Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2672–2676, April 1988
Genetics

Intron conservation across the prokaryote–eukaryote boundary: Structure of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize

(CpG-rich islands/enhancer-like elements/exon shuffling/transit peptide evolution/alcohol dehydrogenase)

FRANÇOISE QUIGLEY*, WILLIAM F. MARTIN†, AND RÜDIGER CERFF*‡

*Laboratoire de Biologie Moléculaire Végétale, Centre National de la Recherche Scientifique UA 1178, Université Joseph Fourier, B.P. 53, F-38041, Grenoble Cedex, France; and †Max-Planck-Institut für Züchtungsforschung, D-5000 Cologne 30, Federal Republic of Germany

Communicated by Winslow R. Briggs, December 11, 1987 (received for review September 14, 1987)

ABSTRACT **The nuclear gene encoding chloroplast glyceraldehyde-3-phosphate dehydrogenase (GAPDH) from maize has been cloned and sequenced. The gene is G+C rich in its coding sequences and, in addition, contains a CpG-rich region surrounding the promoter. Further upstream several enhancer-like repetitions have been identified that may control the light- and phytochrome-mediated expression of this gene. The gene is interrupted by three introns. Introns 1 and 2 are located within the sequence encoding the transit peptide, dividing it into three parts, each containing one of the three major homology blocks typical for transit peptides of nucleus-encoded chloroplast proteins. Intron 3 is located at codon 166 (glycine) at the same nucleotide position as intron 1 in the GAPDH gene from the nematode *Caenorhabditis elegans*, suggesting that this intron was present in the parental GAPDH gene from which these two modern descendants originated. Intron 3 divides the GAPDH protein into its two constituent domains, the NAD-binding and the catalytic domain, immediately after helix α1 at a position homologous to that of intron 9 in the gene for maize alcohol dehydrogenase, thereby confirming the prediction of Brändén *et al.* on the basis of gene–protein structure correlations in maize alcohol dehydrogenase for the placement of introns in the GAPDH gene [Brändén, C.-I., Eklund, H., Cambillau, C. & Pryor, A. J. (1984) *EMBO J.* 3, 1307–1310]. These results suggest that intron 3 is an archetypical relic of early GAPDH and alcohol dehydrogenase evolution, whereas introns 1 and 2 were implicated in the evolution of chloroplast transit peptides.**

Introns have existed in the genomes of eukaryotes for at least one billion years (1, 2), but were they also present in genes before the divergence of prokaryotes and eukaryotes? Structural information from nuclear genes of higher and lower eukaryotes (for reviews see refs. 1, 3, and 4) indicates that there are two major mechanisms responsible for the differences observed in number and position of introns between homologous genes of separate species: (*i*) selective loss of preexisting introns rather than selective addition of new ones and (*ii*) intron sliding. This is in agreement with the suggestion that introns were present in the most primitive genes, where they accelerated the evolution of new functions by enhancing the combination of exons encoding small stably folding polypeptides to yield mosaic (sub-)genes encoding functional domains. These domains could then again be "shuffled" to form genes encoding the basic enzymes of primary metabolism found in all cells (3, 5–7). Whereas exon shuffling has been demonstrated in the evolution of the genes for the serine proteases (8) and the low density lipoprotein receptor (9), only proteins belonging to the most ancient metabolic pathways, such as glycolysis or photosynthesis, are likely to be encoded by genes old enough to provide evidence of exon shuffling in the primordial assembly of structures conserved in the proteins of extant species.

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) of higher plants is involved in glycolysis [NAD-dependent; D-glyceraldehyde-3-phosphate:NAD^+ oxidoreductase (phosphorylating); EC 1.2.1.12] and photosynthesis [NADP-dependent; D-glyceraldehyde-3-phosphate:$NADP^+$ oxidoreductase (phosphorylating); EC 1.2.1.13]. The photosynthetic GAPDH of the chloroplast shares more amino acid homology with the GAPDHs of thermophilic eubacteria than it does with the cytosolic enzyme encoded within the same nucleus (10–12), suggesting that the gene for chloroplast GAPDH is of prokaryotic origin and that its present nuclear location reflects a gene transfer from the endosymbiotic plastid ancestor into the nucleus of a primitive eukaryotic "host." Since the genes of present day bacteria as a rule do not contain introns (any more), a "fossil prokaryotic gene" encoding chloroplast GAPDH in the nucleus of higher plants represents an interesting test system for the hypothesis that introns existed before the divergence of prokaryotes and eukaryotes.

Here we discuss the nucleotide sequence and intron–exon structure of the gene for chloroplast GAPDH from *Zea mays*. The gene contains three introns. Introns 1 and 2 are located in the region of the gene encoding the N-terminal transit peptide; the third intron is located at codon 166 at the same nucleotide position as intron 1 in the GAPDH gene from the nematode *Caenorhabditis elegans* (13).

MATERIALS AND METHODS

Plant Material. The mRNA and genomic DNA used for cloning originate from two genetic stocks of P. A. Peterson (Ames, Iowa). The maize plants were grown in Cologne [the accession numbers 81 4822Y-2 (cDNA library, see ref. 14) and 893 (genomic library)].

Cloning of the Maize Chloroplast GAPDH Gene. Preparation of the plant DNA and cloning of the *Mbo* I partial digestions thereof into the EMBL4 vector were carried out as described by Schwarz-Sommer *et al.* (14). Recombinant clones containing the chloroplast GAPDH gene were identified by plaque hybridization, by using the nick-translated 1.4-kilobase (kb) *Eco*RI fragment of pZm57 (12). To obtain genomic clones specifically corresponding to cDNA pZm57, a 150-base-pair (bp) *Eco*RI–*Hae* III fragment from the 3′-untranslated region of pZm57 was labeled and used to screen positive recombinants during rounds of purification. *Eco*RI fragments from positive genomic clones were subcloned into pBR322.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: GAPDH, glyceraldehyde-3-phosphate dehydrogenase; ADH, alcohol dehydrogenase; PGK, phosphoglycerate kinase; SSU, small subunit of ribulose bisphosphate carboxylase.

‡To whom reprint requests should be addressed.

Isolation of Further GAPDH cDNAs. From the maize cDNA library cloned into the *Eco*RI site of NM1149 that had been used to isolate clone pZm57 encoding chloroplast GAPDH (12), 10 additional clones were isolated that hybridized to the 150-bp *Eco*RI–*Hae* III fragment from the 5′ end of pZm57. The three longest cDNAs were subcloned into pUC18, and their 5′ and 3′ ends were sequenced.

Nucleic Acid Sequencing. The nucleotide sequence of the chloroplast GAPDH gene was determined by the dideoxy chain-termination method (Amersham), after subcloning fragments into suitable sites of the replicative form of M13mp10, mp11, mp18, and mp19 following the protocol supplied by Amersham. The 3′- and 5′-terminal sequences of the three cDNA subclones were determined by the plasmid-sequencing protocol provided by Boehringer Mannheim.

RESULTS AND DISCUSSION

Cloning and Sequence Analysis of the Maize Chloroplast GAPDH Gene. From a genomic library of 2×10^6 recombinant phages, 24 clones were isolated that hybridized to the 1.4-kb *Eco*RI fragment of pZm57 (12). Of these 24 only 2, λgapA1 and λgapA2, hybridized to the 150-bp *Eco* RI-*Hae* III fragment from the 3′-untranslated region of pZm57. Restriction analysis of clones λgapA1 and λgapA2 revealed that they overlapped. Clone λgapA1 contained the larger hybridizing 6.4-kb *Eco*RI fragment that was subcloned into the vector pBR322 and sequenced. The reported sequence is 4060 bases long and contains the entire chloroplast GAPDH gene (*gapA*) and 1.5 kb of the 5′ upstream region. This sequence has been entered in the EMBL DNA library and is available from the authors.

Intron positions were determined by comparison to the cDNA sequence: introns 1 (134 bp) and 2 (182 bp) are located within the region encoding the transit peptide, and intron 3 (119 bp) is at codon 166 (glycine) (see below). The 5′ and 3′ junctions of these introns exhibit no significant deviations from the consensus sequences derived from other plant nuclear genes (15).

Since cDNA clone pZm57 (12) does not encode a full-length transit peptide, additional cDNA clones extending farther in the 5′ direction were identified. The 5′ end of the longest clone maps to Met-5, the second methionine of the transit peptide. The remaining four amino acids of the transit peptide were deduced from the genomic sequence (see Fig. 2*A* and below).

At 62 bp upstream from the initiation codon a putative "TATA box" (TATATA) can be found, and 121 bp further upstream, at position −183, the 13-mer CAGCCATTCGTCG is present that shows 11 of 13 matches to a sequence corresponding to the "CAAT box" of the chicken GAPDH gene (16). The sequence surrounding the start codon GGCCAUGGC conforms well to the consensus sequence AACAAUGGC determined for translation initiation in plants (17). Within 60 bp on the 5′ side of the putative CAAT box GGGC is repeated five times. Between base positions −1485 and −1347 the 24-base sequence AAGTCCCCGTGGTGTTTTTTCGAT is repeated twice. Its central GTGGTGTTTTTT is similar to the enhancer-like GTGG(T/A)(T/A)(T/A)(T/A/G) sequences that have been noted for possible roles in the expression of the 35S promoter of cauliflower mosaic virus, of chalcone synthase, the pea small subunit of ribulose bisphosphate carboxylase (SSU) gene family, and of the genes for barley aleurain and α-amylase (reviewed in ref. 18), for sucrose synthase (19), and for the A1-locus enzyme from maize (14). The CCCCGTGG of this duplication is repeated twice further downstream at base positions −709 and −678. Whether or not these enhancer-like repeats are implicated in the light control of chloroplast GAPDH expression (20) or in any other regulatory function is not known. At the 3′ end there are two potential polyadenylylation signals AATATA 30 and 118 bp upstream of the poly(A) addition site.

The maize chloroplast GAPDH mRNA is very G+C-rich and shows an extreme bias for guanosine or cytidine (97%) at the third base position of codons. It has been suggested (12) that this bias in the maize chloroplast GAPDH gene and several other cereal genes may be the result of a translational constraint exerted by the relative abundancies of various isoaccepting tRNA species, analogous to the cases demonstrated for yeast and *Escherichia coli*. However, it may be argued that the G+C richness of this plant mRNA simply reflects the G+C content of the surrounding genomic sequences, as has been shown for a number of G+C-rich animal genes (reviewed in ref. 12). Yet, as shown in Table 1, this is clearly not the case. The gene for maize chloroplast GAPDH and its surrounding sequences have disproportionate G+C contents: 40% in flanking sequences (excluding the region surrounding the promoter, see below), 53% in introns, 67% in the coding parts of the exons, and 97% in the third base position of triplets (12). Similar distributions in G+C content have been noted (12) for other maize genes. For the maize chloroplast GAPDH gene, the G+C content is also high in the region surrounding the promoter (60%, Table 1), suggesting that selection exerted by other unknown factors, perhaps the association of nuclear proteins with a CpG-rich island (see below and ref. 21), in addition to codon usage may influence the nucleotide distribution within this plant nuclear gene.

An analysis of the distribution of CpG and GpC dinucleotides (Fig. 1 and Table 1) shows that there are some deviations from the expected frequencies. The doublet GpC is overrepresented in intron 2 (12.7% GpC, 8.3% expected; data not shown in Table 1), where it is part of an 11-fold GCT repetition. The doublet CpG, the site of DNA methylation, is underrepresented in the flanking sequences (2.4% vs. 4.0% expected) and somewhat overrepresented in the region surrounding the promoter (11.3% vs. 8.9% expected, Table 1). This contrast between promoter and flanking regions is somewhat analogous to the situation found in vertebrate genes, where CpG doublets are highly underrepresented except within the so-called CpG-rich islands surrounding the transcription start sites of housekeeping genes (for review see ref. 21; for CpG distribution in the chicken GAPDH gene see ref. 16). Thus the gene of maize chloroplast GAPDH as a whole, because of its extreme codon bias (12), may be considered a CpG-rich region (12.4% CpG in codons, see Table 1) with three interruptions representing the introns. Since maize DNA is highly methylated (2), our work (ref. 12 and the present results) raises the interesting questions of whether or not the biased maize genes, rich in G+C and hence CpG, are protected against methylation and its effects on the gene activity and mutation rate, as has been demonstrated for the CpG-rich islands of active vertebrate genes (reviewed in ref. 21).

Intron-Mediated Evolution of the Chloroplast GAPDH Transit Peptide. The mature catalytic subunit of mustard

Table 1. G+C content and frequencies of CpG and GpC dinucleotides in the maize chloroplast GAPDH gene

Sequence(s)	Observed, %					Expected, %
	G	C	G+C	CpG	GpC	CpG and GpC
3′ and 5′ sequences (2023 bases)	18.2	21.8	40.0	2.4	4.1	4.0
Introns (435 bases)	23.9	29.0	52.9	7.4	9.5	6.9
Promoter region (bases −1 to −390)	25.6	34.6	60.2	11.3	8.7	8.9
Exons (404 codons)	29.4	37.6	67.0	12.4	10.1	11.1

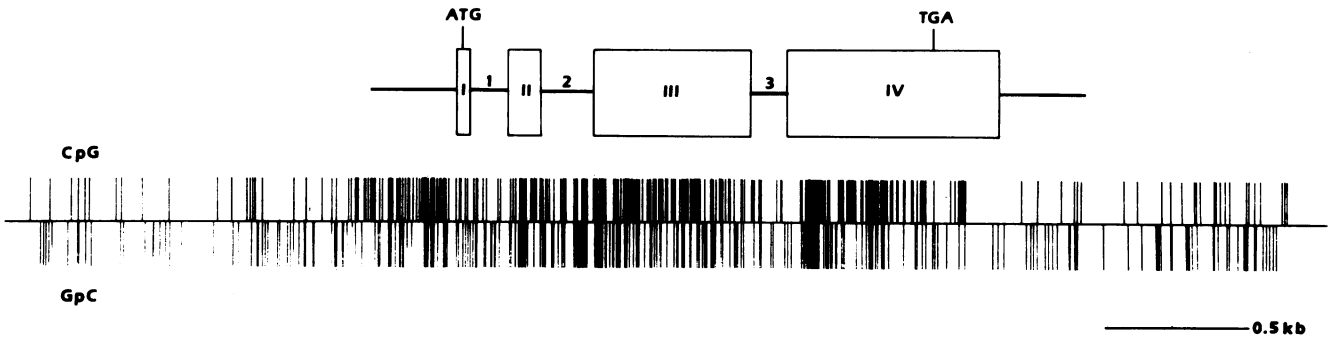

FIG. 1. Distributions of CpG and GpC dinucleotides within the maize chloroplast GAPDH gene and its surrounding sequences. Each vertical line represents a single CpG or GpC dinucleotide. CpG doublets are slightly underrepresented in flanking sequences and somewhat overrepresented in the region surrounding the promoter (see Table 1 and ref. 21). Exons are indicated by boxes and Roman numerals; introns are indicated by horizontal lines and Arabic numerals.

chloroplast GAPDH starts with Met-0, corresponding to lysine in the maize enzyme (Fig. 2*B* and ref. 12). The four N-terminal residues Met-Ala-Ser-Ser of the chloroplast GAPDH precursor (20) from maize precisely match the N-terminal consensus sequence determined for the transit peptides of SSU (Fig. 2*A* and see below). The transit peptide of maize chloroplast GAPDH precursor defined by these criteria contains 66 amino acids and has a calculated molecular mass of 7.2 kDa, which is in excellent agreement with the reported value of 7 kDa (20).

It has become apparent that transit peptides of nuclear-encoded chloroplast proteins are composed of three functional domains (reviewed in ref. 25), and a common amino acid framework of three major homology blocks has been identified by Karlin-Neumann and Tobin (22). As shown in Fig. 2*A*, homology blocks I, II, and III are well conserved in the maize chloroplast GAPDH precursor transit peptide especially when compared to the SSU consensus sequences.

A common amino acid framework in chloroplast transit peptides poses interesting evolutionary problems. The central question is, how various sequences representing descendants of separate evolutionary lineages have come to possess similar N-terminal transit peptides. According to the exon shuffling hypothesis (5), transit peptides may have

A. TRANSIT SEQUENCE

```
                intron 1                                  intron 2
        maize chloroplast                         maize chloroplast
              gln/gly                                   ala/val
   Block I          Block II                          Block III
1. MA+SM+SS         PF+G+K     P                      SG+GRV      general consensus
2. MASSM+SS         PF+GLK     P                      SNGGRV      SSU consensus
3. MASSMLSATTVPLQQGGGLSEFSGLRSSASLPMRRNATSDDFMSAVSFRTHAVGTSGGPRRAPTEA GAPDH chloroplast
        |         |         |         |         |         |
       -60       -50       -40       -30       -20       -10
```

B. GAPDH MATURE SUBUNITS

```
   0         10        20  25ABC    30        40        50  54A    60   66AB 68A69 70        80        90        100       110
   |         |         |    |||    |         |         |    |     |      ||  | ABC|        |         |         |         |
1.GKIKIGINGFGRIGRLVARVALQSEDV---ELVAVNDPFITTDYMTYMFKYDTVHGHWKHSDITLKDSKTLL--FGDK---PVTVFGIRNPEEIPWGEAGAEYVVESTGVFTDKDKAAAHLK
2.S ANV          L A VEKDT ---QV         I   V L    ST  QF -GTV YDGDFLIVQKD - SSHKIK  NSKD AA A  SVK DF          E  S   Q
3.  V V V        T A VL GK ---QV  I     DLN  V      ST   F -GTVKAENG LVI--N -H---AI I QE D SN K AD             TME  G
4. KL VA       NFL CWHGRG ASPLDVI I  TG-GVKQASHLL   STL IFD-A VKPVGDNAISV-D - ---VIK VSD   SNL    L IDL I G    V REG GK IQ
    <<<<<< oooooooooooooo       <<<<<<  ooooooooooo                              <<<<   oooo  oooo   <<<<<       oooooooooo
      βA       αB                 βB       αC                                    βC    αD    αD     βD             αE
```

```
                                                intron 3 maize chloroplast
                                                    intron 1 nematode
                                                        g ly 166
          120 123A   130     138A140       150       160       170       180       190       200       210       220       230
            |   |    |         |  |        |         |         |         |         |         |         |         |         |
1.GGAKKVVISAPSK-DAPMFVVGVNEDKYT-SDVNIVSNASCTTNCLAPLAKVIHDNFGIVEGLMTTVHAITATQKTVDGPSAKDWRGGRAASFNIIPSSTGAAKAVGKVLPDLNGKLTGMSF
2.      I     A-    Y     HE  DA NDHV                  N     I         V          G L  D  G GQ    A           I E         A
3.    R I     A-      M   HE  D-KSLK                                               G L  DD G AQ    A           I E         A
4.A     L T  G G I TY     A Q NPDEP- I           FV  LDQK    IK T   T SY GD RLL A- HR L RA    AL  V T        SL   N K   N IAL
  oo  <<<<        <<<<              <<<< oooooooooooooooooo   <<<<<<<<<                          <<<<<<<      ooo        <<<<<<
       βE                          βF        α1              β1                                  β2                    β3
```

```
                                                                  intron 2 nematode
                                                                intron 11 chicken
                                                                    tr p 310
         240       250       260       270       280       290       300                 320       330
           |         |         |         |         |         |         |                   |         |
1.RVPTVDVSVVDLTVRIEKGASYEDIKKAIKAASEGPLKGIMGYVEEDLVSTDFLGDSRSSIFDAKAGIALNDHFVKLVSWYDNEWGYSNRVVDLIRHMFKSQ* maize cyt. --
2.    P         L  P  MD    VV   AD  M   LA T DQV     VS PH      G C S  PN          Y          GYIATRG* nematode   67% --
3.    PN      C L  P K D   RVV   AD      L  T DQV  C  N   H  T   G                   F         MV  ASKE* chicken    71% 74% --
4.    PN      V QVS KTLA EVNQ FRD AANE T  LEVCDVP   V  RCSDV  TI  SLTMVMG DM  VI              Q     ADICANQWK*maize chl. 45% 46% 45%
  <<<   <<<<<<<<<<    oooooooooooooooo   <<<<<             <<<<   <<<<   <<<<<<<<oooooooooooooooooooooo
          β4                 α2          β5                β6    β7     β8          α3
```

FIG. 2. (*A*) Comparison of the transit peptide of maize chloroplast GAPDH (line 3) with the general and the SSU consensus sequences of homology blocks I, II, and III (lines 1 and 2; see ref. 22). Positions of introns 1 and 2 of the gene for maize chloroplast GAPDH are indicated by arrows. (*B*) Amino acid sequence alignment of the mature GAPDH proteins from maize cytosol (line 1, ref. 12), nematode (line 2, ref. 13), chicken (line 3, ref. 16), and maize chloroplast (line 4) by the standard numerical order (10). The sequences are compared to maize cytosolic GAPDH, which is the only sequence written in full. For the other three enzymes only amino acids not identical to the reference sequence are indicated, and the sequence similarities for all six pairwise comparisons are given at the bottom of the figure. In the present alignment the following three insertions have been shifted (previous positions are given in parentheses, see refs. 10 and 12): maize cytosolic GAPDH, His-54A (Lys-53A); chloroplast GAPDH, Val-66A (Ile-64A) and Gly-123A (Lys-122A). Sequence elements forming helices (OOOOO) and β-structures (<<<<<) are indicated (see refs. 23 and 24). The positions of the conserved introns are indicated by arrows.

been combined with existing protein functions through intron-dependent recombination and the location of introns near the cleavage site of higher plant SSU transit peptides (for review see ref. 26) is consistent with this idea. The present discovery of two internal introns separating homology blocks I, II, and III within the chloroplast GAPDH transit peptide (Fig. 2*A*) suggests that introns have also been implicated in the (convergent) evolution of the common amino acid framework. Assembly of the common sequence elements may have been achieved by intron-mediated recombination, while the individual spacing of these elements and sequence variation within interblock regions could have been accomplished by intron sliding (27).

According to this hypothesis all nuclear genes encoding chloroplast proteins such as the SSU (reviewed in ref. 26), the light-harvesting chlorophyll a/b complex (28), and plastocyanin (29) would have lost their introns from the region encoding the transit peptide. However, the alternative hypothesis that introns have been inserted specifically into the transit peptide region of the chloroplast GAPDH gene can also not be discarded completely at the present time, although characteristics of transposable elements, as found for the intervening sequence of the light-harvesting chlorophyll a/b complex gene from *Lemna gibba* (28), are absent from introns 1 and 2.

Is Intron 3 of the Chloroplast GAPDH Gene from Maize Older Than Eukaryotic Cells? The intron–exon structures of the genes for maize chloroplast GAPDH and the glycolytic enzymes from chicken (11 introns, refs. 16 and 30) and nematode (2 introns, ref. 13) are compared in Fig. 3 and the corresponding amino acid sequences in Fig. 2*B*. The position of intron 2 in the nematode GAPDH gene is identical to that of intron 11 in the chicken gene, as reported (13). It interrupts the coding sequence within the triplet for the conserved residue Trp-310 and separates the terminal helix $\alpha 3$ (residues 312–332 in Fig. 2*B*; box α' in Fig. 3) from the rest of the catalytic domain. Intron 3 of the chloroplast GAPDH gene is located between the first and second base of the triplet for the conserved Gly-166, separating helix $\alpha 1$ (residues 148–165 in Fig. 2*B*; box α in Fig. 3) and strand $\beta 1$ (residues 169–177 in Fig. 2*B*) of the catalytic domain (for details of the folding pattern see refs. 23 and 24). Its base position in maize coincides precisely with that of intron 1 of the nematode GAPDH gene (see Figs. 2*B*, 3*C*, and 4 and ref. 13), suggesting that it was also present in the parental GAPDH gene from which these two modern descendants originated. The two introns are 119 (maize) and 52 (nematode) bases long. They are both relatively A+T-rich and share a certain sequence similarity at their 5′ ends (17 of 22 matches, see Fig. 4). The fact that intron 7 in the chicken gene does not fall within Gly-166 but is located six codons further downstream (Fig. 3*A*) may be due to intron sliding (27).

Extensive intron conservation between plants and animals has been described by Marchionni and Gilbert (2) for the triosephosphate isomerase gene, where five introns were found at identical positions. Yet in this report, the glycolytic triosephosphate isomerase from maize was compared to its counterpart from chicken. Although a chloroplast-specific, photosynthetic triosephosphate isomerase exists (32), neither cDNA nor genomic sequences for this enzyme have yet been described. In contrast, the present contribution permits the comparison of gene structures for a photosynthetic enzyme and its glycolytic homologues (32). Nematode GAPDH and maize chloroplast GAPDH have only 46% of amino acid sequence identity, whereas the glycolytic GAPDHs from these species are 67% conserved (Fig. 2*B*). These values convert, by using the methods described (11, 12), to divergence times of 2.2 billion years, corresponding to the separation of GAPDH genes into the prokaryotic and eukaryotic subfamilies (10, 12), and to ≈1 billion years corresponding to the plant–animal division (1, 2). Hence, the age of 2 billion years for intron 3 of chloroplast GAPDH is about twice that of the most ancient intron reported (1, 2) and compares intriguingly with the appearance of the earliest eukaryotic cells some 1.5 billion years ago.

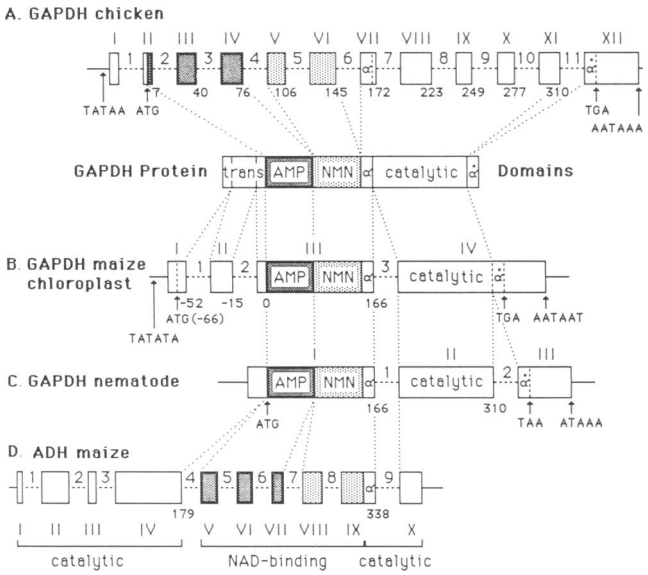

FIG. 3. Schematic comparisons of the exon–intron structures for the genes encoding chicken GAPDH (*A*; refs. 16 and 30), maize chloroplast GAPDH (*B*), nematode GAPDH (*C*; ref. 13) and maize ADH (*D;* ref. 31). Exons are indicated by boxes and Roman numerals and introns by broken horizontal lines and Arabic numerals. Numbers below the exons indicate their terminal codons. The lengths of all exons are true to scale. Trans, transit peptide; AMP and NMN, AMP and NMN subdomains of the NAD-binding domain; α and α', helices $\alpha 1$ and $\alpha 3$ of the catalytic domain (see Fig. 2*B*); catalytic, central part of the catalytic domain.

Is Intron 3 in the Gene for Chloroplast GAPDH Homologous to Intron 9 in the ADH Gene? The three-dimensional structures of GAPDH (23, 24), alcohol dehydrogenase (ADH), and lactate dehydrogenase have long been known. It was recognized that the nucleotide-binding domains of NAD-dependent dehydrogenases (reviewed in ref. 33) and other related enzymes [e.g., phosphoglycerate kinase (PGK), ref. 34], although dissimilar in sequence, represent highly homologous structures at the level of conformation: they all fold into a six-stranded parallel pleated sheet surrounded by helices that is subdivided into two symmetrical β–α–β–α–β elements, the AMP and NMN mononucleotide subdomains

```
                  Phe G                                                                ly Ile
Nematode:         TTC G ----GTAT-GATCATCATCAAATCAATTTGTTTTTTTCTAATTTATTTTATTTTCAG GT ATC
Maize chloroplast: TTC G GTGC····C····T··T···-C····C·C·AGCGCCG·A··C··GA·GCTCCC··CT-

                  TGTGATTATTGTTGCAAATCGGACAAACCTATAATAACGGCGTATATATATGCATGGATCAAG GC ATC
```

FIG. 4. Sequence alignment of intron 1 (52 bases; 42 A+T) and intron 3 (119 bases; 67 A+T) of the genes for nematode and maize chloroplast GAPDH, respectively. Both introns coincide precisely (except for the additional bases in the maize chloroplast intron) and split the highly conserved Gly-166 between the first and the second bases of its codon (see Figs. 2*B* and 3). Dots indicate identical bases in intron 3. Dashes specify deletions.

(see linear arrangements of GAPDH domains in Fig. 3). On the basis of these crystallographic findings, it was suggested (33) that the early assembly of the genes for NAD-dependent dehydrogenases occurred through the separate fusions of a region encoding a nucleotide-binding domain onto various unrelated genic regions encoding polypeptides conferring substrate specificity (the catalytic domains). If introns facilitated such recombination events (3, 5–7), then some introns that bound the nucleotide-binding domain and its subdomains might have been conserved. GAPDH and ADH provide a stringent test for this hypothesis because the nucleotide-binding domains are situated in different positions in the two proteins, in the *N*-terminal region in GAPDH and inserted in the catalytic domain in ADH (Fig. 3), this latter arrangement being also present in PGK (34).

Correlations between exon–intron junctions and the structural arrangement of the nucleotide-binding domain have been reported for chicken GAPDH (30), maize ADH (31), and human PGK (35). In each case it was shown that introns occur at the junction between the AMP and NMN subdomains (introns 4 and 7 for GAPDH and ADH, see Fig. 3 *A* and *D*; intron 8 for PGK, see ref. 35) and at one of the boundaries between the nucleotide-binding and catalytic domains (introns 6 and 4 for GAPDH and ADH, see Fig. 3 *A* and *D*; intron 10 for PGK, see ref. 35).

These gene–protein structure correlations between various nucleotide-binding proteins can now be extended for GAPDH and ADH to the helix element connecting in both enzymes the NMN subdomain with the catalytic domain (helix α1 for GAPDH, corresponding to residues 148–165 in Fig. 2*B*; residues 324–338 in ADH, see ref. 31; box α for GAPDH and ADH in Fig. 3). The junction between this helix and the catalytic domain coincides with intron 9 in the ADH gene (codon 338, see Fig. 3*D*), which led Brändén *et al.* (31) to suggest that introns may be found at corresponding positions in the genes for GAPDH (codon 165) and lactate dehydrogenase (codon 180). The present results confirm this prediction for GAPDH. They show that intron 3 in the maize gene [represented by intron 1 in *C. elegans* (13)] is located one nucleotide away from the predicted site. This striking experimental verification of the prediction of Bränden *et al.* (31), unrecognized by Yarbrough *et al.* (13) in their paper on the nematode GAPDH gene, provides further support for the proposed evolution of nucleotide-binding proteins (33) and the suggested role of introns (exon shuffling) in this early constructional pathway (31, 35). The absence of amino acid homology between the NAD-binding domains of GAPDH and ADH, although of common ancestry, places their time of divergence long before that of the GAPDH dichotomy (see above and refs. 10–12), possibly dating their most recent common progenitor and hence the appearance of their homologous introns in the early period of molecular evolution, perhaps >3 billion years ago (36).

While the gene–protein structure correlations found for various nucleotide-binding proteins agree with the hypothesis that most introns are relics of primordial genes, the GAPDHs from maize chloroplast and nematode present convincing proof of an identical intron position in two genes displaying a sequence divergence corresponding to the prokaryote–eukaryote separation. Indeed, precisely such a comparison has been suggested (37) as an experimental test for the early intron hypothesis. According to this hypothesis the nuclear genes encoding the GAPDHs from *C. elegans* and plant chloroplasts have lost most of their introns—e.g., 9 and 10 introns, respectively, if compared to the chicken gene (16, 30). It is intriguing that this apparent massive loss of introns along two separate evolutionary pathways led to the conservation of an identical intron in both modern descendants. Rather than coincidence, this may be the result of a gene-specific selective pressure distinguishing between more and less important introns. Finally, the present discovery of introns within the transit peptide region indicates the possibility that not all introns are necessarily relics of primordial genes. Under the conditions of transit peptide evolution, where the appropriate spacing of certain topogenic signals appears more important than stringent amino acid sequence conservation, a spontaneous mutational origin of introns seems even conceivable.

The excellent technical assistance of M. Rocipon is gratefully acknowledged. This work was funded by grants from the Centre National de la Recherche Scientifique (UA 1178), the Ministère de la Recherche et Technologie, and the Ministère de l'Education Nationale.

1. Gilbert, W., Marchionni, M. & McKnight, G. (1986) *Cell* **46,** 151–154.
2. Marchionni, M. & Gilbert, W. (1986) *Cell* **46,** 133–141.
3. Blake, C. C. F. (1983) *Nature (London)* **306,** 535–537.
4. Gilbert, W. (1985) *Science* **228,** 823–824.
5. Gilbert, W. (1978) *Nature (London)* **271,** 501.
6. Blake, C. C. F. (1978) *Nature (London)* **273,** 267.
7. Doolittle, W. F. (1978) *Nature (London)* **272,** 581–582.
8. Rogers, J. (1985) *Nature (London)* **315,** 458–459.
9. Südhoff, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228,** 815–822.
10. Martin, W. & Cerff, R. (1986) *Eur. J. Biochem.* **159,** 323–331.
11. Shih, M.-C., Lazar, G. & Goodman, H. M. (1986) *Cell* **47,** 73–80.
12. Brinkmann, H., Martinez, P., Quigley, F., Martin, M. & Cerff, R. (1988) *J. Mol. Evol.* **26,** 320–328.
13. Yarbrough, P. O., Hayden, M. A., Dunn, L.A., Vermersch, P. S., Klass, M. R. & Hecht, R. M. (1987) *Biochim. Biophys. Acta* **908,** 21–33.
14. Schwarz-Sommer, Zs., Shepherd, N., Tacke, E., Gierl, A., Rohde, W., Leclercq, L., Mattes, M., Berndtgen, R., Peterson, P. A. & Saedler, H. (1987) *EMBO J.* **6,** 287–294.
15. Brown, J. W. S. (1986) *Nucleic Acids Res.* **14,** 9549–9559.
16. Stone, E. M., Rothblum, K. N., Alevy, M. C., Kuo, T. M. & Schwartz, R. J. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1628–1632.
17. Lütcke, H. A., Chow, K. C., Mickel, F. S., Moss, K. A., Kern, H. F. & Scheele, G. A. (1987) *EMBO J.* **6,** 43–48.
18. Whittier, R. F., Dean, D. A. & Rogers, J. C. (1987) *Nucleic Acids Res.* **15,** 2515–2535.
19. Werr, W., Frommer, W.-B., Maas, C. & Starlinger, P. (1985) *EMBO J.* **4,** 1373–1380.
20. Cerff, R. & Kloppstech, K. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 7624–7628.
21. Bird, A. P. (1986) *Nature (London)* **321,** 209–213.
22. Karlin-Neumann, G. A. & Tobin, E. M. (1986) *EMBO J.* **5,** 9–13.
23. Buehner, M., Ford, G. C., Moras, D., Olsen, K. W. & Rossmann, M. G. (1974) *J. Mol. Biol.* **90,** 25–49.
24. Biesecker, G., Harris, J. I., Thierry, J. C., Walker, J. E. & Wonacott, A. J. (1977) *Nature (London)* **266,** 328–333.
25. Schmidt, G. W. & Mishkind, M. L. (1986) *Annu. Rev. Biochem.* **55,** 879–912.
26. Goldschmidt-Clermont, M. & Rahire, M. (1986) *J. Mol. Biol.* **191,** 421–432.
27. Craik, C. S., Rutter, W. J. & Fletterick, R. (1983) *Science* **220,** 1125–1129.
28. Karlin-Neumann, G. A., Kohorn, B. D., Thornber, J. P. & Tobin, E. M. (1985) *J. Mol. Appl. Genet.* **3,** 45–61.
29. Rother, C., Jansen, T., Tyagi, A., Tittgen, J. & Herrmann, R. G. (1986) *Curr. Genet.* **11,** 171–176.
30. Stone, E. M., Rothblum, K. N. & Schwartz, R. J. (1985) *Nature (London)* **313,** 498–500.
31. Brändén, C.-I., Eklund, H., Cambillau, C. & Pryor, A. J. (1984) *EMBO J.* **3,** 1307–1310.
32. Weeden, N. F. (1981) *J. Mol. Evol.* **17,** 133–139.
33. Rossmann, M. G., Liljas, A., Brändén, C.-I. & Banaszak, L. J. (1975) in *The Enzymes,* ed. Boyer, P. D. (Academic, New York), 3rd Ed., Vol. 11, pp. 61–102.
34. Banks, R. D., Blake, C. C. F., Evans, P. R., Haser, R., Rice, D. W., Hardy, G. W., Merrett, M. & Phillips, A. W. (1979) *Nature (London)* **279,** 773–777.
35. Michelson, A. M., Blake, C. C. F., Evans, S. T. & Orkin, S. H. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 6965–6969.
36. Doolittle, W. F. (1980) *Trends Biochem. Sci.* **5,** 146–149.
37. Cornish-Bowden, A. (1985) *Nature (London)* **313,** 434–435.