteria (0.651 vs. 0.551, *P*<0.05 by t-test, Table S7 and S8). But the difference drastically decreased when the mammals were excluded (o/$e_{5C/G}$ = 0.574 for non-mammalian eukaryotes, Table S7).

Synonymous codon pairs showed similar patterns (Figure S3), and the elimination of synonymous codon pairs from sense:sense codon pairs resulted in virtually the same patterns (Figure S4). The negative relationships between the o/e ratio and the number of mononucleotide runs in both sense:sense and synonymous codon pairs further support the consistent avoidance of long mononucleotide runs in codon pairs in the genome evolution of prokaryotes and eukaryotes.

Effect of prokaryotic GC content on the o/e ratio

The results above revealed a stronger avoidance of codon pairs with long C/G runs relative to A/T runs. Because genomes with higher GC content would contain more C/G pairs under random expectation, we evaluated the effect of genomic GC content on the tolerance of genomes to the C/G mononucleotide runs. The wide range of GC content in prokaryotic genomes (29.9-72.8% in bacteria and 27.6-68.0% in archaea) provided an opportunity for investigating a correlation between o/e ratio and GC content.

When the 53 bacterial genomes were classified into three groups depending on GC content — group I (GC%<50%), group II (50%≤GC%<60%) and group III (GC%≥60%)— their o/e ratios for sense:sense pairs were quite different (Figure 2A). The o/e ratios for C/G pairs in group I were significantly higher than those in group III (P<0.05, t-test). For C/G pairs, the o/e ratios for group II were between those of group I and III. Notably, o/$e_{6C/G}$ for the high GC content, group III decreased to a very low value (0.223, Figure 2A). This observation suggested that the avoidance of long C/G mononucleotide runs was much stronger in genomes with higher GC content. This propensity was also shown through the negative correlation between GC content of individual genomes and their o/e ratios for C/G codon pairs, e.g., o/$e_{6C/G}$ in Figure 2B (R=-0.550, P<0.0001) and the other ratios in Figure S9B (P<0.05). For A/T pairs, on the other hand, the negative relationship between the o/e ratio and GC content was much weaker (Figure S5 and S9A). All the patterns observed in bacteria were also present in three groups of the archaeal genomes (Figure S6A-C; group I with GC% <40%, group II with GC% ranging from 40% to 50% and group III with GC%≥50%).

The normalized offset value (r) measures the difference between the observed and the expected counts of a certain codon pair (BOYCHEVA *et al.* 2003). Our calculations showed that the number of C/G codon pairs, in which the observed counts are significantly less than expected ($r \leq -2$), is positively correlated with the genomic GC content (Figure 2C; $P<0.0001$), reflecting the strong propensity of genomes with higher GC contents to avoid C/G pairs. According to the linear regression, in the bacterial genomes with GC content of 70%, the proportion of significantly underrepresented C/G pairs with mononucleotides in pair junctions was 55.2% (274/496), wheareas it was only 28.0% (139/496) in the genomes with low GC content (30%). The number of A/T pairs, which are significantly less than expected, was weakly correlated with genomic GC content in prokaryotes (Figure S5A and S7A).

The o/e ratios in conserved coding sequences

Given that long mononucleotide runs have a greater potential to produce indels, a lower o/e ratio was expected in more conserved CDS. Since 6Ns had the smallest o/e ratios in analyzed codon pairs with mononucleotides in pair junctions in eukaryotes (Figure 1 and Table S7), o/e_{6N} was analyzed in conserved regions in the four alignments of the eight genomes (see Methods).

Indeed, o/e_{6N} was significantly smaller in highly-conserved than in less-conserved regions in all alignments of non-mammalian eukaryotes (Figure 3A, $P<0.01$ by chi-square test). Moreover, the 6N codons appeared less frequently in highly-conserved regions in those three comparisons (Figure 3B). In the mammalian sequences, no such differences were observed. Although o/e_{6N} was slightly smaller in less-conserved regions in the human-mouse comparisons, the difference was not significant.

DISCUSSION

The biased usage of codons or codon-pairs is a common phenomenon in a wide range of species (GUTMAN and HATFIELD 1989; BUCHAN *et al.* 2006). A variety of factors, selective or non-selective, might be responsible for such bias. For example, the synonymous codons decoded in the ribosomal A-site by the same tRNA exhibit significantly similar ribosomal P-site pairing preference (BUCHAN *et al.* 2006). In other words, the codon pair preference is primarily determined by the interplay between nucleotides cP3 (the third nucleotide of the codon positioned at ribosomal P-site) and cA1/cA2 (the first/second nucleotide of the codon positioned at ribosomal A-site) (BUCHAN *et al.* 2006). Our results suggest that the avoidance of mononucleotide repeats in pair junctions is an additional explanation. For codon pairs encoding a certain dipeptide, nucleotides cP3 and cA3 are degenerate, and cP3 is more important in determining the mononucleotides run size. The interplay between cP3 and cA1/cA2 largely determines the mononucleotide run size in the degenerate codon pairs. Thus, the deleterious effect of indels and the consequent avoidance of long mononucleotide repeats in CDS can contribute to the close connection between cP3 and cA1/cA2 as well.

In this study, we confirmed a deficit of codon pairs with long mononucleotide runs relative to those with short mononucleotide runs in a variety of species by analyzing two kinds of codon pairs. This result is also consistent with previous studies, such as Tats *et al.* (2008) in which certain mononucleotide repeats are identified as avoided codon pairs among several other kinds (e.g., nnTAnn). In addition, we revealed three additional patterns: higher o/e ratio for A/T codon pairs than that for C/G pairs; negative correlation between

GC-content of individual genomes and their o/e ratios, particularly for C/G codon pairs; lower o/e ratio for codon pairs with long mononucleotide runs in conserved coding sequences. These patterns cannot be explained by the simplest tRNA geometry hypothesis. In *E. coli*, for example, the deficit of long mononucleotide A/T runs in codon pairs cannot be elucidated by tRNA geometry because the synonymous codons, AAG and AAA, TTC and TTT, are recognized by the same tRNA.

Natural selection on mutability of codons or codon-pairs might be an alternative explanation. The high frequency of indel occurrence has been confirmed to be closely associated with simple nucleotide repeats (STRAUSS 1999). The previous investigation of the mutability of mononucleotide runs in yeast showed that the mutation rate of 6N mononucleotide repeats was about 10-fold of that of 2N or 3N (GREENE and JINKS-ROBERTSON 1997). The high indel-mutability of long mononucleotide runs and the severely detrimental effect of coding indels can enforce the choice of the codon or codon pair usage. Consequently, the avoidance of long mononucleotide runs in coding sequences will minimize the change of coding function, particularly in highly conserved regions. Thus, this model can explain a scarcity of long runs in codon pairs, and why the conserved genes have a less codon pairs with long mononucleotide runs.

Under this scenario, long A/T mononucleotides can be better tolerated in codon pairs than in C/G runs, which exhibit higher mutability. For example, in a frameshift reversion assay in *S. cerevisiae* (GREENE and JINKS-ROBERTSON 1997), the mutation events in a 4C run were as many as in a 6A run, consistent with our observation that $o/e_{4C/G}$ was similar to $o/e_{6A/T}$ (0.841 *vs*. 0.811 in *S. cerevisiae*, Table S10). It has been known that C/G mononucleotides are more prone to produce indels in *E. coli* and yeast (GREENE and JINKS-ROBERTSON 1997; SAGHER *et al.* 1999; HARFE and JINKS-ROBERTSON 2000), and the frameshift instability of mononucleotide C or G runs may be due to stabilization of a stacked intermediate (SAGHER *et al.* 1999). Both the DNA polymerase fidelity (primarily avoidance of slippage) and the efficiency of the removal of frameshift intermediates by the MMR system are affected by the composition of mononucleotide runs; DNA polymerase slippage occurs more often while the MMR system removes frameshift intermediates less efficiently in C/G than in A/T mononucleotide runs (GRAGG *et al.* 2002). Considering the greater ability to produce indel, long C/G runs are less favored in those regions sensitive to frameshifts and their appearance would be underrepresented. In contrast, A/T mononucleotide runs would exert less influence on the maintenance of sequence stability. In higher eukaryotes, there is no experimental data on the mutability of mononucleotides, but it has been reported that the mutation rate of G_{17} repeat sequences was much higher than those of A_{17} and $(CA)_{17}$ in mismatch repair proficient embryonic mouse fibroblasts (BOYER *et al.* 2002).

If the avoidance of slippages from long mononucleotide runs contributes to the biased usage of codon pairs, it is understandable that there is a negative correlation between GC content of individual genomes and their o/e ratios because the genomes with higher GC content are expected to have a higher possibility of forming of long mononucleotide sequences. In the three bacterial groups with GC%<50%, 50%<GC%<60% and GC%>60%, the expected number of six C/G pairs were 466, 745, 1406 (*P*=0.057, *P*<0.01, *P*<0.05 for comparison of group 1&2, 1&3 and 2&3 respectively, t-test), whereas the observed counts were roughly the same, 285, 329, 361, respectively (*P*=0.524-0.787 for comparison of groups by t-test). The same tendency was observed in archaea as well. Therefore, prokaryotes may have evolved a mechanism to control the mutability of their genomes.

All analyzed genomes have o/$e_{6C/G}$ less than one except *G. uraniireducens*. This bacterium was isolated from subsurface sediment undergoing uranium bioremediation. This species reduces metals including uranium with acetate and other organic acids serving as the electron donor (SHELOBOLINA *et al.* 2008). It is generally accepted that uranium induces DNA damage and subsequent high mutation rate through a combination of chemical and radiological effects (STEARNS *et al.* 2005). *G. uraniireducens* may be able to tolerate more 6 C/G codon pairs, due to its higher tolerance of mutational constraints or the advantage of rapid evolution for adaptation to its harsh environment.

A more frequent occurrence of indels in longer C/G mononucleotides may partly explain why some amino acids have more synonymous codons than others. It has been shown that the genetic code is not a random assignment of codons to amino acids, and that the code minimizes the effects of point mutation or mistranslation (FREELAND and HURST 1998). Therefore, a good strategy to avoid mutation would be a reduced usage of codons with higher GC content, potentially to minimize the risk of longer C/G runs. To achieve this goal, such codons would have evolved as synonymous codons which are used less often. In fact, this hypothesis can be tested by a GC analysis for all codons. There are eight amino acids with four or more synonymous codons. In these codons, the average GC content is as high as 68.1%, which is significantly larger than 33.3% (P<0.001, t-test), the GC content of the other 12 amino acids with <3 synonymous codons. Notably, only six codons out of 23 for these 12 amino acids have two GC each, while 26/38 such codons are found for the eight amino acids with >3 synonymous codons. In addition, the start and stop codons have only one or no G. Thus, almost all AT-rich codons are used for the amino acids with limited synonymous codons or stop/start codons. Clearly, these codons have a little chance of forming long C/G mononucleotides, and therefore a higher opportunity to maintain stable gene function.

The indel-mutability model can shed light on the usage of codons and codon-pairs. Our results suggest that the avoidance of long mononucleotides can maintain the conserved gene function by preventing indel occurrence in coding sequences. This may be the best way to minimize mutation by constructing an appropriate gene composition, e.g., the choice of GC content and specific nucleotide combination. In highly conserved genes, on which mutations are supposed to be highly deleterious, the maximal avoidance of long mononucleotide repeats might be essential. Our results on the conserved regions (Figure 3) are well consistent with this scenario.

One exception was the case of mammalian genomes, which showed no consistent pattern as the other comparisons (Figure 3). Further study is required to understand the cause of the different patterns in mammals, although less efficient natural selection due to the smaller effective population size in mammals relative to invertebrates and prokaryotes, typically at the magnitude of one or two orders (LYNCH and CONERY 2003), might explain the phenomenon.

Results from recent studies suggest either up or down-regulation of the mutability level. The gene composition might be the first step in controlling the mutability level. If a severely detrimental effect would result from the occurrence of any particular indel in CDS, the indel would be removed efficiently (CHEN *et al.* 2009). When a region can tolerate indels, the result could be induction of more mutations (TIAN *et al.* 2008; ZHU *et al.* 2009), promotion of ectopic recombination (SUN *et al.* 2008), or reduction of recombination for the surrounding regions to maintain additional mutations (DU *et al.* 2008). This process would be an efficient way to regulate the mutation level and suggests that the mutability in a gene is self-regulated, at least to some extent. From this point of view, the mechanism for indel due to slippage might be a consequence of adaptive evolution, which can explain why this mechanism works well for long C/G runs but less well for the same long A/T mononucleotide sequences. The o/e ratio, particular the $o/e_{C/G}$ ratio, could be used as a measure of the mutation potential for individual or multiple genes in a species. Therefore, these ubiquitous and selectively-maintained mononucleotide runs can greatly contribute to the high genetic diversities and to the molecular evolution. Analysis of the distribution of long mononucleotide runs will provide information for the evolution of genes and genomes. With recent works that have revealed the possible causes for codon bias, our study suggests that the role played by mononucleotide runs in such bias can be very important in shaping genetic evolution.

ACKNOWLEDGEMENTS

We thank Dr. Gary Stormo and two anonymous reviewers for helpful comments on the earlier version of this manuscript. This study was supported by National Science Foundation of China (30930049) to D. T. and the Swiss National Science Foundation (31003A_125213) to H. A.

LITERATURE CITED

AKASHI, H., 1994 Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics **136:** 927-935.

AKASHI, H., 2001 Gene expression and molecular evolution. Curr Opin Genet Dev **11:** 660-666.

BARANOV, P. V., A. W. HAMMER, J. ZHOU, R. F. GESTELAND and J. F. ATKINS, 2005 Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. Genome Biol **6:** R25.

BOSSI, L., and J. R. RUTH, 1980 The influence of codon context on genetic code translation. Nature **286:** 123-127.

BOYCHEVA, S., G. CHKODROV and I. IVANOV, 2003 Codon pairs in the genome of Escherichia coli. Bioinformatics **19:** 987-998.

BOYER, J. C., N. A. YAMADA, C. N. ROQUES, S. B. HATCH, K. RIESS *et al.*, 2002 Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum Mol Genet **11:** 707-713.

BUCHAN, J. R., L. S. AUCOTT and I. STANSFIELD, 2006 tRNA properties help shape codon pair preferences in open reading frames. Nucleic Acids Res **34:** 1015-1027.

BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897-907.

CHEN, J. Q., Y. WU, H. YANG, J. BERGELSON, M. KREITMAN *et al.*, 2009 Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. Mol Biol Evol **26:** 1523-1531.

CRICK, F. H., L. BARNETT, S. BRENNER and R. J. WATTS-TOBIN, 1961 General nature of the genetic code for proteins. Nature **192:** 1227-1232.

DOS REIS, M., R. SAVVA and L. WERNISCH, 2004 Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res **32:** 5036-5044.

DU, J., T. GU, H. TIAN, H. ARAKI, Y. H. YANG *et al.*, 2008 Grouped nucleotide polymorphism: a major contributor to genetic variation in Arabidopsis. Gene **426:** 1-6.

FREELAND, S. J., and L. D. HURST, 1998 The genetic code is one in a million. J Mol Evol **47:** 238-248.

GATCHEL, J. R., and H. Y. ZOGHBI, 2005 Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet **6:** 743-755.

GRAGG, H., B. D. HARFE and S. JINKS-ROBERTSON, 2002 Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in Saccharomyces cerevisiae. Mol Cell Biol **22:** 8756-8762.

GREENE, C. N., and S. JINKS-ROBERTSON, 1997 Frameshift intermediates in homopolymer runs are removed efficiently by yeast mismatch repair proteins. Mol Cell Biol **17:** 2844-2850.

GROSJEAN, H., and W. FIERS, 1982 Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18:** 199-209.

GURVICH, O. L., P. V. BARANOV, J. ZHOU, A. W. HAMMER, R. F. GESTELAND *et al.*, 2003 Sequences that direct significant levels of frameshifting are frequent in coding regions of Escherichia coli. EMBO J **22:** 5941-5950.

GUTMAN, G. A., and G. W. HATFIELD, 1989 Nonrandom utilization of codon pairs in Escherichia coli. Proc Natl Acad Sci U S A **86:** 3699-3703.

HARFE, B. D., and S. JINKS-ROBERTSON, 2000 Sequence composition and context effects on the generation and repair of frameshift intermediates in mononucleotide runs in Saccharomyces cerevisiae. Genetics **156:** 571-578.

IRWIN, B., J. D. HECK and G. W. HATFIELD, 1995 Codon pair utilization biases influence translational elongation step times. J Biol Chem **270:** 22801-22806.

JUKES, T. H., and C. R. CANTOR, 1969 Mammalian protein metabolism, pp. 21-132 in *Evolution of protein molecules*. Academic Press, New York.

KROUTIL, L. C., K. REGISTER, K. BEBENEK and T. A. KUNKEL, 1996 Exonucleolytic proofreading during replication of repetitive DNA. Biochemistry **35:** 1046-1053.

LAWRENCE, J. G., and H. OCHMAN, 1998 Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci U S A **95:** 9413-9417.

LYNCH, M., and J. S. CONERY, 2003 The origins of genome complexity. Science **302:** 1401-1404.

MCLEAN, M. J., K. H. WOLFE and K. M. DEVINE, 1998 Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol **47:** 691-696.

MURGOLA, E. J., F. T. PAGEL and K. A. HIJAZI, 1984 Codon context effects in missense suppression. J Mol Biol **175:** 19-27.

MUTO, A., and S. OSAWA, 1987 The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A **84:** 166-169.

PLOTKIN, J. B., H. ROBINS and A. J. LEVINE, 2004 Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A **101:** 12588-12591.

ROBINSON, M., R. LILLEY, S. LITTLE, J. S. EMTAGE, G. YARRANTON *et al.*, 1984 Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res **12:** 6663-6671.

SAGHER, D., A. HSU and B. STRAUSS, 1999 Stabilization of the intermediate in frameshift mutation. Mutat Res **423:** 73-77.

SHELOBOLINA, E. S., H. A. VRIONIS, R. H. FINDLAY and D. R. LOVLEY, 2008 Geobacter uraniireducens sp. nov., isolated from subsurface sediment undergoing uranium bioremediation. Int J Syst Evol Microbiol **58:** 1075-1078.

STEARNS, D. M., M. YAZZIE, A. S. BRADLEY, V. H. CORYELL, J. T. SHELLEY *et al.*, 2005 Uranyl acetate induces hprt mutations and uranium-DNA adducts in Chinese hamster ovary EM9 cells. Mutagenesis **20:** 417-423.

STRAUSS, B. S., 1999 Frameshift mutation, microsatellites and mismatch repair. Mutat Res **437:** 195-203.

SUN, X., Y. ZHANG, S. YANG, J. Q. CHEN, B. HOHN *et al.*, 2008 Insertion DNA Promotes Ectopic Recombination during Meiosis in Arabidopsis. Mol Biol Evol **25:** 2079-2083.

TANIGUCHI, T., and C. WEISSMANN, 1978 Inhibition of Qbeta RNA 70S ribosome initiation complex formation by an oligonucleotide complementary to the 3' terminal region of E. coli 16S ribosomal RNA. Nature **275:** 770-772.

TATS, A., T. TENSON and M. REMM, 2008 Preferred and avoided codon pairs in three domains of life. BMC Genomics **9:** 463.

TIAN, D., Q. WANG, P. ZHANG, H. ARAKI, S. YANG *et al.*, 2008 Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature **455:** 105-108.

TRAN, H. T., J. D. KEEN, M. KRICKER, M. A. RESNICK and D. A. GORDENIN, 1997 Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. Mol Cell Biol **17:** 2859-2865.

WAGNER, L. A., R. B. WEISS, R. DRISCOLL, D. S. DUNN and R. F. GESTELAND, 1990 Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. Nucleic Acids Res **18:** 3529-3535.

YARUS, M., and L. S. FOLLEY, 1985 Sense codons are found in specific contexts. J Mol Biol **182:** 529-540.

ZHU, L., Q. WANG, P. TANG, H. ARAKI and D. TIAN, 2009 Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. Mol Biol Evol **26:** 2353-2361.

FIGURE LEGENDS

FIGURE 1. o/e ratios for all possible sense:sense codon pairs. Circles represent A/T pairs; triangles represent C/G pairs. The geometric mean of the o/e ratio for each type of sense:sense codon pairs is plotted with standard error.

FIGURE 2. Correlation between biased usage of C/G codon pairs and GC-content in bacteria genomes. A: The geometric mean of the o/e ratio of C/G pairs in bacteria group I with GC%<50%, group II with 50%<GC%<60% and group III with GC%>60%. The geometric mean of the o/e ratio for each type of synonymous codon pairs is plotted with standard error. B: Correlation between o/$e_{6C/G}$ and GC-content. C: Correlation between GC-content and the number of significantly underrepresented C/G codon pairs in individual genomes.

FIGURE 3. Comparison of mononucleotide runs in highly-conserved (white-bar) and less-conserved (shaded-bar) regions. (A) o/e_{6N} in 4 comparisons of eukaryote genomes. (B) Frequency of 6N codon pairs. The four comparisons are: *S. cerevisiae* & *S. paradoxus*, *D. melanogaster* & *D. yakuba*, *C. elegans* & *C. briggsae,* and *H. sapiens* & *M. musculus*). ** *P*<0.01.

TABLE S1 Eukaryotic genome data used in this study.

TABLE S2 Bacterium strains used in this study.

TABLE S3 Archaeal strains used in this study.

TABLE S4 Synonymous codon pairs with mononucleotide repeats.

TABLE S5 Examples of sense:sense codon pairs with mononucleotide repeats.

TABLE S6 The average nucleotide divergence (*D*) and o/e_{6N} in four comparisons of eukaryote genomes.

TABLE S7 o/e ratio of the synonymous codon pairs in eukaryote genomes.

TABLE S8 o/e ratio of the synonymous codon pairs in bacteria genomes.

TABLE S9 o/e ratio of the synonymous codon pairs in archaeal genomes.

TABLE S10 o/e ratio of the sense:sense codon pairs in eukaryote genomes.

TABLE S11 o/e ratio of the sense:sense codon pairs in bacteria genomes.

TABLE S12 o/e ratio of the sense:sense codon pairs in archaeal genomes.

FIGURE S1 Correlation between o/e ratio of A and T, C and G codon pairs in 13 eukaryotic (A) and 53 bacteria (B) and 68 archaea (C) genomes. Dashed line represents y=x.

FIGURE S2 Proportion of genomes with o/e > 1.0 for different number of mononucleotide runs.

FIGURE S3 o/e ratios for synonymous codon pairs.

FIGURE S4 o/e ratios for all possible sense:sense codon pairs after excluding the synonymous pairs.

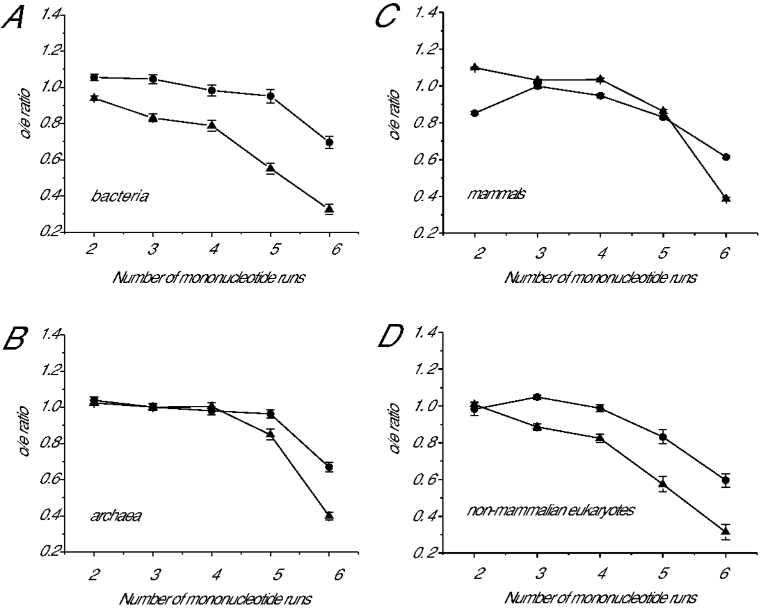
FIGURE S5 Correlation between usage of A/T codon pairs and GC-content in bacteria genomes.

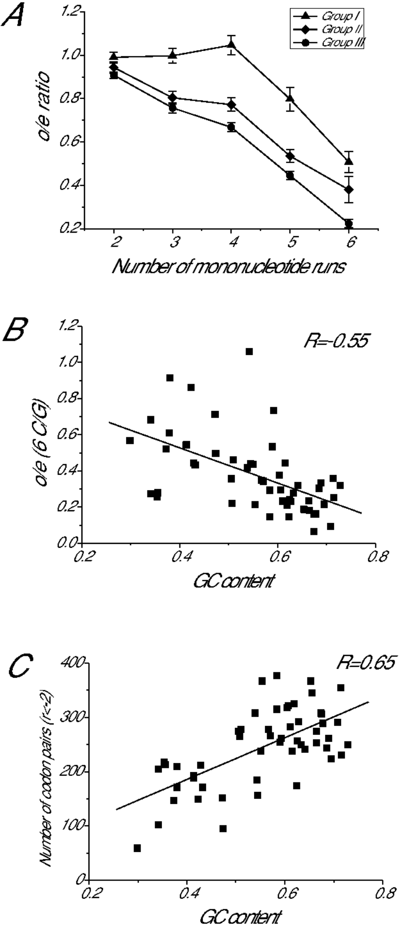
FIGURE S6 Correlation between biased usage of C/G codon pairs and GC-content in archaea genomes.

FIGURE S7 Correlation between usage of A/T codon pairs and GC-content in archaea genomes.

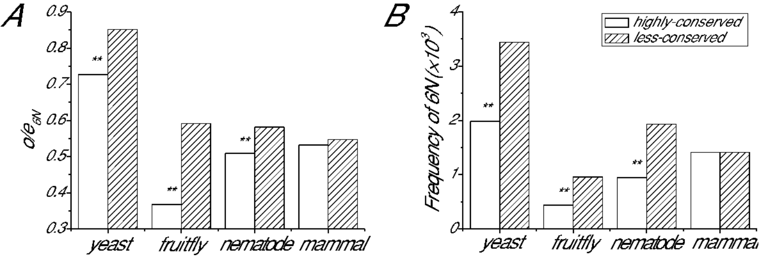
FIGURE S8 Correlation between biased usage of codon pairs and GC content in archaea genomes.

FIGURE S9 Correlation between biased usage of codon pairs and GC content in bacteria genomes.


A
bacteria
B
archaea
C
mammals
D
non-mammalian eukaryotes
o/e ratio
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Number of mononucleotide runs
2
3
4
5
6



A
1.2
1.0
0.8
0.6
0.4
0.2
o/e ratio
Group I
Group II
Group III
2
3
4
5
6
Number of mononucleotide runs
B
R=-0.55
1.2
1.0
0.8
0.6
0.4
0.2
0.0
o/e (6 C/G)
0.2
0.4
0.6
0.8
GC content
C
R=0.65
400
300
200
100
0
Number of codon pairs (r<-2)
0.2
0.4
0.6
0.8
GC content


A
0.9
0.8
0.7
0.6
0.5
0.4
0.3
o/e_{6N}
**
**
**
yeast
fruitfly
nematode
mammal
B
4
3
2
1
0
Frequency of 6N ($\times 10^3$)
highly-conserved
less-conserved
**
**
**
yeast
fruitfly
nematode
mamma