Chloroplast genes transferred to the nuclear plant genome have adjusted to nuclear base composition and codon usage

J.L.Oliver*, A.Marín[1] and J.M.Martínez-Zapater[2]
Unidad de Genética, Facultad de Ciencias, Universidad de Granada, E-18071-Granada and [1]Departamento de Genética y Biotecnia, Facultad de Biología, Universidad de Sevilla, Aptdo. 1095, E-41080 Seville and [2]Departamento de Protección Vegetal, CIT-INIA, Carretera Gral. de La Coruña Km 7, E-28040-Madrid, Spain

Received October 17, 1989; Revised and Accepted November 30, 1989

ABSTRACT

During plant evolution, some plastid genes have been moved to the nuclear genome. These transferred genes are now correctly expressed in the nucleus, their products being transported into the chloroplast. We compared the base compositions, the distributions of some dinucleotides and codon usages of transferred, nuclear and chloroplast genes in two dicots and two monocots plant species. Our results indicate that transferred genes have adjusted to nuclear base composition and codon usage, being now more similar to the nuclear genes than to the chloroplast ones in every species analyzed.

INTRODUCTION

The existence of adjustment of base composition to different genomic G+C contents was shown in homologous genes and noncoding sequences of microorganisms and mitochondrial genomes[1]. Demonstration of the existence of such adjustment has been hindered in higher eukaryotes because of the compartmentalization of their genomes. Only recently, it has been shown that there is a compositional adjustment for *Alu* repetitive sequences that are located in different human genome compartments[2].

Coding sequences that have moved between different genomes can be good candidates to probe the existence of compositional adjustment and to analyze the involved mechanisms. Such gene movements have recurrently occurred along plant evolution and most of the plastid genes have been transferred to the nuclear genome[3,4,5,6]. Since in most plant species chloroplast and nuclear genomes have different GC contents[7], we think that plastid genes transferred to the plant nuclear genome can be an excellent model system to analyze if such compositional adjustment exists and, if so, how it works.

We compared nucleotide composition and codon usage of nuclear, chloroplast and nuclear genes encoding chloroplast proteins (transferred genes) in two dicots (pea and tobacco) and two monocots (wheat and maize) species. The distributions of some relevant dinucleotides were also studied. Results indicate that — at the level of base composition, dinucleotide distribution and codon usage — transferred genes are more similar to nuclear genes than to chloroplast ones. We analyzed how they have adjusted their base composition and codon usage to that of the nuclear environment.

DATA AND METHODS

Gene sequences

Sequences from nuclear and chloroplast genes were retrieved from the GenBank genetic sequence data bank[8] (release 57), or directly taken from original publications. We selected the four species with higher numbers of nuclear and chloroplast genes sequenced: *Nicotiana tabacum* (*Solanaceae*), *Pisum sativum* (*Leguminosae*), *Triticum aestivum* and *Zea mays* (*Poaceae*). Nuclear genes encoding chloroplast proteins can be considered as transferred chloroplast genes[5,6]. Although this seems to be a common situation, two exceptions of nuclear encoded chloroplast proteins that probably evolved from nuclear genes have already been described[9,10]. Table 1 shows the genes we have identified as transferred genes. A list of the remaining nuclear and chloroplast genes used in this study is shown in the Appendix.

Nucleotide composition data

Before nucleotide composition and codon usage were analyzed, introns of all genes and the sequence coding for the signal peptide present in transferred genes were removed. Nucleotide sites subject to silent changes ('silent sites') are calculated according to reference 1 [N = A, C, G, or T(U); R = A or G; Y = C or T(U)]: A, third positions of all codons, plus A in first positions of AGR codons; C, third positions of all codons, plus C in first positions of CTR and CGR codons; G, third positions of all codons, minus G in third positions of ATG and TGG codons; T, third positions of all codons, plus T in first positions of TTR codons.

* To whom correspondence should be addressed

Codon usage data

The following strategy was used to study codon usage in nuclear and chloroplast genes. First, we define as codon groups the sets of synonymous codons differing only in the third nucleotide. There is a single codon group for each amino acid, except for arginine, leucine and serine, each with two codon groups. We exclude from the analysis termination codons and single-codon groups (methionine and tryptophan). This leaves 21 codon groups with a total of 59 codons. Second, we count codon appearances in each gene and compute the relative frequency of each codon in each of the codon groups (the count of that codon divided by

Table 1. Chloroplast genes transferred to the nucleus of pea (PEA), tobacco (TOB), wheat (WHT) and maize (MZE) used in this study. Sequences were retrieved from GenBank (Release 57) or, when no GenBank LOCUS name is specified, directly from the indicated source.

GENE SYMBOL	GenBank LOCUS	PROTEIN
PEA:		
cab15	PEACAB15	Chlorophyll a/b-binding protein
cab80	PEACAB80	Major light harvesting protein AB80
rubp15	PEARUBP15	RuBisCo small subunit
fnr	(1)	Ferredoxin-NADP+ induced protein
elip	(2)	Early light-induced protein
rps18	(3)	Chloroplast ribosomal protein (CL18)
rps24	"	" " " (CL24)
rps25	"	" " " (CL25)
rps9	"	" " " (CL9)
TOB:		
gapA	TOBGAPA	Chloroplast GAPDH-A
gapB	TOBGAPB	Chloroplast GAPDH-B
rbpco	TOBRBPCO	RuBisCo, small subunit
als	(4)	Acetolactate synthase
WHT:		
cab	WHTCAB	Major chlorophyll a/b-binding protein
rbca	WHTRBCA	RuBisCo, small subunit
MZE:		
rbcs	(5)	RuBisCo, small subunit
gapB	(6)	Glyceraldehide-3-phosphate dehydrogenase B
cab	(7)	Chlorophyll a/b-binding protein

(1) Newman, B.J. and Gray, J.C. (1988) Plant Mol. Biol. 10, 511–520. (2) Kolanus, W., Scharnhorst, C., Kühne, U. and Herzfeld, F. (1987) Mol. Gen. Genet. 209, 234–239. (3) Gantt, J.S. (1988) Curr. Genet. 14, 519–528. (4) Mazur, B.J., Chui, C.F. and Smith, J. (1987) Plant Physiol. 85, 1110–1117. (5) Matsuoka, M., Kano-Murakami, Y., Tanaka, Y., Ozeki, Y. and Yamamoto, N. (1987) J. Biochem. 102, 673–676. (6) Brinkmann, H., Martínez, P., Quigley, F., Martin, W. and Cerff, R. (1987) J. Mol. Evol. 26, 320–328. (7) Matsuoka, M., Kano-Murakami, Y. and Yamamoto, N. (1987) Nucl. Acids Res. 15, 6302.

Table. 2 Nucleotide composition and GC content in replacement (RS) and silent (SS) sites (see text) of chloroplast (CP), transferred (TF) and nuclear (NUC) genes. Species abbreviations are as in Table 1.

GENOME	TOTAL G+C	RS					SS				
		A	T	C	G	G+C	A	T	C	G	G+C
PEA:											
NUC	.43	.31	.21	.19	.29	.48	.29	.37	.20	.14	.34
TF	.44	.30	.22	.19	.29	.48	.29	.34	.19	.18	.37
CP	.41	.25	.28	.20	.27	.47	.27	.44	.18	.11	.29
TOB:											
NUC	.45	.27	.23	.21	.29	.50	.24	.40	.23	.13	.36
TF	.48	.28	.22	.19	.31	.50	.21	.36	.25	.18	.43
CP	.41	.27	.24	.20	.29	.49	.30	.42	.17	.11	.28
WHT:											
NUC	.57	.28	.20	.27	.25	.52	.20	.13	.42	.25	.67
TF	.59	.25	.24	.20	.31	.51	.07	.19	.48	.26	.74
CP	.38	.31	.24	.19	.26	.45	.32	.41	.15	.12	.27
MZE:											
NUC	.57	.26	.22	.25	.27	.52	.14	.19	.40	.27	.67
TF	.67	.26	.22	.21	.31	.52	.02	.02	.67	.29	.96
CP	.42	.30	.21	.19	.30	.49	.33	.38	.18	.11	.29

the total of codons in the pertinent codon group); this method draws attention to the specific choices made by the organism among different options (the synonymous codons) regardless of the frequencies of the different amino acids in its proteins. Third, we compute the overall difference in codon usage between any two genes through a distance algorithm which is a version of the 'Manhattan metric' often used by numerical taxonomists. The codon usage distance between genes A and B is simply the sum of the absolute values of the differences in codon frequencies:

$$D\ (A,B) = \sum_{i=i}^{i=59} \mid x(i,A) - x(i,B) \mid$$

where x(i,A) and x(i,B) are the frequencies of the ith codon in genes A and B, respectively.

RESULTS

Similarity estimates between non-homologous genes

Measurement of similarity between nuclear and chloroplast genes requires comparisons of non-homologous sequences. Grantham[11] proposed the combination of four indexes, based on codon usage and GC content, to estimate the similarity among non-homologous genes from very different sources. When applied to our data, this method was unable to differentiate among the nuclear, transferred, and chloroplast gene sets from each species (data not shown); only the GC content of the third codon position, taken individually, was able to clearly differentiate chloroplast from nuclear genes in all four species analyzed (data not shown). Chloroplast genes transferred to the nuclear genome were always grouped among the nuclear genes using this index.

Nucleotide composition analysis

Changes in nucleotide composition of transferred genes after relocation in the nuclear genome were analyzed by studying base

Table 3. Distributions of CpG and TpA doublets in nuclear (NUC), transferred (TF) and chloroplast (CP) genes of different species. The ratio of the observed to the expected frequencies for these dinucleotides are given. Deviations from expectations were tested by Chi-square. Gene abbreviations are as in Table 1 and the Appendix.

TOB			PEA		
GENE	CpG	TpA	GENE	CpG	TpA
NUC:					
gapC	0.26 ***	0.45 ***	abn2	0.86	0.65 **
pr-1a	0.59 *	0.84	legJ	0.59 ***	0.48 ***
pr-1b	0.55 *	0.80 gs1	0.40 ***	0.71 **	
pr-1c	0.60 *	0.86 gs2	0.25 ***	0.59 ***	
ech	0.56 ***	0.62 **	gs3 0.39 ***	0.55 **	
pox	0.42 ***	0.75 *	lecA 0.52 **	0.72 *	
thaur	0.39 ***	0.67 *	vic 0.45 ***	0.57 ***	
			adh-1	0.39 ***	0.51 ***
			hsp	0.43 **	0.55 ***
TF:					
gapA	0.52 ***	0.35 ***	cab15	0.43 ***	0.71
gapB	0.48 ***	0.52 ***	cab80	0.44 ***	0.53 **
rbpco	0.48 *	0.63	rubp15	0.50 *	0.39 **
als	0.65 ***	0.74 **	fnr	0.50 **	0.54 ***
			elip	0.15 ***	0.60 **
			rps18	0.43	0.29 ***
			rps24	0.56	0.75
			rps25	0.95	0.36 **
			rps9	0.38 **	0.59 **
CP:					
psaA	0.59 ***	0.87	atpA	0.96	0.95
psaB	0.66 ***	0.85 *	cyf	0.66 *	0.82
psbA	0.64 *	0.96	psbD	0.71 *	0.87
psbC	0.60 **	0.95	psbC	0.60	1.03
psbD	0.80	0.82			
atpA	0.97	0.93			
atpB	0.99	0.94			
atpE	0.73	0.84			
atpF	1.35	0.66 *			
atpH	0.85	1.01			
atpI	0.60 *	0.97			
rps2	0.94	0.82			
rps4	1.39	0.90			
rps14	1.02	0.46 **			
rps16	1.35	0.64			
rpoB	0.81 *	0.81 ***			
rbcl	0.85	0.88			
petA	0.83	0.84			
petB	1.10	1.07			

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.0001$

Table 3 (continued)

MZE			WHT		
GENE	CpG	TpA	GENE	CpG	TpA
NUC:					
act1G	0.58 ***	0.59 ***	glgB	0.37 ***	0.31 ***
adh1F	0.69 **	0.40 ***	gliABA	0.42 ***	0.49 ***
ant	0.46 ***	0.62 **	glumrA	0.51 *	1.01
eg2R	0.81	0.50 *	h3	1.00	0.44
h3	1.08	0.21 *	h4	1.19	0.97
h4	1.15	0.76	gir	0.92	0.57 ***
susysG	0.73 ***	0.51 ***	amy	0.98	0.59 **
ze19A	0.34 ***	0.68 *	em	0.96	0.37
ze22A	0.46 ***	0.79			
ze22B	0.57 **	0.75			
zea20M	0.34 ***	0.75			
zea30M	0.34 ***	0.72			
b32	0.88	0.47 **			
gapA	0.70 **	0.46 ***			
pepC	0.92	0.48 ***			
gst	1.09	0.57			
cat	1.11	0.68 *			
c1	1.14	0.41 **			
TF:					
rbcs	0.98	0.91	cab	0.75 *	0.41 ***
gapB	1.17	0.25 ***	rbca	1.22	0.59
cab	1.05	0.48 *			
CP:					
atbB	0.92	0.86	atp	0.45	1.03
atbE	0.70	0.79	atps	1.19	0.70 *
atpB	0.64	0.94	cytf	0.85	0.74 *
rps4	1.24	0.85	cytb	1.29	1.04
rubp	0.93	0.89	rps2	0.96	0.71 *
rps14	1.32	0.87	frxB	0.90	0.79
rps8	0.96	0.91			

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.0001$

composition of silent and replacement sites. Table 2 shows the average base frequencies in silent and replacement sites of nuclear, transferred and chloroplast genes of each species. Total G+C of the analyzed genes is higher in the nucleus than in the chloroplast for all four species. This difference is specially high in the two monocots species. Table 2 also shows that transferred genes have reached similar GC content than nuclear genes by increasing their GC content mainly in silent sites. Note that in our sample of genes, GC content at replacement sites is very similar among dicot and monocot species but both groups clearly differ in their GC content at silent sites: GC content at silent sites is lower than at replacement sites in dicots but higher in monocots.

Dinucleotide distributions

Nuclear plant DNA has a high content of 5-methylcytosine in both the CG dinucleotide and also in the C(A/T)G trinucleotides[12] and CpG methylation is not found in the chloroplast genome[13]. Since chloroplast genes transferred to the nucleus show an increase in GC content (Table 2), we analyzed the distribution of methylation sites in transferred genes to find out if the increase in GC parallels an increase in the methylation sites available.

CTG and CAG trinucleotides were found at expected frequencies in most genomes (data not shown). The gene to gene distributions of the other methylation target, CpG, are shown in Table 3. Chloroplast genes generally show the expected frequencies of this dimer in the four species; on the contrary, most of the nuclear and transferred genes show significant deficiencies. This is specially true for dicots, while in monocots some nuclear genes show the CpG expected frequencies or even an excess of this dinucleotide.

TpA is other dinucleotide of special relevance, since a general avoidance of this dimer has been reported in most genomes[14,15,16,17]. Table 3 shows its distribution for every gene. Nuclear and transferred genes show very variable TpA ratios while chloroplast genes are more uniform. Significant avoidances of TpA were more often found in nuclear and transferred genes of the four species than in chloroplast genes.

Codon usage analysis

Codon usage distances. A triangular matrix containing all pairwise comparisons of codon usage distances among all genes was computed for each species by means of the distance algorithm described in the Data and Methods section; all pairwise distances were then categorised into six groups (nuclear vs nuclear, nuclear vs transferred, etc.) and the average distance for each group computed (Table 4). Codon usage distances between chloroplast and nuclear genes are small in dicots and higher in monocot species. In monocots this is paralleled by higher distances between chloroplast and transferred genes. This method gives a global idea of codon usage in genes from the three different gene sets but does not consider the variation in codon usage within genes from the same group.

Correspondence analysis. To get a better representation of the variability for codon usage within the different groups of genes,

Table 4. Codon usage distances (S.E.) among nuclear (NUC), transferred (TF) and chloroplast (CP) genes. Species abbreviations are as in Table 1.

NUC	9.75 ± 0.28			PEA
TF	13.05 ± 0.31	15.25 ± 0.69		
CP	12.44 ± 0.36	15.77 ± 0.49	11.81 ± 0.90	
	NUC	TF	CP	
NUC	12.61 ± 0.53			TOB
TF	12.77 ± 0.66	11.99 ± 1.45		
CP	13.12 ± 0.24	14.52 ± 0.46	10.46 ± 0.22	
	NUC	TF	CP	
NUC	15.38 ± 0.93			WHT
TF	13.59 ± 0.84	11.22		
CP	23.73 ± 0.76	23.81 ± 0.90	13.94 ± 0.73	
	NUC	TF	CP	
NUC	13.06 ± 0.32			MZE
TF	13.57 ± 0.73	4.65 ± 1.11		
CP	22.18 ± 0.38	29.84 ± 0.51	13.15 ± 0.97	
	NUC	TF	CP	

we performed a factorial correspondence analysis on a data matrix containing the n-1 codon frequencies at each synonymous codon group for each gene (see ref. 18 for methodology involved in correspondence analysis of codon usage). Differences in codon usage cluster chloroplast genes separately from the nuclear ones in monocot (Fig. 1) but not in dicot species (data not shown). In all the species nuclear and transferred genes are always mixed together, being more widely scattered than the chloroplast ones.

DISCUSSION

Nucleotide composition of transferred genes

GC content is lower in the chloroplast genomes than in the nuclear ones[7]; Table 2 shows that at gene level the differences are more extreme in maize and wheat than in the dicot species; this is due to the higher GC content of monocot nuclear genes. Since there is a correlation between genomic GC content and GC level in the three codon positions of genes[1,19,20], it would be interesting to investigate if an adjustment of base composition occurred in transferred genes adapting them to the high GC content of the nucleus. Table 2 shows that chloroplast genes relocated into the nuclear genome have now similar GC content to nuclear genes. Increase in GC has been more pronounced at the silent sites than at the replacement sites. At silent sites, transferred genes show higher GC content than nuclear genes, while at replacement sites GC contents are very similar between nuclear and transferred

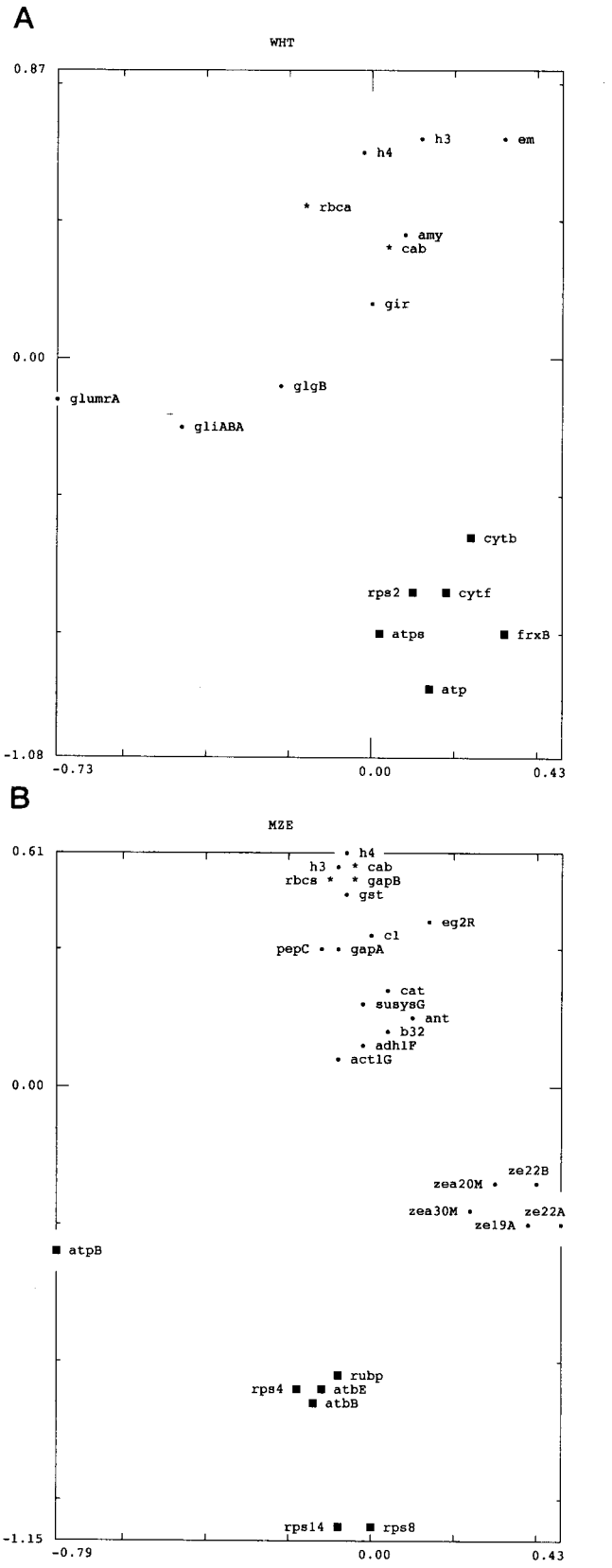

Figure 1. Correspondence analysis of codon use in different genes of wheat and maize. The n−1 codon frequencies in each synonymous group were used for each gene. Gene and species abbreviations are as in Table 1 and the Appendix. The plots of F1 (vertical) × F2 (horizontal) factors explain the 59% of the variability in codon usage of WHT and the 60% in MZE. (● = nuclear gene; ■ = chloroplast gene; * = transferred gene).

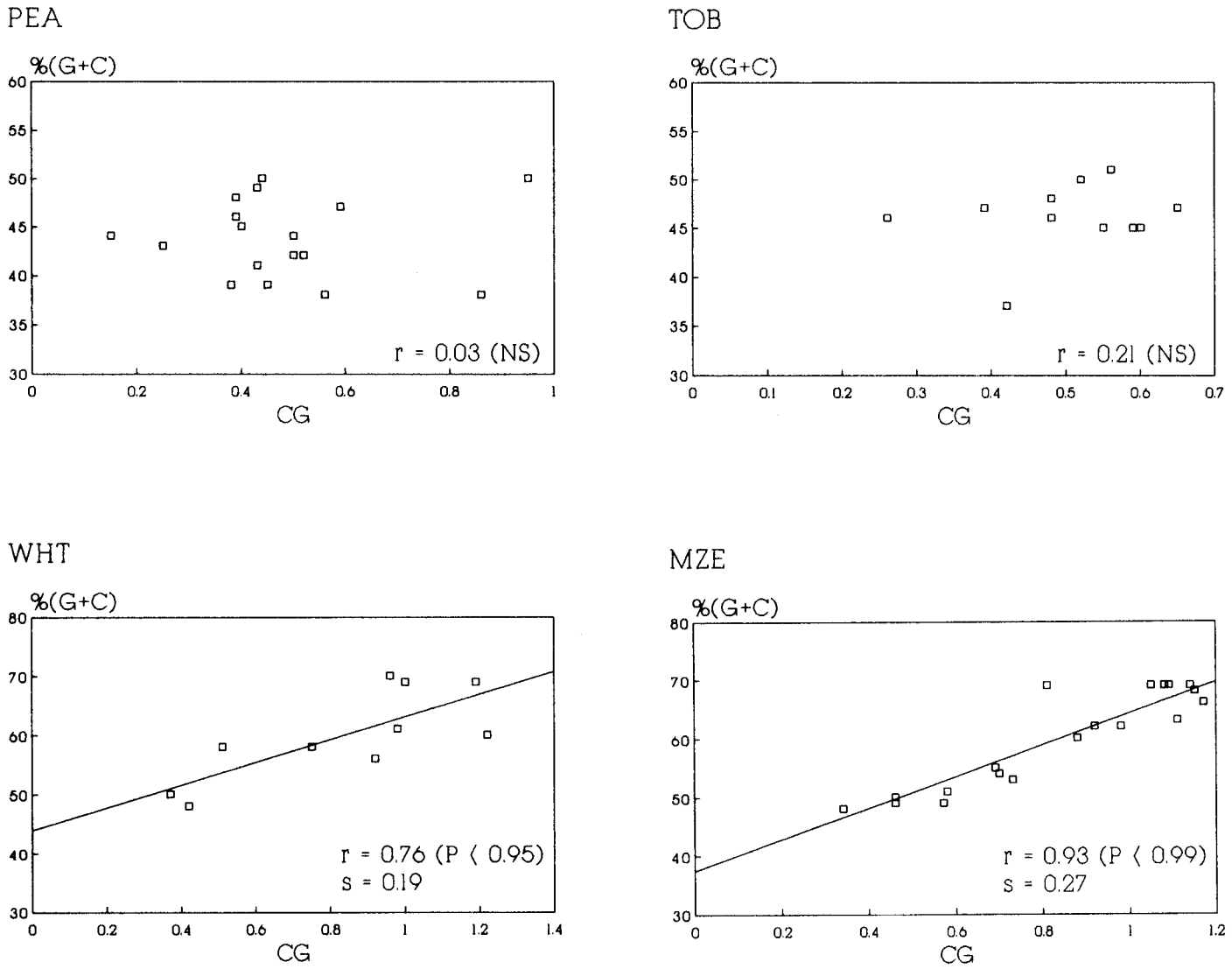

Figure 2. Plots of %(G+C) and CG (Obs/Exp ratio) of nuclear and transferred genes in each species.

genes. Similar adjustments at silent and replacement sites provoked by GC pressure have been previously observed when homologous genes and noncoding sequences are compared in bacterial and mitochondrial genomes with different GC contents (see reference 1 and references therein). The differences in GC content at silent sites between monocots and dicots (Table 2) are probably indicating the existence of a higher GC pressure in the two monocots studied.

The rise in GC content of transferred genes of dicots is due to similar increments in both bases G and C (Table 2); however, in the two monocots the rise in GC content of transferred genes has been due preferentially to increases in C. This is a similar situation to what was found earlier in the genomes of warm-blooded vertebrates[19,21].

Distributions of TpA and CpG dinucleotides

The compositional adjustment of transferred genes to the nuclear environment is also reflected by a stronger avoidance of the doublet TpA. This avoidance is commonly found in most eukaryotic genomes[14,15], including the nuclear genomes of plants[16,17].

The higher CpG avoidance in nuclear than in chloroplast genes is probably due to the existence of CpG methylation in the nuclear genome[12] but not in the chloroplast[13,22]. Transferred genes also follow a similar pattern in CpG avoidance as nuclear genes, thus adjusting their base composition to that of the nuclear genome.

Nuclear genes of dicot species analyzed here always show avoidances of TpA and CpG dinucleotides. However, a different situation was found in the two monocot species analyzed, where some genes do not show any avoidance. We think that these differences between species are reflecting the different compositional organization of their genomes. The genomes of tobacco and pea are far more homogenous in base composition than the genomes of wheat and maize[20,23]. Having this in mind, the lack of CpG shortage in four out of five transferred genes of monocots could be due to the location of these genes in GC rich chromosome regions with a decreased discrimination against CpG doublets. Bernardi et al.[24] have shown that CpG shortage decreases in degree when increasing genomic GC level in both vertebrates and their viruses. This could also be happening in the nucleus of wheat and maize, since in these two species — but not in pea and tobacco — we found that the CpG doublet

level is strongly correlated with overall GC content of different genes (Fig. 2).

Codon usage comparisons

In all four species the codon usage of transferred genes has consistently become undistinguishable from that of the nucleus where they are integrated. Table 4 shows that the codon usage of transferred genes is more distantly related to the chloroplast genome from which they derive than to the nuclear genome in which they are now located. This is particularly clear in the two monocots species. Since the GC contents of the dicot nuclear genes analyzed are more similar to the chloroplast ones (Table 2), codon usage does not differentiate so well genes from one or the other compartment.

Multivariate analyses shown in Fig. 1 allow to visualize the heterogeneity present for codon usage within the nuclear genomes. The dispersion found for nuclear genes contrasts with the homogeneity found for the chloroplast ones. Since constraints on codon usage imposed by the aminoacid composition of proteins can be discarded due to the method we used to compute codon frequencies, this is probably a reflection of the compositional organization of the plant genomes[20]. Transferred genes always appeared mixed with the nuclear genes by this analyses (Fig. 1), which again supports their adjustment to the nuclear environment.

A strong bias in codon usage for the nuclear encoded chloroplast GAPDH of maize has been reported[6]. These authors hypothesize that the strong codon bias found could be a consequence of a selection for higher expressitivity. However, the same bias is not found for the same gene in other species. A similar pattern of stronger biases in codon usage in monocots than in dicot species have also been reported[17] for two transferred genes, *rbcs* (maize) and *cab* (wheat). Our more global results indicate that this bias could be the consequence of the increase in GC content that transferred genes have suffered to reach the level of the new host genome. Since this increase is mainly supported by changes in silent sites (Table 2), it produces a strong bias in the codon usage of these genes. Bias is extremely high in species like maize where differences in GC content between chloroplast and nuclear genome are very high.

As a general conclusion, chloroplast genes transferred to the nucleus seem to have adjusted their base composition, dinucleotide distribution and codon usage according to the characteristics prevailing in their new host genomes, and thus they behave as polite DNA[25,26]. Table 2 reveals the clear trend of transferred genes to achieve similar GC content as the nuclear genomes where they are integrated. Consequently codon usage is affected and since changes in nucleotide sequence affect, almost exclusively, to the silent sites, protein sequences can remain mainly unmodified. Therefore, the GC increase can occur without changing the coding capacity of transferred genes, which at the aminoacid level are still homologous to their prokaryotic counterparts[5]. Because the mosaic organization of the eukaryotic genome[19,20,24,27], a certain level of variability should be expected in base composition and codon usage among genes from the nuclear genome and this is in fact found (Fig. 1). As expected from the isochore organization in these four species, variation is higher in wheat and maize that show a wider compositional heterogeneity[20]. The distributions of TpA and CpG doublets in transferred genes also reflect the conditions of every nuclear genome or genome compartment.

It seems clear from these results that there is an evolution towards compositional homogenization within the different compartments of the nuclear plant genome. If this evolution is based on a selective advantage due to improved expressitivity or to other compositional modifying mechanisms not related with gene expression, such as different mutational bias of DNA polymerases in germline cells[28] or variation in mutation patterns along the replication timing of different chromosomal regions in the germline[29], is not known at this moment. Experiments testing the expressitivity of coding sequences with different base composition and codon usage are required to elucidate the presence of any selective advantage.

ACKNOWLEDGEMENTS

We are most grateful to Drs. M. Ruiz Rejón and J. Salinas by the critical reading of the manuscript. This work was partially supported by the DGICYT (PB87-0881) and the INIA (#7556) of the Spanish Government.

REFERENCES

1. Jukes, T.H. and Bhushan, V. (1986) J. Mol. Evol. 24,39–44.
2. Filipski, J., Salinas, J. and Rodier, F. (1989) J. Mol. Biol. 206,563–566.
3. Palmer, J.D. (1985) Ann. Rev. Genet. 19,325–354.
4. Martin, W. and Cerff, R. (1986) Eur. J. Biochem 159:323–331.
5. Shih, M.-C., Lazar, G. and Goodman H.M. (1986) Cell 47,73–80.
6. Brinkmann, H., Martínez, P., Quigley, F., Martin, W. and Cerff, R. (1987) J. Mol. Evol. 26, 320–328.
7. Boudraa, M. (1987) Génet. Sél. Evol. 19,143–154.
8. Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.L., Rindone, W.P., Swindell, C.D. and Tung, C.S. (1986) Nucl. Acids Res. 14,1–4.
9. Tingey, S.V., Tsai, F.-Y., Edwards, J.W., Walker, E.L. and Coruzzi, G.M. (1988) J. Biol. Chem. 263:9551–9657.
10. Vierling, E., Nagao, R.T., DeRocher, A.E. and Harris, L.M. (1988) EMBO J. 7, 575–581.
11. Grantham, R. (1978) FEBS Letters 95(1),1–11.
12. Gruenbaum, Y., Naveh-Many, T., Cedar, H. and Razin, A. (1981) Nature 292, 860–862.
13. Tewari, K.K. and Wildman, S.G. (1966) Science 153,1269–1271.
14. Grantham, R., Greenland, T., Louail, S., Mouchiroud, D., Prato, J.L., Gouy, M. and Gautier, C. (1985) Bull. Inst. Pasteur 83, 95–148.
15. Ohno, S. (1988) Proc. Natl. Acad. Sci. USA 85:9630–9634.
16. Boudraa, M. and Perrin, P. (1987) Nucl. Acids Res. 15,5729–5743.
17. Murray, E.E., Lotzer, J. and Eberle, M. (1989) Nucl. Acids Res. 17,477–498.
18. Holm, L. (1986) Nucl. Acids Res. 14,3075–3087.
19. Bernardi, G. and Bernardi, G. (1986) J. Mol. Evol. 24,1–11.
20. Salinas, J., Matassi, G., Montero, L.M. and Bernardi, G. (1988) Nucl. Acids Res. 16,4269–4285.
21. Marín, A., Bertranpetit, J., Oliver, J.L. and Medina, J.R. (1989) Nucl. Acids Res. 17,6181–6189.
22. Ngernprasirtsiri, J., Kobayashi, H. and Akazawa, T. (1988) Proc. Natl. Acad. Sci. USA 85:4750–4754.
23. Matassi, G., Montero, L.M., Salinas, J. and Bernardi, G. (1989) Nucl. Acids Res. 17,5273–5290.
24. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science 228,953–958.
25. Zuckerkandl, E. (1986) J. Mol. Evol. 24, 12–27.
26. Holmquist, G.P. (1989) J. Mol. Evol. 28, 469–486.
27. Aota, S. and Ikemura, T. (1986) Nucl. Acids Res. 14,6345–6355.
28. Filipski, J. (1987) FEBs Lett. 217,184–186.
29. Wolfe, K.H., Sharp, P.M. and Li, W.-H. (1989) Nature 337,283–285.

APPENDIX

List of the remaining nuclear and chloroplast genes from pea (PEA), tobacco (TOB), wheat (WHT) and maize (MZE) used in this study. Sequences were retrieved from GenBank (Release 57) or, when no GenBank LOCUS name is specified, directly from the indicated source.

GENE SYMBOL	GenBank LOCUS	PROTEIN
PEA (nucleus):		
abn2	PEAABN2	Albumin 2
legJ	(1)	'Minor' legumin polypeptide
gs1	PEAGSR1	Glutamine synthase 1
gs2	"	" " 2
gs3	"	" " 3
lecA	PEALECA	Seed lectin A
vic	(2)	Vicilin
adh-1	(3)	Adh-1
hsp	(4)	Chloroplast hsp
PEA (chloroplast):		
atpA	PEACPATPG	ATP synthase subunit a (aa 1–247)
cyf	PEACPCYF	Cytochrome f propeptide
psbD	PEACPD2	PSII D2 protein
psbC	PEACPD2	PSII 44kDa reaction center protein
TOB (nucleus):		
gapC	TOBGAPC	Cytosolic GAPDH-C
pr-1a	(5)	Pathogenesis related protein 1a
pr-1b	"	Pathogenesis related protein 1b
pr-1c	TOBPR1CR	Pathogenesis related protein 1c
ech	TOBECH	Endochitinase
pox	TOBPXDLF	Lignin-forming peroxidase
thaur	TOBTHAUR	TMV induced protein homologous to thaumatin
TOB (chloroplast):		
psaA	TOBCPCG	PSI P700 apoprotein A1
psaB	"	" " " A2
psbA	"	PSII 32kD protein
psbC	"	" 44kD protein
psbD	"	" D2 protein
atpA	"	ATPase alpha subunit
atpB	"	" beta subunit
atpE	"	" epsilon subunit
atpF	"	" I subunit
atpH	"	" III subunit
atpI	"	" a subunit
rps2	"	Ribosomal protein S2
rps4	"	" protein S4
rps14	"	" protein S14
rps16	"	" protein S16
rpoB	"	RNA polymerase beta subunit
rbcl	"	RuBisCo large subunit
petA	"	Cytochrome f
petB	"	" b
WHT (nucleus):		
glgB	WHTGLGB	Gamma-gliadin B
gliABA	WHTGLIABA	Alpha-beta-gliadin A-II
glumrA	WHTGLUMRA	High-M-r gluten polypeptide
h3	WHTH3	H3 histone
h4	WHTH4	H4 histone
gir	WHTGIR	Gibberellin responsive wheat gene
amy	WHTAMYA	Alpha-amylase
em	WHTEMR	EM protein
WHT (chloroplast):		
atp	WHTCPATP	ATP synthase proton-translocating subunit
atps	WHTCPATPS	ATP synthase CF-0 subunit I prepeptide
cytf	WHTCPCYF	Cytochrome f
cytb	WHTCPCYTB	Cytochrome b-559 (aa 1–83)
rps2	(6)	Ribosomal protein S2
frxB	(7)	frxB gene
MZE (nucleus):		
act1G	MZEACT1G	Actin 1
adh1F	MZEADH1F	Alcohol dehydrogenase (ADH1-F)

APPENDIX (continued)

GENE SYMBOL	GenBank LOCUS	PROTEIN
ant	MZEANT	ATP/ADP translocator
eg2R	MZEEG2R	Endosperm glutelin-2
h3	MZEH3C2	Histone H3
h4	MZEH4C14	Histone H4
susysG	MZESUSYSG	Sucrose synthase
ze19A	MZEZE19A	Zein 19 kD protein
ze22A	MZEZE22A	" 22 kD protein
ze22B	MZEZE22B	" 26.99 kD protein
zea20M	MZEZEA20M	" (clone a20)
zea30M	MZEZEA30M	" (clone a30)
b32	(8)	b-32 protein
gapA	(9)	Glyceraldehide-3-phosphate dehydrogenase A
pepC	MZEPEPCR	Phosphoenolpyruvate carboxylase
gst	MZEGST3A	Glutation-S-transferase GSTIII
cat	(10)	Catalase
c1	(11)	Regulatory c1 locus
MZE (chloroplast):		
atbB	MZECPATBE	Coupling factor complex, beta subunit
atbE	"	" " " , epsilon subunit
atpB	MZECPATPB	ATPase, beta subunit (aa 1–25)
rps4	MZECPRPS4	Ribosomal protein S4
rubp	MZECPRUBP	RuBisCo, large subunit
rps14	(12)	Ribosomal protein L14
rps8	(12)	Ribosomal protein S8

REFERENCES

1. Gatehouse, J.A., Bown, D., Gilroy, J., Levasseur, M., Castleton, J. and Ellis, T.H.N. (1988) Biochem. J. 250, 15–24.
2. Watson, M., Lambert, N., Delauney, A., Yarwood, J.N., Croy, R.R.D., Gatehouse, J.A., Wright, D.J. and Boulter, D. (1988) Biochem. J. 251, 857–864.
3. Llewellyn, D.J., Finnegan, E.J., Ellis, J.G., Dennis, E.S. and Peacock, W.J. (1987) J. Mol. Biol. 195, 115–123.
4. Vierling, E., Nagao, R.T., DeRocher, A.E. and Harris, L.M. (1988) EMBO J. 7, 575–581.
5. Matsuoka, M., Yamamoto, N., Kano-Murakami, Y., Tanaka, Y., Ozeki, Y., Hirano, H., Kagawa, H., Oshima, M. and Ohashi, Y. (1987) Plant Physiol. 85, 942–946.
6. Höglund, A.S. and Gray, J.C. (1987) Nucl. Acids Res. 15, 10590.
7. Dunn, P.P.J. and Gray, J.C. (1988) Nucl. Acids Res. 16, 348.
8. Di Fonzo, N., Hartings, H., Brembilla, M., Motto, M., Soave, C., Navarro, E., Palau, J., Rhode, W. and Salamini, F. (1988) Mol. Gen. Genet. 212, 481–487.
9. Brinkmann, H., Martínez, P., Quigley, F., Martin, W. and Cerff, R. (1987) J. Mol. Evol. 26, 320–328.
10. Bethards, L.A., Skadsen, R.W. and Scandalios, J.G. (1987) Proc. Natl. Acad. Sci. USA 84, 6830–6834.
11. Paz-Ares, J., Ghosal, D., Wienand, U., Peterson, P.A. and Saedler, H. (1987) EMBO J. 6, 3553–3558.
12. Markmann-Mulisch, U. and Subramanian, R. (1988) Eur. J. Biochem. 170, 507–514.