VI.4 Maize Chloroplast Gene Transfer to Nucleus

J.L. OLIVER', A. Marín2, and J.M. Martínez-Zapater3

1 Introduction

The chloroplast genome of higher plants is a small circular chromosome, whose length generally varies between 120 and 160 kb among different species. Size differences are mainly due to variations in the length of a large inverted duplication (IO-76 kb) while the two single-copy regions are highly conserved in size and gene content. Chloroplast DNA encodes proteins involved in photosynthesis, several of the protein components of the transcription and translation apparatus, four kinds of rRNAs associated with the 70S bacteria-like chloroplast ribosomes, and all the tRNAs involved in chloroplast protein synthesis. There are a total of 60-100 chloroplast encoded proteins, the remaining SO-90% of the plastid proteins being encoded in the nucleus and post-translationally imported into the chloroplast (for reviews see Palmer 1985, 1990; Umesono and Ozeki 1987; Gray 1989; Sugiura 1989; Shimada and Sugiura 1991).

Chloroplasts probably originated roughly lo9 years ago as eubacteria-like endosymbionts, whose closest contemporaries are cyanobacteria. The organelle DNAs are the remnants of a previously complete and autonomous genome of the endosymbiont progenitor (Penny and O'Kelly 1991). According to the endosymbiotic theory, there would have been a transfer of genetic material from the chloroplast to the nuclear genome along the process of endosymbiosis. Since most of the chloroplast genomes analyzed encode the same set of proteins, it has been suggested that most of the gene transfer occurred soon after endosymbiosis (Palmer 1985). This genetic material would have evolved to give rise to nuclear genes whose products are transported into the chloroplast. Several cases of nuclear encoded chloroplast proteins have been analyzed for homology with their prokaryotic counterparts. The results support the predictions of the endosymbiotic theory (Shih et al. 1986; Brinkmann et al. 1987; Baldauf and Palmer 1990). Moreover, small differences in the gene content of different chloroplast genomes indicate that secondary transfer events have taken place in the ancestors of different plant taxa (Baldauf and Palmer 1990; Smooker et al. 1990). The

' Departamento de Genktica e Instituto de Biotecnologia, Facultad de Ciencias, Universidad de Granada, 1807 1 -Granada, Spain
*Departamento de GenCtica y Biotecnia, Facultad de Biologia, Universidad de Sevilla, Aptdo 1095, 41080-Sevilla, Spain
' Departamento de Proteccibn Vegetal, CIT-INIA, Carretera de La Corufia Km 7, 28040-Madrid, Spain

Biotechnology in Agriculture and Forestry, Vol. 25
Maize (ed. by Y.P.S. Bajaj)

movement of DNA sequences from chloroplast to nucleus still continues today (Baldauf et al. 1990).

Subsequent to the transfer, and according to the observation that genes generally show base content similar to the genomes or genomic regions where they are integrated (Bernardi et al. 1985; Filipski et al. 1989), a process of compositional homogenization can be predicted for transferred sequences, that would finally achieve similar compositional features (mainly G+C content) as their surroundings. The main mechanisms underlying this so-called process of compositional adjustment (Jukes and Bhushan 1986) are both mutational bias induced along the DNA replication or repair processes (Filipski 1987; Wolfe et al. 1989) and selective advantage due to improved expressitivity (Brinkmann et al. 1987).

Although the maize chloroplast genome is not yet entirely sequenced (only the entire sequences of tobacco, rice, and liverwort chloroplasts have been determined to date), there are a sufficient number of maize chloroplast DNA sequences available in nucleotide databases to allow a comparative compositional analysis with the nuclear genome. Such study may be particularly interesting due to the fact that the differences in G+C content commonly found between nuclear and chloroplast genomes are specially pronounced in maize. Thus, maize transferred genes can be a good system to follow the compositional changes suffered by the chloroplast sequences moved to the nuclear genomic environment. We have identified five maize nuclear genes that most likely evolved from transferred chloroplast genes, and use them both to prove the existence of compositional adjustment to the nuclear genomic conditions and to identify the main factors determining this adjustment. We compared nucleotide composition, dinucleotide preference, CpG distribution and codon usage of 39 nuclear, 5 transferred, and 33 chloroplast genes. Results indicate that transferred genes in maize have adjusted their base composition, dinucleotide preference, and codon usage to the features of the nuclear genome. This adjustment has mainly been the consequence of nucleotide changes at codon silent sites that do not alter the amino acid composition of the encoded proteins.

2 Transferred Genes

The release 26 (February, 1991) of the EMBL nucleotide sequence database (Hamm and Cameron 1986) contains 57 maize chloroplast entries with a total of 66.4 kb. From this release, we selected 33 protein genes encompassing a total of 5 1.7 kb and also retrieved 44 maize nuclear genes for comparison (Table 1).

Nuclear genes encoding chloroplast proteins can generally be considered as transferred chloroplast genes. However, a few exceptions of nuclear encoded chloroplast proteins that probably evolved from nuclear genes have already been described in pea (Tingey et al. 1988; Vierling et al. 1988). In maize, the chloroplast NADP-dependent malic enzyme (ZMNDMEX entry in Table 1) is more closely related to the eukaryotic cytosolic malic enzyme (49% similar) than to the

Table 1. Nuclear, transferred, and chloroplast maize genes used in this study. Both the percentages of G+C at the entire (Tot), replacement (RS) and silent (SS) codon positions and the ratio of the observed to the expected frequencies of CpG dinucleotides are given. Deviations from expectations were tested by Chi-square. (After Oliver et al. 1990)

EMBL entry	Protein	%G+C			CpG (Obs/Exp)	
		Tot	RS	SS		
Nucleus						
N1 ZMAUX	Auxin-binding protein	51.5	52.5	49.5	0.82	NS
N2 ZMACT1	Actin 1	51.4	49.3	55.7	0.58	***
N3 ZMADH1 FA	Alcohol dehydrogenase 1	54.8	49.9	64.4	0.68	**
N4 ZMADH2NR	Alcohol dehydrogenase 2	60.3	49.5	81.5	0.94	NS
N5 ZMW64CWG	Cell wall glycoprotein	66.8	64.1	12.1	1.17	NS
N6 ZMANTG2	Adenine nucleotide transloc.	48.5	48.0	49.4	0.49	***
N7 ZMALPTUB	α 1 tubulin	56.7	49.9	70.1	0.59	***
N8 ZMB1TUB	β1 tubulin	65.0	49.1	96.2	0.99	NS
N9 ZMALBB32	Albumin b-32 protein	60.4	56.1	68.6	0.88	NS
N10 ZMALD	Aldolase	58.1	51.3	71.3	0.49	***
N11 ZMGLB1SA	Embryo globulin S (7S-like)	68.8	56.8	89.8	1.09	NS
N12 ZMGRP	Glycine-rich prot. ABA-ind.	70.3	66.0	78.6	1.26	NS
N13 ZMPLTP	Phospholipid transfer prot.	72.5	61.2	93.7	0.98	NS
N14 ZMKD18	Oleosin	74.3	64.6	91.6	0.92	NS
N15 ZMMPL3	Major prot. lipid bodies	71.6	59.7	94.2	0.98	NS
N16 ZMMYBAA	cl locus myb homologue cDNA	68.9	62.6	80.6	1.15	NS
N17 ZMNAR	NADH:nitrate reductase	61.9	49.9	86.1	1.09	NS
N18 ZMNDMEX	NADP-dependent malic enzyme	52.9	50.7	56.9	0.67	***
N19 ZMRAB17G	RAB-17 gene	68.6	59.8	86.0	0.92	NS
N20 ZMREGG	Lc regulatory protein	62.2	56.6	75.4	0.98	NS
N21 ZMSOD2A	Superoxide dismutase 2	55.5	58.0	50.3	0.55	**
N22 ZMBZMCC	UDPglucose flav. glyc.-tran.	75.1	65.5	93.3	1.12	NS
N23 ZMCAT1	Catalase 1	51.2	48.0	57.3	0.59	***
N24 ZMCAT2	Catalase 2	64.4	53.6	84.8	1.07	NS
N25 ZMCAT3	Catalase 3	65.8	52.0	90.7	1.14	NS
N26 ZMGPC1	Glyceraldeh-3-phosph. deh.	54.7	48.1	68.1	0.71	*
N27 ZMGST1	Gluth. S-transferase I	57.5	49.8	72.2	0.71	*
N28 ZMGST3	Gluth. S-transferase III	69.8	57.9	91.9	1.09	NS
N29 ZMH3C2	Histone H3 gene	68.6	56.1	92.3	1.08	NS
N30 ZMH4C14	Histone H4 gene	67.6	54.7	91.7	1.15	NS
N31 ZMOPA2	Opaque-2 gene	57.2	55.2	60.9	0.83	NS
N32 ZMPEP	Phosphoenolpyruvate carbox.	62.1	51.0	82.4	0.93	NS
N33 ZMGLUT2E	Endosperm glutelin-2 gene	69.2	64.8	76.8	0.82	NS
N34 ZMPML1	pML1 gene for zein	48.6	52.0	42.1	0.57	*
N35 ZMZC1	14kDa zein	67.2	59.4	81.2	0.82	NS
N36 ZMZE19B1	19kDa zein	47.7	50.5	43.0	0.35	***
N37 ZMZEI10K	10 kDa zein	51.9	50.3	55.6	0.53	**
N38 ZMSUCSI	Sucrose synthase	52.8	45.4	66.6	0.73	***
N39 ZMTPI	Triosephosphate isom. 1	50.9	52.1	41.4	0.35	***
Transferred						
T1 ZMCAB1	Chloroph. a/b-binding prot.	68.3	55.4	95.3	0.97	NS
T 2 ZMG3PD	Cp glyc.-3-phos. dehidrog.	67.2	52.8	95.4	1.16	NS
T3 ZMRBCS	RuBisCo small subunit	64.5	48.7	95.9	1.01	NS
T4 ZMPOD	Pyruv. orthoph. dikinase	58.4	53.6	67.5	0.74	***
T5 CHZMMDH	NADP-malate dehydrogenase	52.4	52.6	52.2	0.90	NS

Table 1. (*Contd.*)

EMBL entry	Protein	%G+C			CpG (Obs/Exp)	
		Tot	RS	SS		
Chloroplast						
C1 CHZM02	RuBisCo large subunit	44.2	51.9	29.6	0.93	NS
C2 CHZMATBE	Coupling factor β subunit	42.4	49.4	30.1	0.92	NS
C3 CHZMATBE	Coupling factor ε subunit	41.1	47.3	30.3	0.70	NS
C4 CHZMATPA	atpA CF (I) α subunit	42.2	49.4	29.4	0.96	NS
C5 CHZMATPH	atpH CF (0) subunit III	44.3	58.7	19.8	0.52	NS
C6 CHZMPETE	Cytochrome b6-f subunit 5	39.5	43.5	33.3	1.20	NS
C7 CHZML23	Ribosomal protein L23	38.7	43.7	29.3	0.96	NS
C8 CHZMR14	Ribosomal protein S14	39.1	45.6	28.9	1.09	NS
C9 CHZMR15	Ribosomal protein S15	33.0	38.0	25.2	1.10	NS
C10 CHZMRP14	Ribosomal protein L16	42.7	52.0	25.0	1.03	NS
Cl 1 CHZMRPl4	Ribosomal protein L14	38.7	48.3	22.1	1.32	NS
Cl2 CHZMRP14	Ribosomal protein S8	36.7	42.9	27.0	0.96	NS
Cl3 CHZMRP19	Ribosomal protein S19	39.4	43.5	31.6	1.38	NS
Cl4 CHZMRP22	Ribosomal protein L22	39.6	39.5	39.9	1.27	NS
C15 CHZMRPL2	Ribosomal protein L2	45.0	52.6	32.1	0.73	NS
Cl 6 CHZMRPS3	Ribosomal protein S3	33.3	38.0	24.9	1.01	NS
Cl7 CHZMSECX	Ribosomal protein secX	39.5	45.7	29.6	1.64	NS
Cl8 CHZMSIIA	Ribosomal protein S1 1	43.1	54.4	23.4	1.07	NS
Cl9 CHZMNDHD	ndhE gene	34.0	37.3	28.3	1.02	NS
C20 CHZMNDHD	psaC gene	43.1	52.2	25.9	0.97	NS
C2 1 CHZMNDHD	ndhD gene	35.1	41.0	25.0	0.73	NS
C22 CHZMNPOO	ndhC gene	39.1	44.2	29.3	0.82	NS
C23 CHZMNPOO	PSII-G gene	38.3	44.3	27.6	0.88	NS
C24 CHZMPSB1	psbB gene	44.2	53.3	21.2	0.86	NS
C25 CHZMPSB2	psbF gene	40.1	46.8	28.4	1.15	NS
C26 CHZMPSB3	petB gene	40.3	41.5	26.2	0.88	NS
C27 CHZMPSB4	petD gene	40.8	50.0	23.9	0.96	NS
C28 CHZMPSI 1	Photosystem I psIA1	43.4	50.0	31.0	0.70	**
C29 CHZMPSI2	Photosystem I psIA2	41.5	48.8	28.1	0.65	***
C30 CHZMRPOA	RNA polymerase a-subunit	36.1	42.0	26.1	0.84	NS
C31 CLZMRPB	RNA polymerase B-subunit	39.7	45.6	29.4	0.91	NS
C32 CHZMATPI	rps2 gene	31.3	44.5	24.3	0.86	NS
C33 CHZMATPI	atpI gene	31.9	44.2	26.3	0.79	NS

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

prokaryotic ones (less than 20% similar). Thus, a nuclear evolutionary origin has been assumed for it (Rothermel and Nelson 1989). Except for this gene, the remaining five nuclear genes encoding chloroplast proteins have been considered as descendants from chloroplast transferred genes (Table 1).

3 Base Composition

It is known that GC content is lower in the chloroplast genome than in the nuclear one; the differences are more extreme in species as maize due to the high GC content of the Gramineae nuclear genes. Since there is a correlation between

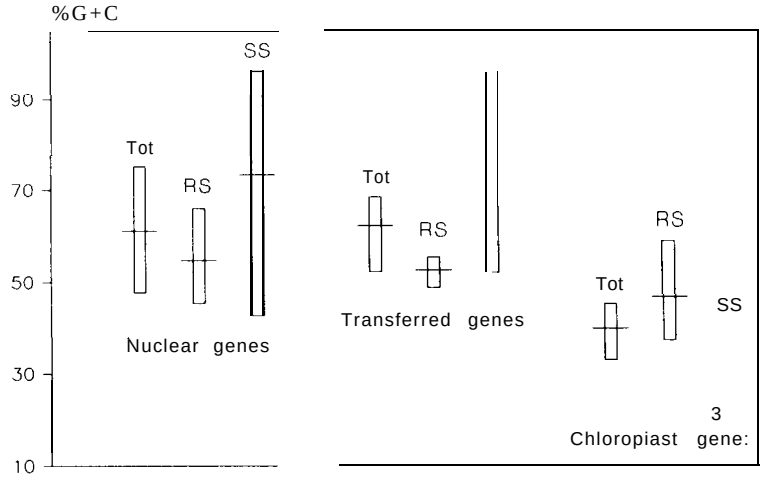

Fig. 1. Range of G+C content values in nuclear, transferred, and chloroplast maize genes. G+C levels at the entire sequence (*Tot*), replacement (*RS*) and silent (*SS*) codon positions are shown. The ***vertical bar*** expands from the minimum to the maximum gene G+C content in the corresponding set, the mean value being indicated by a *horizontal* ***bar***

genomic GC content and GC level in the three codon positions of genes (Bernardi et al. 1985; Jukes and Bhushan 1986; Matassi et al. 1989), it was interesting to investigate if an adjustment of base composition occurred in transferred genes adapting them to the high GC content of the maize nucleus. This chapter is restricted to the coding regions of the genes.

The G+C content values in total, replacement and silent codon positions (see Jukes and Bhushan 1986, for details about these computations) for each of the maize genes analyzed here are shown in Table 1. The ranges of gene G+C content in nuclear, transferred and chloroplast maize gene sets are summarized in Fig. 1. Only slight differences in G+C content were found in the total sequence and at replacement sites. However, at silent sites, G+C content values show strong differences among the different gene sets. Most nuclear genes show higher G+C content at silent than at replacement positions, while in chloroplast genes G+C contents at silent sites are slightly lower than at replacement sites. Thus, maize nuclear genes are under a GC biased mutational pressure, while chloroplast genes seem to be under an AT biased one. All maize transferred genes, as the remaining nuclear ones, are clearly under GC mutational pressure. A similar situation was found in other plant species (Oliver et al. 1990).

Chloroplast genes relocated into the nuclear genome have reached GC content similar to nuclear genes. Increase in GC has been more pronounced at silent than at replacement sites. At silent sites, transferred genes show higher GC content than chloroplast genes, while at replacement positions, GC contents are very similar among the three different gene sets. The rise in GC content of transferred genes has been preferentially due to increases in C at the expenses of T (results not shown), similarly to what was found earlier in the genomes of warm-blooded vertebrates (Bernardi and Bernardi 1986; Marín et al. 1989). Base

composition adjustments at silent and replacement sites provoked by GC pressure have also been observed when homologous genes and noncoding sequences were compared among bacterial and mitochondrial genomes with different GC contents (Jukes and Bhushan 1986, and references therein).

4 CpG Doublets

The CpG dinucleotide is a target for cytosine methylation, which makes the cytosine a mutational hotspot (Bird 1980; Shpaer and Mullins 1990; Sved and Bird 1990). For this reason, this dinucleotide is found at lower frequencies than expected in eukaryotic nuclear genomes. Due to the existence of CpG methylation in the plant nucleus but not in the chloroplast (Gruenbaum et al. 1981), a higher CpG avoidance in nuclear than in chloroplast genes can be expected.

Most of the 33 maize chloroplast genes analyzed here show the expected frequencies for this dimer, the exceptions being psIA 1 and psIA2 genes, encoding proteins of photosystem-I, which show significant avoidances of CpG (Table 1). On the contrary, 15 out of 39 nuclear maize genes show significantly lower frequencies than expected for this doublet. Additionally, only one out of the five maize transferred genes shows CpG shortage. These results are probably related to both the compositional heterogeneity of the maize genome and the respective integration sites of transferred genes (see below).

CpG avoidance in nuclear and transferred genes is homogeneously distributed-at different codon positions (Fig. 2). This figure shows the ratios of the

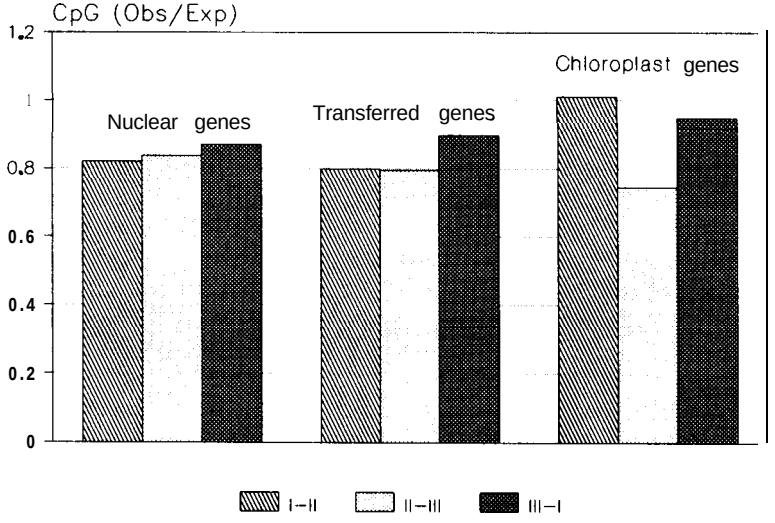

Fig. 2. Ratios of observed versus expected number of CpG dinucleotides in three codon-defined positions of maize genes. Data are presented for 39 pooled nuclear genes, 5 pooled transferred genes, and 33 pooled chloroplast genes. Details in pooling gene sets for these computations can be found in Shpaer and Mullins (1990)

observed versus expected number of CpG dinucleotides, pooled for the different maize gene sets, in different codon-defined positions. The most conspicuous difference among the three gene sets was found at positions I-II; both nuclear and transferred genes show CpG avoidance, while in chloroplast genes no avoidance at all was observed. Due to the presence of CpG methylation in the nuclear genome, genes evolved from chloroplast genes transferred to the nucleus have developed patterns of CpG avoidance similar to the nuclear ones.

5 Codon Usage and Dinucleotide Preference

Patterns of codon-choice have been shown to be genome-specific in noncompartmentalized (Grantham et al. 1980, 1981) and isochore-specific in compartmentalized (Bernardi et al. 1985; Salinas et al. 1988) genomes. Patterns of dinucleotide preference also differ among distinct organisms (Subak-Sharpe et al. 1966; Russell et al. 1976; Nussinov 1981). We have studied the differentiation that the three maize gene sets analyzed here show for these important compositional features.

As an example to illustrate this differentiation, let us consider the usage of leucine codons and the frequencies of TpA doublets. The number of leucine residues encoded by the chloroplastic, transferred, and nuclear sets of genes are 883, 195 and 1152, respectively. Most of the leucines in chloroplastic dictated proteins are encoded for by the duet of codons beginning with UU (56.3%), this proportion is lowered to 7.8% in the set of transferred genes, which is very close to the proportion found in nuclear genes (10.8%). The same effort in the accommodation of transferred genes to the nuclear environment is revealed when considering the frequencies of TpA doublets in codon positions II–III. Deficiency in this dimer seems to be a common rule for most of eukaryotic sequences, and thus the low frequencies found in nuclear (1.3%) and transferred (1.6%) set of genes in comparison with chloroplastic ones, where this frequency reaches 9.7%.

The global differentiation that the three maize gene sets show for codon usage and dinucleotide preference has been also analyzed. Gramineae genomic regions show a strong heterogeneity in G+C content (Salinas et al. 1988), and thus it is hard to predict what G+C content we would expect in maize transferred genes if the existence of plant nuclear constraints on transferred sequences is to be supposed. A technique which simultaneously relates base composition of all chloroplast and nuclear genes as a whole is needed. The method of choice is correspondence analysis, a multivariate technique widely used to analyze global codon usage differentiation among and within gene sets (Grantham et al. 1980; Shields et al. 1988; Oliver et al. 1990). We have also used this multivariate data-reduction method to study gene differentiation in dinucleotide preference. Figure 3 shows the results obtained with the codon-usage frequencies of nuclear, transferred and chloroplast genes, and Fig. 4 those obtained when dinucleotide frequencies of the same gene sets were used. The vertical axis (factor 1) of Figs. 3 and 4 roughly corresponds to G+C content at silent codon positions, which agrees with the results found in the analysis of other gene sets (Grantham et al.

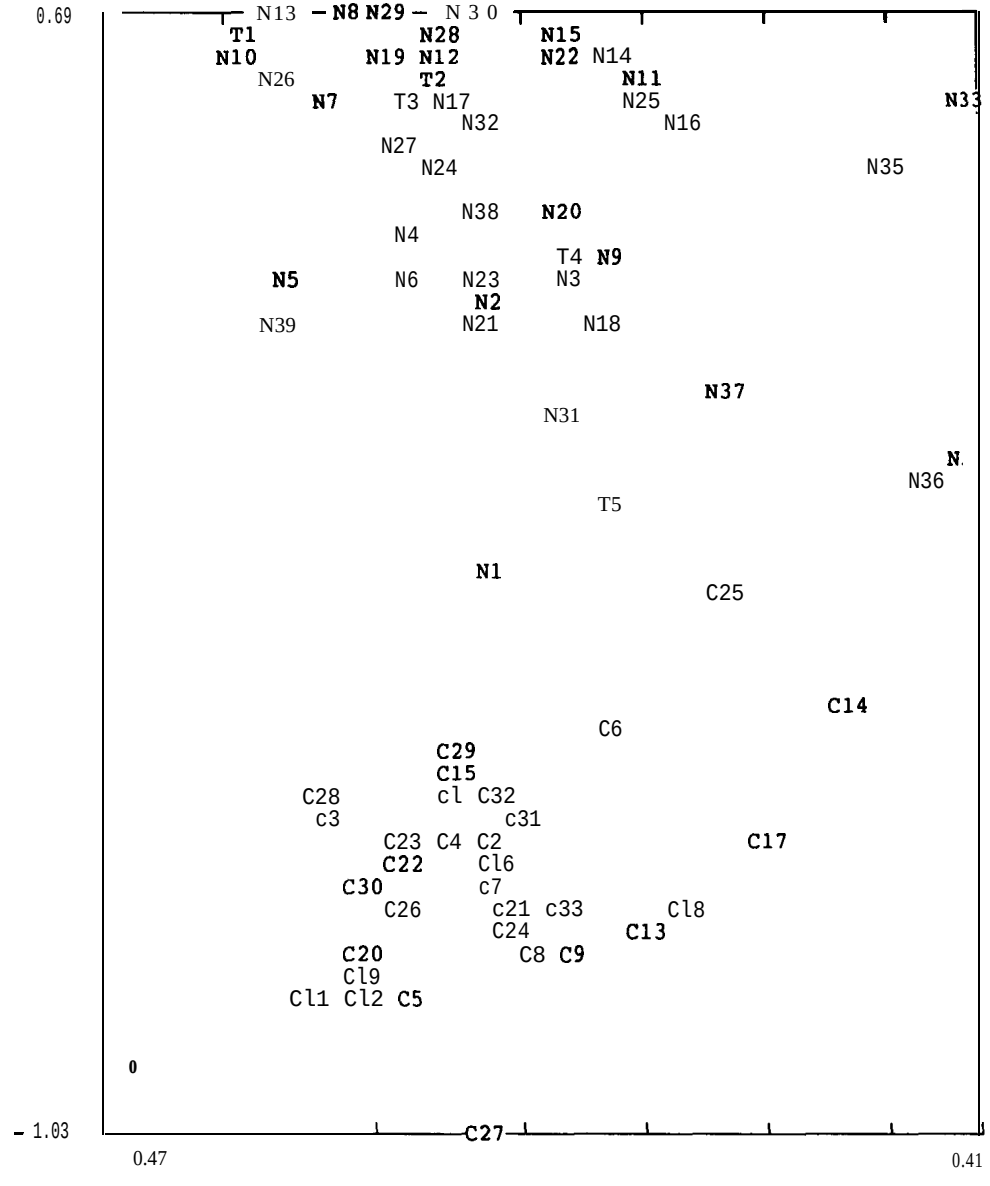

Fig. 3. Correspondence analysis on codon-usage frequencies of nuclear (*N*-), transferred (*T*-), and chloroplast (C-) maize genes. Relative codon-usage frequencies at 21 codon groups were computed as described elsewhere (Oliver et al. 1990). Only the n-l (independent) frequencies of each codon group were used. Factor l (vertical axis) explains the 51% of the variability in codon usage exhibited by these genes and factor 2 (horizontal axis) the 7%. Gene symbols are given in Table l. (Oliver et al. 1990)

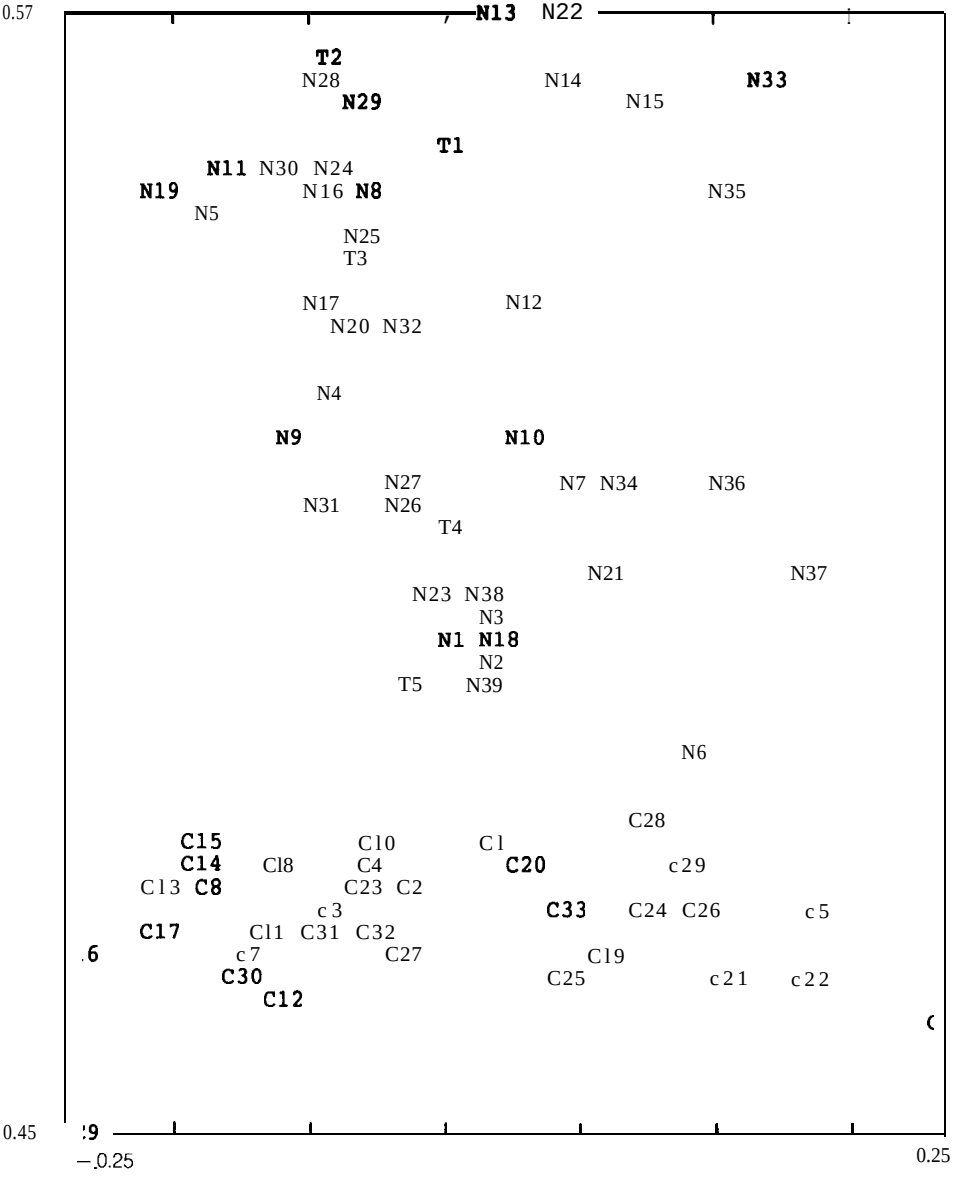

Fig. 4. Correspondence analysis on the dinucleotide-preference frequencies of nuclear (*N*-), transferred (*T*-), and chloroplast (*C*-) maize genes Dinucleotide frequencies were computed as follows. The overlapping dinucleotide appearances in each gene were counted; the 16 dinucleotides were then grouped in 4 dinucleotide groups, according to the nucleotide in the first dinucleotide position; we then computed the relative frequency of each dinucleotide within its dinucleotide group (the count of that dinucleotide divided by the total of dinucleotides in the pertinent group). Only the n-l (independent) frequencies of each dinucleotide group were used for correspondence analysis. Factor 1 (*vertical axis*) explains the 64.5% of the variability in dinucleotide preference exhibited by these genes and factor 2 (*horizontal axis)* the 10%. Gene symbols are given in Table 1

1980). The uppermost genes of both correspondence graphs often have G and C at silent sites, while the AT-richest genes are located in the lower part of both figures. For example, CHZMPSB4 (C27) has only a 23.9% of G+C at silent positions, while ZMB1TUB (NE) shows a 96.2%. In both, the codon-usage (Fig. 3) and the dinucleotide-preference (Fig. 4) correspondence graphs, transferred genes are grouped with nuclear ones, whereas chloroplast genes stand apart, indicating that global codon usage and dinucleotide preference are much more similar between the nuclear sets than among the nuclear and the chloroplast ones.

Therefore, both the codon usage and the dinucleotide preference of transferred genes have consistently become undistinguishable from those of the nucleus where they are integrated. Figures 3 and 4 show that the codon usage of transferred genes is more distantly related to the chloroplast genome from which they derive than to the nuclear genome within which they are now located. In addition, multivariate analyses allow to visualize the compositional heterogeneity within the maize nuclear genome. The dispersion found for nuclear genes contrasts with the homogeneity shown by the chloroplast ones. Since constraints on compositional features imposed by the amino acid composition of proteins can be discarded due to the method we used to compute both codon (Oliver et al. 1990) and dinucleotide (see the legend to Fig. 4) frequencies, this is probably a reflection of the compositional heterogeneity of the maize genome (Salinas et al. 1988). Transferred genes appeared always mixed with the nuclear genes by these analyses, which proves their adjustment to the nuclear environment.

6 Mutational Bias or Expression Optimization?

The homogenization of base composition found in maize transferred genes with respect to the nuclear genomic conditions could be explained either by mutational bias or as the result of selection for optimization of gene expression. Mutational bias produced by differences in the performance of replication and repair mechanisms in distinct genomic regions have been proposed by Filipski (1987) and Wolfe et al. (1989) to explain compositional differences along the genomes. Alternatively, a relationship between codon usage bias and gene expressitivity have been found in noncompartmentalized genomes, as in the case of *E.coli* (Ikemura 198 1; Gouy and Gautier 1982) and yeast (Bennetzen and Hall 1982; Ikemura 1982; Sharp and Li 1986).

The existence of selection for optimization of gene expression has also been proposed in maize to explain the strong codon bias of the nuclear gene encoding the chloroplast glyceraldehyde-3-phosphate dehydrogenase (Brinkmann et al. 1987). However, the analysis of a higher number of dicot and monocot transferred genes lead us to hypothesize that this bias could be the consequence of the increase in GC content that all transferred genes suffer to reach the level of the new host genome (Oliver et al. 1990). Since this increase is mainly supported by changes at silent sites (Table 1 and Fig. 1), it results in a strong bias in the codon usage of those genes. This bias can become extremely high in species like maize where, as mentioned above, differences in GC content between chloroplast and

nuclear genomes are very high. Because transferred genes are sharing with the nuclear genes the same replication, repair, and expression environment, they cannot be properly used to discriminate between mutational or selective compositional pressures (i.e., it would not be possible to decide whether the codon-usage changes suffered by transferred genes are due to mutational bias or expression optimization). Only the analysis of genes in which replication/repair and expression functions are physically separated, can allow to assess the relative roles that mutational bias and expression optimization play in the homogenization of DNA base composition (Martinez-Zapater, Marín and Oliver, 1993).

7 Preference for Integration Sites

It has been suggested that interorganellar transfer of genetic information may occur via RNA and subsequent local reverse transcription and integration (Schuster and Brennicke 1987). A general mechanism for gene movement in the plant genome has been proposed by Pichersky (1990) and the target integration sequence has been determined at least for a transferred gene (Pichersky and Tanksley 1988). However, to our knowledge, no systematic studies have been directed to locate the chromosomal regions or genome compartments in which integration occurs. The patterns of CpG shortage in a monocot species like maize can provide insights about the location of transferred genes at particular genomic regions with a given G+C content.

In almost all the nuclear dicot genes analyzed in a previous work (Oliveret al. 1990), CpG shortage was observed. However, a different situation was found in Gramineae species, where some nuclear genes do not show CpG avoidance. The higher number of nuclear genes analyzed here confirms this observation: Table 1 shows that many maize nuclear genes do not show CpG shortage. The differences between dicot and Gramineae species are probably reflecting the different compositional organization of their genomes. The genomes of dicots are far more homogeneous in base composition than the genomes of Gramineae (Salinas et al. 1988; Matassi et al. 1989). Bernardi et al. (1985) have shown that CpG shortage decreases when increasing genomic GC level in both vertebrates and their viruses. This could also be happening in the nucleus of maize, since in this species – but not in dicots (Oliver et al. 1990) – we found that the CpG doublet level is strongly correlated with overall GC content ofdifferent nuclear genes (Fig. 5). Having this in mind, the lack of CpG shortage in four out of five maize transferred genes (Table 1) could be due to the location of these genes in GC rich genomic regions with a decreased discrimination against CpG doublets. Similar results have also been found for the chloroplast transferred genes from wheat (Oliver et al. 1990), and thus this behavior could be characteristic of the Gramineae transferred genes. This analysis provides information about the actual location of these genes in specific plant genomic regions but does not allow to discriminate if there were preferential sites of insertion in the ancestral plant genome or if insertion took place randomly with later selection of the genes that become expressed.

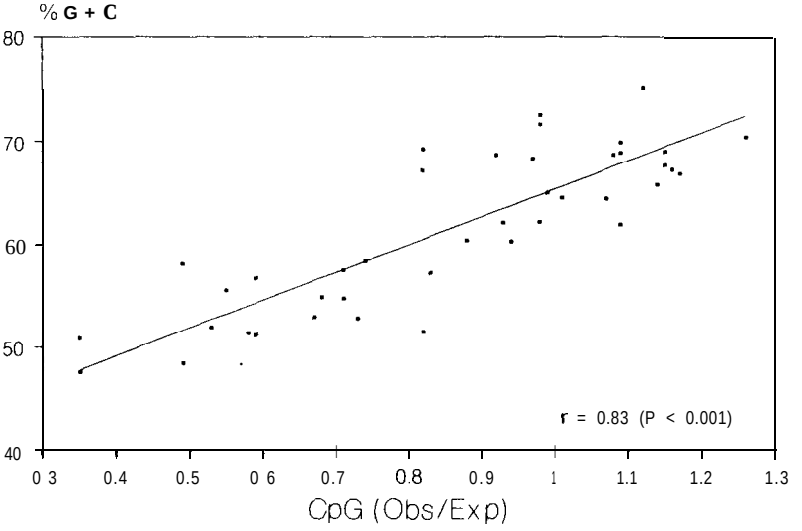

Fig. 5. Plot of CpG (Obs/Exp) versus %G+C in nuclear genes from maize. (Oliver et al. 1990)

8 Conclusions

The use of maize to analyze the evolution of chloroplast genes transferred to the nuclear genome indicates that transferred genes have adjusted their base composition, dinucleotide distribution, and codon usage according to the characteristics prevailing in the nuclear host genome. Due to codon usage redundancy, the GC increase can occur without changing the coding capacity of transferred genes, which at the amino acid level are still homologous to their prokaryotic counterparts (Shih et al. 1986; Brinkmann et al. 1987). Because of the mosaic organization of the maize genome (Salinas et al. 1988), a certain level of variability should be expected in base composition, dinucleotide preference, and codon usage among maize nuclear genes and this was in fact found (Figs. 3 and 4). The distribution of CpG doublets in maize transferred genes also reflect the conditions of every nuclear genomic compartment.

These results support the idea that there is an evolution towards compositional homogenization within the different compartments of the nuclear plant genome. The relative roles played by selective advantage due to improved expressitivity, or by othercompositional modifying mechanisms not related with gene expression, are not known at this moment. Experiments testing the expressitivity of coding sequences with different base composition and codon usage, or the analysis of appropriate genetic systems in which replication/repair and gene expression are physically separated will be required to elucidate this subject.

Compositional differences might be a barrier to gene transfer experiments between phylogenetically remote species, resulting in low expression levels of the transferred genes. Perlak et al. (1991) have recently shown that certain

compositional modifications of the coding sequence enhance the expression level of the engineered gene in manipulated plants. Since chloroplast genes, naturally transferred to the nuclear genome, have successfully crossed the prokaryote-eukaryote boundary, their analysis can help to identify some of the factors responsible for correct expression of foreign engineered sequences in different genomic environments. Thus, the compositional analysis reported here on maize transferred genes may be useful in designing the strategies to genetically manipulate the maize genome.

References

Baldauf SL, Palmer JD (1990) Evolutionary transfer of the chloroplast *tufA* gene to the nucleus. Nature 344: 262-265

Baldauf SL, Manhart JR, Palmer JD (1990) Different fates of the chloroplast *tufA* gene following its transfer to the nucleus in green algae. Proc Natl Acad Sci USA 87: 5317-5321

Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257: 30263031

Bemardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Bemardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228: 9533958

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucl Acids Res 8: 1499–1504

Brinkmann H, Martinez P, Quigley F, Martin W, Cerff R (1987) Endosymbiotic origin and codon bias of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. J Mol Evol 26: 320–328

Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett 217: 184186

Filipski J, Salinas J, Rodier F (1989) Chromosome localization-dependent compositional bias of point mutations in *Alu* repetitive sequences. J Mol Biol 206: 5633566

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressitivity. Nucl Acids Res IO: 705557074

Grantham R (1978) Viral, prokaryote and eukaryote genes contrasted by mRNA sequence indexes. FEBS Lett 95: l-l 1

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucl Acids Res 8: r49–r62

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressitivity. Nuccl Acids Res 9: r34–74

Gray MW (1989) The evolutionary origins of organelles. Trends Genet 5: 294-299

Gruenbaum Y, Naveh-Many T, Cedar H, Razin A (1981) Sequence specificity of methylation in higher plant DNA. Nature 292: 860–862

Hamm GH, Cameron GN (1986) The EMBL data library. Nucl Acids Res 14: 5–9

Hanai R, Wada A (1990) Doublet preference and gene evolution. J Mol Evol 30: 109-l 15

Ikemura T (1981) Correlation between the abundance of *E.coli* transfer RNAs and the occurrence of the respecitve codons in its protein genes. J Mol Biol 146: 1–21

Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J Mol Biol 158: 573-597

Jukes TH, Bhushan V (1986) Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J Mol Evol 24: 3944

Marín, A, Bertranpetit J, Oliver JL, Medina JR (1989) Variation in G+C content and codon choice: differences among synonymous codon groups in vertebrate genes. Nucl Acids Res 17: 61816189

Martinez-Zapater JM, Marin A, Oliver JL (1993) Evolution of base composition in T-DNA genes from *Agrobacterium*. Mol Biol Evol 10 (2): 437448

Matassi G. Montero LM, Salinas J, Bemardi G (1989) The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. Nucl Acids Res 17: 527335290

Nussinov R (1981) Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J Mol Biol 149: 125–131

Oliver JL, Marín A, Martinez-Zapater JM (1990) Chloroplast genes transferred to the nuclear plant genome have adjusted to nuclear base composition and codon usage. Nucl Acids Res 18: 65573

Palmer JD (1985) Comparative organization of chloroplast genomes. Annu Rev Genet 19: 3255354

Palmer JD (1990) Contrasting modes and tempos ofgenome evolution in land plant organelles. Trends Genet 6: 115–120

Palmer JD, Baldauf SL, Calie PJ, dePanphilis CW (1990) Chloroplast gene instability and transfer to the nucleus. In: Clegg MT, O'Brien SJ (eds) Molecular Evolution. Alan R Liss, New York, pp 97–106

Penny D, O'Kelly CJ (1991) Seeds of a universal tree. Nature 350: 106107

Perlak FJ, Fuchs RL, Dean DA, McPherson SL, Fischhoff DA (1991) Modification of the coding sequence enhances plant expression of insect control protein genes. Proc Natl Acad Sci USA 88: 332443328

Pichersky E (1990) Nomad DNA a model for movement and duplication of DNA sequences in plant genomes. Plant Mol Biol 15: 437448

Pichersky E, Tanksley SD (1988) Chloroplast DNA sequences integrated into an intron of a tomato nuclear gene. Mol Gen Genet 215: 65568

Rothermel BA, Nelson T (1989) Primary structure of the maize NADP-dependent malic enzyme. J Biol Chem 264: 19587719592

Russell GJ, Walker PMB, Elton RA, Subak-Sharpe JH (1976) Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol 108: 1–23

Salinas J, Matassi G, Montero LM, Bernardi G (1988) Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucl Acids Res 16: 42694285

Schuster W, Brennicke A (1987) Plastid, nuclear and reverse transcriptase sequences in the mitochondrial genome of *Oenothera*: is genetic information transferred between organelles via RNA? EMBO J 6: 285772863

Sharp PM, Li W-H (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. Nucl Acids Res 14: 77347749

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5: 7044716

Shih M-C, Lazar G, Goodman HM (1986) Evidence in favor of the symbiotic origin of chloroplasts: primary structure and evolution of tobacco glyceraldehyde-3-phosphate dehydrogenases. Cell 47: 73380

Shimada H, Sugiura M (1991) Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucl Acids Res 19: 983–995

Shpaer EG, Mullins JI (1990) Selection against CpG dinucleotides in lentiviral genes: a possible role of methylation in regulation of viral expression. Nucl Acids Res 18: 5793 5797

Smooker PM, Kruft V, Subramanian AR (1990) A ribosomal protein is encoded in the chloroplast DNA in a lower plant but in the nucleus in angiosperms. J Biol Chem 256: 16699–16703

Subak-Sharpe H, Burk RR, Crawford LV, Morrison JM, Hay J, Keir HM (1966) An approach to evolutionary relationship of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequence. Cold Spring Harbor Symp Quant Bio13 1:737–748

Sugiura M (1989) The chloroplast chromosomes in land plants. Annu Rev Cell Biol 1989 5: 51–70

Sved J, Bird A (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci USA 87: 46924696

Tingey SV, Tsai F-Y, Edwards JW, Walker EL, Coruzzi GM (1988) Chloroplast and cytosolic glutamine synthetase are encoded by homologous nuclear genes which are differentially expressed in vivo. J Biol Chem 263: 9551–9657

Umesono K, Ozeki H (1987) Chloroplast gene organization in plants. Trends Genet 3: 281–287

Vierling E, Nagao RT, DeRocher AE, Harris LM (1988) A heat shock protein localized to chloroplasts is a member of a eukaryotic superfamily of heat shock proteins. EMBO J 7: 575–581

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337: 283–285

Biotechnology in Agriculture and Forestry 25

Maize

Edited by Y.P.S. Bajaj

With 189 Figures

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Professor Dr. Y .P.S. BAJAJ
A-137
New Friends Colony
New Delhi 110065, India

ISBN 3-540-56392-X Springer-Verlag Berlin Heidelberg New York
ISBN O-387-56392-X Springer-Verlag New York Berlin Heidelberg

Library of Congress Cataloging-in-Publication Data. Maize / edited by Y.P.S. Bajaj. p. cm. (Biotechnology in agriculture and forestry ; 25) Includes bibliographical references and index. ISBN 3-540-56392-X (Berlin : acid-free paper). – ISBN O-387-56392-X (New York : acid-free paper) 1. Corn – Micropropagation. 2. Corn- Biotechnology. I. Bajaj, Y.P.S., 1936- II. Series. SB191.M2M3247 1994 631.5'23 – dc20 94-14372

Typesetting: Thomson Press (India) Ltd., New Delhi

SPIN: 10064307 31/3130/SPS – 5 4 3 2 1 0 – Printed on acid-free paper