

# Compositional dynamics of guanine and cytosine content in prokaryotic genomes

Jianfei Hu<sup>a,b,1</sup>, Xiaoqian Zhao<sup>b,d,1</sup>, Zhang Zhang<sup>b,c,d,1</sup>, Jun Yu<sup>b,c,e,\*</sup>

<sup>a</sup> College of Life Sciences, Peking University, Beijing 100871, China

<sup>b</sup> Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China

<sup>c</sup> Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China

<sup>d</sup> Graduate School of Chinese Academy of Sciences, Beijing 100039, China

<sup>e</sup> James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310007, China

Received 3 November 2006; accepted 15 February 2007

Available online 6 March 2007

## Abstract

Nucleotide compositional analyses of disparities in genomic guanine and cytosine (gGC) content directly relate to the amino acid composition, through the union of the genetic code. Here we analyzed 229 prokaryotic genomes to address the intricate relationships between gGC, amino acids and their codons in the context of genes. First, we not only confirmed the universal rule that the average GC content at codon position 1 (GC1) is always higher than that at codon position 2 (GC2), but also extended the rule to show that it holds true even when codon-position-related GC contents are calculated on a per gene basis. The “GC1 > GC2 rule” is attributable essentially to a few dominant amino acids that have GC at one of these two codon positions or the intermediate-GC group of amino acids. Second, we found that gGC fluctuations were largely compensated for at the codon level, when codons belonging to high-GC and low-GC amino acid groups varied accordingly. Finally, we found that prokaryotic genes also have a GC content gradient (Gd) distributed along their transcripts. The gradients at three codon positions (Gd1, Gd2 and Gd3) all correlated with gGC in two different directions: Gd3 was positive, whereas the other two were negative.

© 2007 Elsevier Masson SAS. All rights reserved.

**Keywords:** Genomic GC content (gGC); Codon positions (GC1, GC2 and GC3); GC content gradient (Gd)

## 1. Introduction

The genomic guanine and cytosine (gGC) content of different species varies greatly, from 0.2 to 0.8, especially among prokaryotes [16,25]. However, for a particular species, its gGC tends to be unique [5,6]. For a given genome, gGC variations measured along a DNA strand can be defined as GC skews that are often used to determine replication origins or

termini of prokaryotic genomes [10,17]. GC content alterations reveal adaptive advantages at least in thermophilic prokaryotes, where the GC content of rRNA and tRNA genes exhibits a strong positive correlation with their optimal growth temperatures (OGTs) [2,8,18]. Another class of cellular mechanisms that evoke GC content variations is that of DNA repair, and some of their compositional signatures, especially at the transcript level, are biologically significant. For instance, such signatures have been observed as a negative gradient along the 5'-portion of a transcript, most notably in *Gramineae* genes [30,32] and later confirmed in mammalian genes [9]. A particular mechanism of DNA repair was believed to be the causative factor, i.e. transcription-coupled DNA repair (TCR) that is universal and is found in both prokaryotes and eukaryotes [26,27]. A usage gradient of codons and nucleotide

\* Corresponding author. Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Airport Industrial Zone B-6, Beijing 101300, China. Tel.: +86 10 80481455; fax: +86 10 80498676.

E-mail address: junyu@genomics.org.cn (J. Yu).

<sup>1</sup> These authors contributed equally to this work.

combinations had been previously noted in *Escherichia coli* and was related to the expressiveness of genes and translation efficiency [13].

In this study, we first collected 229 prokaryotic genomes from public databases and focused our analyses on correlations of gGC with other genomic parameters such as genome sizes, average GC content at three codon positions (GC1, GC2 and GC3) and amino acid compositions of protein-coding genes. We then looked at all individual codons and their relationships in order to pinpoint what codons of an amino acid with variable GC compositions are actually mitigating the pressure imposed by gGC. Finally, we identified the effect of gGC on GC compositional gradients of prokaryotic transcripts, and postulated that a causative factor in generating such gradients is the interplay of sequence errors made in replication and transcription-coupled DNA repair.

## 2. Materials and methods

The genome sequences and annotations of 229 prokaryotes were retrieved from GenBank (May 29, 2005) [19]. Open reading frames (ORFs) and related information were extracted from \*.gbk files with Perl script.

To compute dynamic changes in nucleotide composition along transcripts, ensemble averages across ORFs were calculated with a 51-bp window size and a 3-bp step. In most cases, we observed that the length of a sequence exhibiting Gd was less than 500 bp so that the gradient was defined only for the first 500 bases of an ORF. To ensure precise identification of codons, only ORFs possessing an ATG as start codon were included in our analysis, despite the fact that some ORFs in prokaryotes use ATA as a start codon.

We used the relative usage count (RUC) of codons and amino acids (relative to the expected usage count), rather than the frequency per hundred or thousand codons or amino acids to show biases. RUC values for codons and amino acids were calculated using the following two equations:

$$\text{RUC}(\text{codon}) = \frac{n(\text{codon}) \times 61}{n_T}$$

$$\text{RUC}(\text{aa}) = \sum_{\text{codon}} \text{RUC}(\text{codon})$$

where  $n$  (codon) is the number of a given codon for all ORFs of a bacterium and  $n_T$  is the total number of codons for all ORFs of a bacterium. RUC (aa) and RUC (codon) are relative usage counts of amino acids and their corresponding codons, respectively.

## 3. Results

### 3.1. The variable gGC among sequenced prokaryotes

The gGC among prokaryotes is known to vary broadly; in the current data set, it ranged from 0.225 to 0.721. For instance, *Wigglesworthia brevipalpis* has the lowest gGC [1],

and the highest gGC belongs to *Streptomyces coelicolo* [3]. Prokaryotic gGC diverges widely, even for genomes in the same genera and species. Our selective collection of 229 prokaryotic genomic sequences belonged to 114 genera and 184 species; within the data set, 47 genera had more than one genome sequenced. Despite the fact that the genomes within a genus had similar gGC, eight genera contained isolates whose gGC content varied over 10%, such as *Corynebacterium* and *Mycoplasma*. The most extreme case was *Prochlorococcus*, the smallest known oxygen-evolving autotroph, in which the difference in gGC between *Prochlorococcus MED4* and *Prochlorococcus MIT9313* was about 0.2 [22]. The two ecotypes were classified together as a single species based solely on the similarity of their rRNA sequences (97%) [12]; the sequence similarity between the two genomes is so low that it is less than 5%. It was speculated that they either originally went through a massive divergence process from the same species [22], or limited horizontal gene transfer involving rRNA genes, which was shown to be possible in the laboratory [21]. Although gGC in general is distributed rather evenly among prokaryotic genomes [24,29], it can be affected by many factors, such as insertions of bacteriophage lysogens and transposons [14]. We also plotted the genome size of these prokaryotic genomes as a function of gGC, and hypothesized that there was weak correlation between these two major genome parameters, with a correlation coefficient of 0.54 (Fig. 1a). This did not contradict a previous observation [4], but rather cast doubt on it; the distributions may show ranges or boundaries rather than a linear curve. In addition, the less significant correlation coefficient was presumed to be partly due to limited data points (229 genomes). Genome sequences of a few *Archaea* species behaved in a similar way as *Eubacteria* species when their gGC varied toward the extremes. This suggested that genome size may, to a certain extent, serve as a minor variable in balancing mutation pressures in addition to accommodating more genes.

### 3.2. GC content variability at three codon positions

The average GC content at each codon position among prokaryotic genomes varies in certain unique ways [28]. In our collection, GC2 is the least variable position; it changes from 0.268 to 0.512. The next in line is GC1, ranging from 0.307 to 0.724. GC3 alters the most, from 0.087 to 0.933 (Fig. 1b). GC1, GC2 and GC3 are all strongly correlated with gGC, with very significant correlation coefficients of 0.97, 0.92, and 0.99, respectively. The slopes for GC1, GC2 and GC3 appeared different, with values of 0.73, 0.44 and 1.80, respectively, showing different degrees of response to gGC alterations. The GC3 distribution intersected with GC1 and GC2 at values of 0.523 and 0.39. Our results strengthened the rule that GC1 is always higher than GC2 with an adequate data support. We also plotted GC1 against GC2 for all ORFs in our collection to see if this rule extends to single genes (some are predicted ORFs). Among 659,773 ORFs, there were only 17,840 ORFs whose GC1 was lower than GC2 (<3%). We extracted these ORFs and inspected them manually and found

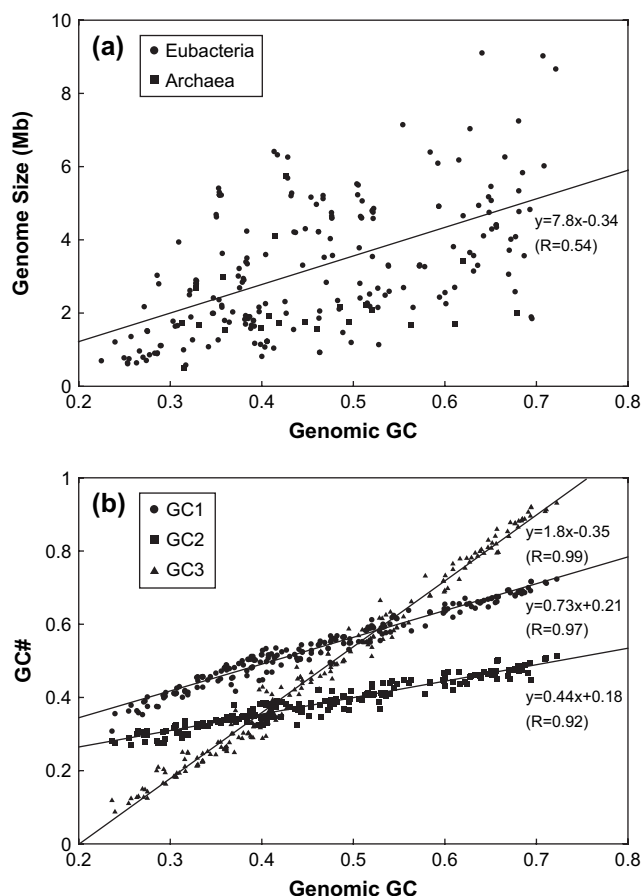


Fig. 1. Correlation of gGC with genome size (a) and codon positions, GC1, GC2 and GC3 (b). Genome sizes of individual prokaryotes (solid squares and circles depict *Archaea* and *Eubacteria* species, respectively) correlated weakly with gGC ( $R = 0.54$ ). GC1, GC2 and GC3 are indicated by solid circles, squares and triangles, respectively; they are highly correlated with gGC, with correlation coefficients of 0.97, 0.92, and 0.99, respectively. GC3 was most strongly altered, from 0.087 to 0.933, with a variance of 0.052. The slopes for three codon positions ranked as cp3 > cp1 > cp2, with values of 1.80, 0.73 and 0.44, respectively. The near linear distribution of GC3 intersected with those of GC2 and GC1 when gGC shifted from low to high. GC1 was always greater than GC2, with a nearly constant odds ratio fluctuating around 0.72 and a variance of just 0.002.

that over 85% of these ORFs were categorized as “hypothetical”, ‘putative’, ‘probable’ and ‘possible’ ORFs. It was clear that these predicted ORFs were largely intergenic sequences or degraded pseudogenes. Therefore, this “GC1 > GC2 rule” provides an excellent filter for prokaryotic gene annotation. However, we should be cautious, since a minuscule number of the putative ORFs might be activated, becoming real genes when proven empirically.

GC2 is always lower than 0.5 even when gGC rises above 0.7. Taking *E. coli* K12 (gGC = 0.508) as an example, we found that nearly 98% of its genes have a GC2 lower than 0.5, with an overall average of 0.406. To examine whether leader peptides (or signal peptides) with a higher content of hydrophobic amino acids may lead to an artifactual relationship with GC2 (hydrophobic amino acids have GC-rich codons), we analyzed ORFs encoding membrane proteins from *E. coli*. The result demonstrated that GC2 of the membrane

proteins (ORFs) is 0.414, slightly higher than the overall average, and this difference is statistically not significant,  $P = 0.171$  (Wilcoxon rank sum test).  $P$  is the probability there is no difference between the two groups of data.  $P < 0.05$  indicates that the difference is significant and  $P < 0.01$  indicates that the difference is very significant.

The trends among GC1, GC2 and GC3 can be explained by the probability of non-synonymous mutations at three codon positions. There are 64 codons and 192 ( $64 \times 3$ ) possible substitutions at each codon position; the number of substitutions that do not change amino acids at cp1 (codon position 1), cp2 (codon position 2) and cp3 (codon position 3) are 8, 2 and 128, respectively. Although only 33% ( $1 - \{128/192\}$ ) of the nucleotide substitutions at cp3 theoretically change amino acids, the observed value is often much lower, since most of the substitutions are transitional changes (mutations within purines or pyrimidines). For amino acids with an even number of codons, 17 out of 20 (except for ATG, ATA and TGG; Table 1) transitions at cp3 do not change amino acids [15,20,31]. In contrast, nucleotide substitutions at cp1 and cp2 frequently change amino acids.

### 3.3. Correlation analysis between gGC and compositional variations in amino acids and their codons

The set of 20 amino acids can be classified into three groups according to their GC content at cp1 and cp2 [7,11,23]. The amino acids in the high-GC group and low-GC group have both cp1 and cp2 occupied by GC or AT (Table S1). In the intermediate-GC group, only one of the two invariable nucleotides in a codon has G or C. The content of high- and low-GC amino acids strongly correlates with gGC; in other words, gGC

Table 1  
Eleven amino acids related to variations in  $\Delta$ GC

Amino acid	$y = bx + a$	Expected RUC	Observed RUC	$\Delta$ RUC
<i>Intermediate-GC-(I)</i>				
Asp	$y = 0.43x + 2.90$	2	3.1	1.1 <sup>b</sup>
Glu	$y = -1.9x + 4.60$	2	3.73	1.73 <sup>b</sup>
Gln	$y = -0.02x + 2.23$	2	2.22	0.22
His	$y = 0.8x + 0.93$	2	1.29	-0.71
Leu <sup>a</sup>	$y = 9.47x - 0.58$	4	3.69	-0.31
Val	$y = 2.62x + 2.92$	4	4.12	0.12
<i>Intermediate-GC-(II)</i>				
Arg <sup>a</sup>	$y = -2.06x + 1.86$	2	0.93	-1.07 <sup>b</sup>
Cys	$y = 0.48x + 0.48$	2	0.7	-1.3 <sup>b</sup>
Ser	$y = -1.51x + 4.64$	6	3.95	-2.05 <sup>b</sup>
Thr	$y = 0.42x + 3.02$	4	3.22	-0.78
Trp	$y = 1.33x + 0.09$	1	0.69	-0.31

Almost all of them are intermediate-GC amino acids. Five of them, Asp, Glu, Arg, Cys and Ser, contribute significantly to the “GC1 > GC2 rule”; their  $\Delta$ RUCs are all larger than 1.

Note that the slope values of leucine and valine are both positive and large enough to have a stronger impact on  $\Delta$ GC.

<sup>a</sup> For arginines and leucines, only those codons belonging to the intermediate-GC group are considered.

<sup>b</sup> Five amino acids contribute decisively to the GC1 > GC2 rule, whose absolute values of  $\Delta$ RUC are all greater than 1.

changes in the mean overall amino acid usage of a genome favor the high-GC group. In most cases, the content of intermediate-GC amino acids correlates weakly with gGC. Two exceptions were found, in the cases of histidines and valines. The increased correlation of these two amino acids with GC richness was expected, since ATN and TTN (N stands for all four nucleotides) both encode hydrophobic amino acids that are readily convertible to valines, especially from isoleucine (ATY; Y stands for pyrimidines), permitted by their approximate chemical properties. In the case of histidines, the likely conversion is from tyrosines (TAY), relating to both hydrophobicity and structure accommodation (histidine vs. benzene). This was confirmed by reduced representation of the TAC codon, one of the two codons encoding tyrosines. Interestingly, the most dominant amino acids in the three groups are all hydrophobic amino acids, alanines, isoleucines and leucines for the high-, low-GC and intermediate groups, respectively. These amino acids provide essential chemical balances among proteins when gGC changes from low to high, driven by mutation biases. An overall trend becomes clear: when gGC changes to its extremities, the proportions of amino acids in high- and low-GC groups approach minimal divergence collectively. This trend has strong implications for the relationship between gGC changes and protein structures. Moderate gGC is best for amino acid diversity that enables proteins to build up complexity and achieve robustness for functions. In addition, as gGC changes over time, amino acids positioned in the protein also have a chance to change accordingly, albeit at a much slower pace.

With increasing gGC, variations in codons and their encoded amino acids are rather distinct. Among codons, the most GC-content-sensitive position is cp3 [7,11,23]. If the nucleotide at cp3 is G or C, the codon composition is positively correlated with gGC. The opposite trend was observed when the nucleotide at cp3 was A or T. It is clear that cp3 provides an essential GC balance within codons of a given amino acid. For example, phenylalanine is a low-GC amino acid (its codons at cp1 and cp2 are both T) so that the content of phenylalanine gradually decreases when gGC increases, while its two codons gradually shift from TTT to TTC.

#### 3.4. Over- and underuse of five intermediate-GC amino acids leads to the “GC1 > GC2 rule”

High- and low-GC amino acids do not contribute to the rule, since they have GC or AT at both cp1 and cp2, respectively. What contributes to the rule is amino acids within the intermediate-GC group. These amino acids can be further classified into two subgroups: intermediate-GC-(I) and -(II). In the first subgroup, the nucleotides at cp1 are G or C and the nucleotides at cp2 are A or T; overuse of these amino acids will contribute to the rule. In the second subgroup, nucleotides at cp1 are A or T and nucleotides at cp2 are G or C; underuse of these amino acids will contribute to the rule. The over- and underuse of amino acids are evaluated by the difference in observed RUC and expected RUC (see Section 2),  $\Delta$ RUC. The observed RUC was computed from the ensemble average of all 229 genomes.

Five amino acids have absolute values of  $\Delta$ RUC higher than 1, despite the fact that most of the  $\Delta$ RUC values of the intermediate-GC group vary to some extent (Table 1). Two of them are charged acidic amino acids, aspartic and glutamic acids. They are the intermediate-GC-(I) amino acids, and their electrolytic effects in proteins appeared neutralized, at least in part, by two basic amino acids, arginine (fourfold degenerate set, CGN only) and lysine (AAR), where lysine belongs to the low-GC group and arginine in part belongs to the high-GC group. The extreme  $\Delta$ RUC value went to serine, an intermediate-GC-(II) amino acid, arguably the most versatile amino acid of all. First, it is uncharged, polar and hydrophilic; it has a residue volume of 89.0 [33] that is very close to 88.6 of the alanine, one of the high-GC group amino acids. Second, it is one of the amino acids possessing six codons, together with leucine and arginine. What is unique about serine is its six codons, TCN and AGY; all have G or C at cp2, and its RUC value was found to be extremely underrepresented ( $\Delta$ RUC = -2.05). This extreme underuse certainly contributes most to the “GC1 > GC2 rule”. The second significant rule-contributor is the duplex codon of arginine, AGR (it has G at cp2), which was underused in our usage counts. The third case concerns cysteine, an intermediate-GC-(II) amino acid. Aside from the fact that it forms disulfide bonds in proteins, so that its abundance must be controlled functionally, it is encoded by TGY and its underuse certainly contributes to the rule. Finally, there is one stop codon (TGA) within the subgroup. It has T at cp1 and G at cp2, and even though its count should lead to a reduction in theoretical value of GC2, it appears at most once per ORF. As a result, its contribution to the rule is negligible. Therefore, we believe that the over- and underuse of these five amino acids lead to the “GC1 > GC2 rule”. We would like to emphasize that we only included AGR for arginine and CTN for leucine in this analysis.

The difference between GC1 and GC2 ( $\Delta$ GC) is not constant, ranging from less than 5% for low-GC genomes to more than 20% for high-GC genomes. When gGC increases, amino acid content variations in the intermediate-GC group, reflected as the slope of the linear-regression formula, is very limited; most of the slope values are less than 2 (Table 1). Therefore, these small variations cannot explain the large differences ( $\Delta$ GC). When looking into individual amino acids that contribute to  $\Delta$ GC, we noticed a single amino acid—leucine (although valine with a slope value of 2.62 may contribute to this effect to some extent). Leucine has six codons, CTN (CTA, CTT, CTG and CTC) and TTR (TTA and CTG); four of them belong to intermediate-GC-(I). When gGC rises rapidly, CTG content also increases dramatically, with a slope value of 8.82, leading to a fast increase in  $\Delta$ GC (Table S1). The elevated CTG is largely compensated for by TTA, another codon of leucine with slope of -9.0. At the same time, the RUC of leucine remains nearly constant.

#### 3.5. The presence of a transcript-based gradient in prokaryotic genomes

The GC content of genes exhibits a gradient along the transcription direction when GC1, GC2 and GC3 are plotted from

the start codon to the full length. The strongest gradient was first discovered in *Gramineae* genes [30] and later noted in genes of warm-blooded vertebrates [9].

Our initial analysis of the gradient along the direction of RNA transcription was limited to the first 1 kb of all genes, since the average length of prokaryotic genes is around 1 kb. When plotting the gradient, we set the start codon as the origin. All ORFs were aligned from their origins and the average across ORFs was calculated with a 51-bp sliding window in a 3-bp increment. We only included ORFs with clear annotations and ATG as start codon, despite the fact that some ORFs in prokaryotes use ATA as start codon. We also set an upper boundary at which the ORF length was limited to greater than 500 bp in length to eliminate the influence of short ORFs. After observing all the plots, we found that the GC gradient in prokaryotic genes was different from those of *Gramineae* plants. The prokaryotic GC gradient in most cases lasted around 300 bp, shorter than that in the rice genome (around 1 kb). Furthermore, the gradient of prokaryotic genes varied among different genomes and only a few of them were found to have average gradient values greater than 0.2, whereas the value for rice was 0.225.

The gradient at three codon positions varied differently among the genomes in our collection when their average gGC fluctuated. According to their gGC, prokaryotic genomes could be classified into three basic categories: low-GC (<0.38), intermediate-GC (0.38–0.55) and high-GC (>0.55) groups. Among the low-GC genomes, GC gradients at cp1 and cp2 (Gd1 and Gd2) were mostly positive and at cp3 (Gd3) they were mostly negative. Among the intermediate-GC genomes, Gd1 and Gd3 were positive, but Gd2 was negative. The high-GC genomes had an opposite trend from the low-GC group genomes: Gd1 and Gd2 were negative and Gd3 was positive (Fig. 2). This indicates that the gradient at three codon positions was strongly affected by gGC. We then plotted the distribution of Gd1, Gd2 and Gd3 against gGC to show that Gd1, Gd2 and Gd3 all correlated strongly with gGC, with absolute values of correlation coefficient larger than 0.7 (Fig. 3). Since the increase in Gd3 ( $R = 0.84$ ) was accompanied by a decrease in Gd1 ( $R = -0.57$ ) and Gd2 ( $R = -0.75$ ), the integral gradient was virtually indistinguishable.

The gradient appeared universal to every prokaryotic genome, but it differed according to gGC and may not be obvious for all genomes, depending on how it is viewed; some may not be easily measured when the gGC varies from 35% to 50%. Our analysis showed that the gradient effect is neither an artifact of 5'-untranslated sequences nor signal sequence interference, although consecutive hydrophobic amino acids (such as transmembrane domains) may make a minor contribution (data not shown).

#### 4. Discussion

We re-examined the plasticity of prokaryotic gGC, leveraging on the enormous public data from many genomic sequencing efforts. The gGC is primarily governed by mutations

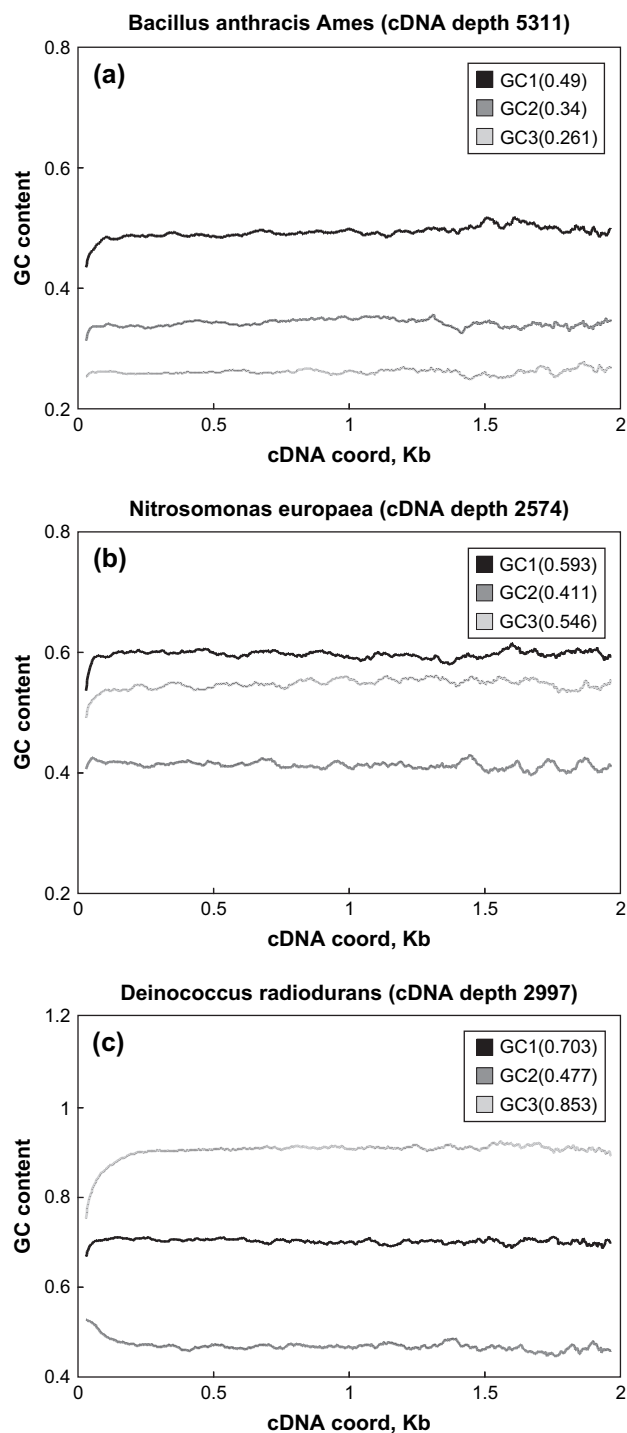


Fig. 2. Examples of gradient in prokaryotic genomes in low-GC, intermediate-GC and high-GC groups. GC content distributions at three codon positions are coded as black, dark gray, and light gray for GC1, GC2 and GC3, respectively. Representatives for the three GC groups are: (a) *Bacillus anthracis* Ames (gGC = 0.3537), (b) *Nitrosomonas europaea* (gGC = 0.5071) and (c) *Deinococcus radiodurans* (gGC = 0.67). GC content distributions at three codon positions were plotted as a function of ORF positions relative to the start codon and averaged over the length of cDNAs with a 51-bp sliding window. The gradient is computed from the first 300 bp only.

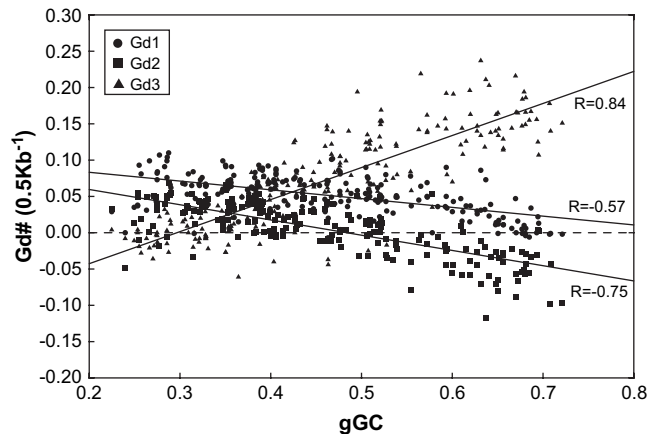


Fig. 3. Correlation of Gd with gGC. Solid circles, squares and triangles represent Gd1, Gd2 and Gd3, respectively. Gd values at three codon positions all correlate with gGC. The correlation coefficient of Gd3 is positive ( $R = 0.84$ ), but those of Gd1 ( $-0.57$ ) and Gd2 ( $-0.75$ ) are both negative. Since Gd3 is sensitive to gGC changes, but not very sensitive to amino acid changes, it is expected to correlate rather positively. The negative correlation of Gd1 and Gd2 indicates that GC gradients can also become more pronounced as gGC moves toward the high-GC end. Increased Gd1 and Gd2 suggest strong amino acid variations at the 5'-portions of the proteins as gGC increases.

originating from replication and repair errors that, in turn, provide an indispensable basis for biological diversity for these most primitive life forms to achieve the best fitness. Indeed, we noted that the genome size of prokaryotes correlates weakly with gGC changes, and the reason for such correlation is currently unknown. We also confirmed a previous observation and showed, with an adequate amount of data, that GC1, GC2 and GC3 are almost perfectly correlated with gGC. The correlation implies that gGC variations are the major force in generating nucleotide compositional dynamics among prokaryotic genomes. The gGC variability at cp1 and cp2, indicated by GC1 and GC2, certainly implies that amino acid compositional changes coincide with gGC fluctuations, providing the basis (though to a lesser extent due to the balancing power of the codon arrangement) for amino acid compositional dynamics.

We went one step further to pinpoint the detailed compositional signatures of such a correlation and possible causative factors. One rule is obvious: GC1 is always greater than GC2. To confirm this at the gene level, we not only plotted all ORFs, but also inspected the “rule-breakers” and concluded that they are most likely non-functional ORFs and have lost protein-coding characteristics. What previous analyses have not done is to simultaneously examine contributing factors at both the amino acid and their codon levels. We found that gGC-sensitive amino acids are those of high- and low-GC groups. The intermediate-GC group is rather gGC-insensitive in general, since codons of this group have either GC or AT at cp1 and cp2, but not both. Furthermore, the gGC-sensitive groups have well-balanced amino acids with respect to their chemical properties when gGC changes approximately from 0.2 to 0.7. First, hydrophobic isoleucines and leucines (UURs) in the low-GC group are replaced by alanines and glycines in the high-GC group. Second, aromatic tryptophan

(encoded by a single codon, TGG, so that it is considered to be one of the high-GC amino acids) is compensated for by two other aromatic amino acids, tyrosine and phenylalanine, as well as histidine, to a lesser extent. The contribution of histidine can be verified by its increased abundance when gGC increases. Third, the basic amino acids, arginines (CGN) and lysines (AAR), are also self-balanced between the two groups as gGC varies. Finally, the intermediate-GC, the largest group, serves as GC content “buffers” to cope with gGC changes, providing vital flexibility for achieving greater compositional plasticity. For instance, the electrolytic charges appeared adjusted between the GC-sensitive groups and the GC-tolerant group (the intermediate-GC amino acids); all basic amino acids are in the former and all acidic amino acids are in the latter. We concluded that the compositional dynamics of prokaryotes is mainly achieved at the codon level, governed by the intricate relationship between their nucleotide compositions. We argue that this intricate relationship is a natural commandment which has evolved over billions of years and is followed primarily by prokaryotes since its fixation. Although exceptions do exist, its power extends to eukaryotes, since compartmentalization reduced the mutation pressure (for example, by separating replication and translation in different compartments and regulating replication and transcription differently in cell cycles) to become largely indirect among genomes of these biologically more complex organisms.

We demonstrate that the “GC1 > GC2 rule” is primarily contributed to dominant amino acids. The major contributors are five amino acids (or a subset of their codons, such as AGR of arginine) of the intermediate-GC group. Their over- or underuse plays a decisive role in setting the rule. Since they are not as sensitive as the other two groups to gGC changes, the selective usage of these amino acids mainly reflects natural selection acting on functions and structures of proteins, governed by chemical properties as well as topological characteristics. Therefore, the mutation pressure is relieved at different codon positions and coped within an orderly way by the intrinsic relationship among codons and thus their encoded amino acids. The first “defense line” to gGC changes is cp3, making them largely synonymous. The second “defense line” is the distinct yet related chemical properties of amino acids, providing compensations and tolerances by their relatedness through codon arrangement. The third “defense line” is amino acids of the intermediate-GC group, where the abundance of amino acids varies to accommodate both gGC variation and protein function (or in other words, mutation drive against natural selection). A current conjecture for explaining the mechanistic foundation of these mechanisms relies on evolutionary doctrines—the universal codon arrangement is a product of the genomes’ compositional variations that are selected based on fitness of early organisms, especially those which built their biological processes and mechanisms upon a single cellular compartment (largely microbes). For instance, alanine and glycine are both high-GC amino acids and have four codons. The slope of GC correlation to gGC changes for alanine is 11.83, about twofold greater than that of glycines (5.4). We

believe that alanines are more amenable to being converted to serines than glycines when residue size becomes a decisive factor, since the residue value of alanines is identical to that of serines, but that of glycines is smaller. This serine to alanine conversion is best seen when dominant amino acids of the *Arabidopsis* and rice genomes were compared; serine and alanine are the dominant amino acids in *Arabidopsis* and rice, respectively [10]. In addition, we separately inspected the differences between biological classification schemes such as Archaea vs. Eubacteria, thermophiles vs. mesophiles and halobacteria vs. others, and no obvious relationship was detected.

The gradient effect was first found in *Gramineae* (grass family) genes [30]. It has been suggested that it may correlate with transcription-related mutation bias and translation-related selection. The negative trend of the *Gramineae*-specific GC gradient is believed to be a signature of repair enzymes that tend to make errors toward GC-rich, as opposed to replication that is less biased between transcribed and non-transcribed sequences. When the repair process aborts or stalls more frequently than transcription itself, a compositional gradient may be generated along the transcript. In the case of prokaryotes, a similar mechanism may be at work. The positive gradients found in Gd3 suggested that DNA repair enzymes in these organisms make repair errors toward AT-rich, in contrast to *Gramineae* species and warm-blooded vertebrates. This notion is supported by two facts. First, as gGC increases, suggesting that the replication system is making errors toward GC-rich, the gradient effect becomes strong and more obvious, thus showing a positive correlation. Second, the differential effects between Gd1 and Gd2 indicate that Gd1 is more permissive at reducing the GC content of cp1 due to repair errors toward AT-rich. This argument is supported by the codon relationship among codon groups, where several changes either involve the same amino acids (arginine: CGR to AGR; leucine: CTR to TTR) or occur between amino acids that have very similar chemical properties (such as GTY to ATY for valine to isoleucine in hydrophobicity; GGY to AGY for glycine to serine in residue volume). In the case of Gd2, changes from GC to AT are difficult, since most of them are non-synonymous so that the biases we have seen are arguably to balance mutation forces. The gGC increase results in stronger replication-driven mutations toward GC-rich, whereas TCR-driven mutations toward AT-rich appear strong in upstream sequences and weaken toward downstream sequences of a transcript; Gd2 thus becomes negative.

Compositional dynamics of prokaryotic genomes are measurable at both the nucleotide and amino acid levels. They are connected in a very intricate yet ultimately explainable way by the genetic codes that are arranged in a logical way related to the biochemical characteristics of the amino acids. The amino acid dynamics in a functional context of proteins keeps these primitive yet powerful organisms alive and evolving for billions of years. The overall poorer protein sequence conservations among prokaryotes than eukaryotes, as well as many better conserved protein domains over the rest of the sequences, all reflect such a mechanistic relationship.

## Acknowledgements

We thank three anonymous reviewers for their constructive comments on this manuscript. This work is supported by grants from the Chinese Academy of Sciences (KSCX2-SW-223), Chinese Natural Science Foundation (30270748), and Ministry of Sciences and Technology (2005AA235110) awarded to JY.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.resmic.2007.02.007.

## References

- [1] L. Akman, A. Yamashita, H. Watanabe, K. Oshima, T. Shiba, M. Hattori, S. Aksoy, Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*, Nat. Genet. 32 (2002) 402–407.
- [2] Q. Bao, Y. Tian, W. Li, Z. Xu, Z. Xuan, S. Hu, W. Dong, J. Yang, Y. Chen, Y. Xue, Y. Xu, X. Lai, L. Huang, X. Dong, Y. Ma, L. Ling, H. Tan, R. Chen, J. Wang, J. Yu, H. Yang, A complete sequence of the *T. tengcongensis* genome, Genome Res. 12 (2002) 689–700.
- [3] S.D. Bentley, K.F. Chater, A.M. Cerdeno-Tarraga, G.L. Challis, N.R. Thomson, K.D. James, D.E. Harris, M.A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C.W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabinowitsch, M.A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B.G. Barrell, J. Parkhill, D.A. Hopwood, Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2), Nature 417 (2002) 141–147.
- [4] S.D. Bentley, J. Parkhill, Comparative genomic structure of prokaryotes, Annu. Rev. Genet. 38 (2004) 771–792.
- [5] E. Chargaff, Structure and function of nucleic acids as cell constituents, Fed. Proc. 10 (1951) 654–659.
- [6] E. Chargaff, How genetics got a chemical education, Ann. N.Y. Acad. Sci. 325 (1979) 344–360.
- [7] P.G. Foster, L.S. Jermini, D.A. Hickey, Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria, J. Mol. Evol. 44 (1997) 282–288.
- [8] N. Galtier, N. Tourasse, M. Gouy, A nonhyperthermophilic common ancestor to extant life forms, Science 283 (1999) 220–221.
- [9] P. Green, B. Ewing, W. Miller, P.J. Thomas, E.D. Green, Transcription-associated mutational asymmetry in mammalian evolution, Nat. Genet. 33 (2003) 514–517.
- [10] A. Grigoriev, Analyzing genomes with cumulative skew diagrams, Nucleic Acids Res. 26 (1998) 2286–2290.
- [11] X. Gu, D. Hewett-Emmett, W.H. Li, Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria, Genetica 102–103 (1998) 383–391.
- [12] A. Hagstrom, T. Pommier, F. Rohwer, K. Simu, W. Stolte, D. Svensson, U.L. Zweifel, Use of 16S ribosomal DNA for delineation of marine bacterioplankton species, Appl. Environ. Microbiol. 68 (2002) 3628–3633.
- [13] S.D. Hooper, O.G. Berg, Gradients in nucleotide and codon usage along *Escherichia coli* genes, Nucleic Acids Res. 28 (2000) 3517–3523.
- [14] F. Kunst, N. Ogasawara, I. Moszer, A.M. Albertini, G. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S.C. Brignell, S. Bron, S. Brouillet, C.V. Bruschi, B. Caldwell, V. Capuano, N.M. Carter, S.K. Choi, J.J. Codani, I.F. Connerton, A. Danchin, et al., The complete

- genome sequence of the gram-positive bacterium *Bacillus subtilis*, Nature 390 (1997) 249–256.
- [15] W.H. Li, Unbiased estimation of the rates of synonymous and nonsynonymous substitution, J. Mol. Evol. 36 (1993) 96–99.
- [16] J.R. Lobry, Properties of a general model of DNA evolution under non-strand-bias conditions, J. Mol. Evol. 40 (1995) 326–330.
- [17] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (1996) 660–665.
- [18] R. Musto, M.G. Bigotti, C. Travaglini-Allocatelli, M. Brunori, F. Cutruzzola, Folding of *Aplysia limacina* apomyoglobin involves an intermediate in common with other evolutionarily distant globins, Biochemistry 43 (2004) 230–236.
- [19] NCBI. <<ftp.ncbi.nlm.nih.gov/genomes/Bacteria/>>.
- [20] M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, Mol. Biol. Evol. 3 (1986) 418–426.
- [21] M. Nomura, Engineering of bacterial ribosomes: replacement of all seven *Escherichia coli* rRNA operons by a single plasmid-encoded operon, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 1820–1822.
- [22] G. Rocap, F.W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N.A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W.R. Hess, Z.I. Johnson, M. Land, D. Lindell, A.F. Post, W. Regala, M. Shah, S.L. Shaw, C. Steglich, M.B. Sullivan, C.S. Ting, A. Tolonen, E.A. Webb, E.R. Zinser, S.W. Chisholm, Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation, Nature 424 (2003) 1042–1047.
- [23] G.A. Singer, D.A. Hickey, Nucleotide bias causes a genomewide bias in the amino acid composition of proteins, Mol. Biol. Evol. 17 (2000) 1581–1588.
- [24] N. Sueoka, Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein, Proc. Natl. Acad. Sci. U.S.A. 47 (1961) 1141–1149.
- [25] N. Sueoka, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, J. Mol. Evol. 40 (1995) 318–325.
- [26] J.Q. Svejstrup, Mechanisms of transcription-coupled DNA repair, Nat. Rev. Mol. Cell Biol. 3 (2002) 21–29.
- [27] K.S. Sweder, R.A. Verhage, D.J. Crowley, G.F. Crouse, J. Brouwer, P.C. Hanawalt, Mismatch repair mutants in yeast are not defective in transcription-coupled DNA repair of UV-induced DNA damage, Genetics 143 (1996) 1127–1135.
- [28] A. Wada, A. Suyama, R. Hanai, Phenomenological theory of GC/AT pressure on DNA base composition, J. Mol. Evol. 32 (1991) 374–378.
- [29] A. Wada, H. Tachibana, O. Gotoh, M. Takanami, Long range homogeneity of physical stability in double-stranded DNA, Nature 263 (1976) 439–440.
- [30] G.K. Wong, J. Wang, L. Tao, J. Tan, J. Zhang, D.A. Passey, J. Yu, Compositional gradients in Gramineae genes, Genome Res. 12 (2002) 851–856.
- [31] Z. Yang, R. Nielsen, Synonymous and nonsynonymous rate variation in nuclear genes of mammals, J. Mol. Evol. 46 (1998) 409–418.
- [32] J. Yu, S. Hu, J. Wang, G.K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, H. Yang, A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), Science 296 (2002) 79–92.
- [33] A.A. Zamyatnin, Protein volume in solution, Prog. Biophys. Mol. Biol. 24 (1972) 107–123.