RESEARCH ARTICLE

Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species

QINGPO LIU* and QINGZHONG XUE*

Department of Agronomy, College of Agriculture and Biotechnology, Zhejiang University, 310 029, Hangzhou, China

Abstract

A detailed comparison was made of codon usage of chloroplast genes with their host (nuclear) genes in the four angiosperm species *Oryza sativa*, *Zea mays*, *Triticum aestivum* and *Arabidopsis thaliana*. The average GC content of the entire genes, and at the three codon positions individually, was higher in nuclear than in chloroplast genes, suggesting different genomic organization and mutation pressures in nuclear and chloroplast genes. The results of Nc-plots and neutrality plots suggested that nucleotide compositional constraint had a large contribution to codon usage bias of nuclear genes in *O. sativa*, *Z. mays*, and *T. aestivum*, whereas natural selection was likely to be playing a large role in codon usage bias in chloroplast genomes. Correspondence analysis and chi-test showed that regardless of the genomic environment (species) of the host, the codon usage pattern of chloroplast genes differed from nuclear genes of their host species by their AU-richness. All the chloroplast genomes have predominantly A- and/or U-ending codons, whereas nuclear genomes have G-, C- or U-ending codons as their optimal codons. These findings suggest that the chloroplast genome might display particular characteristics of codon usage that are different from its host nuclear genome. However, one feature common to both chloroplast and nuclear genomes in this study was that pyrimidines were found more frequently than purines at the synonymous codon position of optimal codons.

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

Introduction

Non-random codon usage of codons from degenerate codon families has been observed in a wide range of biological systems from prokaryotes to eukaryotes. Due to different genomes having their own characteristic patterns of synonymous codon usage (Grantham *et al*. 1980), it has not been easy to provide a satisfactory explanation for the particular pattern that is found in a given genome. For example, among prokaryotes such as *Escherichia coli* and *Bacillus subtilis*, it is generally agreed that codon usage is attributable to the equilibrium between natural selection and compositional mutation bias (Bulmer 1988; Sharp *et al*. 1993). This paradigm has been successfully applied to several eukaryotes such as *Saccharomyces cerevisiae* (Sharp *et al*. 1986), *Drosophila melanogaster* (Shields *et al*. 1988) and *Caenorhabditis elegans* (Stenico *et al*. 1994). However, in some prokaryotes with extremely high A + T or G + C contents (Sharp *et al*. 1993), and in human genes (Karlin and Mrázek 1996), mutation bias is known as the major factor accounting for the variation in codon usage. In the case of plant species, it was found that the patterns of codon usage greatly differed between monocot and dicot species (Murray *et al*. 1989; Campbell and Gowri 1990; Fennoy and Bailey-Serres 1993; Chiapello *et al*. 1998; de Amicis and Marchetti 2000). In addition, a significant relationship between codon usage bias and gene expression level was observed (Iannacone *et al*. 1997; Rouwendal *et al*. 1997), which suggested stronger natural selection constraints on highly expressed genes to optimize translation efficiency by the use of major codons (Bulmer 1988).

*E-mail: liuqp@genomics.org.cn; qzhxue@hotmail.com.

Keywords. codon usage; bioinformatics; chloroplast; plant; genome.

As an important organelle of plants, the chloroplast has its own genomic environment and relatively stable genetic system, and plays a crucial role in the process of photosynthesis. The translation machinery in chloroplasts, e.g. 70 S ribosomes, tRNAs and basal translation factors, is known to be structurally similar to those in prokaryotes, leading to the suggestion that the translation mechanism and patterns of codon usage in chloroplasts might be similar to that in *Escherichia coli* (Sugiura 1992). However, most studies in plants mainly focused on codon bias in nuclear genomes. Nevertheless, Morton (1999) found that the asymmetry of two strands of the chloroplast genome from *Euglena gracilis* was the major factor contributing to codon bias. In addition, it was considered that context-dependent mutation played some roles in shaping codon usage of the chloroplast genomes of grass species (Morton 2003), although there was also evidence that the codon usage of certain chloroplast genes was influenced by selection (Morton 1998). One reason for interest in possible differences in codon usage bias between chloroplast and nuclear genomes is that such information may shed some light on the interactions between the chloroplast and its host nuclear genome. In this study, we investigated the general patterns of codon usage bias in the nuclear and chloroplast genomes of four plant species.

Materials and methods

Sequence data

The complete chloroplast genome sequences of four plant species (Accession Numbers: NC_001320 for *Oryza sativa*; NC_001666 for *Zea mays*; NC_002762 for *Triticum aestivum*; and NC_000932 for *Arabidopsis thaliana*) and full-length nuclear cDNA sequences were obtained from the GenBank database (release 140.0), with the exception of the cDNA sequence for *Oryza sativa* being downloaded from the KOME database (Kikuchi *et al*. 2003). A PERL script developed by us was used to retrieve the coding sequences. To minimize sampling errors (Wright 1990), we analysed only those CDS sequences that were 100 codons or more in length and had correct initial and termination codons.

Measures of codon usage bias

GC content of the entire gene, first, second, and third codon positions (GCall, GC1, GC2, and GC3 respectively) were calculated after excluding the tryptophan, methionine, and three stop codons. GC12 is the average of GC1 and GC2, and was used for neutrality plot analysis. GC3s value is the frequency of G + C at the third synonymously variable coding position. Relative synonymous codon usage (RSCU) values were calculated to normalize codon usage within datasets of differing amino acid compositions (Sharp and Li 1986). RSCU values greater than 1.0 indicate that the corresponding codons are used more frequently than the expected frequency whereas the reverse is true for RSCU values less than 1.0. We also calculated the effective number of codons (ENC), a commonly used measure of the magnitude of codon bias for an individual gene, yielding values ranging from 20, for a gene with extreme bias using only one codon per amino acid, to 61 for a gene with no bias using synonymous codons equally (Wright 1990). Among the well-characterized host genes, we identified the weakly and highly expressed genes for each species. The sequences in which ENC was less than 30 and those in which ENC was greater than 55 were considered to correspond to highly and weakly expressed genes respectively. This was justified, because codon bias is positively correlated with gene expression level in these four species (Fennoy and Bailey-Serres 1993; Chiapello *et al*. 1998; Duret and Mouchiroud 1999; Liu *et al*. 2004).

Correspondence analysis (COA)

Multivariate statistical techniques have successfully been used to investigate the variation of RSCU values among genes (Morton 1999; Musto *et al*. 2001; Grocock and Sharp 2002; Singer and Hickey 2003; Peixoto *et al*. 2003; Romero *et al*. 2003; Gupta *et al*. 2004). The most commonly used method is correspondence analysis (Greenacre 1984), in which all genes are plotted in a 59-dimensional hyperspace according to their usage of the 59 sense codons. This method can detect the difference in codon usage between gene sequences and identify the codons involved. Sequences in which a given codon is used in a similar fashion lie close to each other on the graph. CodonW 1.4, an integrated codon bias and correspondence analysis program, was used to perform the COA on our data.

Results and discussion

Base composition of chloroplast and nuclear genes

Table 1 shows the average per cent of GC for the first, second, and third codon positions of coding sequences of chloroplast genomes, and of the host genes, according to the four species. The global GC content and the per cent of GC for the three codon positions were all clearly higher in nuclear genes than in chloroplast sequences. Differences in GC content were largest at the third codon position, followed by the first and second codon position. These observations are consistent with the results reported by Salinas *et al*. (1988) and Kawabe and Miyashita (2003). Notably, it was found that chloroplast genes tended to have higher incidence of AU, especially at the third codon position, regardless of the GC-richness of the host species. In addition, the per cent of GC in the chloroplast genes of the GC-rich genomes of *O. sativa*, *Z. mays*, and *T. aestivum* was nearly as high as in the chlo-

roplast genes of the AU-rich genome of *A. thaliana*. In the analysis of mononucleotide composition, we found that A and U were more common in chloroplasts and *A. thaliana*, whereas C and G were more common in the three monocot species (table 1).

The GC content could be one of the most important factors in the evolution of genomic structures (Bellgard *et al*. 2001). Ikemura (1985) demonstrated that the correlation between codon usage bias and GC content in surrounding noncoding region could be taken as a support for directional mutation pressure. Codon usage bias of human genes was related to location in the genome because of the mosaic patterns of GC content (Bernardi 1993). However, in *Chlamydomonas reinhardtii* (Naya *et al*. 2001) and *Echinococcus spp*. (Fernandez *et al*. 2001) genomes that are GC-rich, no clear relationship between codon usage bias and GC content was found. In some unicellular eukaryotes, such as *Entamoeba histolytica* (Romero *et al*. 2000) and *Streptococcus pneumoniae* (Hou and Yang 2003) genomes, nucleotide compositional pressure played minor roles in codon usage variation, whereas it was the major factor in shaping codon usage in human genes (Karlin and Mrázek 1996). As discussed by Sharp and Matassi in their review (1994), codon usage in the mammalian genome could reflect the physical location of the genes, which in turn may simply reflect difference in mutation patterns. Thus, the difference in GC contents of entire genomes and of the three-codon positions in the present study suggests that chloroplasts and nuclear genomes might possess different genomic organization, in part due to different mutational pressures.

Variation in codon usage of chloroplast and nuclear genes

A plot of ENC against GC3s (Nc-plot) was used to detect the codon usage variation among the genes. Wright (1990) argued that the comparison of actual distribution of genes, with the expected distribution under no selection could be indicative if codon usage bias of genes is influenced by some factor (s) other than compositional constraints. If a particular gene is subject to G + C compositional constraints, it will lie on or just below the GC3s curve. Nc-plots of the four species (figure 1) showed that the patterns of Nc-plot of nuclear genes differed significantly from that of chloroplast genes. It is interesting to note that although there were a small number of genes lying on the continuous Nc-plot curve, a majority of the points with low ENC values were lying well below the expected curve in every genome, suggesting that apart from the compositional constraints other factors might have influences in dictating codon usage variation among genes (Hou and Yang 2003; Gupta *et al*. 2004).

We also calculated (ENCexp-ENCobs)/ENCexp, defined as "ENC ratio", to estimate the difference between observed and expected ENC values (Kawabe and Miyashita 2003). Figure 2 shows the frequency distributions of "ENC ratio". We observed that there was a single peak located in 0.1–0.2 of "ENC ratio" values in chloroplast genomes, while two peaks located in 0–0.1 and 0.1–0.2 of "ENC ratio" values in nuclear genomes. Most chloroplast CDS sequences have 0.1–0.3, whereas nuclear genes have 0–0.2 of "ENC ratio" values. This result indicated that most CDS sequences have ENCs smaller than expected ENC values from their GC3s, and suggested that the difference in codon usage bias between chloroplasts and their host genomes is dependent factors other than compositional constraints, which is consistent with the results of Nc-plot.

To examine the relationship between mutation bias and codon usage bias, neutrality plots (GC12 vs GC3) (Sueoka 1988) were performed (figure 3). It was observed that plant nuclear genomes had a wide range of GC3 (figure 1; 0.15–0.99 for *O. sativa*, 0.27–0.99 for *Z. mays*, 0.24–1.00 for *T. aestivum*, 0.15–0.71 for *A. thaliana*), while the chloroplast genomes had relatively narrow GC3 distributions (figure 1; 0.26–0.49 for *O. sativa*, 0.17–0.51 for *Z. mays*, 0.20–0.36 for *T. aestivum*, 0.15–0.34 for *A. thaliana*). More importantly, there existed four signifi-

Table 1. Base composition of chloroplast and host genes.

	Number of sequences	GC First position	GC Second position	GC Third position	GC all	A	U	C	G
Chloroplast genes									
O. sativa	63	0.4933	0.4046	0.2929	0.3969	0.2926	0.3106	0.1871	0.2097
Z. mays	66	0.4912	0.4051	0.2936	0.3966	0.2943	0.3100	0.1872	0.2085
T. aestivum	55	0.4902	0.3991	0.2718	0.3870	0.2990	0.3113	0.1820	0.2077
A. thaliana	59	0.4846	0.3946	0.2446	0.3746	0.3100	0.3189	0.1739	0.1972
Nuclear genes									
O. sativa	26935	0.6060	0.4445	0.6300	0.5628	0.2313	0.2165	0.2703	0.2819
Z. mays	766	0.6086	0.4455	0.6623	0.5721	0.2353	0.2179	0.2654	0.2814
T. aestivum	411	0.6074	0.4395	0.6932	0.5800	0.2345	0.2089	0.2749	0.2817
A. thaliana	53020	0.5309	0.4108	0.4098	0.4505	0.2839	0.2692	0.2061	0.2408

cant correlations in neutrality plots of the nuclear genomes (r = 0.540, 0.506, 0.365, and, – 0.013, $P < 0.01$ in all cases), which indicated the similar effect of intragenomic GC mutation bias on the GC content among all positions of codons. On the contrary, there was no clear relationship between the two factors in these four chloroplast genomes (r = 0.164, 0.188, 0.076, and 0.130, $P > 0.05$ in all cases), suggesting low mutation bias or high conservation of GC contents level throughout the whole genome. Kawabe and Miyashita (2003) discussed the possibility that selection against mutational bias could cause narrow distribution of GC contents and no correlation between GC12 and GC3. Accordingly, the above result may indicate high levels of selection against directional mutation in chloroplast genomes.

Synonymous codon usage bias is a very complex phenomenon that is correlated with many factors, such as base compositional mutation bias (Karlin and Mrázek 1996; Hou and Yang 2003), natural selection (Sharp and Li 1986; Duret and Mouchiroud 1999; Peixoto *et al*.

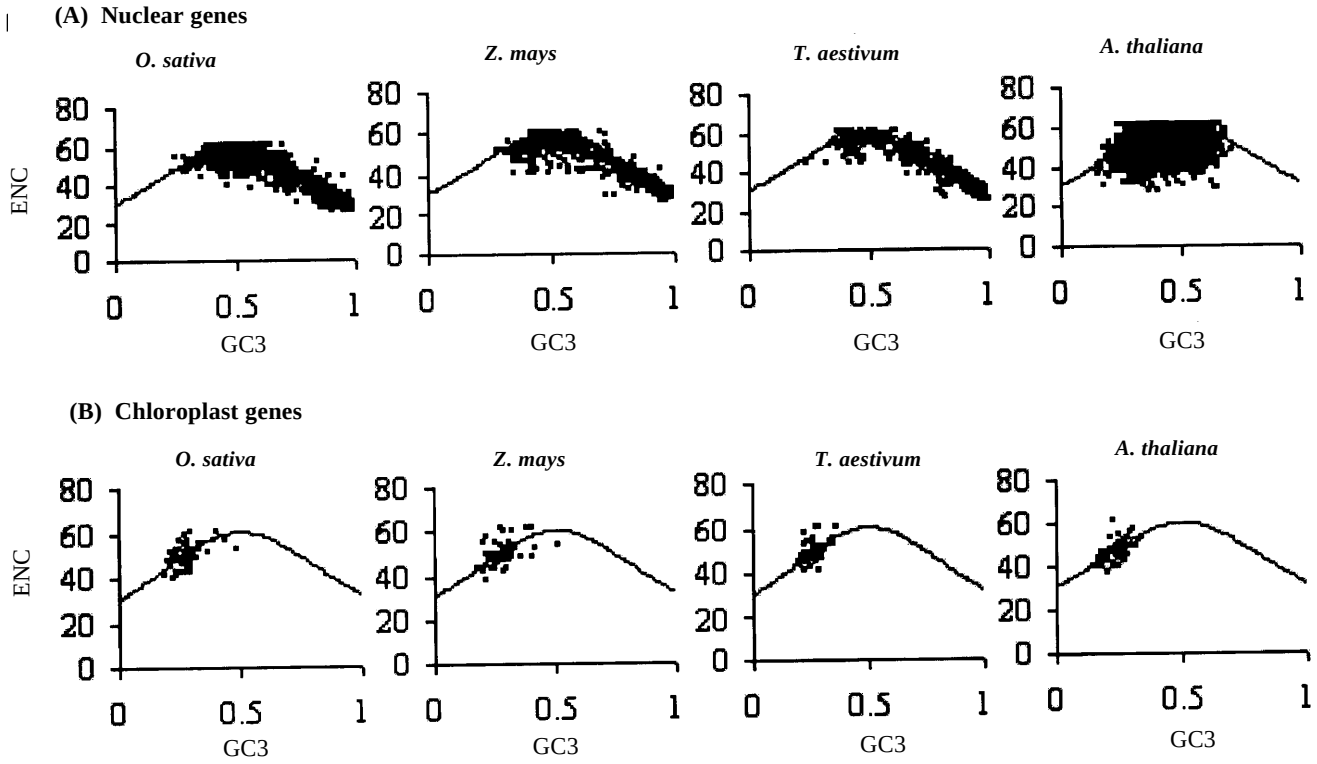

Figure 1. Nc-plots (ENC values vs GC3s) for nuclear genes and chloroplast sequences of four plant species. The continuous curve represents the expected curve between GC3s and ENC under random codon usage.

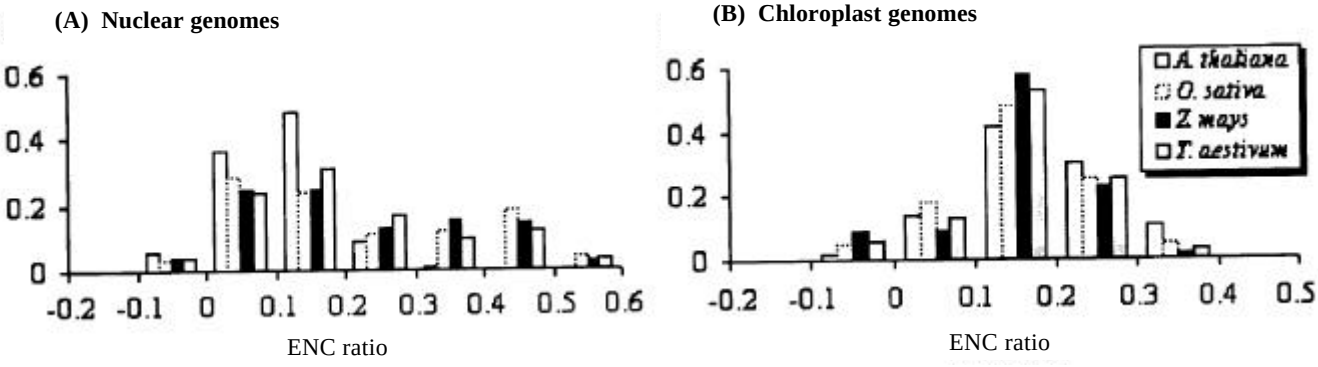

Figure 2. Frequency distribution of (ENCexp-ENCobs)/ENCexp for nuclear genes and chloroplast sequences of four plant species.

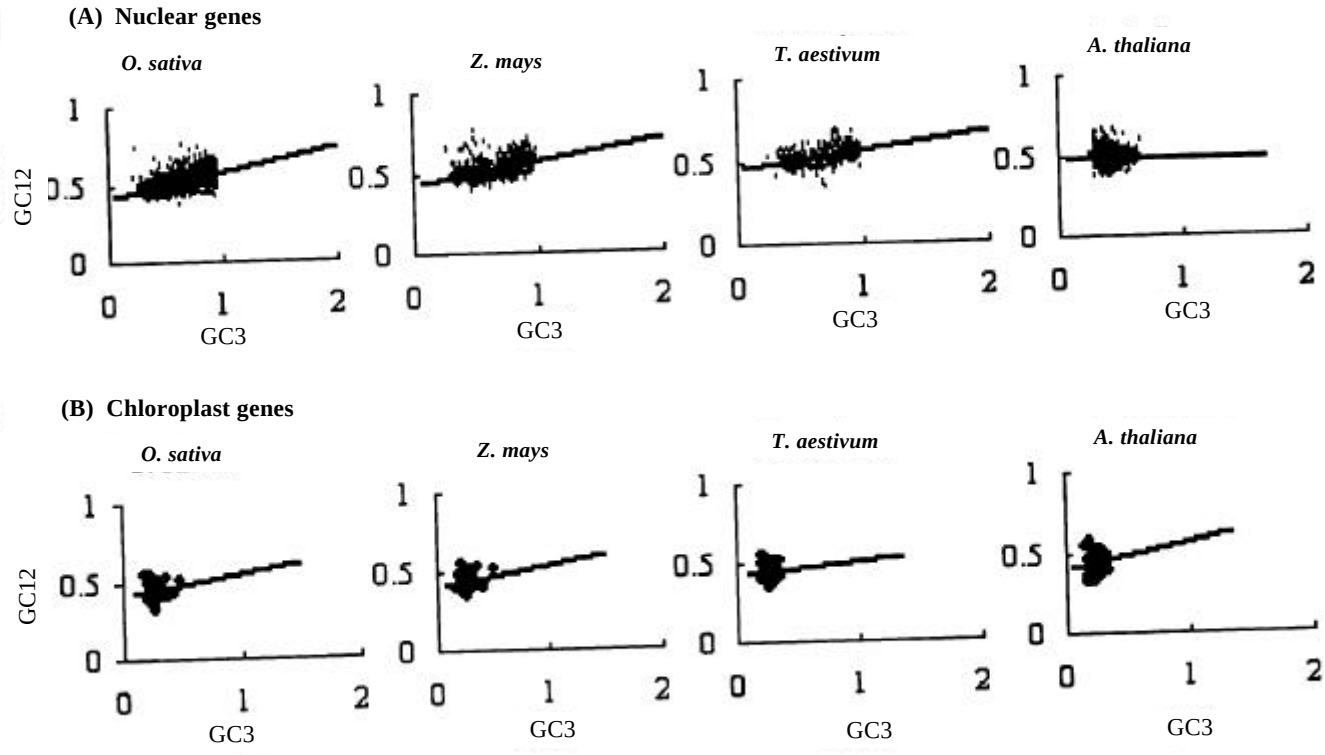

Figure 3. Neutrality plots (GC12 against GC3) for nuclear genes and chloroplast genes of four plant species.

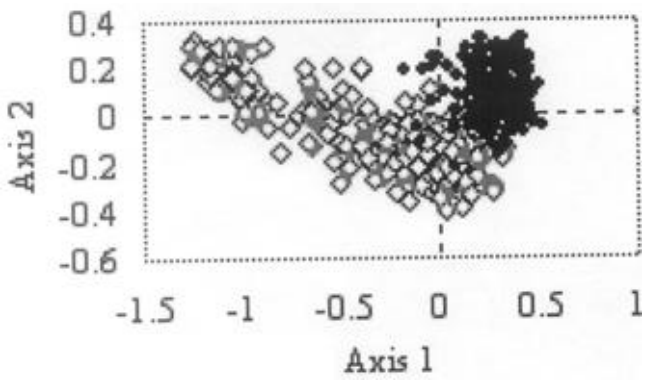

Figure 4. Projection of nuclear genes (white open diamonds) and chloroplast sequences (black circles) on the first two axes of COA, which represent 37.7% and 5.9% of the total variance, respectively.

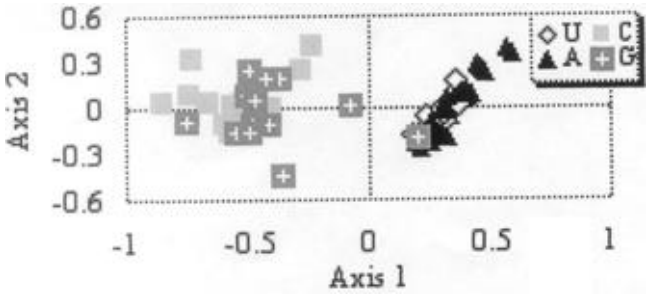

Figure 5. Projection of the 59 degenerate codons on the first two axes of the COA.

2003; Romero *et al*. 2003), gene length (Moriyama and Powell 1998), tRNA abundance (Percudani *et al*. 1997; Duret 2000), mRNA secondary structure (Gu *et al*. 2004), codon-anticodon interaction (Shi *et al*. 2001), the hydropathy level of each protein, and the amino acid conservation (Romero *et al*. 2000). However, the main contributors to codon bias appear to be compositional mutation bias and natural selection, with varying relative importance in different species (Singer and Hickey 2003; Peixoto *et al*. 2003; Romero *et al*. 2003; Gupta *et al*. 2004). In this study, we inferred that many factors contributed to codon usage variation in chloroplast and nuclear genomes, according to the results of Nc-plot and neutrality plot, and that the relative importance of different factors varied across chloroplast and nuclear genomes.

Correspondence analysis

Since there were many more host genes than chloroplast sequences, we randomly selected 63 nuclear genes for *O. sativa*, 66 for *Z. mays*, 55 for *T. aestivum*, and 59 for *A. thaliana* so as to avoid statistical bias in the correspondence analysis (COA). Thus, we analysed two gene groups,

Table 2. Average relative synonymous codon usage (RSCU) of the 59 degenerate codons for highly and weakly expressed host and chloroplast genes according to species.

Amino acids	Codons	*O. sativa*			*Z. mays*			*T. aestivum*			*A. thaliana*		
		High	Weak	Chloro-plast	High	Weak	Chloro-plast	High	Weak	Chloro-plast	High	Weak	Chloro-plast
Ala	GCA	0.10	1.14	1.17	0.09	1.04	1.16	0.14	1.15	1.21	0.71	1.10	1.11
	GCC	**2.30***	0.92	0.60	**2.35***	0.98	0.63	**2.55***	0.94	0.62	0.31*	0.67	0.61
	GCG	1.49*	0.70	0.50	1.34	0.75	0.46	1.22*	0.75	0.43	**1.88**	0.58	0.41
	GCU	0.11	**1.24**	**1.73**	0.22	**1.24**	**1.75**	0.09	**1.17**	**1.73**	1.10	**1.64**	**1.87**
Cys	UGC	**1.95***	**1.16**	0.58	**1.96***	**1.11**	0.55	**2.00***	**1.12**	0.46	0.11*	0.87	0.51
	UGU	0.05	0.84	**1.42**	0.04	0.89	**1.45**	0.00	0.88	**1.54**	**1.89**	**1.13**	**1.49**
Asp	GAC	**1.90***	0.80	0.46	**1.95***	0.87	0.45	**1.93***	0.84	0.43	0.97*	0.68	0.37
	GAU	0.10	**1.20**	**1.54**	0.05	**1.13**	**1.55**	0.07	**1.16**	**1.57**	**1.03**	**1.32**	**1.63**
Glu	GAA	0.06	0.87	**1.47**	0.03	0.83	**1.50**	0.06	0.93	**1.49**	0.86	**1.03**	**1.52**
	GAG	**1.94***	**1.13**	0.53	**1.97***	**1.17**	0.50	**1.94***	**1.07**	0.51	**1.14***	0.97	0.48
Phe	UUC	**1.97***	**1.02**	0.74	**1.90***	**1.08**	0.77	**1.97***	**1.01**	0.68	0.55*	0.98	0.64
	UUU	0.03	0.98	**1.26**	0.10	0.92	**1.23**	0.03	0.99	**1.32**	**1.45**	**1.02**	**1.36**
Gly	***GGA***	0.21	**1.08**	**1.52**	0.15	1.04	**1.56**	0.22	**1.11**	**1.57**	1.54	**1.46**	**1.66**
	GGC	**2.94***	1.02	0.44	**3.08***	**1.08**	0.44	**2.81***	1.05	0.46	0.62*	0.61	0.37
	GGG	0.69	0.84	0.81	0.64	0.84	0.79	0.48	0.82	0.68	0.09	0.66	0.64
	GGU	0.16	1.07	1.24	0.13	1.04	1.21	0.48	1.02	1.29	**1.76***	1.27	1.33
His	CAC	**1.93***	0.87	0.51	**1.85***	0.95	0.53	**1.95***	0.85	0.53	**1.44***	0.81	0.49
	CAU	0.07	**1.13**	**1.49**	0.15	**1.05**	**1.47**	0.05	**1.15**	**1.47**	0.56	**1.19**	**1.51**
Ile	AUA	0.04	0.75	0.90	0.01	0.75	0.97	0.06	0.72	0.94	0.50	0.74	0.96
	AUC	**2.90***	1.04	0.63	**2.92***	1.10	0.57	**2.88***	1.04	0.54	1.07*	1.06	0.53
	AUU	0.06	**1.21**	**1.48**	0.07	**1.15**	**1.45**	0.06	**1.24**	**1.52**	**1.43**	**1.20**	**1.51**
Lys	***AAA***	0.06	0.76	**1.44**	0.04	0.68	**1.43**	0.02	0.80	**1.46**	0.79	0.99	**1.54**
	AAG	**1.94***	**1.24**	0.56	**1.96***	**1.32**	0.57	**1.98***	**1.20**	0.54	**1.21***	**1.01**	0.46
Leu	CUA	0.02	0.61	0.86	0.02	0.61	0.88	0.01	0.64	0.87	0.09	0.67	0.81
	CUC	**3.90***	1.17	0.47	**2.94***	1.22	0.44	**3.65***	1.08	0.42	0.97*	1.06	0.40
	CUG	1.79*	1.22	0.33	2.87*	1.28	0.34	2.12	1.24	0.31	0.37	0.68	0.37
	CUU	0.10	**1.36**	1.32	0.07	**1.30**	1.35	0.09	**1.38**	1.25	0.83	**1.48**	1.25
	UUA	0.01	0.53	**1.92**	0.00	0.50	**1.88**	0.01	0.53	**2.05**	1.06	0.80	**2.03**
	UUG	0.17	1.11	1.09	0.10	1.08	1.10	0.12	1.12	1.09	**2.68**	1.30	1.13
Asn	AAC	**1.94***	0.96	0.57	**1.95***	0.99	0.52	**1.96***	0.97	0.52	**1.25***	0.99	0.48
	AAU	0.06	**1.04**	**1.43**	0.05	**1.01**	**1.48**	0.04	**1.03**	**1.48**	0.75	**1.01**	**1.52**
Pro	CCA	0.18	**1.29**	1.09	0.20	1.25	1.03	0.18	**1.45**	1.07	0.37	1.30	1.12
	CCC	1.62*	0.70	0.87	**1.75***	0.78	0.96	**2.06***	0.69	0.88	0.19*	0.48	0.76
	CCG	**2.06***	0.74	0.48	1.72*	0.76	0.48	1.58	0.67	0.46	1.40	0.75	0.52
	CCU	0.14	1.27	**1.57**	0.33	**1.20**	**1.53**	0.19	1.19	**1.60**	**2.05**	**1.46**	**1.60**
Gln	CAA	0.08	0.86	**1.52**	0.08	0.87	**1.53**	0.02	**1.02**	**1.56**	0.26	**1.11**	**1.57**
	CAG	**1.92***	**1.14**	0.48	**1.92***	**1.13**	0.47	**1.98***	0.98	0.44	**1.74***	0.89	0.43
Arg	AGA	0.04	1.25	1.80	0.06	1.24	**1.81**	0.09	1.31	**1.81**	1.29	**1.90**	**1.76**
	AGG	1.20	**1.57**	0.63	1.23	**1.40**	0.63	1.03	**1.43**	0.63	1.04*	1.24	0.63
	CGA	0.06	0.58	1.18	0.06	0.60	1.24	0.05	0.55	1.23	0.35	0.76	1.40
	CGC	**3.66***	0.94	0.58	**3.58***	1.12	0.52	**3.39***	1.03	0.54	0.26	0.48	0.46
	CGG	0.81	0.88	0.49	0.98	0.91	0.46	1.16	0.78	0.43	**1.64**	0.61	0.46
	CGU	0.24	0.78	1.32	0.08	0.73	1.33	0.27	0.90	1.36	1.42*	1.00	1.30
Ser	AGC	1.76*	1.08	0.41	1.91*	1.07	0.47	1.62*	1.14	0.47	**1.71***	0.79	0.35
	AGU	0.03	0.89	1.18	0.04	0.85	1.17	0.05	0.88	1.20	1.33	0.93	1.21
	UCA	0.16	**1.21**	1.02	0.03	1.10	0.94	0.05	**1.23**	1.03	0.81	1.23	1.20
	UCC	**2.76***	1.04	1.24	**2.91***	1.08	1.23	**3.38***	1.02	1.11	0.38*	0.79	0.89
	UCG	1.15*	0.64	0.53	0.91*	0.72	0.58	0.77	0.59	0.48	1.19	0.66	0.61
	UCU	0.14	1.14	**1.63**	0.20	**1.17**	**1.62**	0.14	1.13	**1.70**	0.57	**1.60**	**1.73**
Thr	ACA	0.06	**1.28**	1.09	0.10	**1.22**	1.09	0.06	**1.25**	1.16	0.99	1.20	1.25
	ACC	**2.69***	0.99	0.80	**2.54***	0.97	0.78	**2.83***	1.05	0.71	0.80*	0.82	0.72
	ACG	1.21*	0.55	0.43	1.26*	0.66	0.51	1.05*	0.62	0.45	**1.93**	0.66	0.41
	ACU	0.05	1.18	**1.67**	0.10	1.15	**1.62**	0.06	1.08	**1.68**	0.28	**1.32**	**1.62**
Val	GUA	0.02	0.55	1.44	0.02	0.53	1.46	0.03	0.60	**1.54**	0.22	0.61	1.41
	GUC	**2.05***	0.91	0.49	1.82*	0.93	0.49	**2.06***	1.00	0.49	0.53*	0.80	0.50
	GUG	1.84*	1.22	0.56	**2.10***	1.26	0.56	1.85*	1.14	0.50	1.59	1.04	0.57
	GUU	0.09	**1.32**	**1.51**	0.05	**1.28**	**1.49**	0.06	**1.25**	1.47	**1.66**	**1.55**	**1.51**
Tyr	UAC	**1.94***	**1.04**	0.46	**1.96***	**1.02**	0.43	**1.94***	0.95	0.43	**1.21***	0.98	0.36
	UAU	0.06	0.96	**1.54**	0.04	0.98	**1.57**	0.06	**1.05**	**1.57**	0.79	**1.02**	**1.64**

Note: The highest RSCU of codons for each amino acid is in boldface. Asterisks indicate optimal codons as found by Liu *et al*. (2004), Kawabe and Miyashita (2003), and Chiapello *et al*. (1998).

namely nuclear gene group and chloroplast gene group, both containing an equal number of genes. The random selection of genes was repeated 10 times (Lerat *et al*. 2002). Because the results obtained by the COA were almost similar for all randomizations, we only used the data from the first randomization.

Correspondence analysis of chloroplast and host genes identified a single major trend in codon usage: the first axis accounted for 37.7% of all variation among genes, whereas the next three axes only accounted for 5.9%, 4.1% and 3.0% respectively, confirming that the first axis was the main axis in explaining codon usage variation among genes in the eight genomes. The position of each gene on the plane defined by the first two axes is displayed in figure 4. Axis 1 coordinates were strongly negatively correlated with GC content and GC3s (r = – 0.939 and – 0.991, $P < 0.01$), which indicated that the discrimination on the first axis resulted mainly from differences in the overall GC composition of the GC-rich *O. sativa*, *Z. mays*, *T. aestivum* and the AU-rich chloroplast genes. Accordingly, it was easy to distinguish the A- and/or U-ending codons from that G- and/or C-ending ones (figure 5). From the positions of 59 codons projected onto the plane (figure 5), we speculated that A- and/or U-ending codons were more frequent in chloroplast genes than in nuclear genes for all species.

On the other hand, the results of chi-square test of codon biases indicated that codon usage in nuclear and chloroplast genes was distinct in each species (c^2 = 4321.7 for *O. sativa*, 5831.5 for *Z. mays*, 6929.9 for *T. aestivum*, and 2599.7 for *A. thaliana*, $P < 0.001$ in all cases). Table 2 shows the relative synonymous codon usage (RSCU) values of the 59 synonymous codons for the highly and weakly expressed host genes and the chloroplast gene sequences for each species. It was found that the codons found more often in chloroplast genes were similar to those used more frequently in weakly expressed nuclear genes. Highly expressed nuclear genes typically had G- and/or C-ending codons, whereas U- and/or A- ending codons were more frequently seen in chloroplast genes.

Conclusions

In the present study, there were three different and two similar findings among chloroplasts and their host nuclear genes. The first difference was in the GC contents at all codon positions, although the difference between chloroplasts and *A. thaliana* nuclear genes was smaller than that between chloroplasts and nuclear genes of the three monocot species (table 1). A second difference was that codon usage bias of chloroplast genomes and their host nuclear genomes appeared to be shaped to different degrees by factors such as selection and compositional constraints (figures 1, 2). For instance, the codon usage bias of rice nuclear genes was affected by nucleotide compositional constraint to a high degree, whereas natural selection was implicated as a major factor in shaping codon usage bias of the rice chloroplast genome (data not shown). The last difference we observed was that the codon usage of chloroplast genome significantly differed from its host nuclear genes according to species. All the chloroplast genes had greater proportion of A- and/or U-ending codons, whereas G- and/or C-ending codons were used more often in host nuclear genes. The fact that chloroplast genomes had similar biases of codon usage regardless of the genomic environment of their hosts strongly suggests that chloroplasts in general may have particular characteristics of codon usage.

On the other hand, we observed that pyrimidines (C and U) were found more frequently at the synonymous codon position than purines (G and A) in both chloroplast and nuclear genomes (C: 13/20 amino acids for *O. sativa*, *Z. mays*, and *T. aestivum*; C + U: 15/20 for *A. thaliana*; U: 12/20 for chloroplasts). Another similarity between chloroplast and nuclear genomes was the relatively higher conservation of GC content levels at the second position of codons (table 1), a pattern found in almost all investigated genomes. As more chloroplast genome data became available, further studies to confirm these results could be performed to investigate the differential codon usage patterns and the relationship between the chloroplast genomes and their hosts which has implications for our understanding of parallel gene transfers and nuclear-organelle gene interactions.

Acknowledgments

We are grateful for valuable input from Dr Amitabh Joshi and the anonymous referees. This work was supported by grants from the National Natural Science Foundation of China (3987042) and the Key Research Project of Zhejiang Province (2003C2007).

References

Bellgard M., Schibeci D., Trifonov E. and 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.

Bernardi G. 1993 The isochore organization of the human genome and its evolutionary history – a review. *Gene* **135**, 57–66.

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

Campbell W. H. and Gowri G. 1990 Codon usage in higher plants, green algae, and cyanobacteria. *Plant Physiol.* **92**, 1–11.

Chiapello H., Lisacek F., Caboche M. and Hénaut A. 1998 Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209**, GC1–GC38.

de Amicis F. and Marchetti S. 2000 Intercodon dinucleotides affect codon choice in plant genes. *Nucleic Acids Res.* **28**, 3339–3345.

Duret L. 2000 tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet*. **16**, 287–289.

Duret L. and Mouchiroud D. 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **96**, 4482–4487.

Fennoy S. L. and 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.

Fernandez V., Zavala A. and Musto H. 2001 Evidence for translational selection in codon usage in *Echinococcus spp*. *Parasitology* **123**, 203–209.

Grantham R., Gautier C. and Gouy M. 1980 Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res*. **8**, 1893–1912.

Greenacre M. J. 1984 *Theory and applications of correspondence analysis*. Academic Press, London.

Grocock R. J. and Sharp P. M. 2002 Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* **289**, 131–139.

Gu W., Zhou T., Ma J., Sun X. and Lu Z. 2004. The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *BioSystems* **73**, 89–97.

Gupta S. K., Bhattacharyya T. K. and Ghosh T. C. 2004 Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. *J. Biomol. Struct. Dyn*. **21**, 1–9.

Hou Z. C. and Yang N. 2003 Factors affecting codon usage in *Yersinia pestis*. *Acta Biochimica et Biophysica Sinica* **35**, 580–586.

Iannacone R., Grieco P. D. and Cellini F. 1997 Specific sequence modifications of a cry3B endotoxin gene result in high levels of expression and insect resistance. *Plant Mol. Biol*. **34**, 485–496.

Ikemura T. 1985 Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol*. **2**, 13–34.

Karlin S. and Mrázek J. 1996 What drives codon choices in human genes? *J. Mol. Biol*. **262**, 459–472.

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

Kikuchi1 S., Satoh K., Nagata T., Kawagashira N., Doi K., Kishimoto N., Yazaki J., Ishikawa M., Yamada H. and Ooka H. 2003 Collection, mapping, and annotation of over 28000 cDNA clones from *japonica* rice. *Science* **301**, 376–379.

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

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

Moriyama E. N. and Powell J. R. 1998 Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res*. **26**, 3188–3193.

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

Morton B. R. 1999 Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis*. *Proc. Natl. Acad. Sci. USA* **96**, 5123–5128.

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

Murray E. E., Lotzer J. and Eberle M. 1989 Codon usage in plant genes. *Nucleic Acids Res*. **17**, 477–498.

Musto H., Cruveiller S., Onofrio G. D., Romero H. and Bernardi G. 2001 Translational selection on codon usage in *Xenopus laevis*. *Mol. Biol. Evol*. **18**, 1703–1707.

Naya H., Romero H., Carels N., Zavala A. and Musto H. 2001 Translational selection shapes codon usage in the GC-rich genomes of *Chlamydomonas reinhardtii*. *FEBS Lett*. **501**, 127–130.

Peixoto L., Zavala A., Romero H. and Musto H. 2003 The strength of translational selection for codon usage varies in the three replicons of *Sinorhizobium melioti*. *Gene* **320**, 109–116.

Percudani R., Pavesi A. and Ottonello S. 1997 Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J. Mol. Biol*. **268**, 322–330.

Romero H., Zavala A. and Musto H. 2000 Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res*. **28**, 2084–2090.

Romero H., Zavala A., Musto H. and Bernardi G. 2003 The influence of translational selection on codon usage in fishes from the family Cyprinidae. *Gene* **317**, 141–147.

Rouwendal G. J. A., Mendes O., Wolbert E. J. H. and de Boer A. D. 1997 Enhanced expression in tobacco of the gene encoding green fluorescent protein by modification of its codon usage. *Plant Mol. Biol*. **33**, 989–999.

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

Sharp P. M. and Li W. H. 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol*. **24**, 28–38.

Sharp P. M. and Matassi G. 1994 Codon usage and genome evolution. *Curr. Opin. Genet. Dev*. **4**, 851–860.

Sharp P. M., Stenico M., Peden J. F. and Lloyd A. T. 1993 Codon usage: mutational bias, translational selection, or both? *Biochem. Soc. Trans*. **21**, 835–841.

Sharp P. M., Tuohy T. M. and Mosurski K. R. 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*. **14**, 5125–5143.

Shi X. F., Huang J. F., Liang C. R., Liu S. Q., Xie J. and Liu C. Q. 2001 Is there a close relationship between synonymous codon bias and codon-anticodon binding strength in human genes? *Chinese Sci. Bulletin* **12**, 1015–1019.

Shields D. C., Sharp P. M., Higgins D. G. and Wright F. 1988 "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol*. **5**, 704–716.

Singer G. A. C. and Hickey D. A. 2003 Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317**, 39–47.

Stenico M., Lloyd A. T. and Sharp P. M. 1994 Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res*. **22**, 2437–2446.

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

Sugiura M. 1992 The chloroplast genome. *Plant Mol. Biol*. **19**, 149–168.

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

Received 6 September 2004; in revised form 11 January 2005