FEBS Letters 581 (2007) 1015–1021

Evidence of selectively driven codon usage in rice: Implications for GC content evolution of *Gramineae* genes

Xingyi Guo[a], Jiandong Bao[a], Longjiang Fan[a,b,*]

[a] *Institute of Bioinformatics, Zhejiang University, Hangzhou 310029, China*
[b] *Institute of Crop Science, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China*

Received 10 December 2006; revised 27 January 2007; accepted 31 January 2007

Available online 8 February 2007

Edited by Takashi Gojobori

Abstract **Two gene classes characterized by high and low GC content have been found in rice and other cereals, but not dicot genomes. We used paralogs with high and low GC contents in rice and found: (a) a greater increase in GC content at exonic fourfold-redundant sites than at flanking introns; (b) with reference to their orthologs in *Arabidopsis*, most substitution sites between the two kinds of paralogs are found at 2- and 4-degenerate sites with a T → C mode, while A → C and A → G play major roles at 0-degenerate sites; and (c) high-GC genes have greater bias and codon usage is skewed toward codons that are preferred in highly expressed genes. We believe this is strong evidence for selectively driven codon usage in rice. Another cereal, maize, also showed the same trend as in rice. This represents a potential evolutionary process for the origin of genes with a high GC content in rice and other cereals.**

Keywords: Codon usage; Selection; GC content evolution; *Oryza sativa*

1. Introduction

Large-scale analysis of the GC-content distribution at the gene level in the *Gramineae* (grass) genome has revealed two gene classes, high-GC or GC-rich genes and low-GC or AT-rich genes, which have been found in monocot, but not dicot genomes [1,2]. This large-scale variation in base composition was discovered in the bovine genome over 30 years ago by Bernardi and colleagues [3]. Based on current genomic sequences, wide differences in GC content have been observed both between entire genomes and among genes within genomes [4]. For example, mammalian genomes can be divided into discrete blocks with distinct GC contents that occur over scales of hundreds of kilobases to megabases, the so-called isochors [5].

There has been long-standing interest in explaining why there is such large-scale variation in base composition along chromosomes. It has been suggested that this variation could be a consequence of mutation bias, natural selection, or biased gene conversion [6]. These hypotheses can be grouped into two categories according to whether or not natural selection is involved [5]. The important observation by D'Onofrio et al. [7] that the GC content at exonic fourfold-redundant sites (GC4) tends to be greater than that at flanking introns (GCi) has been put forward to distinguish between selection and neutral hypotheses [1,6,8]. This observation is consistent with a selection model according to Eyre-Walker [6]. Under the supposition of neutral evolution and base composition of a sequence maintained solely by mutation bias, mutations equally affect introns and exons, and the base composition should be expected to be equal in all silent sites. For example, in rodents, divergence is approximately equal at fourfold-degenerate sites in codons and intronic sequences [8]. Based on substitution at synonymous and non-synonymous sites, and on GC content along the direction of transcription, mutational bias or natural selection has been proposed to explain the evolutionary divergent pattern of GC content in rice (*Oryza sativa*) [9–11]. However, no conclusive evidence has been presented.

Accumulating evidence has suggested that neither fourfold-degenerate sites nor introns are entirely free from constraint and selections were trailed at silent sites. For example, recent direct evidence has shown that synonymous mutations could be highly deleterious [12]; codon usage bias has been found more often in highly expressed genes under selection [13,14]; and constitutively expressed exons have a higher GC content than those that are alternatively expressed [15]. In humans, it is estimated that as many as 40% of fourfold-degenerate sites are possible under selection [16].

Codon usage bias tends to make GC content change drastically, particularly at silent sites, which affects synonymous substitution rates significantly, and therefore has been extensively examined in many species, such as Drosophila [17,18]. Selectively driven codon usage has been detected in many organisms, including bacteria, yeast, fly and nematode. In mammals, several years ago it was believed that there is no selectively driven codon usage [19]. However, recent studies have provided evidence of such selectively driven codon usage [20]. To our knowledge, no evidence for selectively driven codon usage in plants has been reported to date, particularly based on current genomic sequences.

To address whether selection at synonymous sites exists, or more specifically, whether we can detect selectively driven codon usage in rice, here we used the best reciprocal paralogs with high and low GC contents and found at least three points of evidence to support selectively driven codon usage in rice: (a) a much greater increase in GC4 than in GCi; (b) with reference to their best reciprocal orthologs in *Arabidopsis*, most substitution sites between the two kinds of paralogs occur at

*Corresponding author. Address: Institute of Crop Science, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China. Fax: +86 571 86971117.
E-mail address: fanlj@zju.edu.cn (L. Fan).

0014-5793/$32.00
doi:10.1016/j.febslet.2007.01.088

2- and 4-degenerate sites with T → C mode, while A → C and A → G play major roles at 0-degenerate sites; and (3) the effective number of codons (N_c) and the codon adaptation index (CAI) indicate that high-GC genes have much greater bias and codon usage is skewed toward codons that are preferred in highly expressed genes.

2. Materials and methods

2.1. Sources of sequence data

A total of 61 250 protein-coding sequences from rice (*Oryza sativa*) genome, 2653 protein-coding genes in BAC sequences of maize (*Zea mays*) and 28 581 protein-coding sequences from *Arabidopsis* were downloaded from TIGR (The International Genome Research, ftp://ftp.tigr.org).

2.2. Identification of homologs

All rice coding sequences were divided into two groups according to their GC content: (1) GC content > 65% (total 13 620) and (2) GC content < 60% (total 16 288) (Fig. 1). The sequences in the two groups were further filtered as follows: (1) less than 75 codons, (2) transposable element-related. Paralogous protein pairs between high-GC and low-GC were identified by performing BlastP searches [21] with an *E*-value of <1e−7 and best reciprocals. A total of 1022 paralogous pairs were identified and further searched against *Arabidopsis* proteins with the same criteria. Pairs with the same *Arabidopsis* protein reciprocal best hit were retained. This step left a data set of 321 entries that each contained a paralogous pair of high- (H_{gene}) and low-GC (L_{gene}) genes from rice, and an *Arabidopsis* ortholog. Translated amino acid sequences from the final data set were then performed by multi-alignment using ClustalW. The resulting alignment was used as a guide to align the codons of nucleotide sequences. GC content and degenerate substitution change modes were calculated using perlscript. Similarly, 382 reciprocal best paralog pairs of high-and low-GC genes in maize were generated.

2.3. Codon usage and statistical analysis

To normalize codon usage within datasets of different amino acid compositions, relative synonymous codon usage (RSUC) values were calculated by dividing the observed codon usage by that expected when all codons for the same amino acids are used equally [17]. The effective number of codons (N_c) was used to measure the magnitude of codon bias for an individual gene, and ranged from 20 (extreme bias using only one codon per amino acid) to 61 (no bias toward codons). The codon adaptation index (CAI) was used to estimate the extent of bias toward codons that are known to be preferred in highly expressed genes. A correspondence analysis was used to detect differences in codon usage between genes and to identify the codons involved, where all genes are plotted in a 59-dimensional hyperspace according to their use of the 59 sense codons. CodonW (version 1.4) was used for the above analysis.

A χ^2 test was used to examine the significance of differences in codon usage between high- and low-GC genes [22]. For each of the 59 sense codons, the χ^2 test involved a 2×2 table that gave one degree of freedom, where the first row contains the values observed for the codon being analyzed, and the second row shows the total number of synonymous alternatives. Significance was examined at a level of 5% (χ^2 value of 3.84).

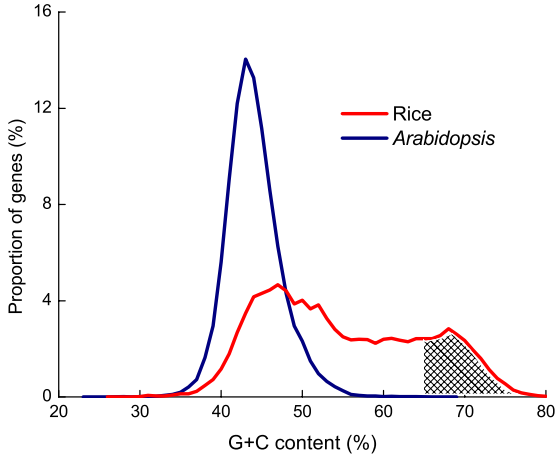

Fig. 1. Distribution of genes with different GC contents in rice and *Arabidopsis* genome. The portions of genes with GC content over 65% (high-GC genes in this study) are shaded.

3. Results

3.1. Structural differences

A wide variation was observed in GC content among genes from rice genome, but not *Arabidopsis* genome (Fig. 1). *Arabidopsis* genes have relatively low GC contents (almost no gene has a GC content of over 65%) and give a unimodal distribution. In contrast, rice genes have both high and low GC content and present a multimodal distribution. This observation is similar to previous reports [1,10].

As our primary effort to understand the evolution of GC content in rice, high-GC genes (GC content > 65%) and their paralogs with low GC content (<60%) were first identified using a reciprocal BLAST search. A total of 1022 paralogous pairs were retained. To determine the substitution pattern of nucleotide sequences between them, a reference ortholog needed to be used. The dicot *Arabidopsis*, the genome of which is available and almost none of the genes have a GC content of 65%, is an ideal candidate for such a reference. As a result, a total of 321 gene pairs were isolated from rice and *Arabidopsis*. Of the 321 pairs, the average GC contents of high-GC genes (H_{gene}) and low-GC genes (L_{gene}) from rice and of genes from *Arabidopsis* were 69 ± 0.13%, 54 ± 2.6% and 45 ± 0.16%, respectively.

The H_{gene} and L_{gene} of rice showed significant differences in gene structure, for example, with regard to gene length and intron numbers. H_{gene} (1018 ± 30 bp) was significantly shorter than L_{gene} (1405 ± 45 bp). Furthermore, a large difference in intron numbers was found between H_{gene} and L_{gene}. There is a twofold difference in the percentages of single-exon genes in H_{gene} and L_{gene} genes (35.8% and 17.1%, respectively) (Table 1). Orthologous *Arabidopsis* genes were similar to L_{gene} in the above features (1362 ± 42 bp; 17.4%).

In summary, our results suggested that the divergence of some duplicate genes of rice seems to be very special and, in addition to the GC content, they can be characterized by short length and a high proportion of single-exon genes.

3.2. Codon usage bias

Regarding the important role of codon usage in GC-content evolution, codon usage variation between H_{gene} and L_{gene} was analyzed. The results indicated that N_c values tended to be lower for H_{gene}(Fig. 2). N_c values, ranging from 20 to 61, are usually used to measure the magnitude of codon usage bias. A lower value indicates high bias, with only one codon used per amino acid. In general, N_c values of most L_{gene} ranged from 45 to 60 (average 52 ± 0.27), suggesting that these genes are not very biased with regard to codon usage. However, most of the N_c values for H_{gene} were between 25 and 45 (average

Table 1
Distribution of intron numbers of 321 high-GC genes (H_{gene}) and paralogous low-GC genes (L_{gene}) in rice and their orthologs in *Arabidopsis*

Genes	Intron number per gene								
	0	1	2	3	4	5	6	>6	Total
H_{gene}	115	82	57	24	21	12	3	7	321
L_{gene}	55	61	53	36	31	28	19	40	321
Arabidopsis	56	75	49	35	32	18	19	37	321

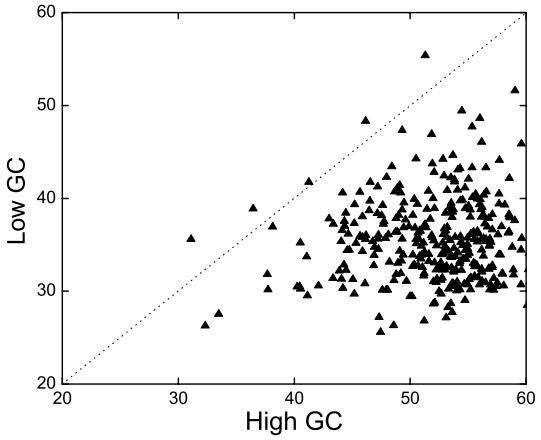

Fig. 2. Relationship between effective number of codons (N_c) of high-GC genes and paralogous low-GC genes in rice. The line indicates equal values.

35 ± 0.24). These results suggest that there is a bias in codon usage in high-GC genes.

A χ^2 test was performed to evaluate whether there were significant differences in codon usage between the two categories of genes. The results showed that, except for GGG (Gly), 58 of the total 59 codons exhibited significant differences in usage between the two groups ($P < 0.05$) (Table 2). Interestingly, UUG (Leu) and AGG (Arg) were used less in H_{gene} than in L_{gene}, although codons that end in G and C are predominant in H_{gene}.

3.3. Selectively driven codon usage

We wish to elucidate the mechanism (selection or neutral model) that is responsible for the GC divergence between H_{gene} and L_{gene}. GC4 and GCi were calculated in the two groups and used to distinguish the two models. The results indicated that ΔGC4 (difference in GC content at fourfold-redundant sites in exons between paralogous pairs of H_{gene} and L_{gene}) was generally greater than ΔGCi (difference in GC content of introns between paralogous pairs of H_{gene} and L_{gene}) ($R^2 = 0.18$, $P < 0.05$) (Fig. 3). Furthermore, 67.5% of the total paralogous pairs gave slightly higher values of ΔGC12 (difference in GC content at the first two positions in codons between paralogous pairs of H_{gene} and L_{gene}) than of ΔGCi ($R^2 = 0.20$, $P < 0.01$) (Fig. 3). These results indicated that the divergence at fourfold-degenerate site in codons is significantly higher than that at intronic sequences. This is strong evidence for selectively driven codon usage in rice, and it is difficult to use mutational bias to explain the difference between GC4 and GCi.

The substitution patterns of degenerate sites in H_{gene} and L_{gene} were checked using *Arabidopsis* as a reference. We divided substitution sites into 4 groups ($T \rightarrow C + C \rightarrow T$, $A \rightarrow C + C \rightarrow A$, $T \rightarrow G + G \rightarrow T$, $A \rightarrow G + G \rightarrow A$). Only these site changes could increase the GC content. By observing site changes at 0-, 2- and 4-degenerate sites, we found that the substitution rates from A or T to G or C in H_{gene} were significantly higher than those in L_{gene} at all of the degenerate sites, although the extent of the change varied (Fig. 4). At 4-degenerate sites, there were differences in the percentages in the four change modes to G or C, from near 10% to near 20%, in H_{gene}. Of these modes, $T \rightarrow C$ plays the largest role and accounted for 18.3% ± 0.4% of the total nucleotide mutations for H_{gene}; $T \rightarrow G$ accounted for about 15% of the substitutions. All of these change rates are higher than the average substitution rates under neutral evolution. In L_{gene}, there was a similar trend as in H_{gene}. For example, $T \rightarrow C$ mode also played the biggest role and accounted for 10.7% ± 0.4% of substitution for L_{gene}. At 2-degenerate sites, the four change modes to G or C also showed large differences. The $T \rightarrow C$ mode is again a major contributor for H_{gene} (22.3% ± 0.5%), but $A \rightarrow G$ rather than the $T \rightarrow G$ change mode played an important role for H_{gene}. Interestingly, the percentages of the two change modes at 2-degenerate sites, which are constrained, were higher than those at 4-degenerate sites. At 0-degenerate sites, the $A \rightarrow G$ change mode was a major contributor: 1.9% ± 0.2% in L_{gene} and 4.6 % ± 0.1% in H_{gene}. However, the $T \rightarrow C$ change mode accounts for only 0.8% ± 0.01% in L_{gene} and 1.6% ± 0.1% in H_{gene}. Overall, our observations indicated that the mode of mutation affects both synonymous and non-synonymous sites, and different sites face different selection forces. The nucleotide "T" seems to face the strongest selection force to make it change to "C" at silent sites during rice evolutionary processes. Meanwhile, the above results also suggested that, in addition to purifying selection, positive selection may also serve to shape codon usage in rice.

In summary, the evolutionary rates in GC4 were significantly higher than those in GCi based on a set of genes with high GC content and their paralogous low-GC genes in rice. The non-random use of substitution sites is seen universally among 0-, 2- and 4-degenerate sites, according to a reference genome the genes of which have low GC contents. Transcription-coupled mutational processes and biased gene conversion cannot explain these results, since they should affect introns and flanking exons equally. Therefore, by excluding these effects, we believe that the present results are strong evidence for selectively driven codon usage in rice.

This conclusion raised the important issue of whether this mechanism is universal in cereals. Therefore, maize, for which some high-quality genomic sequences are available, was used to test the hypothesis. Interestingly, a similar trend of GC4 and GCi was also observed in maize (see Supporting Information, Fig. S1B). These results suggested that maize and rice, and likely *Gramineae*, may have experienced similar evolutionary selection processes to shape their codon usage.

Table 2
Differences in codon usage between high-GC genes (H_{gene}) and paralogous low-GC genes (L_{gene}) in rice genome

AA	Codon	RCSU	
		H_{gene}	L_{gene}
Ala	GCA+	0.178	0.928
	GCC*	1.823	1.206
	GCG*	1.767	0.92
	GCU+	0.231	0.946
Cys	UGC*	1.867	1.349
	UGU+	0.133	0.651
Asp	GAC*	1.732	0.982
	GAU+	0.268	1.018
Glu	GAA+	0.184	0.691
	GAG*	1.816	1.309
Phe	UUC*	1.875	1.271
	UUU+	0.125	0.729
Gly	GGA+	0.304	0.816
	GGC*	2.517	1.418
	GGG	0.93	0.894
	GGU+	0.249	0.872
His	CAC*	1.686	1.037
	CAU+	0.314	0.963
Ile	AUA+	0.208	0.61
	AUC*	2.519	1.382
	AUU+	0.273	1.008
Lys	AAA+	0.167	0.613
	AAG*	1.833	1.387
Leu	CUA+	0.127	0.483
	CUC*	3.104	1.663
	CUG*	2.145	1.469
	CUU+	0.259	1.081
	UUA+	0.037	0.345
	UUG+	0.327	0.958
Asn	AAC*	1.759	1.084
	AAU+	0.241	0.916
Pro	CCA+	0.353	1.147
	CCC*	1.043	0.794
	CCG*	2.264	1.005
	CCU+	0.34	1.054
Gln	CAA+	0.236	0.742
	CAG*	1.764	1.258
Arg	AGA+	0.253	1.065
	AGG+	1.39	1.673
	CGA+	0.243	0.455
	CGC*	2.172	1.246
	CGG*	1.679	0.978
	CGU+	0.263	0.583
Ser	AGC*	1.774	1.296
	AGU+	0.132	0.708
	UCA+	0.251	1.083
	UCC*	1.82	1.165
	UCG*	1.779	0.77
	UCU+	0.245	0.977
Thr	ACA+	0.24	1.085
	ACC*	1.698	1.26
	ACG*	1.848	0.74
	ACU+	0.214	0.915
Val	GUA+	0.081	0.412
	GUC*	1.758	1.164
	GUG*	1.963	1.42
	GUU+	0.199	1.004
Tyr	UAC*	1.871	1.199
	UAU+	0.129	0.801

An asterisk (*) after the codon indicates that this codon is used significantly more often in H_{gene} than L_{gene} according to χ^2 test (5% level). Conversely, a plus (+) indicates that the codon is used significantly more often in L_{gene}. The absence of any symbol after a codon means there is no significance of difference in usage of that particular codon. AA, the amino acid; RCSU, the relative synonymous codon usage.

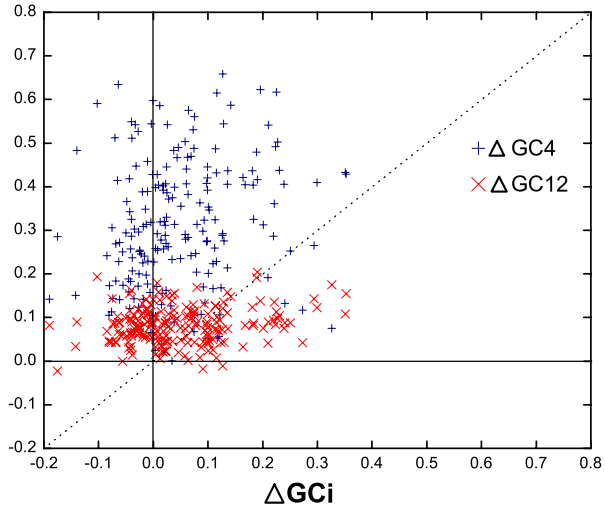

Fig. 3. Relationship between the variation in GC content at degenerate sites of exons and introns in rice. ΔGC4, ΔGC12 and ΔGCi: differences in GC content at fourfold-degenerate sites, the first two positions in codons, and introns between high-GC genes and paralogous low-GC genes.

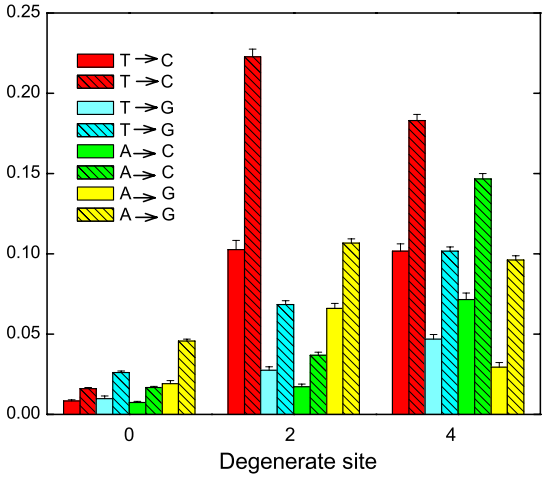

Fig. 4. Substitution modes from A or T to G or C in 0-, 2- and 4-degenerate sites of high-GC genes and paralogous low-GC genes in rice. Orthologous *Arabidopsis* genes were used as references. High-GC genes are shaded.

3.4. Gene expression level

The important observation that highly expressed genes have greater bias and skewed codon usage was used to elucidate the potential function of selectively driven codon usage bias [20]. A codon adaptation index (CAI), which is usually used to estimate the extent of bias toward codons that are known to be preferred in highly expressed genes, and a multivariate statistical analysis were used to measure the expressivities of the two kinds of genes in rice. Multivariate statistical analysis has been widely used to study codon usage variation among genes [23]. A correspondence analysis is a type of multivariate statistical analysis in which the data are plotted in a multidimensional space of 59 axes (excluding Met, Trp and stop Codons), and the most prominent axes that contribute to codon usage variation among the genes are then determined. As a reference set for CAI calculation, genes encoding translational elongation factor eF-Tu, heat-shock synthetase and methyonyl-tRNA synthetase were used [24].

The results of the multivariate analysis indicated that the first axis accounted for 12.4% and 22.5% of the total inertia of the 59-dimensional space in H_{gene} and L_{gene}, respectively, whereas the next three axes accounted for 7.4%, 4.8% and 4.2% in H_{gene}, and 6.8%, 5.7% and 4.2% in L_{gene}. These results indicate that the first axis was the major explanatory axis for interpreting codon usage bias. Statistical tests showed that there were more significant differences between the two groups in both the first axis and CAI (non-parametric Wilcoxon test, $P < 0.01$). Furthermore, a regression analysis showed that both H_{gene} and L_{gene} significantly correlated with their CAI values along the first major axis ($R^2 = 0.162$, $P < 0.001$ and $R^2 = 0.126$, $P < 0.001$, respectively) (Fig. 5A). More importantly, CAI values of H_{gene} were generally higher than those of L_{gene} ($R^2 = 0.162$, $P < 0.001$) (Fig. 5B). The above results suggested that H_{gene} has both higher codon usage bias and a higher expression level than L_{gene}.

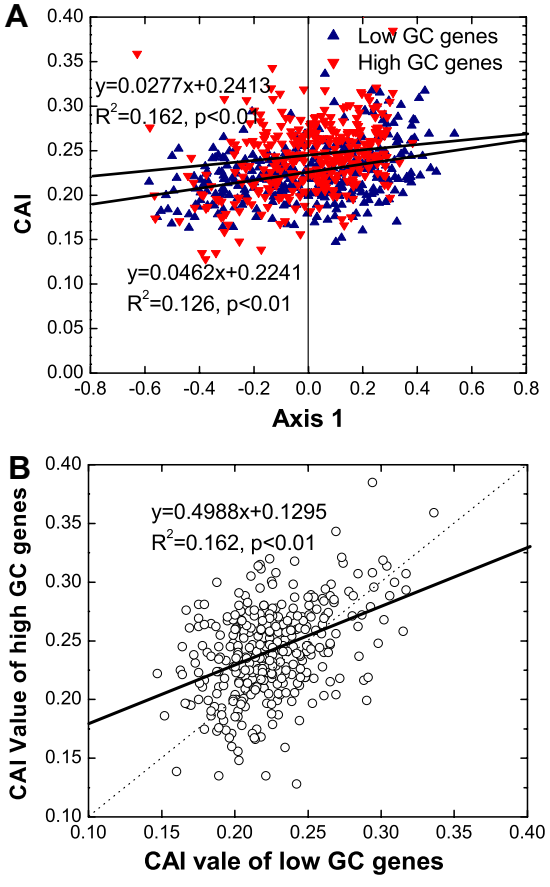

Fig. 5. Relationship between gene expression level and codon usage bias in high- and low-GC genes in rice. (A) Relationship between codon adaptation index (CAI) and first axis of codon usage variation by multivariate statistical analysis. (B) Relationship between CAI values in high- and low-GC genes.

4. Discussion

Our findings show that the rate of evolution in GC4 is significantly higher than that in GCi based on differences in the GC content of some genes in rice. Considering other observations (such as substitution model, expression level, etc.), we propose that this is strong evidence for selectively driven codon usage in rice. Potential alternative explanations, such as transcription-coupled mutational processes and biased gene conversion, which are expected to affect introns and flanking exons equally, cannot explain the observations in rice and therefore can be eliminated. Under the assumption that these silent sites (GC4 and GCi) evolve neutrally, their rates of evolution have been used to measure the mutation rate [6]. However, our data indicate that they show heterogeneous rates of evolution in rice, and possibly in *Gramineae*. Our observations are not the result of a compensation effect, which might favor selection of a particular GC content at different positions: since the first two positions in exons are constrained, selection might favor an increased GC content at the third site to provide local compensation [6]. To our knowledge, this is the first successful genome-scale attempt to uncover evidence of selection regarding codon usage that would promote the GC content in rice. This represents a potential evolutionary process for genes with a high GC content in rice and other cereals. Genes with a high GC content are unique to monocot *Gramineae*, and are absent in dicots. This implies that the selection of *Gramineae* genes may be species-specific.

To test our hypothesis, several controls were added to the data set used in this study. First, a GC content of 65% was used as a threshold for the so-called high-GC genes. These genes are unique to rice relative to *Arabidopsis* (Fig. 1). This provides a unique perspective for our study of GC-content evolution. Since all of the high- and low-GC genes (including the "outgroup" *Arabidopsis* genes) diverged from their common ancestral sequences at the same start point in evolutionary time, we can perform a comparative study on the two sets of rice genes. Second, a previous report showed that transposable elements (TE) may influence intron base composition, which could reduce intron GC content [25]. However, this should not affect our results, since TE-related genes were filtered from our data set. In addition, a long coding sequence should be more likely to be affected by TE than short coding sequences, which should increase the ΔGCi value and decrease the extent of the difference between ΔGC4 and ΔGCi, i.e. it should not affect our conclusion. Third, a reciprocal best BLAST search was used to identify homologous genes within rice and among rice and *Arabidopsis*. It is important to guarantee that our pairs of rice and *Arabidopsis* arose from a common ancestral sequence. The present method is a promising

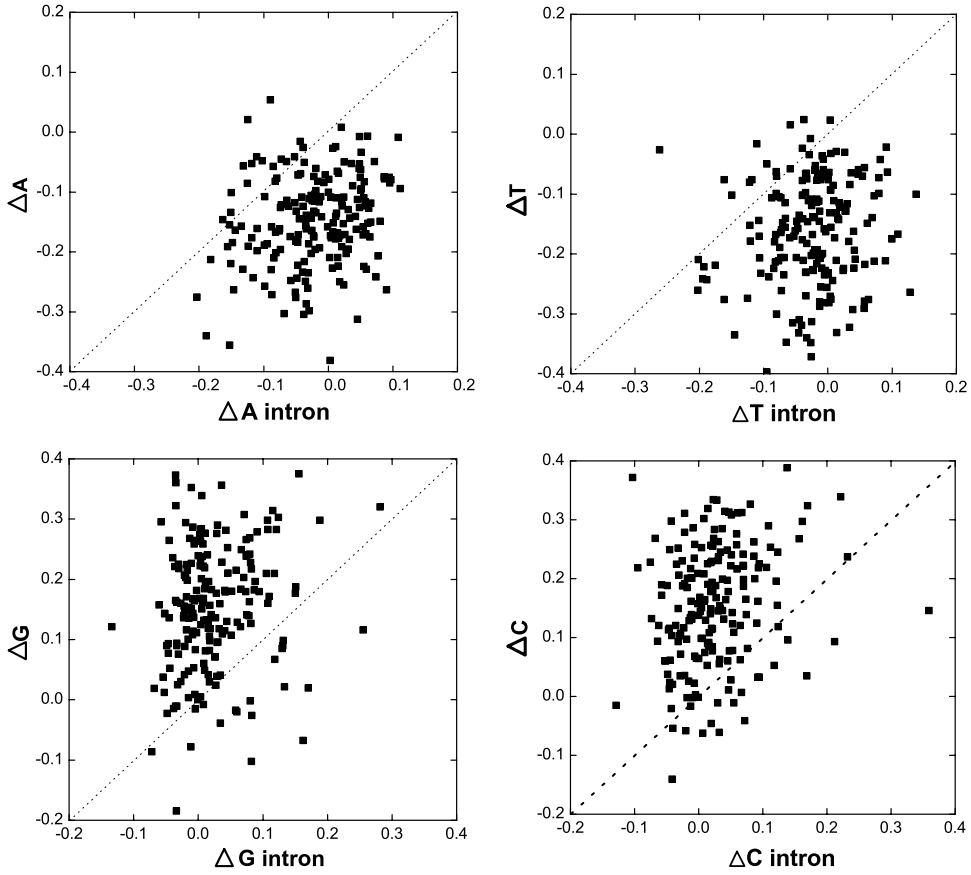

Fig. 6. Relationship between nucleotide content in introns and at fourfold-degenerate sites in flanking exons. The line indicates equal content.

approach to finding such data and has been used in many previous studies (such as [26]).

Our data set (321 pairs, or 642 rice genes) is not a large proportion of the total number of rice genes, but is representative. We also investigated GC4 and GCi using the 1022 paralogous pairs identified at our first step in the rice genome (see Section 2). Similar results were obtained (see Supporting Information, Fig. S1A and Fig. S2).

The strand-specific enrichment of nucleotides (such as C) has been reported in rodents [20]. If this effect was not strand-specific, we would expect that both the G and C content would be higher in exons than in introns. However, in rice, such an effect was not observed, and both the G and C contents in their exons were higher (Fig. 6).

Our results indicated that, except with regard to codon usage bias, there is a significant difference in the gene expression level between H_{gene} and L_{gene} as assessed by CAI values. These results are consistent with previous reports that highly expressed genes have greater bias. The classical model for selectively driven codon usage bias suggests that its function is to increase the efficiency (rate or accuracy, [19]) of mRNA translation, and this is supported by the above observation and copies of tRNA genes (frequently used codons tend to have more copies of tRNA genes). Co-adaptation between codon usage, expression rate, and/or tRNA abundance has been described in the worm, fruitfly, yeast and bacteria [20]. Our results suggest that translation-related selection might have been involved in the codon usage variation of high- and low-GC genes, particularly high-GC genes, in rice.

High-GC genes seem to be characterized by a short length and a high proportion of single-exon genes, according to our study. This result is consistent with previous surveys [1,9,27]. These studies suggested that genes with shorter coding sequences and fewer exons have a greater tendency to have a higher GC content. This observation was suggested to be related to RNA splicing. Longer genes may encode more complex transcripts and proteins that have a greater chance of being functionally disrupted and, in contrast, shorter genes have a smaller risk for mutations and therefore are subject to less selective constraint [9].

Acknowledgements: This work was supported by the Ministry of Sciences and Technology (2006CB101700/2006AA10A102) and the National Natural Science Foundation of China (30270810/30471067).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2007.01.088.

References

[1] Carels, N. and Bernardi, G. (2000) Two classes of genes in plants. Genetics 154, 1819–1825.
[2] Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J.,

Chen, W.H., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, J., Yang, G., Zhang, Y., Xiong, Z., Li, L., Mao, C., Zhou, Z., Zhu, R., Chen, B., Hao, S., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L. and Yang, H. (2003) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296, 79–92.

[3] Filipski, J., Thiery, J.P. and Bernardi, G. (1973) An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation. J. Mol. Biol. 80, 177–197.

[4] Gautier, C. (2000) Compositional bias in DNA. Curr. Opin. Genet. Dev. 10, 656–661.

[5] Eyre-Walker, A. and Hurst, L.D. (2001) The evolution of isochores. Nat. Rev. Genet. 2, 549–555.

[6] Eyre-Walker, A. (1999) Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152, 675–683.

[7] D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. and Bernardi, G. (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. 32, 504–510.

[8] Hughes, A.L. and Yeager, M. (1997) Coordinated amino acid changes in the evolution of mammalian defensins. J. Mol. Evol. 44, 675–682.

[9] Wang, H.C., Singer, G.A. and Hickey, D.A. (2004) Mutational bias affects protein evolution in flowering plants. Mol. Biol. Evol. 21, 90–96.

[10] Wong, G.K., Wang, J., Tao, L., Tan, J., Zhang, J., Passey, D.A. and Yu, J. (2002) Compositional gradients in Gramineae genes. Genome Res. 12, 851–856.

[11] Shi, X., Wang, X., Li, Z., Zhu, Q., Tang, W., Ge, S. and Luo, J. (2006) Nucleotide substitution pattern in rice paralogues: implication for negative correlation between the synonymous substitution rate and codon usage bias. Gene 376, 199–206.

[12] Duan, J., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J. and Gejman, P.V. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. 12, 205–216.

[13] Urrutia, A.O. and Hurst, L.D. (2003) The signature of selection mediated by expression on human genes. Genome Res. 13, 2260–2264.

[14] DeBry, R.W. and Marzluff, W.F. (1994) Selection on silent sites in the rodent H3 histone gene family. Genetics 138, 191–202.

[15] Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13–34.

[16] Hellmann, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B. and Paabo, S. (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. 13, 831–837.

[17] Sharp, P.M. and Li, W.H. (1989) On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. 28, 398–402.

[18] Moriyama, E.N. and Hartl, D.L. (1993) Codon usage bias and base composition of nuclear genes in Drosophila. Genetics 134, 847–858.

[19] Duret, L. (2002) Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12, 640–649.

[20] Chamary, J.V. and Hurst, L.D. (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol. Biol. Evol. 21, 1014–1023.

[21] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

[22] McInerney, J.O. (1998) Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc. Natl. Acad. Sci. USA 95, 10698–10703.

[23] Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. and Ikemura, T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. 53, 290–298.

[24] Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M. and Claverie, J.M. (2004) The 1.2-megabase genome sequence of Mimivirus. Science 306, 1344–1350.

[25] Duret, L. and Hurst, L.D. (2001) The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. 18, 757–762.

[26] The Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493–521.

[27] Xia, X., Xie, Z. and Li, W.H. (2003) Effects of GC content and mutational pressure on the lengths of exons and coding sequences. J. Mol. Evol. 56, 362–370.