Comparative Analysis of Codon Usage Patterns Among Mitochondrion, Chloroplast and Nuclear Genes in *Triticum aestivum* L.

Wen-Juan Zhang[1], Jie Zhou[1], Zuo-Feng Li[1], Li Wang[1], Xun Gu[2] and Yang Zhong[1, 3*]

([1]*Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering*, *School of Life Sciences*, *Fudan University*, Shanghai 200433, China;
[2]*Department of Genetics and Cell Biology*, *Center for Bioinformatics and Biological Statistics*, *Iowa State University*, Ames, IA 50011, USA;
[3]*Shanghai Center for Bioinformation Technology*, Shanghai 201203, China)

Abstract

In many organisms, the difference in codon usage patterns among genes reflects variation in local base compositional biases and the intensity of natural selection. In this study, a comparative analysis was performed to investigate the characteristics of codon bias and factors in shaping the codon usage patterns among mitochondrion, chloroplast and nuclear genes in common wheat (*Triticum aestivum* L.). GC contents in nuclear genes were higher than that in mitochondrion and chloroplast genes. The neutrality and correspondence analyses indicated that the codon usage in nuclear genes would be a result of relative strong mutational bias, while the codon usage patterns of mitochondrion and chloroplast genes were more conserved in GC content and influenced by translation level. The Parity Rule 2 (PR2) plot analysis showed that pyrimidines were used more frequently than purines at the third codon position in the three genomes. In addition, using a new alterative strategy, 11, 12, and 24 triplets were defined as preferred codons in the mitochondrion, chloroplast and nuclear genes, respectively. These findings suggested that the mitochondrion, chloroplast and nuclear genes shared particularly different features of codon usage and evolutionary constraints.

Key words: chloroplast; codon usage; mitochondrion; neutral mutation; nuclear gene; selection constraint; *Triticum aestivum*.

Zhang WJ, Zhou J, Li ZF, Wang L, Gu X, Zhong Y (2007). Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in *Triticum aestivum* L. *J. Integr. Plant Biol.* **49**(2), 246–254.

Available online at www.blackwell-synergy.com/links/toc/jipb, www.jipb.net

The vast majority of prokaryotic and eukaryotic species have non-random codon usage. Alternative synonymous codons in most genes are used with unequal frequency, i.e., certain synonymous codons are significantly preferred over others. It has been widely accepted that the biased codon usage is a result of natural selection (Sharp and Li 1986; Akashi 1994) or mutation pressure (Jukes and Bhushar 1986; Osawa et al. 1988; Sueoka 1988; Kano et al. 1991). In plants, most studies on codon bias have focused on nuclear genes. Chiapello et al. (1998) investigated the association between codon usage and gene functions in *Arabidopsis thaliana*, and identified two classes of genes: one was highly biased to G/C and the other had a weak preference for A/T-biased codons. Kliman and Henry (2005) also inferred the preferred codons in *A. thaliana* using the factor analysis and found that these codons corresponded to the commonly represented tRNA anticodons. For poaceous plants that include a number of important crops, a correspondence analysis of 101 genes in maize (*Zea mays*) revealed that the major pattern in codon usage was a result of differences among the genes in the frequencies of G+C in the third

Received 8 May 2006 Accepted 12 May 2006

Supported by the Sate Key Basic Research and Development Plan of China (2003CB715904) and the National Science Foundation for Overseas Distinguished Young Scholar (30428003).

*Author for correspondence.
Tel: +86 (0)21 5566 4436;
Fax: +86 (0)21 6564 2468;
E-mail: <yangzhong@fudan.edu.cn>.

doi: 10.1111/j.1672-9072.2007.00404.x

silent nucleotide position (Fennoy and Bailey-Serres 1993). Liu et al. (2004a, 2005a) also analyzed a large dataset of rice (*Oryza sativa*) genes and found that synonymous codon usage and gene functions were strongly correlated. Kawabe and Miyashita (2003) compared features of codon choice among three dicot and four monocot plant species and figured out that despite the difference in GC contents, the codon usage tendencies were similar among most complete coding sequences (CDSs).

As important organelles of plants, mitochondrion and chloroplast have their own stable genomic systems and environment. Mitochondrion is the energy provider. In a particular organism, the mitochondrion genome often accumulates nucleotide substitutions at faster rates than the nuclear genome (Brown et al. 1979; Martin et al. 1992; Ramirez et al. 1993), especially for mammalian mitochondrial DNA (Brown et al. 1979). The codon bias in animal mitochondrion has been considered as the result of strong AT nucleotide and genetic drift. However, there is a great of difference between plant and animal mitochondria. The mitochondria of higher plants exhibit a number of unique features, such as large size (200–2 400 kb), slow rates of evolutionary change, and the usage of the universal genetic codes (Schuster and Brennicke 1994; Notsu et al. 2002). In particular, it was reported that the expression level was the major factor in shaping the codon usage pattern in rice mitochondrion genes (Liu 2004b). On the other hand, chloroplast that contains two photo-systems plays a crucial role in the process of photosynthesis. Although there was evidence that the codon usage patterns of certain chloroplast genes were influenced by selection (Morton 1998), the context-dependent mutation was considered as the primary factor in shaping codon usage patterns of the chloroplast genomes of certain grass species (Morton 2003).

In poaceous species, the mitochondrion and chloroplast were relative AT-rich genomes, while the nuclear one was GC-rich (Salinas et al. 1988). Currently, the complete genomic sequences of mitochondrion and chloroplast, as well as a number of CDSs of nuclear genes in wheat (*Triticum aestivum* L.), have been released in GenBank. In order to better understand the variation in codon usage and the evolutionary constraints influencing them among different genes within a single plant species, we presented a detailed comparison of codon usage patterns among the mitochondrion, chloroplast and nuclear genes in *T. aestivum* and explored the key factor in shaping their codon choice in this study.

Results

Codon usage patterns of mitochondrion, chloroplast and nuclear genes

The global G+C content and the percentage of GC at three codon positions were all relatively higher in nuclear genes than in mitochondrion and chloroplast genes (Table 1). Differences in GC contents were greatest at the third codon position where usually reflected the neutral mutation bias leading to different codon choice within the same protein sequence. The neutrality plots (*GC12* vs *GC3s*) were analyzed for the mitochondrion, chloroplast and nuclear genes (Figure 1). The nuclear genes in *T. aestivum* had a wide range of *GC3s* (0.212–1.000), while the mitochondrion and chloroplast genes had relatively narrow *GC3s* distributions (0.223–0.560 and 0.197–0.364, respectively). In the nuclear genes of *T. aestivum*, there was a significant correlation ($P < 0.01$) and the regression coefficient was approximately 0.301. This significantly positive correlation in neutrality plots indicated that the effect on the GC contents by the intragenomic GC mutation bias was similar at all three codon positions. In contrast, although there existed weak correlations between two *GC12* and *GC3s* in mitochondrion and chloroplast genes, they were insignificant ($P > 0.05$).

Association between effective number of codons (*ENC*) and *GC3s*

The standard curve shown in Figure 2 showed the relationship between *ENC* and *GC3s* under H_0 (no selection). Organelle and nuclear genes presented different patterns of *ENC*-plot according to their different GC composition (Figure 2). Notably, whatever organelle and nuclear, most genes with lower *ENC* values than expected ones were lying well below the curve, suggesting there were other factors, particularly selective expression level affecting the codon usage combined with composition mutation.

The frequency distributions of the *ENC* ratio are shown in Figure 3. Each dataset displayed a single peak. The most genes with location of 0.0–0.2 of *ENC* ratios indicated that they had *ENC* values slightly smaller than expected ones. It suggested

Table 1. Base composition of the mitochondrion, chloroplast and nuclear genes in *Triticum aestivum*

Genes	No. of sequences	GC (%)	AT (%)	GC First position (%)	GC Second position (%)	GC Third position (%)
Mitochondrion genes	36	42.4	57.6	47.7	41.4	38.0
Chloroplast genes	51	39.1	60.9	47.8	39.5	29.9
Nuclear genes	1 235	55.5	44.5	59.7	42.7	64.2

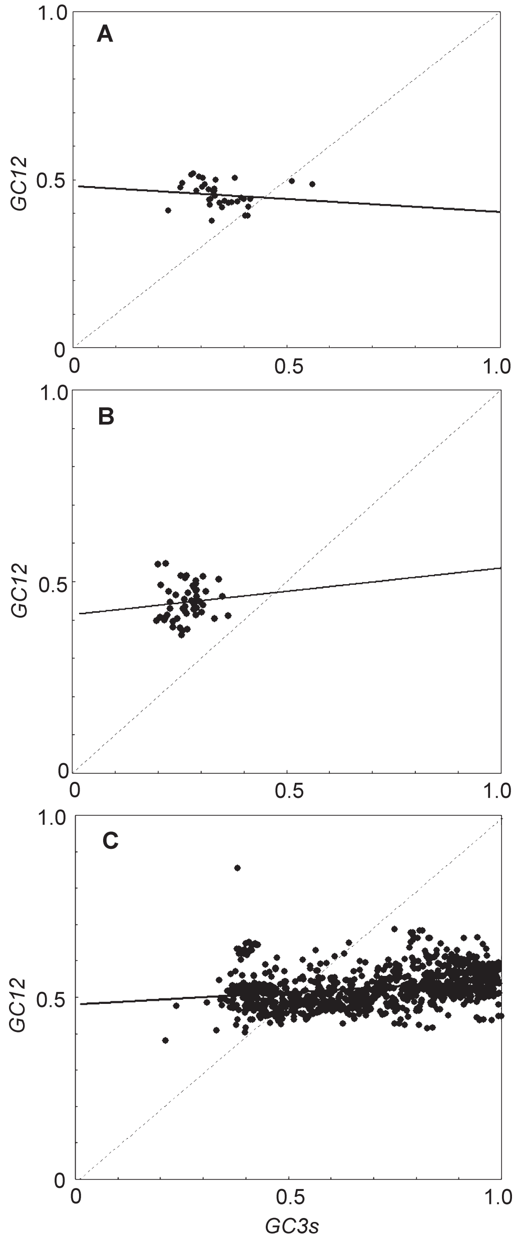

Figure 1. Neutrality plots (*GC12 vs. GC3s*).

(A) For mitochondrion genes, the regression line: y = –0.077 5x + 0.481 8, r2 = 0.019 5, OP (optimal point) = 0.447.

(B) For chloroplast genes: y = 0.122 2x+0.413 6, r2 = 0.010 5, OP = 0.471.

(C) For nuclear genes: y = 0.066 2x+0.481, r2 = 0.079 7, OP = 0.515. OP indicates the point at which the regression line crossed the diagonal line.

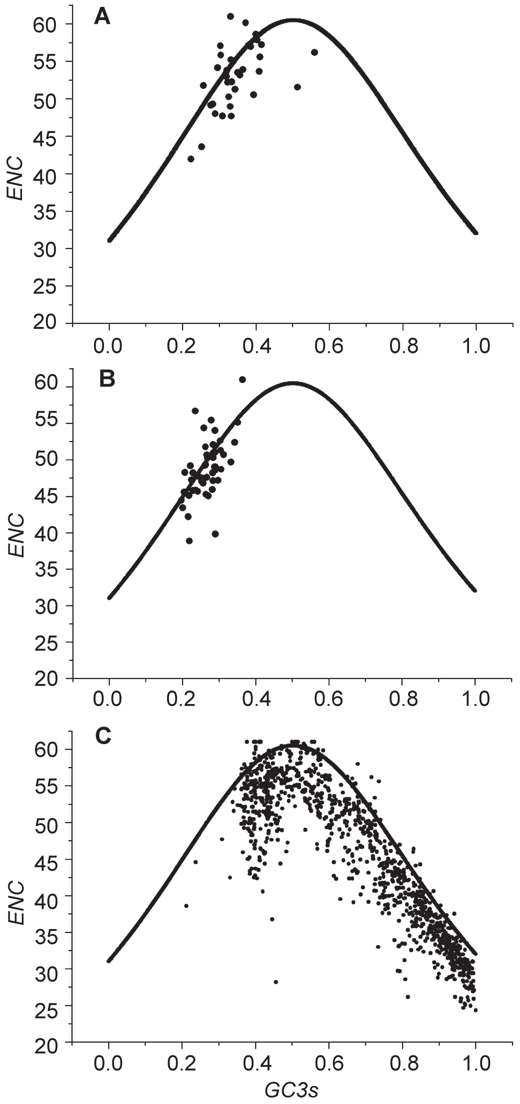

Figure 2. Effective number of codons (*ENC*)-plot showing relationship between *ENC* and *GC3s*.
Each expected *ENC* from *GC3s* is shown as a standard curve.

(A) Mitochondrion genes.

(B) Chloroplast genes.

(C) Nuclear genes.

that besides other factors influencing the codon bias, *T. aestivum* used codons predicted from *GC3s*, i.e., neutral mutation played a role in shaping the codon usage.

Correspondence analysis

On the basis of the *RSCU* values, our correspondence analysis revealed the major trends in codon usage of mitochondrion, chloroplast and nuclear genes in *T. aestivum*. In the nuclear genes, Axis 1 accounted for 39.16% of the overall variance among genes in codon usage, whereas the next three axes accounted for 7.36%, 5.60%, and 3.64% respectively, confirming that Axis 1 was the primary axis and the factor it represented was the main factor in explaining the data. Axis 1 was significantly correlated to *ENC*, *GC* and *GC3s* (r = 0.888, –0.941, –0.985, *P* < 0.01), implying that the nucleotide composition mutation bias could be considered as the relatively major factor in shaping the codon usage.

In the mitochondrion genes, Axis 1 accounted for 12.97% of the overall variance, whereas the next three axes accounted for 10.61%, 9.07%, and 8.64%, respectively. Axis 1 was significantly correlated to *ENC*, *GC* and *GC3s* (r = –0.665, –0.339, –0.800, *P* < 0.01). This result indicated that there was no clearly single trend in codon bias in the dataset and several factors could be the cause. In other words, although neutral mutation press would be the relative major factor in shaping the codon usage, the translation level coupled factors also played a critical role in mitochondrion evolution.

In the chloroplast genes, Axis 1 accounted for 9.25% of the overall variance, whereas the next three axes accounted for 8.32%, 7.44%, and 6.62%, respectively. Similar to the result in mitochondrion, there was no independent single trend in codon bias in the dataset. Axis 1 was significantly correlated to *ENC* (r = 0.470, *P* < 0.01), but not to *GC* and *GC3s* (r = 0.045, 0.223, *P* > 0.05). The correlation analysis suggested that the expression level reflecting selective constraint would be the major factor in shaping the codon bias in chloroplast genome.

In this paper, we only showed the result of the combined correspondence analysis (CA) from the first randomization because the results obtained from the 10 independent correspondence analyses were similar (Figure 4A). Axis 1 accounted for 41.73% of the total variance, whereas the next three axes accounted for 4.90%, 4.50%, and 3.34%, respectively. Meanwhile, Axis 1 was significantly correlated to *ENC*, *GC* and *GC3s* (r = –0.527, –0.948, –0.988, *P* < 0.01), indicating that the difference in GC compositions between the GC-rich nuclear genes and the GC-poor organelle genes was the cause of gene discrimination on the first axis. It was also easy to distinguish the G- or C-ending codons from A- or U-ending ones (Figure 4B).

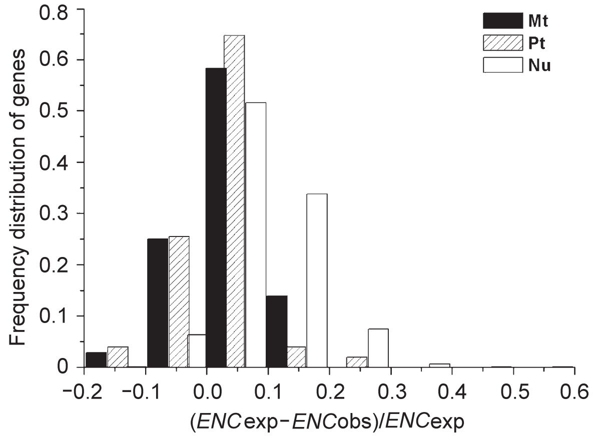

Figure 3. The frequency distribution of effective number of codons (*ENC*) ratio.

Mt, Pt and Nu denote mitochondrion, chloroplast and nuclear genes respectively.

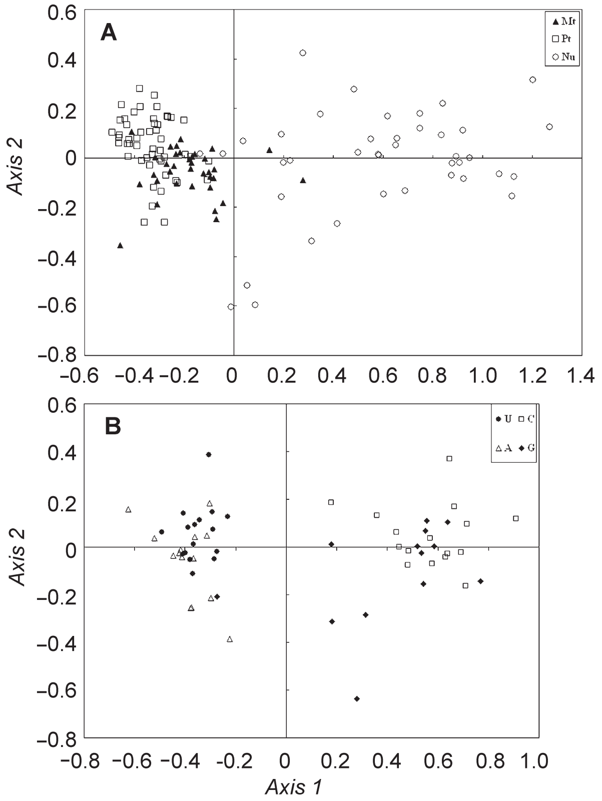

Figure 4 The results of correspondence analysis from combined data.

(A) Distribution of wheat mitochondrion, chloroplast and nuclear genes on the plot defined by the two main axes in the correspondence analysis, which represented 41.73% and 4.90% of the total variance.

In the plot, Mt (▲), Pt (□) and Nu (○) denote mitochondrion, chloroplast and nuclear genes respectively.

(B) Projection of the 59 degenerate codons on the first two axes in the correspondence analysis where U (●), C (□), A (△) and G (◆) denote codons ending with U, C, A, and G respectively.

Preferred codons

The *RSCU* values of genes within high and low bias datasets were compared. In particular, for mitochondrion

Table 2. “Preferred” codons of the mitochondrion, chloroplast and nuclear genes in *Triticum aestivum* based on their *RSCU* values

		Mt	Pt	Nu			Mt	Pt	Nu			Mt	Pt	Nu			Mt	Pt	Nu
Phe	TTT				Ser	TCT			*	Tyr	TAT	*			Cys	TGT		*	
	TTC		*	*		TCC		*	*		TAC			*		TGC			*
Leu	TTA					TCA				Ter	TAA				Ter	TGA			
	TTG	*				TCG		*			TAG				Trp	TGG			
	CTT				Pro	CCT				His	CAT				Arg	CGT		*	
	CTC			*		CCC					CAC			*		CGC			*
	CTA			*		CCA			*	Gln	CAA	*		*		CGA		*	
	CTG	*		*		CCG					CAG					CGG			
Ile	ATT		*		Thr	ACT			*	Asn	AAT				Ser	AGT			
	ATC			*		ACC		*	*		AAC		*	*		AGC			
	ATA					ACA				Lys	AAA				Arg	AGA			
Met	ATG					ACG					AAG			*		AGG	*		*
Val	GTT				Ala	GCT				Asp	GAT	*			Gly	GGT	*	*	
	GTC			*		GCC	*		*		GAC		*	*		GGC			
	GTA	*	*			GCA				Glu	GAA					GGA	*		*
	GTG	*				GCG					GAG			*		GGG			*

Asterisk (*) denotes codons that have been putatively identified as preferred codons.

and chloroplast genes, the codon usage pattern of highly expressed genes, ribosomal protein coding genes, was used to compare with that in low-biased datasets. To do so, we identified a somewhat new and different subset of preferred codons (Table 2). Overall 11, 12 and 24 triplets whose Δ*RSCU* were larger than 0.08 were defined as preferred codons in mitochondrion, chloroplast and nuclear genes, respectively. These codons were significantly correlated with translation levels and would be helpful for the design of degenerate primers and investigation of evolutionary history of *T. aestivum*.

PR2 plot analysis

The association between purines (A and G) and pyrimidines (C and T) in four-codon amino acid families was analyzed by PR2 plot (Figure 5). For organelle genes, pyrimidines were relatively used more frequently than purines at the third codon position (T: 26/36 in mitochondrion and 36/51 in chloroplast; C: 29/36 in mitochondrion and 40/51 in chloroplast). However, in nuclear genes, usage variation between purines and pyrimidines was primarily shown on difference between G and C contents, while that not significantly between A and T contents (C: 1017/1235, T: 692/1235). This observation indicated that despite the different GC contents between organelle and nuclear genes, the patterns of codon choice were similar at a certain extent. In other words, a higher proportion of pyrimidines (C and T) were observed against purines (A and G) in wheat though there was a relatively weak selection pressure on nuclear genes.

Discussion

Codon usage bias in genes is an important evolutionary phenomenon and has been increasingly documented in a wide range of organisms from prokaryotes to eukaryotes. The neutral theory and natural selection theory have been used to explain the origin of codon usage bias (Sharp and Li 1986; Bulmer 1999; Nakamura et al. 1999). If mutation occurs on the third codon position neutrally, it would result in random synonymous codon choice where GC and AT be used proportionally among the degenerate codon groups in a gene. In contrast, if translational selection pressure influences the shaping of codon usage, the bias would be significantly positive correlated to expression levels and some translation-preferred codons appear to be used more frequently than others. Previous studies demonstrated that genes within one species often shared a single codon usage pattern, except for a few species. For instance, synonymous codon usage in *Bacillus subtilis* was reflected by both translation selection and neutral mutational biases (Shields and Sharp 1987). In this study, we focused on the features of codon usage patterns among mitochondrion, chloroplast and nuclear genes within *T. aestivum*, and our results indicated that the nucleotide composition mutation bias could be considered as the relatively major factor in shaping the codon usage. It was consistent with the conclusion in *O. sativa* (Liu et al. 2004a). However, after comparative analyses of codon usage patterns among mitochondrion, chloroplast and nuclear genes, we also found that the constraints as the

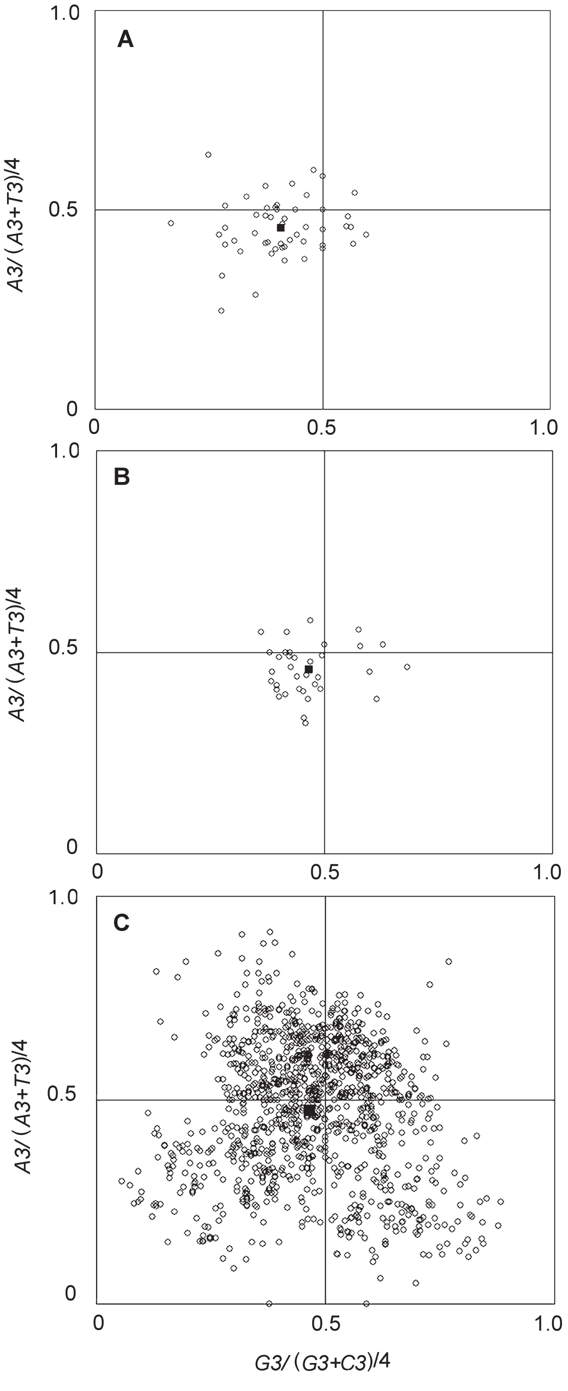

Figure 5. The PR2-bias plot of the third codon position of four-codon amino acids. The average position is shown as ■.

(A) Mitochondrion genes. AT-bias: 0.456 ± 0.061 *SD* and GC-bias: 0.465 ± 0.077 *SD*.

(B) Chloroplast genes. AT-bias: 0.455 ± 0.073 *SD* and GC-bias: 0.409 ± 0.093 *SD*.

(C) Nuclear genes. AT-bias: 0.474 ± 0.158 *SD* and GC-bias: 0.419 ± 0.108 *SD*.

primary factors in shaping the codon usage among mitochondrion and chloroplast genes were different from that in nuclear ones.

The G+C content could be one of the most important factors during the evolution of genomic structures (Bellgard et al. 2001). The GC content levels were much lower in mitochondrion and chloroplast genes than that in nuclear ones in wheat. Unlike *GC3*, *GC1* and *GC2* are subject to functional constraints against change because a mutation at these positions usually leads to an amino acid change, except between some codons of arginine, leucine, or serine (Sueoka 1988). Our neutrality plot revealed the relationship between *GC12* and *GC3s*, which may be helpful to examine the mutation-selection equilibrium in shaping the codon usage in the mitochondrion, chloroplast and nuclear genes. In plotting *GC12* against *GC3*, each point represented an individual gene. If *GC12* was as neutral as *GC3* against selection, the points should be distributed along the diagonal line (slope of unity). In contrast, if *GC12* was completely non-neutral, the points should be on the parallel lines of abscissa (slope of zero). Thus, the regression coefficient (slope) provided a measure of relative neutrality of *GC12* to *GC3*. The extent of the slope smaller than unity indicated the extent of neutrality of *GC12* less than that of *GC3* (Sueoka 1988, 1999a). Both the neutral mutation and selective constraint play roles in shaping codon usage pattern of nuclear DNA sequences according to Sueoka's mutation-selection equilibrium (Sueoka 1988). Selection against mutational bias could cause narrow GC distributions and no correlation between *GC12* and *GC3s* (Kawabe and Miyashita 2003). In our analysis, there were weak correlations between two *GC12* and *GC3s* in mitochondrion and chloroplast genes, suggesting that there were low mutation biases or high conservation of GC content levels throughout the organelle genomes. The correspondence analysis also confirmed that the composition mutation was not the major factor in shaping codon usage patterns in mitochondrion and chloroplast genes. As the organelles that could self-code, mitochondrion and chloroplast were considered to be from prokaryotes and there were divergent evolutionary constraints on their genomes and nuclear genome (Mcintosh et al. 1998). The genes located on the + or – strand in mitochondrion and chloroplast genomes are different from genes on both the two complementary strands in the nuclear genome. These features in the organelle genomes might correlate to high conservation, RNA editing, and strong purifying selection to reduce error copy in replication.

The *ENC* is often used to measure the overall codon bias for an individual gene in population genetics without the knowledge of the optimal codons or a reference set of highly expressed genes. In *EN*C-plot, a comparison of the observed distributions of genes with the expected ones from *GC3s* could reveal whether codon biases of genes were influenced by factors other than composition constraint (Wright 1990). If a given gene is only subject to G+C composition mutation constraint, it will lie above or just below the standard curve. On the other hand, if a particular gene is under pressure from direct expression selection, its *ENC* value will decrease from the expected one and it will lie significantly below the curve. In our *ENC*-plot of nuclear genes, at *GC3s* around the level of 0.4, there were some genes displaying a more biased codon usage than the expected one according to the respective *GC3s*. Most of these genes belong to two seed storage protein (SSP) gene families, i.e., gliadin and glutenin. These important proteins are primary components of wheat seed and determine wheat product quality. During seed germination, as a nutritional source, SSPs provide abundant amino acids. Thus, it requires that the SSPs are proteins keeping high expression levels in seeds during the late stages of seed development. For human diets, SSPs in poaceous plants are the major food source. However, SSPs do not necessarily own ideal compositions of essential amino acids for human needs (Fujiwara et al. 2002).

We also demonstrated an alterative strategy to redefine the preferred codons to make it more restrictively correlated to the translation level. As codon bias is considered a benefit to translation efficiency and expression level, preferred codons, which were those codons used frequently in each degenerate codon family in highly biased gene groups were determined. Previous methods to infer these codons were based on the data grouping results from multivariate exploratory analysis such as CA, which indicated the major trend and factor in shaping codon usage. If natural selection favors the usage of a subset of codons in highly expressed genes, it should be possible to infer the translation-preferred codons by positive correlations of their usages with expression (Kliman and Henry 2005). When the principal trend is assumed to correlate to gene expression, then two groups of datasets, from the extreme values at two ends of the first axis, are defined as high or low bias datasets (Liu and Xue 2005b). Here, three conditions must be satisfied: (i) the Axis 1 should explain a large proportion (> 15%) of the total variation in the data and should ideally explain approximately two times as much of the variation as the second and subsequent axes; (ii) the ordination of the genes on the principal axis should be significantly correlated with some independent measure of codon bias such as *ENC*; and (iii) to reduce the GC composition effect, particularly where there was a significant correlation between the principal axis and some measure of base composition (such as *GC3s* or *GC*) (Peden 1999). However, in this case, the method dependent on the CA result was not suitable because of a complicated factor combined neutral mutation and translational selection pressure affected on codon usage in *T. aestivum*. Particularly for mitochondrion and chloroplast genes, a single trend of codon usage could not be identified. Therefore, we adopted an alterative strategy to define the preferred codons in the present study.

Currently, only five complete mitochondrion genomes of vascular plants are available in GenBank. Obviously, as more plant genomic sequences are released, further comparative analyses are necessary for investigating different genetic and environmental constraints influencing the codon usage patterns at intra- and inter-genome levels.

Materials and Methods

Sequence data

The complete mitochondrion and chloroplast genome sequences of *Triticum aestivum* (accession numbers: NC_007579 and NC_002762, respectively) were obtained from GenBank (release 152.0). The complete CDSs of *T. aestivum* nuclear genes were downloaded according to the CUTG (Codon Usage Tabulated GenBank) database records (Nakamura et al. 1997). All the annotated CDSs were extracted with a PERL script developed by us that integrated with Coderet program in the EMBOSSwin software package (Rice et al. 2000). A rapid check of the CDSs applicability was performed based on the following considerations: (i) the presence of a START codon beginning and a STOP codon ending in each CDS was required; (ii) each CDS was greater than 300 nucleotides in length, because the *ENC* could not be calculated correctly and statistical analysis of codon usage would not be significant with shorter coding sequences (Wright 1990); and (iii) exact duplicated sequences were detected and excluded from the dataset. The final dataset contained 36 sequences for mitochondrion, 51 sequences for chloroplast and 1 235 sequences for nuclear genes.

Data analysis

Measures of codon usage bias

Two single codons for methionine (ATG), tryptophan (TGG), and three stop codons (TAA, TAG, TGA) were excluded from our analysis. *GC3s* value was defined as the proportion of G+C at the third variable coding position of synonymous codons. Similarly, *A3s*, *G3s*, *C3s*, and *T3s* could also be deduced by analogy to quantify the usage of each base at synonymous third codon positions. G+C content of each entire gene, first and second third codon positions (*GC*, *GC1*, and *GC2*, respectively) were also calculated. *GC12* was the average of *GC1* and *GC2* and was used for neutrality plot analysis.

ENC-plot (a plot of *ENC* vs *GC3s*) was used to provide a visual display of the main features of codon usage patterns for a number of genes. The values of *ENC* were always between 20 (when only one codon was effectively used for each amino acid) and 61 (when codons were used randomly). If the observed *ENC* value exceeded 61 (i.e., codon usage was more evenly distributed than expected), it would be adjusted to 61

(Wright 1990). The expected *ENC* values from *GC3s* under H_0 (null hypothesis, i.e., no selection) were calculated according to Equation 1, where *s* denoted *GC3s* (Wright 1990):

$$Nc = 2+S+\frac{29}{S^2+(1=S)^2} \qquad (1)$$

The difference between actual and expected *ENC* values could be estimated as (*ENC*exp–*ENC*obs) / *ENC*exp (*ENC* ratio), which excludes the effect of difference in GC content caused by neutral mutation (Kawabe and Miyashita 2003).

To determine the preferred codon for each synonymous codon group, we also calculated the ‘relative synonymous codon usage’ (*RSCU*); an index to normalize the dataset of codon usage from samples of different amino acids compositions which was defined as Equation 2:

$$RSCU = xi \,/\, \bar{x} \qquad (2)$$

where $\bar{x}$ was the mean number of use of all synonymous codons coding a given amino acid, and X_i was the number of occurrences of the *i*th codon for the amino acid (Sharp et al. 1986).

PR2-bias plot analysis

Parity rule 2 (PR2) is an intra-strand rule where A=T and G=C are expected if there is no bias in mutation and selection between the two complementary strands of DNA. The types of bias from PR2 are expressed as AT-bias and GC-bias defined as A/(A+T) and G/(G+C), respectively (Sueoka 1995). PR2-bias plots were particularly informative when PR2 biases at the third codon position of the four-codon amino acids of an individual gene (Sueoka 1999b). In the present study, “A3/(A3+T3)|4” and “G3/(G3+C3)|4” were plotted as the ordinate and abscissa respectively. The center of the plot, where both coordinates were 0.5, was the place where A=T and G=C, holding PR2. A vector from the center represented the extent and direction of the PR2-bias. In our study, the four-codon amino acids were threonine, proline, arginine (CGA, CGT, CGG, CGC), leucine (CTA, CTT, CTG, CTC), alanine, glycine, valine, and serine (TCA, TCT, TCG, TCC).

Correspondence analysis

Correspondence analysis (CA) is an ordination technique that identifies major trends in the variation of data and distributed genes along continuous axes in accordance with these trends. In particular, the dispersion of the genes on the two main axes identifies the trend that explains the major variation (Greenacre 1984). To investigate the variation of *RSCU* values among different genes, the CA was performed using CodonW version 1.4.2 (John Peden, Oxford University, available from http://bioweb.pasteur.fr/seqana/interfaces/codonw.html). According to the usage of 59 sense codons, all genes used in this study were plotted in a 58-dimensional hyperspace. Correlation analysis was performed using SPSS version 12.0, Origin 7.5 and Microsoft Excel.

Three datasets from mitochondrion, chloroplast and nuclear genes were also combined for further CA. Since there were many more nuclear genes than mitochondrion and chloroplast ones, we randomly selected 40 nuclear genes so as to avoid statistical bias in the CA. After the random selection was repeated 10 times (Lerat et al. 2002; Liu and Xue 2005b), we obtained 10 new mixed datasets that each contained 36 mitochondrion genes, 51 chloroplast genes, and 40 nuclear genes.

Identification of preferred codons

Despite the difference in GC contents among genes, *ENC* ratio was used to estimate the difference between observed and expected *ENC* values. This independent index could be used as a standard to distinguish the high and low bias datasets that translation-coupled. In this study, we calculated an *ENC* ratio for each CDS then ordered the sequences by their *ENC* ratio values. Using 5% sequences from the two ends of the ordered dataset (excluding the sequences with negative values), we formed two subsets called high and low bias datasets, respectively. The high-bias dataset was assigned to the genes with the higher overall *ENC* ratios, suggesting that their observed *ENC* values were far from the expected ones predicted from the GC contents and the codon usage was strongly biased. We compared the *RSCU* values between the two subsets. If the difference between them (Δ*RSCU*) was larger than 0.08, then this corresponding codon would be defined as the preferred codon (Liu et al. 2004a).

Acknowledgements

We thank the Bioinformatics Working Group in School of Life Sciences at Fudan University for helpful discussion.

References

Akashi H (1994). Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics* **136**, 927–935.

Bellgard M, Schibeci D, Trifonov E, Gojobori T (2001). Early detection of G + C differences in bacterial species inferred from the comparative analysis of the two completely sequenced *Helicobacter pylori* strains. *J. Mol. Evol.* **53**, 465–468.

Brown WM, George M, Wilson AC (1979). Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **76**, 1967–1971.

Bulmer M (1988). Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *J. Mol. Biol.* **1**, 15–26.

Chiapello HF, Lisacek F, Caboche M, Henaut A (1998). Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209**, 1–38.

Fennoy SL, Bailey-Serres J (1993). Synonymous codon usage in *Zea mays* L. nuclear genes is varied by levels of C and G-ending codons. *Nucleic Acids Res*. **21**, 5294–5300.

Fujiwara T, Nambara E, Yamagishi K, Goto DB, Naito S (2002). Storage Proteins. In: Somerville CR, Meyerowitz EM, eds. *The Arabidopsis Book*. American Society of Plant Biologists, Rockville, MD. pp. 1–12.

Greenacre MJ (1984). *Theory and applications of correspondence analysis*. Academic Press, London.

Jukes TH, Bhushar V (1986). Silent nucleotide substitution and G content of some mitochondria and bacterial genes. *J*. *Mol*. *Evol*. **24**, 39–44.

Kano A, Andachi Y, Ohama T, Osawa S (1991). Novel anticodon composition of transfer RNAs in Micrococcus luteus, a bacterium with a high genomeic G content. *J*. *Mol*. *Biol*. **221**, 387–401.

Kawabe A, Miyashita NT (2003). Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet*. *Syst*. **78**, 343–352.

Kliman RM, Henry AN (2005). Inference of codon preferences in *Arabidopsis thaliana*. *Int*. *J*. *Plant Sci*. **166**, 3–11.

Lerat E, Capy P, Biémont C (2002). Codon usage by transposable elements and their host genes in five species. *J*. *Mol*. *Evol*. **54**, 625–637.

Liu Q, Dou S, Ji Z, Xue Q (2005a). Synonymous codon usage and gene function are strongly related in *Oryza sativa*. *BioSystems* **80**, 123–131.

Liu Q, Feng Y, Xue Q (2004b). Analysis of factors shaping codon usage in the mitochondrion genome of *Oryza sativa*. *Mitochondrion* **4**, 313–320.

Liu Q, Feng Y, Zhao X, Dong H, Xue Q (2004a). Synonymous codon usage bias in *Oryza sativa*. *Plant Sci.* **167**, 101–105.

Liu Q, Xue Q (2005b). Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J*. *Genet*. **84**, 55–62.

Martin AP, Naylor GJP, Palumbi SR (1992). Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. *Nature* **357**, 153–155.

McIntosh MT, Srivastava R, Vaidya AB (1998). Divergent evolutionary constraints on mitochondrial and nuclear genomes of malaria parasites. *Mol*. *Biochem*. *Parasit*. **95**, 69–80.

Morton BR (1998). Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J*. *Mol*. *Evol*. **46**, 449–459.

Morton BR (2003). The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. *J*. *Mol*. *Evol*. **56**, 616–629.

Nakamura Y, Gojobori T, Ikemura T (1997). Codon usage tabulated from the international DNA sequence database. *Nucleic Acids Res*. **25**, 244–245.

Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M et al. (2002). The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: Frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol*. *Genet*. *Genomics* **268**, 434–445.

Osawa S, Ohama T, Yamao F, Muto A, Jukes TH, Ozeki H et al. (1988). Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **85**, 1124–1128.

Peden JF (1999). Anlysis of Codon Usage. PhD thesis. University of Nottingham, Nottingham.

Ramirez V, Savoie P, Morais R (1993). Molecular characterization and evolution of a duck mitochondrial genome. *J*. *Mol*. *Evol*. **37**, 296–310.

Rice P, Longden I, Bleasby A (2000). EMBOSS: The european molecular biology open software suite. *Trends Genet*. **16**, 276–277.

Salinas J, Matassi G, Montero LM, Bernardi G (1988). Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. *Nucleic Acids Res*. **16**, 4269–4285.

Schuster W, Brennicke A (1994). The plant mitochondrial genome: physical structure, information content, RNA editing, and gene migration to the nucleus. *Annu*. *Rev*. *Plant Physiol*. *Mol*. *Biol*. **45**, 61–78.

Sharp PM, Li WH (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J*. *Mol*. *Evol*. **24**, 28–38.

Sharp PM, Tuohy TM, Mosurski KR (1986). Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*. **14**, 5125–5143.

Shields DC, Sharp PM (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res*. **15**, 8023–8040.

Sueoka N (1988). Directional mutation pressure and neutral molecular evolution. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **85**, 2653–2657.

Sueoka N (1995). Intra-strand parity rules of DNA base composition and usage biases of synonymous codons. *J*. *Mol*. *Evol*. **40**, 318–325.

Sueoka N (1999a). Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra- strand rule of *A=T* and *G=C*. *J*. *Mol*. *Evol*. **49**, 49–62.

Sueoka N (1999b). Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene* **238**, 53–58.

Wright F (1990). The “effective number of codons” used in a gene. *Gene* **87**, 23–29.

(Handling editor: Yong-Biao Xue)