Evidence for a chimeric nature of nuclear genomes: Eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes

(endsymbiosis/lateral gene transfer/paralogous genes/Anabaena variabilis/purple bacteria)

WILLIAM MARTIN*,†, HENNER BRINKMANN*, CATHERINE SAVONNA‡, AND RÜDIGER CERFF*

*Institut für Genetik, Technische Universität Braunschweig, Postfach 3329, D-38023 Braunschweig, Germany; and †Laboratoire de Biologie Moléculaire Végétale, Université Joseph Fourier, B.P. 53X, F-38041 Grenoble Cedex, France

Communicated by Robert K. Selander, July 1, 1993

ABSTRACT Higher plants possess two distinct, nuclear gene-encoded glyceraldehyde-3-phosphate dehydrogenase (GAPDH) proteins, a Calvin-cycle enzyme active within chloroplasts and a glycolytic enzyme active within the cytosol. The gene for the chloroplast enzyme was previously suggested to be of endosymbiotic origin. Since the ancestors of plastids were related to cyanobacteria, we have studied GAPDH genes in the cyanobacterium Anabaena variabilis. Our results confirm that the nuclear gene for higher plant chloroplast GAPDH indeed derives from the genome of a cyanobacterium-like endosymbiont. But two additional GAPDH genes were found in the Anabaena genome and, surprisingly, one of these sequences is very similar to nuclear genes encoding the GAPDH enzyme of glycolysis in plants, animals, and fungi. Evidence that the eukaryotic nuclear genes for glycolytic GAPDH, as well as the Calvin-cycle genes, are of eubacterial origin suggests that eukaryotic genomes are more highly chimeric than previously assumed.

Plastids were once free-living prokaryotes and must have possessed all genes necessary for photoautotrophic growth at the time of endosymbiosis. Yet higher plant chloroplast DNA encodes at least an order of magnitude fewer genes than the genomes of free-living prokaryotes. The majority of higher plant genes involved in photosynthesis, a metabolic pathway surely possessed by the endosymbiont, are currently located in the nucleus. Complete sequences for three plastid genomes have revealed that the known Calvin-cycle enzyme other than the large subunit of ribulose-bisphosphate carboxylase/oxygenase is encoded by chloroplast DNA (for review see ref. 1). Under the gene-transfer corollary to the endosymbiotic theory, plant nuclear genes for those proteins essential to photoautotrophy in cyanobacteria (2) were ultimately derived from the endosymbiont's genome (3). They were transferred into the nucleus and their products were then reimported into the organelle of their origin with the help of a transit peptide (4, 5).

In higher plants, glycolytic and Calvin-cycle pathways possess a number of enzymatic reactions in common which are catalyzed by distinct enzymes unique to each (6). In our previous studies of plant nuclear genes encoding glyceraldehyde-3-phosphate dehydrogenase (GAPDH) enzymes of the Calvin cycle (GAPA and GAPB; EC 1.2.1.13) and glycolysis (GAPC; EC 1.2.1.12), our working hypothesis predicted that GapA and GapB genes should reflect the evolution of the endosymbiont, whereas GapC genes should reflect that of the eukaryotic host nucleus (refs. 5, 7–10; see also ref. 11). To test this prediction, we have screened a gene library of Anabaena variabilis with a degenerate oligonucleotide de
dsigned against a highly conserved region of GAPDH amino acid sequences. We have found three GAPDH genes in Anabaena (8). Our comparative analyses revealed that this cyanobacterium possesses the expected homologue of Calvin-cycle GAPDH genes of plants and, surprisingly, a GAPDH gene closely related to GapC homologues from plants, animals, and fungi. Here we present evidence which strongly suggests that all eukaryotic GAPDH genes studied to date were derived by lateral (endosymbiotic) gene-transfer events early in eukaryotic evolution.

MATERIALS AND METHODS

Molecular Methods. DNA from light-grown (12) axenic cultures of A. variabilis (ATCC 29413) was prepared through cesium chloride gradients. HindIII fragments were cloned into ANML144 (13) and screened by plaque hybridization at 32°C in hybridization buffer (900 mM NaCl/60 mM sodium phosphate, pH 7.4/6 mM EDTA/0.1% SDS/0.02% polyvinylpyrrolidone/0.02% Ficoll 400) with probe at 10 ng/ml (5 × 10⁶ cpm/ml). The probe was an end-labeled 16-fold degenerate 16-mer oligodeoxynucleotide constructed against the conserved amino acid motif WYDNE(W/Y/F). Three classes of positively hybridizing clones could be differentiated on the basis of inserts which correspond to the three HindIII bands found in Southern blots of A. variabilis genomic DNA when either the oligonucleotide or the respective inserts were used as probes (data not shown). Hybridizing HindIII restriction fragments for each class were subcloned into pBluescript SK (Stratagene) vectors and sequenced on both strands. Cloning and Southern hybridization experiments were performed for two separate cultures of A. variabilis which were independently obtained and independently grown. In both Southern and cloning experiments, the three GAPDH genes were present in equimolar amounts, indicating that all three genes are endogenous to the A. variabilis genome.

Phylogenetic Data Analysis. Deduced amino acid sequences were aligned with the Genetics Computer Group package (14). Phylogenetic trees were constructed by the neighboring method (15), which has been shown to be very efficient in recovery of the correct topology under a variety of sequence parameters (16). Distances were measured as nucleotide divergence at nonsynonymous sites (dn; ref. 17) calculated on the basis of the average of 990 homologous nucleotide positions (average of 760 nonsynonymous sites). Reliability of the distance matrix tree was estimated through

Abbreviation: GAPDH, glyceraldehyde-3-phosphate dehydrogenase.
†To whom reprint requests should be addressed.
§The sequences from Anabaena variabilis reported in this paper have been deposited in the GenBank data base [accession nos. L07497 (gap1), L07498 (gap2), and L07499 (gap3)].
comparison to bootstrap parsimony trees (DNABOOT of PHYLIP; ref. 18) performed on a data set consisting of first and second codon positions for the 27 sequences (734 sites each). This procedure approximates the removal of synonymous sites and improves the parsimony result when divergence between sequences is great (19).

**Nomenclature of the GAPDH Genes from Plants and Eubacteria.** Gene products (mRNAs, cDNAs, proteins) encoding or corresponding to subunits A and B of chloroplast GAPDH (Calvin cycle) and subunit C of cytosolic GAPDH (glycolysis) are specified as GAPA, GAPB, and GAPC; products from different members of the same gene family are numbered consecutively: GAPA1, GAPA2, GAPB1, GAPB2, etc. The corresponding genes are designated GapA1, GapA2, GapB1, GapB2, etc. GAPDH genes from eubacteria are numbered consecutively: gap1, gap2, gap3, etc. This nomenclature corresponds to that previously proposed (5) except that the plant genes are now designated by the same four letters as their corresponding products and that the last letter, designating the subunit type, has been capitalized (e.g., GapA1 instead of GpA).

**RESULTS AND DISCUSSION**

*A. variabilis* Contains Three Divergent GAPDH Genes. Each of the three *A. variabilis* GAPDH genes occur as a single copy as determined by Southern blot experiments (data not shown) using specific probes. The gene encoding *Anabaena*-na's homologue of the higher plant chloroplast enzyme has been designated gap2. This derived protein is more similar to its higher plant Calvin-cycle counterparts (roughly 65% amino acid identity) than it is to any other GAPDH enzyme including the glycolytic enzymes of the eukaryotic cytosol (roughly 48% amino acid identity; Table 1). These data confirm the endosymbiotic origin of nuclear GapA and GapB genes.

We were surprised to find a second type of GAPDH gene in *Anabaena*, gap1, which is more similar to eukaryotic GAPC (Table 1) than to any eubacterial enzyme (except gap1 of *Escherichia coli*; see below). Like its eukaryotic homologues, gap1 of *Anabaena* is an NAD+-specific GAPDH as indicated by the presence of proline-188 in the polypeptide chain, a residue which confers NAD+-specificity on GAPDH homologues (21). gap1 of *Anabaena* is located within an operon which also encodes pyruvate kinase and transaldolase, as revealed by data base searches with two open reading frames immediately 5' to the gap1 gene. Pyruvate kinase and transaldolase are not involved in the Calvin cycle (23) yet are integral to the oxidative pentose phosphate cycle, the major pathway of carbohydrate breakdown in cyanobacteria (24, 25). *Anabaena* 's gap1 is therefore probably involved in carbohydrate catabolism. The third GAPDH gene isolated from *A. variabilis* has been termed gap3.

**Nuclear GAPDH Genes Are of Eubacterial Origin.** The thrust of our interest focused upon gap1 of *Anabaena*. This gene shows a similar degree of identity to eukaryotic GapC genes as *Anabaena* gap2 does to higher plant GapA and GapB (Table 1). Indeed, this finding initially led us to believe that our *A. variabilis* cultures may have been contaminated with some eukaryotic organism which had escaped both microscopic and axenic detection; for this reason, cloning and Southern hybridizations were performed from two independently grown axenic cultures (see Materials and Methods). Eukomility of the three genes was found in both cultures. Additionally, we have isolated and identified by partial sequencing the specific homologues of *A. variabilis* gap1 and gap2 genes from *Synechocystis* PCC6803 (data not shown), so that the possibility of contamination for the origin of *A. variabilis* gap1 clones can be excluded.

Prior to constructing a gene phylogeny, we examined base composition in GAPDH sequences, since fluctuations in G+C content can influence topology (26). The nucleotide composition at third codon positions, which contain about 95% of all synonymous sites (17, 27), varies drastically in GAPDH sequences from 97% G+C in maize GapA to 15% G+C in *Clostridium*. First and second codon positions contain about 90% of all nonsynonymous sites, and only about 2% of first and second codon positions are synonymous (17, 27). At first and second codon positions, G+C content is remarkably constant across the 27 GAPDH genes surveyed here, ranging only from 44% in *Clostridium* to 52% in yeast. Since we constructed dendrograms on the basis of either (i) divergence at nonsynonymous sites or (ii) bootstrap parsimony analysis at first plus second codon positions, G+C bias at third positions in these genes should have virtually no effect upon our topology.

Phylogenetic inference revealed a very complex picture of GAPDH gene evolution (Fig. 1). The most notable result is that the gap1 gene of *Anabaena* is significantly (99/100 bootstrap parsimony replicates) more closely related to eukaryotic GapC than it is to any eubacterial sequence (except gap1 of *E. coli*; see below). In striking analogy to the plastid origin for nuclear GapA genes in higher plants (Table 1), this finding suggests that GapC genes of eukaryotes, in addition to GapA and GapB of plants (5, 11), were laterally transferred from eubacteria to the nucleus early in eukaryotic evolution. This would not only account for the surprising similarity of *A. variabilis* gap1 to eukaryotic GapC genes but would also explain two otherwise puzzling observations previously reported concerning GAPDH gene evolution: (i) the unusual relationship between archaeabacterial GAPDH genes and those of eubacterial/eukaryotic enzymes and (ii) the surprisingly high similarity between *E. coli* gap1 and eukaryotic enzymes (Fig. 1).

**Anomalous Divergence of Archaeabacterial GAPDH Genes.** GapC genes of eukaryotes were previously assumed to have been present in the nucleus prior to any endosymbiotic events, yet GAPDH comparisons between the urkingdoms (34, 35) reveal an anomalous evolutionary behavior for this gene. GAPDH genes from archaeabacteria are quite distinct from both their eubacterial and their eukaryotic homologues. These enzymes share only about 15% identical residues with eubacterial and eukaryotic sequences (34), whereas homologies of the latter two are roughly 45-65% identity (Table 1). This pattern of similarity for GAPDH is precisely the converse of that observed for other conservatively evolving
Fig. 1. Unrooted phylogenetic tree constructed by the neighbor-joining (N-J) method (15) from a matrix of values for nucleotide divergence at nonsynonymous sites, $d_N$ (16). Branch lengths are drawn to scale. Higher plant and cyanobacterial sequences are boxed for clarity. Compartmental localization of the enzyme products (eukaryotes) and genes, as well as pathway involvement and operon structure (prokaryotes) as known, is indicated. Numbers in ovals indicate the number of times out of 100 bootstrap parsimony replicates that the corresponding branch was detected (see Materials and Methods); ovals without numbers indicate those branches which were detected in <18 replicates. Three branches were detected in >50/100 bootstrap replicates which were not detected in the neighbor-joining tree; these were (i) Zymomonas with E. coli gap2 (81/100), (ii) plant GapC sequences with Saccharomyces and Zygosaccharomyces (64/100), and (iii) Ustilago with Aspergillus (58/100). The close relationship between Ustilago, Aspergillus, and metazoan GapC sequences seen in the figure and in ref. 28 was detected in <18/100 bootstrap replicates. The E. coli gap1 sequence is designated gapA in ref. 29; the other 16 enterobacterial sequences reported are born on the E. coli gap1 branch, diverged from one another within about the last 30 million years (29), and were omitted from analysis here for clarity. The leftmost branch of the unrooted tree was simply “bent” here to permit display of species names in a convenient manner. Abbreviations: OPPP, oxidative pentose phosphate pathway; gap, glyceraldehyde-3-phosphate dehydrogenase; pgk: phosphoglycerate kinase; fda: fructose-1,6-bisphosphate aldolase; g6p: glyceraldehyde-3-phosphate; tal: transaldolase. Sources for sequences other than those given in refs. 28 and 30 are as follows: Anabaena gap1, gap2, and gap3 (this paper); tobacco (11); GapA and GapB of pea (5); GapA and GapC of maize (9); E. coli gap2 (31); Clostridium (32); GapC of pea (33). Operon structures are not necessarily complete and show only those reading frames adjacent to gap genes which have been identified. Unidentified reading frames and regions for which sequences are unavailable are not indicated. Nucleotide and amino acid alignments upon which the figure was based are available upon request.

genes (36-41), as schematically summarized in Fig. 2. A number of explanations for this finding have been put forth (28, 30, 34, 46-49), including the suggestion that archaeabacterial GAPDH may not be at all homologous to other GAPDH genes but, rather, functionally converged from a different gene altogether (46, 48). Convergent molecular evolution for GAPDH genes could account for the anomalous relationship of the archaeabacterial homologues, but not for the similarity of gap1 from Anabaena and E. coli to eukaryotic GapC. A eubacterial origin for GapC could account for both findings.

The "Eukaryotic-Like" GAPDH Gene of E. coli. In 1985, a GAPDH gene was isolated from E. coli (50) which, like Anabaena gap1, was more similar to GAPDH of eukaryotes than to eubacterial counterparts (E. coli gap1; Fig. 1). This eukaryotic-like GAPDH gene of E. coli was initially quite puzzling and has been cited (8, 30, 46) as an example for lateral gene transfer from eukaryotes to prokaryotes. A second GAPDH gene was later found in E. coli (gap2 in Fig. 1; ref. 31). Our results place the E. coli gap1 gene in an entirely new light; they strongly suggest that E. coli gap1 is endogenous to this eubacterium and was not acquired from eukaryotes. Clearly, the gene duplication which gave rise to the gap1 and gap2 genes of A. variabilis preceded the separation of the diverse eubacteria surveyed in Fig. 1. Therefore we would expect some eubacteria other than Anabaena to have retained more than one GAPDH gene. Indeed, the eukaryotic-like GAPDH genes found in E. coli and other enterobacteria (29) appear to represent such endogenous descendants of early eubacterial genomes; their similarity to nuclear GapC simply reflects the eubacterial origin of the latter.

Fig. 1 shows a gene phylogeny of duplicated eubacterial sequences, some of which occur today in eukaryotic nuclei, others of which occur in the genomes of free-living eubac-
The Biological Context of GapC Origin: Endosymbiotic Gene Replacement? The genes for Calvin-cycle GAPDH (GapA and GapB) were transferred from the plastid genome to the nucleus, where they became established with a transit peptide and were subjected to proper regulation by the nuclear transcription machinery (5, 8–11). The transfer scenario summarized in Fig. 2 for GapC of glycolytic function would appear to be more or less identical to that for GapA/GapB, except that no transit peptide was necessary to yield a properly compartmentalized, active cytosolic enzyme. Endosymbiotic gene transfer of GapA led to loss of the organellar gene (53), but transfer of GapC appears to have resulted in loss of both the organellar and the endogenous nuclear genes; GapC would thus be the first documented case of eubacterium-to-eukaryote nuclear gene replacement. This working hypothesis for GapC evolution provides us with several testable predictions. Among these, we would expect that further analyses of GAPDH genes of purple bacteria should eventually reveal true orthologues of cyanobacterial gap2. Furthermore, we might expect that in some very primitive eukaryotes, glycolytic GAPDH genes will be found which are not derived from the gap1 type but, rather, from some other member of the eubacterial GAPDH gene family. *Trichomonas vaginalis* appears to possess such a GAPDH gene (54).

The evidence for eubacterial origin of eukaryotic GapC, an essential glycolytic enzyme, suggests that eukaryotic genomes are more highly chimeric than previously assumed. Whereas most organellar proteins are currently encoded in the nucleus, endosymbionts may have donated many genes to the nucleus without organellar reimport of the protein, thereby enriching the genetic and metabolic potential of the host. Those genes which were present in the DNA of both endosymbionts and hosts should be likely candidates for further endosymbiotic gene-replacement events.

We are grateful to W. Ford Doolittle (Halifax) and coworkers for their valuable suggestions on an earlier version of the manuscript, to the Gesellschaft für Biotechnologische Forschung (Braunschweig) for the use of their computer facilities, to Siegfried Scherer (Universität Konstanz, Konstanz, F.R.G.) for the Anabaena inoculum, and to Jean Peltom (Université Joseph Fourier, Grenoble) and Michael Wetter (Technische Universität, Braunschweig) for help in growing the *Anabaena* cultures. This work was funded by grants from the Centre National de la Recherche Scientifique (URA 1178), the Ministère de l’Education Nationale, and the Deutsche Forschungsgemeinschaft (doctoral grant for H.B.).