J Mol Evol (1987) 26:320–328

Journal of Molecular Evolution

Endosymbiotic Origin and Codon Bias of the Nuclear Gene for Chloroplast Glyceraldehyde-3-Phosphate Dehydrogenase from Maize

Henner Brinkmann,[1] Pascal Martinez,[1] Françoise Quigley,[1] William Martin,[2] and Rüdiger Cerff[1]

[1] Laboratoire de Biologie Moléculaire Végétale, CNRS UA 1178, Université de Grenoble I, B.P. 68, F-38402 Saint Martin D'Hères Cedex, France
[2] Max-Planck-Institut für Züchtungsforschung, D-5000 Köln 30, FRG

Summary. The nuclei of plant cells harbor genes for two types of glyceraldehyde-3-phosphate dehydrogenases (GAPDH) displaying a sequence divergence corresponding to the prokaryote/eukaryote separation. This strongly supports the endosymbiotic theory of chloroplast evolution and in particular the gene transfer hypothesis suggesting that the gene for the chloroplast enzyme, initially located in the genome of the endosymbiotic chloroplast progenitor, was transferred during the course of evolution into the nuclear genome of the endosymbiotic host. Codon usage in the gene for chloroplast GAPDH of maize is radically different from that employed by present-day chloroplasts and from that of the cytosolic (glycolytic) enzyme from the same cell. This reveals the presence of subcellular selective pressures which appear to be involved in the optimization of gene expression in the economically important graminaceous monocots.

Key words: cDNAs — GAPDH evolutionary tree — Horizontal gene transfer — Coding strategies — Monocotyledons — Dicotyledons

Introduction

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was the first enzyme for which sequence data of the Calvin cycle/glycolysis homologues (Weeden 1981) became available. The recent analysis of partial sequences from mustard (Martin and Cerff 1986) and subsequently from tobacco (Shih et al. 1986) showed that the photosynthetic enzyme of higher plants, i.e., its catalytic subunit (Cerff 1982a, Cerff et al. 1986), is more similar to the GAPDH of thermophilic bacteria than it is to the glycolytic enzyme encoded within the same nucleus, thereby providing initial evidence in support of the gene transfer hypothesis for the origin of this nucleus-encoded chloroplast enzyme. Here we report the first comparison of complete sequences for both mature polypeptides of chloroplast and cytosolic GAPDH from a single plant species, *Zea mays*. The two enzymes share only 45% of their amino acid residues and show strikingly different codon choice patterns, with chloroplast GAPDH using 97% G or C in the third base position.

Offprint requests to: R. Cerff

Materials and Methods

The maize cDNA library used in the present study had previously been constructed at the Max-Planck-Institut für Züchtungsforschung at Cologne in the laboratory of H. Saedler by cloning the cDNA into the EcoRI site of the lambda vector NM 1149 (Schwarz-Sommer et al. 1987). The library was screened by plaque hybridization with the heterologous probes from mustard pS198c (cytosolic GAPDH) and pS84b (chloroplast GAPDH, subunit A) (Martin and Cerff 1986). The positive maize clones pZm9 (cytosolic GAPDH) and pZm57 (chloroplast GAPDH) were subcloned into the EcoRI site of pUC9 and submitted to DNA sequence analysis. For clone pZm57 the chemical degradation method of Maxam and Gilbert (1977) was used. For clone pZm9 Sanger's enzymatic technique (Sanger et al. 1977) was employed in combination with the use of the single-stranded DNA bacteriophage M13 (Vieira and Messing 1982).

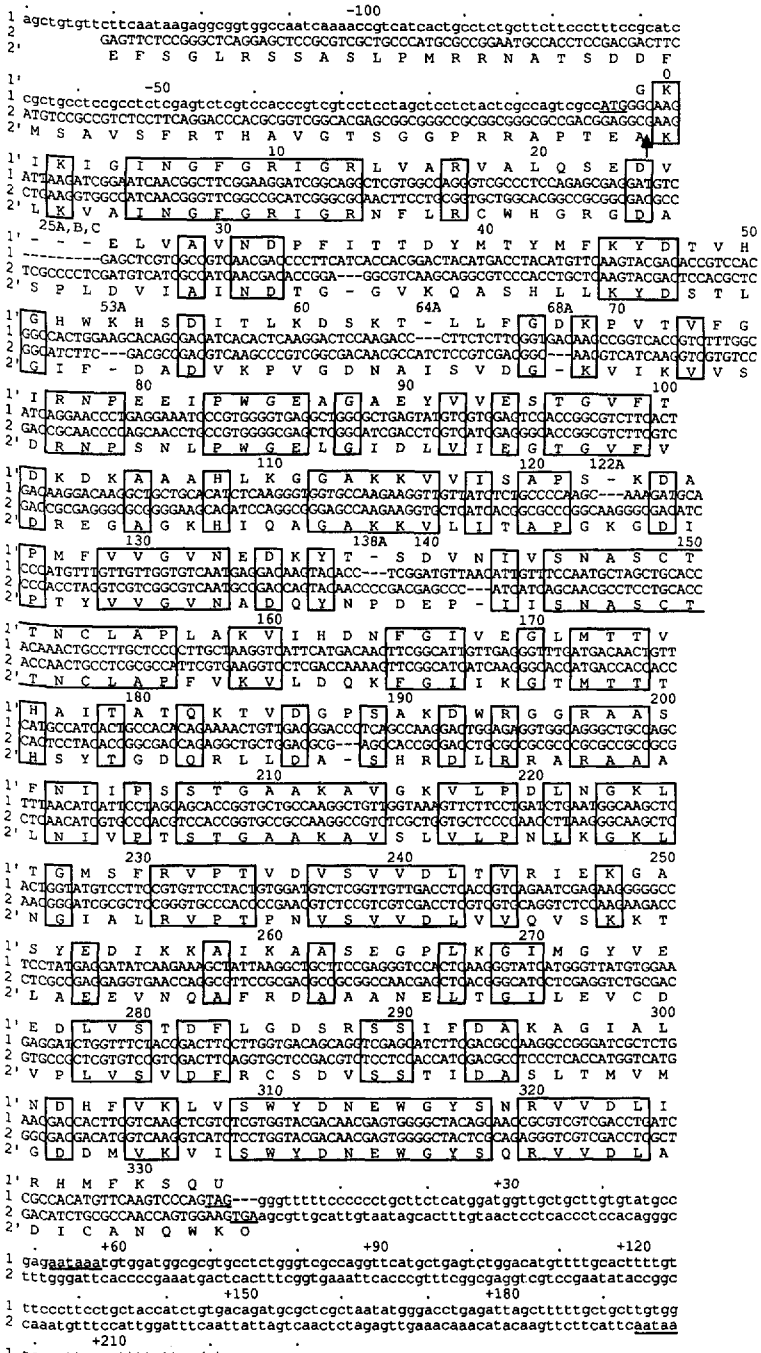

Fig. 1. Nucleotide and deduced amino acid sequences of cDNAs for cytosolic (clone pZm9, lines 1, 1′) and chloroplast GAPDH (clone pZm57, lines 2, 2′) from maize. Amino acid residues of the mature subunits are numbered according to Harris and Waters (1976) and regions of homologies between the two enzymes are boxed. The initiation and termination codons and the presumptive polyadenylation signals aataaa and aataat in the trailers are underlined. The arrow indicates the presumptive processing site of the chloroplast GAPDH precursor corresponding to the amino terminus determined for the mustard enzyme (Cerff and Witt 1983). To obtain the best alignment of the two corresponding coding regions and the most probable positions of the insertions and deletions within the NADP-binding domain, the chloroplast sequence was fitted into the general GAPDH matrix previously established (Martin and Cerff 1986) by maximizing sequence identities and conservative replacements within non-identical stretches.

Results and Discussion

Evidence for a Horizontal Transfer of GAPDH Genes

In Fig. 1 the complementary DNAs encoding the complete sequences of the catalytic subunits of cytosolic (clone pZm9, lines 1 and 1′) and chloroplast GAPDH (clone pZm57, lines 2 and 2′) from maize are aligned. The coding region of clone pZm9 comprises 338 codons, including start and stop codons, and is followed by a trailer of 217 nucleotides plus a polyA tail. The trailer contains a conventional polyadenylation signal aataaa which, however, is located 165 nucleotides upstream of the polyadenylation site (underlined in Fig. 1).

Clone pZm57 (chloroplast GAPDH, lines 2 and 2′ in Fig. 1) starts within the sequence coding for the transit peptide 46 codons upstream of the presumptive processing site. It extends 230 nucelotides

A Identities Matrix of Plant GAPDH Sequences

Cytosol	1	2	3	4
1. Maize				
2. Pea	83			
3. Tobacco	83	88		
4. Mustard	85	88	89	
5. E.coli	65	64	63	64

Chloroplast	1	2	3	4
1. Maize				
2. Pea	90			
3. Tobacco GapA	91	89		
4. Tobacco GapB	81	79	79	
5. Bacillus	58	57	56	57

B GAPDH Evolutionary Tree

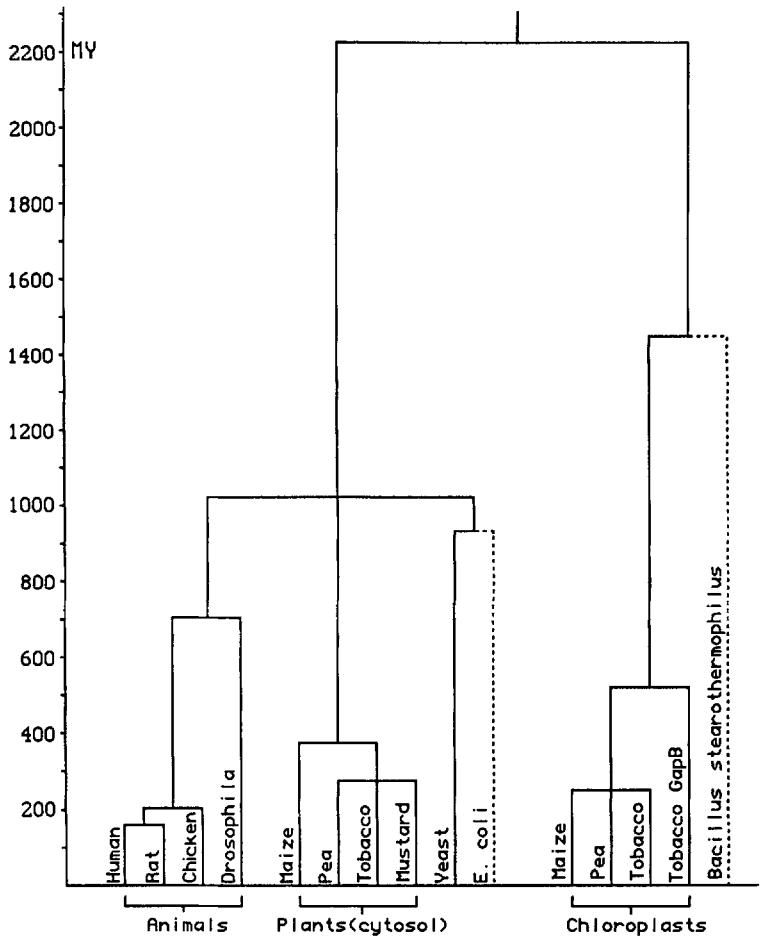

Fig. 2. The evolution of GAPDH enzymes. The GAPDH evolutionary tree (B) is based on the sequence differences calculated from the previously published identities matrix (Martin and Cerff 1986) and that shown in Fig. 2A. These differences were transformed into PAM units (accepted point mutations per 100 residues) by correcting for superimposed mutations according to Dayhoff (1972). The approximate divergence times were calculated by dividing the half PAM values separating two particular sequences with the unit evolutionary rate determined for the animal GAPDH enzymes (2.2 PAM per 100 million years, Dayhoff 1978). For sources of sequence information in Fig. 2A see Table 2 and text.

3′ to the termination codon, has a polyA tail, and 35 nucleotides upstream of the latter the putative polyadenylation signal aataat. The aminoterminus of the purified chloroplast GAPDH from mustard starts with methionine at position 0 (Cerff and Witt 1983). This corresponds to lysine in the maize enzyme leaving 337 amino acids for the polypeptide chain of the mature subunit.

Maize chloroplast GAPDH shows 58% sequence similarity to the enzyme from *Bacillus stearothermophilus* (alignment not shown) but shares only 45% of its amino acids (54% of its nucleotides) with its glycolytic counterpart (amino acids boxed in Fig. 1). This indicates a very distant relationship between the two maize enzymes and suggests that they may have diverged as early as the pro- and eukaryotic lineages. The sequences of chloroplast and cytosolic GAPDHs from maize, pea (Brinkmann and Cerff, publication in preparation), tobacco (Shih et al. 1986), and mustard (cytosolic GAPDH only, Martin and Cerff 1986) together with some sequences from animals, bacteria, and yeast have been used to construct the GAPDH evolutionary tree shown in Fig. 2B (for details see legend of Fig. 2B). This tree clearly demonstrates that chloroplast and cytosolic GAPDH diverged long before the plant–animal division pre-

Table 1. Codon usage of cytosolic (a) and chloroplast (b) GAPDH from maize

	a	b		a	b		a	b		a	b
Phe-TTT	3	0	Ser-TCT	2	0	Tyr-TAT	3	0	Cys-TGT	0	0
Phe-TTC	11	11	Ser-TCC	7	20	Tyr-TAC	6	6	Cys-TGC	2	6
Leu-TTA	0	0	Ser-TCA	1	0	ter-TAA	—	—	ter-TGA	—	1
Leu-TTG	1	0	Ser-TCG	4	4	ter-TAG	1	—	Trp-TGG	5	5
Leu-CTT	5	1	Pro-CCT	4	0	His-CAT	3	0	Arg-CGT	1	0
Leu-CTC	10	20	Pro-CCC	4	11	His-CAC	5	6	Arg-CGC	2	11
Leu-CTA	0	0	Pro-CCA	2	1	Gln-CAA	0	1	Arg-CGA	0	0
Leu-CTG	4	9	Pro-CCG	2	5	Gln-CAG	3	8	Arg-CGG	0	5
Ile-ATT	6	0	Thr-ACT	6	0	Asn-AAT	3	2	Ser-AGT	0	0
Ile-ATC	16	22	Thr-ACC	11	17	Asn-AAC	10	17	Ser-AGC	10	6
Ile-ATA	0	0	Thr-ACA	4	0	Lys-AAA	4	0	Arg-AGA	2	0
Met-ATG	8	7	Thr-ACG	1	6	Lys-AAG	24	21	Arg-AGG	6	5
Val-GTT	15	0	Ala-GCT	14	1	Asp-GAT	7	1	Gly-GGT	14	1
Val-GTC	17	30	Ala-GCC	13	21	Asp-GAC	19	31	Gly-GGC	11	24
Val-GTA	0	0	Ala-GCA	2	0	Glu-GAA	2	0	Gly-GGA	3	2
Val-GTG	4	11	Ala-GCG	0	14	Glu-GAG	13	11	Gly-GGG	2	6

cluding the possibility that chloroplast GAPDH originated by a gene duplication event in the early plant eukaryote. Thus, either animals have lost the gene or plants have acquired it horizontally from the progenitors of chloroplasts after their divergence from animals. The similarity of the chloroplast enzyme to that of *B. stearothermophilus* clearly supports the latter alternative in agreement with the endosymbiotic theory of chloroplast evolution.

The evolutionary tree in Fig. 2B reveals three other major points of interest. First, the enzyme of *Escherichia coli* (Branlant and Branlant 1985) belongs to the eukaryotic part of the tree, which has previously been interpreted in terms of a possible gene transfer event in the opposite, eukaryote to prokaryote, direction (for details see Martin and Cerff 1986). Second, in dicotyledonous (dicot) plants cytosolic and chloroplast GAPDHs seem to evolve at similar rates as shown by the sequence similarities of 88 and 89% for the comparisons tobacco cytosol/pea cytosol and tobacco chloroplast/pea chloroplast, respectively (see values boxed by dashed lines in Fig. 2A). With respect to cytosolic GAPDH it is surprising that 11–17% sequence divergence between maize, pea, tobacco, and mustard (see boxed values in the left half of Fig. 2A) does not lead to significant charge differences at the native enzyme level, as shown by our previous zymogram analysis of cytosolic and chloroplast GAPDHs from higher plants (Cerff 1982b). Third, in maize, as opposed to dicot species, the evolutionary rates of chloroplast and cytosolic GAPDH differ considerably, the chloroplast enzyme being more conserved (90/91% positional identity with pea/tobacco) than its cytosolic counterpart (83/83/85% positional identity with pea/tobacco/mustard, see values boxed by continuous lines in Fig. 2A). This probably reflects a slowdown of chloroplast GAPDH evolution in maize rather than a corresponding acceleration of the evolutionary rate of cytosolic GAPDH. This interpretation is based on our observation (Table 1) that in maize the chloroplast enzyme is subjected to extremely selective constraints at the level of codon utilization.

Codon Bias and Gene Expression in Cereals

It is shown in Table 1 that the codon choice pattern of maize chloroplast GAPDH is highly skewed. The gene uses only 38 amino acid triplets as compared to 51 of the gene for the cytosolic enzyme. This difference is absent in the three other GAPDH pairs from pea (Brinkmann and Cerff, unpublished), mustard (Martin and Cerff 1986), and tobacco (Shih et al. 1986), which use between 50 and 59 different codons (not shown). The codon bias of maize chloroplast GAPDH is expressed as a 97% preference for G (28%) or C (69%) at the third base position. In contrast, cytosolic GAPDH shows 67% third position G+C (20% G and 47% C).

To obtain a more general impression of angiosperm coding strategies, the third position G+C percentage was analyzed for 27 nuclear genes or mRNAs from cereals and dicot plants as well as for some chloroplast genes from two plant species (Table 2). The lowest G+C percentage in the triplet third base position (around 30%) is found for chloroplast genes of both cereals and dicot species (maize and tobacco, see Table 2). With respect to nuclear genes, however, the two taxa exhibit clearly different coding strategies. While in cereals G+C values range between roughly 40 and 100% in each species (see Table 2), values in dicot plants are rather homogeneous, the largest span being found in soybean with 39 and 61% G+C for leghemoglobin and the

Table 2. G+C percentage at the third positions of codons in genes from higher plants. Values in parentheses refer to the G+C percentage within noncoding regions of individual genes (introns and flanking sequences). Coding sequences that are not full length with respect to the mature protein are marked with an asterisk.

Plant	% G+C	Genes or mRNAs	Expression induced by/in/during	EMBL codes and references
Maize	30	3 chloroplast genes		CHZMATBE, CHZMO2; 1, 2
	43 (37)	Zein	Seed development	ZMZEO1; 3
	49 (42)	Triosephosphate isomerase		4
	52	Adenine nucleotide translocator		ZMANT1; 5
	57 (38)	Actin		ZMACT1; 6
	64 (41)	Alcohol dehydrogenase		ZMADH1S; 7
	67	Cytosolic GAPDH		This paper
	67 (44)	Sucrose synthase	Seed development	ZMSUCS2; 8
	78	Glutelin 2	Seed development	9
	84	Chloroplast PEP carboxylase	Light/phytochrome	ZMPEPCR; 10
	84 (45)	A1-locus enzyme	Seed development	11
	93 (53)	Waxy-locus enzyme	Seed development	12
	94	Bronze-locus enzyme	Seed development	13
	95 (52)	Histones H3–H4	Cell cycle/S-phase	14
	97 (45)	Small subunit RuBCase	Light/phytochrome	15
	97	Chloroplast GAPDH	Light/phytochrome	This paper
Wheat	40 (35)	Gliadin	Seed development	TAGLIAA; 16
	41 (41)	HMW-glutenin	Seed development	TAGLU1DG; 17
	54*	HMW gluten	Seed development	TAHGO1; 18
	87*	Small subunit RuBPCase	Light/phytochrome	TARUB2; 19
	97 (52)	H3–H4 histones	Cell cycle/S-phase	TAHIO1, TAHIO2; 20, 21
Barley	38 (41)	B-hordein	Seed development	HVB1HORG; 22
	59*	Cytosolic GAPDH		23
	89	α-amylase	Giberellic acid	HVAMYA; 24
	91	UDP-glucose SGT	Seed development	25
	100*	H3-histone	Cell cycle/S-phase	26
Pea	43	Chloroplast GAPDH	Light/phytochrome	27
	44 (33)	Small subunit RuBPCase	Light/phytochrome	PSRCO1; 28
	46 (29)	Light harvesting protein	Light/phytochrome	PSCAB80; 29
	46	Cytosolic GAPDH		27
Mustard	41*	Chloroplast GAPDH	Light/phytochrome	30
	55	Cytosolic GAPDH		30
Tobacco	31	6 chloroplast genes		31
	38	Chloroplast GAPDH (Gap B)	Light/phytochrome	32
	44*	Cytosolic GAPDH (Gap C)		32
	50	Chloroplast GAPDH (Gap A)	Light/phytochrome	32
	52 (32)	Small subunit RuBPCase	Light/phytochrome	NTRUBSS; 33
Soybean	39 (24)	Leghemoglobin	Root nodules	GMGLO3; 34
	47	7S seed storage protein	Seed development	GM7SAA; 35
	49 (30)	Heat shock protein (hs 6871)	Heat	GMHSP2; 36
	61 (33)	Small subunit RuBPCase	Light/phytochrome	GMRUBP; 37

Abbreviations: PEP, phosphoenolpyruvate; RuBPCase, ribulose bisphosphate carboxylase; waxy-locus enzyme = UDP-glucose SGT, UDP-glucose starch glycosyl transferase; bronze-locus enzyme, UDP-glucose flavonol-3-O-glycosyl transferase. This analysis has been performed by screening the EMBL Nucleotide Sequence Data Library and the sequences of some recent papers with the DNA sequencing program developed by Greaves and Ware (University of Bristol, unpublished) and adapted to the MULTICS computer of the University of Grenoble. Sources of sequence information: 1, Krebbers et al. 1982; 2, McIntosh et al. 1980; 3, Pedersen et al. 1982; 4, Marchionni and Gilbert 1986; 5, Baker and Leaver 1985; 6, Shah et al. 1983; 7, Dennis et al. 1984; 8, Werr et al. 1985; 9, Prat et al. 1985; 10, Izui et al. 1986; 11, Schwarz-Sommer et al. 1987; 12, Klösgen et al. 1986; 13, D. Furtek, personal communication; 14, Chaubet et al. 1986; 15, Lebrun et al. 1987; 16, Anderson et al. 1984; 17, Thompson et al. 1985; 18, Forde et al. 1983; 19, Smith et al. 1983; 20, Tabata et al. 1983; 21, Tabata et al. 1984; 22, Forde et al. 1985; 23, Chojecki 1986a; 24, Rogers and Milliman 1983; 25, Rhode et al. 1986; 26, Chojecki 1986b; 27, Brinkmann and Cerff, unpublished; 28, Coruzzi et al. 1984; 29, Cashmore 1984; 30, Martin and Cerff 1986; 31, Maruyama et al. 1986; 32, Shih et al. 1986; 33, Mazur and Chui 1985; 34, Hyldig-Nielsen et al. 1982; 35, Schuler et al. 1982; 36, Schoeffl et al. 1984; 37, Berry-Lowe et al. 1982

small subunit of ribulose bisphosphate carboxylase, respectively. These findings complement and extend a recent correspondence analysis by Boudraa (1987; for methodological details see Grantham 1980 and references therein) of the codon choice patterns of chloroplast (17 sequences) and nuclear (20 sequences) mRNA sequences from higher plants. This analysis showed a segregation of the two mRNA classes along the horizontal axis which, according to the present data, is due to their overall difference in G+C content at the degenerate codon position. However, while the chloroplast sequences, because of their low and homogeneous G+C content (see Table 2 and Maruyama et al. 1986), form a rather compact cluster, the nuclear mRNAs are horizontally scattered (see Boudraa 1987) and, in view of the present findings, may be separated into two horizontal subclasses: one including all dicot sequences and the "relaxed" monocotyledonous (monocot) sequences maize zein, wheat gliadin, and maize actin (40, 43, and 57% G+C, see Table 2), and the other comprising the highly biased monocot sequences alpha-amylase and histone H3 from barley (89 and 100% G+C) and, interestingly, the sequence prethaumatin (90% G+C, not shown in Table 2) from *Thaumatococcus danielli* (Edens et al. 1982), a member of the monocot plant family Marantaceae (Zingiberales, Liliidae). The high third position G+C content of 90% for prethaumatin, the only noncereal monocot sequence presently available (EMBL release 9, 1986), presumably indicates that the coding strategies of cereals investigated here are a characteristic feature of monocot angiosperms in general.

Large differences in G+C content at the degenerate position of codons have previously been reported for genes of birds and mammals (for review see Ikemura 1985). The original work of Bernardi et al. (1985) and the subsequent investigations by Bernardi and Bernardi (1986) and Aota and Ikemura (1986) suggests that in warm-blooded vertebrates these differences are a consequence of the overall mosaic structure of the genome. According to these findings genes of birds and mammals are embedded in "isochores" (Bernardi et al. 1985), long DNA segments (>300 kilobases) characterized by different G+C levels and fairly homogeneous base compositions. This means that the G+C content at the third base position of a particular mammal gene reflects the G+C level of the surrounding isochore, which comprises introns, flanking sequences, and "spacer DNA." For the plant genes shown in Table 2 the total G+C percentage of introns and flanking regions, where available, are given in parentheses, yet show little if any correlation to third base position values. This observation indicates that the differences in codon choice patterns in monocot genes are maintained independently of the G+C content of the surrounding portions of the genome, in contrast to the situation found in birds and mammals. Since the natural rate of silent substitutions will continuously strive towards randomization of nucleotide frequencies at the third base position (Kimura 1982), an extreme bias for particular nucleotides in certain genes clearly represents the results of selection acting at a level below that of amino acid substitution.

For the cereal genes in Table 2, a rough correlation between increasing codon bias and gene expressivity is observed. The selective mechanism responsible for codon bias, therefore, seems to relate third position G+C preference with the tendency of a gene to show strong expression in response to endogeneous (cell cycle, developmental, hormonal) and exogenous stimuli (such as morphogenetic light mediated by the photoreceptor phytochrome; e.g., chloroplast GAPDH, see Cerff and Kloppstech 1982). Since no such correlation is found for dicot plants, the selective pressure yielding codon bias has either evolved in (graminaceous) monocots or has been lost in dicots since the divergence of these angiosperm classes. Also the cereal storage protein genes (zein, hordein, gliadin) do not appear to conform to the trend just described. However, since storage proteins are encoded by multigene families, their strong expression during seed development may depend on high endogenous levels of multiple mRNAs rather than on the high translational activity of individual mRNA species (see below).

Since it is clear that a broad spectrum of codon bias does exist for cereal genes, one would like to know what selective process is involved. One possibility is that RNA processing efficiency or RNA secondary structure and stability are positively influenced by a high G+C content at the third base position. Although RNA secondary structure and stability can surely not be excluded as points where selection can become effective (see Bernardi and Bernardi 1986), "processability" seems rather improbable as a major pressure since the maize gene for UDP-glucose starch glycosyl transferase (the waxy-locus enzyme) with 13 introns and a polyadenylated mRNA (Klösgen et al. 1986) possesses about the same third position G+C as the maize genes for histones H3 and H4 (93 vs 95%), genes that boast neither polyadenylation nor introns (Tabata et al. 1983, 1984; Chaubet et al. 1986).

A tempting explanation for the occurrence of strongly preferred codons in highly expressed monocot genes is the presence of distinct classes of major isoaccepting tRNAs with anticodons best adapted to only a subset of possible synonymous codons. The existence of such a coding strategy has been demonstrated for *E. coli* (Gouy and Gautier 1982; Grosjean and Fiers 1982; Ikemura 1985) and yeast

(Bennetzen and Hall 1982; Ikemura and Ozeki 1982), where the more heavily a gene is expressed the more extreme its codon bias is, and the codons preferred correspond to the most abundant tRNA species in each isoaccepting subset. Whether such is the case for (graminaceous) monocots remains to be shown. Data for nuclear tRNAs in plants are currently scarce (for a compilation see Sprinzl et al. 1985). The prediction is that when such data become available for cereals, the anticodons of the major species will correspond to those triplets having G or C at the third position. It is interesting in this connection that the highly expressed genes in cereals do not avoid "strong" codons (G/C G/C G/C), in contrast to those in bacteria (for review see Grosjean and Fiers 1982) and yeast (Bennetzen and Hall 1982). In fact, they clearly select for them whenever possible, as shown by the codon families for proline (CCX), alanine (GCX), arginine (CGX), and glycine (GGX) in Table 1 (chloroplast GAPDH, columns b). The gene for maize chloroplast GAPDH is representative in this respect for the other cereal genes with a high G+C content in the third base position (see Table 2). This extends the previous observations by Wells et al. (1986) for the highly expressed histone H3 and actin genes of higher animals, which also show this tendency, although to a lesser extent. Hence, the present data and those of Wells et al. (1986) seem difficult to reconcile with the hypothesis of Grosjean and Fiers (1982) that codons with intermediate binding energies ("intermediate codons"), preferentially used by highly expressed genes of prokaryotes and yeast, are more efficiently translated than "strong" and "weak" codons. The studies of Sharp and Li (1986) and of Holm (1986) suggest that the codon bias of a particular gene does not actively modulate but rather passively reflects its level of expression. This seems reasonable since changes in gene expression may evolve much easier by modifying a single component of the expression system, such as promotor strength (Sharp and Li 1986). The preference of codons recognized by the most abundant tRNAs in *E. coli* genes encoding abundant proteins is explained by Sharp and Li (1986) on the basis of selection against codons that might impair the efficiency of (rapid) translation, and by Holm (1986) in terms of constraints on translation accuracy and on the cost of proofreading at the level of tRNA acylation. If these explanations are correct and if they also apply to the codon bias of genes from monocot angiosperms, the present contribution raises the intriguing question of how dicot plants could have escaped these selection pressures. The striking absence of a marked codon bias in dicot genes is highlighted by a recent publication (Chaboute et al. 1987) showing that in the dicot plant *Arabidopsis thaliana* even the H3 and H4 histone genes, usually G/C-rich in all higher eukaryotes (Wells et al. 1986 and Table 2), use only 54% G+C in the third base position.

Splicing of Transcripts from Highly Biased Cereal Genes in Transgenic Tobacco

Where the manipulative transfer of genes across the monocot–dicot boundary is concerned, the different coding strategies within these two classes of angiosperms could be of considerable importance. Keith and Chua (1986), for example, have recently shown that the gene encoding the small subunit of ribulose bisphosphate carboxylase from wheat is inefficiently spliced in transgenic tobacco plants and the spliced and unspliced products are polyadenylated at multiple novel sites in the wheat 3′ flanking region. In contrast, the introns of the corresponding pea gene are excised efficiently in tobacco suggesting some incompatibility between monocot and dicot splicing mechanisms. It is possible that unknown differences between the intron sequences of monocot and dicot plants play a discriminative role. However, the G+C content of the exons may be equally important. As shown in Table 2, the G+C content at the third base position is 88% in wheat and only 44 and 55% in pea and tobacco small subunit genes, respectively. It seems possible, therefore, that this high G+C percentage may favor secondary structures of wheat primary transcripts that can neither be efficiently spliced nor correctly polyadenylated in tobacco. It should be possible to test this hypothesis by introducing into tobacco a hybrid small subunit gene containing the intron from wheat and the exons from pea and vice versa. In addition, since the splicing apparatus from cereals can cope with large differences in the G+C content of exons, one might expect that primary transcripts from dicot genes can be efficiently spliced in cereals, a testable hypothesis in light of the currently available cereal transformation systems (Fromm et al. 1985, 1986; Lörz et al. 1985; Potrykus et al. 1985; Cocking and Davey 1987).

The amino acid sequence data presented here and elsewhere (Martin and Cerff 1986; Shih et al. 1986) suggest strongly that early in the evolution of plants chloroplasts still possessed the gene for the photosynthetic GAPDH enzyme, but during the course of geologic time relinquished it to the nucleus by an intracellular gene transfer process. With respect to the notably low G+C content at the third base position for present-day chloroplast genes (see Table 2 and Maruyama et al. 1986; see also Morris and Herrmann 1984), the extreme to which the maize chloroplast GAPDH gene has evolved at the third base position in response to selective pressure within the nucleus is striking. By observing the patterns of nature that have evolved in the optimization of

gene expression subsequent to the transfer of genes between subcellular compartments, we should be able to draw conclusions that allow us to optimize the expression of genes transferred between species.

Acknowledgments. We thank H. Saedler, Zs. Schwarz-Sommer, and colleagues for the maize cDNA library. This work was funded by grants from the CNRS (UA 1178), the Ministère de la Recherche et Technology, and the Ministère de l'Education Nationale. H. Brinkmann was supported by a research grant from the Deutsche Forschungsgemeinschaft, FRG.

References

Anderson OD, Litts JC, Gautier MF, Greene FC (1984) Nucleic acid sequence and chromosome assignment of a wheat storage protein gene. Nucleic Acids Res 12:8129–8144

Aota S, Ikemura T (1986) Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res 14:6345–6355

Baker A, Leaver CJ (1985) Isolation and sequence analysis of a cDNA encoding the ATP/ADP translocator of *Zea mays* L. Nucleic Acids Res 13:5857–5867

Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

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

Bernardi 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 228:953–958

Berry-Lowe SL, McKnight TD, Shah DM, Meagher RB (1982) The nucleotide sequence, expression, and evolution of one member of a multigene family encoding the small subunit of ribulose-1,5-bisphosphate carboxylase in soybean. J Mol Appl Genet 1:483–498

Boudraa M (1987) Coding strategy variation in the plant system. Génét Sél Evol 19:143–154

Branlant G, Branlant CH (1985) Nucleotide sequence of the *Escherichia coli* gap gene. Different evolutionary behaviour of the NAD^+-binding domain and of the catalytic domain of D-glyceraldehyde-3-phosphate dehydrogenase. Eur J Biochem 150:61–66

Cashmore AR (1984) Structure and expression of a pea nuclear gene encoding a chlorophyll a/b-binding polypeptide. Proc Natl Acad Sci USA 81:2960–2964

Cerff R (1982a) Separation and purification of NAD- and NADP-linked glyceraldehyde-3-phosphate dehydrogenases from higher plants. In: Edelmann M, Hallick RB, Chua N-H (eds) Methods in chloroplast molecular biology. Elsevier Biomedical Press, Amsterdam, pp 683–694

Cerff R (1982b) Evolutionary divergence of chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases from angiosperms. Eur J Biochem 126:513–515

Cerff R, Kloppstech K (1982) Structural diversity and differential light control of mRNAs coding for angiosperm glyceraldehyde-3-phosphate dehydrogenases. Proc Natl Acad Sci USA 79:7624–7628

Cerff R, Witt J (1983) Evolution of chloroplast glyceraldehyde-3-phosphate dehydrogenase. Amino-terminal amino acid sequence and separation of subunit specific mRNAs. In: Schenk HEA, Schwemmler W (eds) Endocytobiology, vol II. Walter de Gruyter, Berlin, pp 187–193

Cerff R, Hundrieser J, Friedrich R (1986) Subunit B of chloroplast glyceraldehyde-3-phosphate dehydrogenase is related to beta-tubulin. Mol Gen Genet 204:44–51

Chaboute ME, Chaubet N, Philipps G, Ehling M, Gigot C (1987) Genomic organization and nucleotide sequences of two histone H3 and two histone H4 genes of *Arabidopsis thaliana*. Plant Mol Biol 8:179–191

Chaubet N, Philipps G, Chaboute ME, Ehling M, Gigot C (1986) Nucleotide sequences of two corn histone H3 genes. Genomic organization of the corn histone H3 and H4 genes. Plant Mol Biol 6:253–263

Chojeki J (1986a) Identification and characterization of a cDNA clone for cytosolic glyceraldehyde-3-phosphate dehydrogenase in barley. Carlsberg Res Commun 51:203–210

Chojeki J (1986b) Identification and characterization of a cDNA clone for histone H3 in barley. Carlsberg Res Commun 51: 211–217

Cocking EC, Davey MR (1987) Gene transfer in cereals. Science 236:1259–1262

Coruzzi G, Broglie R, Edwards C, Chua NH (1984) Tissue-specific and light-regulated expression of a pea nuclear gene encoding the small subunit of ribulose-1,5-bisphosphate carboxylase. EMBO J 3:1671–1679

Dayhoff MO (1972) Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington DC

Dayhoff MO (1978) Atlas of protein sequence and structure, vol 5, suppl 3, p 3. National Biomedical Research Foundation, Washington DC

Dennis ES, Gerlach WL, Pryor AJ, Bennetzen JL, Inglis A, Llewellyn D, Sachs MM, Ferl RJ, Peacock WJ (1984) Molecular analysis of the alcohol dehydrogenase (Adh 1) gene of maize. Nucleic Acids Res 12:3983–4000

Edens L, Heslinga L, Klok R, Ledeboer AM, Maat J, Toonen MY, Visser C, Verrips CT (1982) Cloning of cDNA encoding the sweet-tasting plant protein thaumatin and its expression in *Escherichia coli*. Gene 18:1–12

Forde J, Forde BG, Fry RP, Kreis M, Shewry PR, Miflin BJ (1983) Identification of barley and wheat cDNA clones related to the high-Mr polypeptides of wheat gluten. FEBS Lett 162:360–366

Forde BG, Heyworth A, Pywell J, Kreis M (1985) Nucleotide sequence of a B1 hordein gene and the identification of possible upstream regulatory elements in endosperm storage protein genes from barley, wheat and maize. Nucleic Acids Res 13:7327–7339

Fromm ME, Taylor LP, Walbot V (1985) Expression of genes transferred into monocot and dicot plant cells by electroporation. Proc Natl Acad Sci USA 82:5824–5828

Fromm ME, Taylor LP, Walbot V (1986) Stable transformation of maize after gene transfer by electroporation. Nature 319: 791–793

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

Grantham G (1980) Workings of the genetic code. Trends Biochem Sci 5:327–331

Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon–anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18:199–209

Harris JI, Waters M (1976) Glyceraldehyde-3-phosphate dehydrogenase. In: Boyer PD (ed) The enzymes, vol 13, ed 3. Academic Press, New York, pp 1–48

Holm L (1986) Codon usage and gene expression. Nucleic Acids Res 14:3075–3078

Hyldig-Nielsen JJ, Jensen E, Paludan K, Wiborg O, Garrett R, Joergensen P, Marker KA (1982) The primary structures of two leghemoglobin genes from soybean. Nucleic Acids Res 10:689–701

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

Ikemura T, Ozeki H (1982) Codon usage and transfer RNA

contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. Cold Spring Harbor Symp Quant Biol 47:1087–1097

Izui K, Ishijima S, Yamaguchi Y, Katagiri F, Murata T, Shigesada K, Sugiyama T, Katsuki H (1986) Cloning and sequence analysis of cDNA encoding active phosphoenolpyruvate carboxylase of the C4-pathway from maize. Nucleic Acids Res 14:1615–1628

Keith B, Chua NH (1986) Monocot and dicot pre-mRNAs are processed with different efficiencies in transgenic tobacco. EMBO J 5:2419–2425

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England

Klösgen RB, Gierl A, Schwarz-Sommer Zs, Saedler H (1986) Molecular analysis of the waxy locus of *Zea mays*. Mol Gen Genet 203:237–244

Krebbers ET, Larrinua JM, McIntosh L, Bogorad L (1982) The maize chloroplast genes for the beta and epsilon subunits of the photosynthetic coupling factor cf-1 are fused. Nucleic Acids Res 10:4985–5002

Lebrun M, Waksman G, Freyssinet G (1987) Nucleotide sequence of gene encoding corn ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit (rbcs). Nucleic Acids Res 15:4360

Lörz H, Baker B, Schell J (1985) Gene transfer to cereal cells mediated by protoplast transformation. Mol Gen Genet 199: 178–182

Marchionni M, Gilbert W (1986) The triosephosphate isomerase gene from maize: introns antedate the plant–animal divergence. Cell 46:133–141

Martin W, Cerff R (1986) Prokaryotic features of a nucleus-encoded enzyme: cDNA sequences for chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases from mustard (*Sinapis alba*). Eur J Biochem 159:323–331

Maruyama T, Gojobori T, Aota S, Ikemura T (1986) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 14:r151–r197

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564

Mazur BJ, Chui CF (1985) Sequence of a genomic DNA clone for the small subunit of ribulose bisphosphate carboxylase-oxygenase from tobacco. Nucleic Acids Res 13:2373–2386

McIntosh L, Poulsen C, Bogorad L (1980) Chloroplast gene sequence for the large subunit of ribulose bisphosphate carboxylase of maize. Nature 288:556–560

Morris J, Herrmann RG (1984) Nucleotide sequence of the gene for the P_{660} chlorophyll a apoprotein of the photosystem II reaction center from spinach. Nucleic Acids Res 12:2837–2850

Pedersen K, Devereux J, Wilson DR, Sheldon E, Larkins BA (1982) Cloning and sequence analysis reveal structural variation among related zein genes in maize. Cell 29:1015–1026

Potrykus I, Saul M, Petruska I, Paszkowski J, Shillito RD (1985) Direct gene transfer to cells of a graminaceous monocot. Mol Gen Genet 199:183–188

Prat S, Cortadas J, Puigdomenech P, Palan J (1985) Nucleic acid (cDNA) and aminoacid sequences of the maize endosperm protein glutelin-z. Nucleic Acids Res 13:1493–1504

Rogers JC, Milliman C (1983) Isolation and sequence analysis of a barley alpha-amylase cDNA clone. J Biol Chem 258: 8169–8174

Rohde W, Barzen E, Marocco A, Schwarz-Sommer Zs, Saedler H, Salamini F (1986) Isolation of genes that could serve traps for transposable elements in *Hordeum vulgare*. Barley Genetics V (in press)

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467

Schoeffl F, Raschke E, Nagao RT (1984) The DNA sequence analysis of soybean heat-shock genes and identification of possible regulatory promoter elements. EMBO J 3:2491–2497

Schuler MA, Ladin BF, Pollaco JC, Freyer G, Beachy RN (1982) Structural sequences are conserved in the genes coding for the alpha, alpha' and beta-subunits of the soybean 7S seed storage protein. Nucleic Acids Res 10:8245–8261

Schwarz-Sommer Zs, Shepherd N, Tacke E, Gierl A, Rohde W, Leclerc L, Mattes M, Berndtgen R, Petersen PA, Saedler H (1987) Influence of transposable elements on the structure and function of the A1 gene of *Zea mays*. EMBO J 6:287–294

Shah DM, Hightower RC, Meagher RB (1983) Genes encoding actin in higher plants: intron positions are highly conserved but the coding sequences are not. J Mol Appl Genet 2:111–126

Sharp PM, Li W-H (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res 14:7737–7749

Shih MC, 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 47:73–80

Smith SM, Bedbrook J, Speirs J (1983) Characterization of three cDNA clones encoding different mRNAs for the precursor to the small subunit of wheat ribulose bisphosphate carboxylase. Nucleic Acids Res 11:8719–8734

Sprinzl M, Moll J, Meissner F, Hartmann T (1985) Compilation of tRNA sequences. Nucleic Acids Res 13:r1–r50

Tabata T, Sasaki K, Iwabuchi M (1983) The structural organization and DNA sequence of a wheat histone H4 gene. Nucleic Acids Res 11:5865–5875

Tabata T, Fukasawa M, Iwabuchi M (1984) Nucleotide sequence and genomic organization of a wheat histone H3 gene. Mol Gen Genet 196:397–400

Thompson RD, Bartels D, Harberd NP (1985) Nucleotide sequence of a gene from chromosome 1D of wheat encoding a HMW-glutenin subunit. Nucleic Acids Res 13:6833–6846

Vieira J, Messing J (1982) The pUC plasmids, an M13 mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19:259–268

Weeden NF (1981) Genetic and biochemical implications of the endosymbiotic origin of the chloroplast. J Mol Evol 17: 133–139

Wells D, Bains W, Kedes L (1986) Codon usage in histone gene families of higher eukaryotes reflects functional rather than phylogenetic relationships. J Mol Evol 23:224–241

Werr W, Frommer WB, Maas C, Starlinger P (1985) Structure of the sucrose synthase gene on chromosome 9 of *Zea mays* L. EMBO J 4:1373–1380

Received May 18, 1987/Revised and accepted August 4, 1987