Curr Genet (1995) 28: 324–332

ORIGINAL PAPER

Yi-Hong Zhou · Mark A. Ragan

The nuclear gene and cDNAs encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase from the marine red alga *Gracilaria verrucosa*: cloning, characterization and phylogenetic analysis

Received: 31 December 1994/15 February 1995

Abstract We have cloned and sequenced the single-copy nuclear gene (*GapC*) encoding the complete 335-amino acid cytosolic glyceraldehyde-3-phosphate dehydrogenase (GAPC) from the red alga *Gracilaria verrucosa*. The proline residue which contributes to the specificity of NAD^+ binding in other GAPC-like proteins is present. Putative regulatory regions, including GC-rich regions, a GATA element, and 11-base T- and T/G-clusters, but excluding TATA- and CCAAT-boxes, were identified upstream. Two types of *GapC* cDNAs differing in polyadenylation site were characterized. An 80-bp phase-two spliceosomal intron was identified in a novel position interrupting the highly conserved cofactor-coding region I. The *G. verrucosa* GAPC was easily aligned with other known GAPC-type sequences. Inferred phylogenetic trees place red algae among the eukaryote crown taxa, although with modest bootstrap support and without stable resolution among related GAPC lineages.

Key words Glyceraldehyde-3-phosphate dehydrogenase · *Gracilaria verrucosa* · Spliceosomal intron · Molecular phylogeny of eukaryotes

Introduction

For many years, red algae were considered to be one of the older groups of eukaryotes, perhaps "intermediate" between cyanobacteria and photosynthetic eukaryotes. This idea, originally based on their cyanobacterium-like pigment composition and the absence of flagella (Ragan and Gutell 1995), appeared to receive support from UPGMA analyses of red algal 5*s* rRNA sequences (Hori and Osawa 1987). However, the basal position of red algae in the 5*s* rRNA tree was eventually recognized as a methodological artifact (Ragan and Gutell 1995), and nuclear-encoded small- and large-subunit rRNA gene sequences clearly indicate that red algae arose comparatively late in eukaryote evolution, at approximately the same time as the divergence of cryptomonads, plants, fungi and animals (Perasso et al. 1989; Bhattacharya et al. 1990; Douglas et al. 1991; Ragan et al. 1994). Analyses of mitochondrial protein-coding gene sequences (Boyen et al. 1994) likewise support a late origin of red algae.

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) catalyses the reversible oxidative phosphorylation of its aldehyde substrate into an acyl phosphate (Harris and Waters 1976; Hardie and Coggins 1986). Gene and deduced amino-acid sequences are known for GAPDHs of eu- and archaebacteria, protists, green plants, animals and fungi. These GAPDHs are clearly homologous, and exhibit high degrees of sequence conservation in regions known to be functionally important. In comparison with other protein-coding genes, GAPDH genes have evolved slowly (5 PAMs per 100 million years; Fothergill-Gilmore and Michels 1993), and thus are potentially useful for the study of organismal phylogeny. In practice, interpretation of GAPDH trees has been complicated by what is now regarded (Martin et al. 1993) as a complex pattern of gene duplication, gene replacement, and selective losses; there are, however, substantial regions within the GAPDH tree which are generally consistent with trees inferred from other molecular sequences (Smith 1989; Doolittle et al. 1990; Michels et al. 1991).

Two kinds of GAPDH isoenzymes exist in investigated green algae and green plants: plastid-localized

GSDB accession L38661
Issued as NRCC no. 38071

Y.-H. Zhou · M.A. Ragan (✉)
Institute for Marine Biosciences, National Research Council of Canada, 1411 Oxford Street, Halifax, Nova Scotia, B3H 3Z1, Canada

Communicated by R.W. Lee

[1]*Present address* Human Genetics Center, University of Texas, P.O. Box 20334, Houston, TX 77225, USA

GAPA (NADP$^+$-GAPDH, EC 1.2.1.13) and cytosolic GAPC (NAD$^+$-GAPDH, EC 1.2.1.12). Genes for both isoforms (*GapA* and *GapC* respectively) are nuclear, although *GapA* is thought to have originated by transfer(s) from endosymbiont(s) during the establishment of plastids (Shih et al. 1986; Brinkmann et al. 1987), and *GapC* may have arisen earlier in the eukaryotic lineage by transfer from an as yet unidentified eubacterium (Markos et al. 1993; Martin et al. 1993). Thus sequence data from GAPDH genes should illuminate both the phylogenetic position of red algae among eukaryotes, and the origin of red-algal plastids. Recently we have reported the sequence of *GapA* cDNAs (Zhou and Ragan 1993) and the single-copy *GapA* gene (Zhou and Ragan 1994) of the florideophycidean marine red algae *Gracilaria verrucosa*, and argued for a single origin of plastids in red algae and green plants. Herein we report the cDNA and gene sequences for *GapC* from *G. verrucosa*, and discuss the position of red algae in phylogenetic trees inferred from GAPDH sequences.

Material and methods

Algal material and libraries. G. verrucosa (Hudson) Papenfuss was collected near Oslo, Norway, by Jan Rueness in December 1984; a sample was obtained from him and cultured at the IMB Aquaculture Research Station, Sandy Cove, Halifax County, N.S. Nuclear DNA and mRNA of *G. verrucosa* were isolated, and λZAP II cDNA and λGEM-11 *Sau*3AI-fragment genomic libraries were constructed as described previously (Zhou and Ragan 1993, 1994). Gene nomenclature follows the recommendations of Cerff et al. (1994).

RT-PCR-mediated generation of an homologous probe. Degenerate oligonucleotides were designed based on an alignment matrix of GAPDH sequences from organisms representing diverse evolutionary lineages, and synthesized on a MilliGen/Biosearch Cyclone Plus instrument. Primers 5′G2 and 3′G1 were based on universally conserved regions corresponding to Cys-156 to Ala-162, and Ser-318 to Trp-324, respectively (Zhou and Ragan 1993). The *GapC*-specific primer 3′GC03 (CATNCCNGTNARYTTNCCRTT) corresponds to the region Asn-230 to Met-236 (numbering according to *Sinapis alba* GAPC; Martin and Cerff 1986). The homologous *G. verrucosa* probe was produced via the polymerase chain reaction (PCR) in two steps using the partially nested primer sets 5′G2/3′G1 and 5′G2/3′GC03 in primary and secondary reactions respectively. The template for the first PCR reaction was single-stranded cDNA (50 ng) that had been generated by poly(T)-primed reverse transcription of *G. verrucosa* mRNA (5 μg) using Superscript II reverse transcriptase (BRL). An aliquot from this reaction was used in the secondary PCR (Table 1), and a product of the expected size was re-amplified and sequenced directly (Bachmann et al. 1990). Identity of the product was confirmed by FASTA (Pearson and Lipman 1988) searches of GenBank. The amplified 240-bp *G. verrucosa GapC* fragment was then labeled with digoxigenin-UTP in PCR (Emanuel 1991) and designated as G6-1.2.

cDNA and gene cloning and genomic Southern hybridization. For cDNA cloning of *GapC*, 87 000 recombinants from the *G. verrucosa* cDNA library were screened using the PCR-generated, digoxigenin-labeled probe G6-1.2 under the same conditions and using the same detection procedures as for *GapA* (Zhou and Ragan 1993). For cloning the *GapC* gene, probe G6-1.2 was initially used to screen the *G. verrucosa* genomic library (180 000 recombinants) by the same procedure. Another screening (480 000 recombinants) was later conducted with an α-^{32}P-random-prime-labeled cDNA fragment from *G. verrucosa GapC* cDNA clone GC6; this 800-bp fragment encodes the *C*-terminal part of the protein, Glu-109 through the last amino acid, and includes the 3′-end non-coding region of the *GapC* mRNA. Hybridization was at medium stringency as in cDNA library screening. The same GC6 cDNA fragment was used as a probe for the Southern hybridization of nuclear DNA.

Generation of a HindIII linking adapter. A *Hin*dIII linking adapter (HLA) was constructed from two partially complementary oligonucleotides, HLP1 (5′GACTAGGTACGAACTAACTG3′) and HLP2 (5′AGCTCAGTTAGTTCGTACCTAGTC3′), with a four-base *Hin*dIII overhang (5′AGCT3′). HLP2 was phosphorylated using T4 polynucleotide kinase (Sambrook et al. 1989). The HLA was generated by mixing 4 μg each of HLP1 and HLP2, warming the mixture to 65 °C, then cooling it slowly to below 30 °C.

Ligation-mediated PCR to recover the GapC unknown region. G. verrucosa nuclear DNA (1 μg) was digested with *Hin*dIII and *Sac*I prior to attachment of the *Hin*dIII adapter. After heat inactivation of the enzymes, DNAs were purified by stepwise phenol/chloroform and chloroform extractions and ethanol precipitation. The *Hin*dIII/*Sac*I nuclear DNA fragments were anchored to the HLA at the *Hind*III end in a 200-μl reaction with purified DNA fragments (1 μg), HLA (10 pmol), ATP (1 mM), 1 x One-Phor-All Buffer Plus (Pharmacia), and T$_4$ DNA ligase (17 units; Pharmacia), at 37 °C for 1 h. Unincorporated HLA was removed by centrifugation using a Centricon-100 filtration device (Amicon). Purified anchored DNA (100 ng) served as template for a first round of PCR with an adapter-specific primer (HLP1) and a gene-specific primer (GCR30, see Fig. 1); the PCR product was used as a template for the second round of PCR with an interior gene-specific primer (GCR32, see Fig. 1) and HLP1 (PCR conditions as in Table I).

PCR-mediated detection of intron positions in GapC. A forward sequencing primer (GCF10 see Fig. 1), annealing to the upstream non-translated region, and two reverse sequencing primers (GCR30, GCR32), annealing to the protein-coding region, were used to PCR-amplify the corresponding cDNA fragment from single-stranded cDNAs of *G. verrucosa*. A stepwise PCR with partially nested primer sets GCF10/GCR30 and GCF10/GCR32 was performed following the secondary reaction conditions described in Table 1. A 540-bp PCR fragment of the expected size was extracted from the agarose

Table 1 Reaction conditions for the amplification of *GapC* sequences

Reaction	Primer pairs	Denaturing step	Annealing step	Extension step	Number of cycles
Primary	5′ G2 3′ G1	94 °C 20 s	40 °C 30 s	72 °C 1 min	40
Secondary	5′ G2 3′ GC03	94 °C 30 s	46 °C 30 s	72 °C 30 s	34

gel using the GENECLEAN II kit (Bio 101). Sequencing was carried out on an ABI automated DNA sequencer, following the manufacturer's protocol for fluorescent-labeled dideoxynucleotides, using GCF10 and GCR32 as primers.

Sequence alignment and phylogenetic analyses. GAPDH protein sequences were aligned using MULTALIN (Corpet 1988) with the parameters of Risler et al. (1988) at gap penalty 50, with subsequent adjustment by eye. A matrix of aligned DNA sequences was produced based on this protein alignment. Trees were inferred using PHYLIP version 3.53c (Felsenstein 1989) on a Sun 10/30 workstation, unless otherwise stated. Protein and DNA parsimony trees were inferred using PROTPARS and DNAPARS respectively, with 500 jumbles (attempts to find the globally optimal result) for non-bootstrapped analyses and at least 100 bootstrap replicates. Weighted parsimony analysis (50 jumbles) was conducted using PAUP version 3.1 (Swofford 1993) on a Macintosh IIsi, with a weighting inversely proportional to the variability at each nucleotide position and the contribution of each position to tree length kept equal. Protein and DNA distance matrixes were calculated using PROTDIST and DNADIST respectively, under Kimura's model for amino-acid substitution and Felsenstein's generalized two-parameter ("maximum likelihood") model for nucleotide substitution. Only first and second codon positions were considered in distance analyses. Neighbor-joining trees were inferred using NEIGHBOR. Edge lengths were re-fitted onto the consensus bootstrapped neighbor-joining tree by a least-squares approach (FITCH). Confidence intervals were estimated by bootstrapping using SEQBOOT, the tree-inference program listed above, and CONSENSE.

Results and discussion

Homologous *GapC* probe for library screening

A DNA fragment (G6-1.2) of the expected size in the absence of introns (240 bp) was generated from single-stranded *G. verrucosa* cDNA by stepwise PCR using degenerate primers corresponding to regions conserved universally (in both *GapC* and *GapA/B*) and specifically (in *GapC* only) in GAPDH genes. DNA sequencing and subsequent FASTA searching confirmed that this fragment had been amplified from a cytosolic GAPDH (*GapC*) cDNA, as evidenced by the higher degree of identity with *GapC* than *GapA/B* sequences of other organisms, and especially by the presence of the *GapC*-specific proline (Pro-188 in *Bacillus stearothermophilus*; Corbier et al. 1990).

cDNA cloning and characterization of *GapC*

Four putative *GapC* cDNA clones (GpC1, GpC2, GpC6, GpC7) were isolated from 87000 phage recombinants using the homologous *GapC*-specific probe G6-1.2. Although these cDNAs were of different lengths due to incomplete first-strand cDNA synthesis, the *GapC* sequence present was identical among all four clones and identical to that of the PCR-generated *GapC* probe. Two of these transcripts (clones GpC1 and GpC2) were 9 bp longer than the other two (clones GpC6 and GpC7) at the 3′ end owing to polyadenylation at different sites in vivo (Fig. 1).

The canonical polyadenylation signal AATAAA, earlier shown to be absent from the *GapA* cDNAs of *G. verrucosa* (Zhou and Ragan 1993), was similarly absent from the *GapC* cDNAs. A probable polyadenylation signal, GATAAA, was found 15 bases from the end of clones GpC6 and GpC7, and 24 bases from the end of clones GpC1 and GpC2 (Fig. 1).

G. verrucosa GAPC is encoded by a single-copy gene

The *GapC* cDNA from clone GpC6 was used to probe a Southern blot of *G. verrucosa* nuclear DNA. Only one band was detected in various digests at medium stringency (Fig. 2). This result is consistent with the cDNA characterizations reported above, and confirms that GAPC is encoded by a single-copy gene in *G. verrucosa.*

Cloning and characterization of the *GapC* gene

Two approaches were taken to recover full-length *GapC* genomic clones from the *G. verrucosa* library. First, by screening with probe G6-1.2 we recovered one *GapC* clone from 180000 recombinants. From this clone, a 2.4-kb *Sac*I fragment (with one *Sac*I site on the λGEM-11 vector) was subcloned into pUC18 and named pNGC1S (Fig. 3 A). Sequence analysis revealed that this insert includes the 3′ end of the gene downstream from the codon for Asp-143.

During genomic Southern hybridization (Fig. 2, lane 3 and Fig. 3 B) the sequence corresponding to *GapC* had been localized to a 3.0-kb *Sac*I/*Hin*dIII fragment. Because the gene is single-copy and the *Sac*I site lies 2.4 kb downstream from the codon for Asp-143, the *Hin*dIII end must lie about 0.6 kb upstream from the Asp-143 codon, i.e., this 0.6-kb region should encompass the entire 5′ end of the *GapC* coding region if no intron is present in the 5′ end of this gene. This *Hin*dIII site was selected for anchoring an *Hin*dIII adapter in order to PCR-amplify the 5′ end of *GapC*. Using this adapter, a PCR product of approximately the expected size was generated (data not shown), and direct sequencing confirmed that it corresponded to *GapC* of *G. verrucosa.*

However, this PCR product did not contain the complete 5′ end of the *GapC* coding sequence, owing to the presence of an insertion sequence interrupting the codon for Ile-14. This insertion sequence was recognized by the appearance of a non-*GapC* sequence in the highly conserved region corresponding to GAPDH cofactor region I. The predicted *Hin*dIII site for anchoring the adapter was located within the insertion sequence (Fig. 1). This was unexpected, as no insertion sequence had been found in the only other known red algal *GapC*, that of the related florideophyte *Chondrus crispus* (Liaud et al. 1994).

```
-300 GC box         GC box                            GC box                T cluster    T/G cluster
ATACCGCCGCTCAACAACTTCCCCGCCGTCGTCCGTTTTTCTGTGCTCGTTTTCTCGCCCCGTCGTGCTGCTGTTTTTTTTTTGCTTGTGTTGTTTCAA    -201

           GATA box                       GC box          GC box        GCF11 primer
ACGGTATGTAAGATACGCGTGCGCTGTGTTTGGACGCTGGCCGCCGTTGTGCATCGCGGGTGAACGCCATCATGTTCTCCCTGTTTTCGACCCGATCTGC    -101

          GC box      Pyrimidine box          GC box                          GC box << GCF10 primer
ACGCGATCGCGCCCCGGCCTTTGTTTCCTCGCCGTCCTTCTTTGCTAACCGCCCCACTCTCCGTTCCAATCGCTGCAGTACCCCCCGTTCCTCAATCATC      -1

+1                                                                           HindIII
ATGACTGTGCCGCAGGTTGGTATCAATGGCTTCGGTCGAATgtaagtagcgtctcaatcgaaatcgcccattagtcaagcttgatccacccatgctaact     100
M  T  V  P  Q  V  G  I  N  G  F  G  R  I                                                                  14

gtttgtgttgcgtttctgtagTGGCCGTCTCGTGCTCCGTGCTGCTATTGAGAAGGATACCATGAGCGTGGTCGCCATCAATGACCCCTTCATCGATCTG     200
                       G  R  L  V  L  R  A  A  I  E  K  D  T  M  S  V  V  A  I  N  D  P  F  I  D  L       40
                                                                          HindIII
GAGTACATGGCGTACATGTTCAAGTTCGACTCTACGCACGGTCGTTACGCAGGTTCCGTTGAGACCAAGGACGGAAAGCTTATCGTCAACGGCAAGTCCA     300
 E  Y  M  A  Y  M  F  K  F  D  S  T  H  G  R  Y  A  G  S  V  E  T  K  D  G  K  L  I  V  N  G  K  S        73
                                                                           < GpC7
TTACCATCTACGGACACCGCGATCCGGCTGAGATCCCGTGGGCCGAAGCCGGTGCCGACTATGTGGTCGAATCTACCGGTGTGTTCACTCTCAAGGAGAA     400
I  T  I  Y  G  H  R  D  P  A  E  I  P  W  A  E  A  G  A  D  Y  V  V  E  S  T  G  V  F  T  L  K  E  K      107
    < GpC6                                      < GpC1
GGCCGAGAAGCATTTCACGGGAGGTGCCAAGAAGGTGATCATCTCTGCGCCATCGAAGGATGCGCCTATGTTTGTGTGCGGTGTGAACGAGGACAAGTAC     500
  A  E  K  H  F  T  G  G  A  K  K  V  I  I  S  A  P  S  K  D  A  P  M  F  V  C  G  V  N  E  D  K  Y       140
                                                                                             < GpC2
ACGCCGGATCTCAACGTGATTTCCAATGCGTCCTGTACCACCAACTGCCTGGCGCCTTTGGTGAAGGTCATTCACGAGAAGTATGGTATTGAGGAAGGTT     600
 T  P  D  L  N  V  I  S  N  A  S  C  T  T  N  C  L  A  P  L  V  K  V  I  H  E  K  Y  G  I  E  E  G       173
           GCR32 primer   >>
TGATGACCACAGTGCATGCCACTACCGCTACGCAGAAGACTGTGGACGGACCGTCGCAGAAGGACTGGCGTGGCGGACGTGGAGCTGGCGCTAACATTAT     700
L  M  T  T  V  H  A  T  T  A  T  Q  K  T  V  D  G  P  S  Q  K  D  W  R  G  G  R  G  A  G  A  N  I  I      207

TCCGTCCAGCACAGGCGCTGCAAAGGCTGTTGGTAAGGTGCTGCCCGAGCTGAACGGAAAGTTGACGGGAATGGCATTCCGTGTTCCCACGTCCGATGTT     800
  P  S  S  T  G  A  A  K  A  V  G  K  V  L  P  E  L  N  G  K  L  T  G  M  A  F  R  V  P  T  S  D  V       240

TCTGTGGTTGACTTGACTGTGCGCCTTGCAACTGAAACTAGCTATGACGATATCAAGGCCACCATGAAGGCTGCGGCTGAGGACTCTATGAAGGGCATCT     900
 S  V  V  D  L  T  V  R  L  A  T  E  T  S  Y  D  D  I  K  A  T  M  K  A  A  A  E  D  S  M  K  G  I       273
                GCR30 primer
TGAAGTACACTGAAGAGGCTGTGGTGAGCACTGACTTTATTCACGAGGAGGCGTCTTGTGTGTTTGATGCGGGAGCCGGTATCATGTTGAATAGCAGGTT    1000
L  K  Y  T  E  E  A  V  V  S  T  D  F  I  H  E  E  A  S  C  V  F  D  A  G  A  G  I  M  L  N  S  R  F      307

TTGCAAGCTGGTGGCTTGGTACGATAATGAGTGGGGATACTCCAACCGCGTTGTGGACCTCATCGCTCATGTTGCCAAGTTGCAGTGAGTCCCACTCTAA    1100
  C  K  L  V  A  W  Y  D  N  E  W  G  Y  S  N  R  V  V  D  L  I  A  H  V  A  K  L  Q  *                   335
                                                                   GpC6, GpC7 >
ACGTCAAGTAGATCGTTTGATATCTTGCTAGGTTTTTTTTATACGACTTGTAGCCGGATAAACATCGGTTAATTTGTGATTTTGTTCAACTTTTCGTTG    1200
                                                         ------
                                                                            GPC1, GpC2 >
AATCCACATGAGATTGGCGGCTGTGCAGACAAGCCGTCTGCACGTCCGCTTGTTCTGGCGATGGTTGGCGCGAGGGGCGGCGGATCAACAGCGAATCGGA    1300

CACGTGCAGCGGACCAATCGCTGTGCGTGCTGTGTGAATATCTCAGCCAGCGCGC                                                 1355
```

Fig. 1 Nucleotide sequence of the *G. verrucosa GapC* gene with upstream and downstream non-transcribed regions, and the derived GAPC translation product. The cloned cDNA regions are shown by *two closed arrowheads* with the names of the respective clones; the PCR-recovered cDNA sequence is in *two closed double arrowheads.* Numbering makes reference to the translation initiation site (+ 1). The intron is in *lower case*; The *Hin*dIII sites are in *italics*; primers selected for PCR of cDNA are *underlined.* Upstream, putative transcriptional *cis*-acting elements are marked. Downstream from the protein-encoding region, functional polyadenylation sites are in *bold*, and the polydenylation signal AGTAAA is shown by *dashed underlines*

Surprisingly, a second *Hin*dIII site (Fig. 3 B, boxed) was found within this amplified region. It was not cleaved by *Hin*dIII digestion of nuclear DNA (e.g., in digestions for genomic Southern hybridizations), implying that it was probably methylated, presumably at the C residue. Consistent with this interpretation, both *Hin*dIII sites can be readily cleaved in the cloned *GapC* sequence (data not shown). DNA methylation within expressed coding sequences is not usual in higher eukaryotes (Bestor 1990; Bird 1992; Zhu et al. 1994).

A second approach was therefore taken to recover the complete *GapC* gene. The *G. verrucosa* genomic library (480000 recombinants) was screened using an α-^{32}P-labeled *GapC* cDNA (GpC6), and four λ clones were recovered. Restriction–enzyme digestion and Southern hybridization indicated that, as expected, the same gene was present in all four clones. Restriction analysis revealed that all four of these clones contained the complete *GapC* gene; one (NGCe) was subcloned into pUC18 (NGCeSl, Fig. 3 A) and sequenced. The complete *GapC* sequence, including a 1068-bp protein-coding region (including the insertion sequence) and upstream and downstream non-coding regions, is shown in Fig. 1.

Although the region upstream of position − 300 was present in clone NGCeSl, the presence of a long G + C-rich region at this point prevented us from obtaining sequence data of acceptable quality. Repeated attempts with various oligonucleotide primers annealing to various positions were likewise unsuccessful. We found no putative TATA-box or CCAAT-box in this 300-bp region (Fig. 1). However, eight GC-rich regions, each 7–12 bp in length, extend upstream from positions −14, −46, −84, −141, −155, −239, −273 and −291. A 26-bp pyrimidine-rich (88% C + T) region, similar to the pyrimidine boxes seen in regions up-

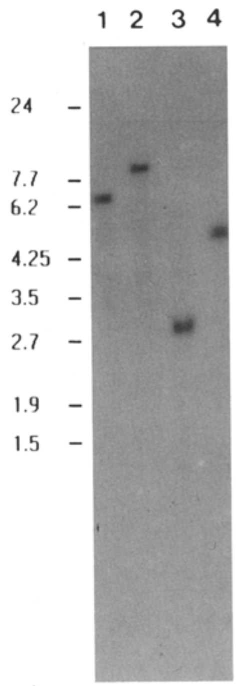

Fig. 2 Southern hybridization of the *G. verrucosa GapC* cDNA to *G. verrucosa* nuclear DNA. Double digestions were applied: *Sac*I/*Eco*RI (*lane 1*), *Sac*I/*Xba*I (*lane 2*), *Sac*I/*Hin*dIII (*lane 3*), *Eco*RI/*Hin*dIII (*lane 4*). The scale to the left is in kilobase pairs

stream from *G. verrucosa GapA*, *UBI6R* and *m-ACN* (Zhou and Ragan 1995), occurs 58 bp upstream from the translational start codon. An apparent GATA element, similar to the GATA box conserved among several light-responsive promoters of plants and CaMV 35S and promoters of animal globin genes (Lam and Chua 1989; Schindler and Cashmore 1990), occurs at position − 186. A striking feature of this region is the existence of an 11-base T-cluster and an 11-base T + G cluster, separated by two nucleotides, extending upstream from positions −217 and −204. The possible involvement of these elements in transcriptional control of *G. verrucosa GapC* requires experimental verification.

PCR-mediated detection of intron position and intron/exon organization in *GapC*

In order to determine whether the insertion sequence (above) is an intron, and to characterize fully the intron/exon structure of *GapC*, it was necessary to compare the gene (Fig. 1) with corresponding cDNAs. As described above, all four cloned *G. verrucosa GapC* cDNAs were incomplete, missing about a third of the gene sequence from their 5′ ends. The 5′ end of the *G. verrucosa GapC* cDNA sequence was recovered by PCR, with single-stranded cDNAs as templates and primer sets calculated from the 5′ end of *GapC*. No PCR product was observed when the GCF11/GCR30 primer set was used, suggesting that the GCF11 primer may be located in the 5′ non-transcribed region. Use of the GCF10/GCR30 primer set yielded an 0.9-kb fragment, as expected, and a second round of PCR with primer set GCF10/GCR32 yielded an 0.6-kb product. Direct sequencing of this product and comparison with the *GapC* sequence revealed the insertion sequence to be an 80-bp phase-two intron located, as expected from the results above, in the codon for Ile-14 of the highly conserved cofactor region I (INGFGRIGR; Fig. 1). This intron, the only one in *G. verrucosa GapC*, is of the GT-AG (putatively spliceosomal) type, and its neighbouring nucleotides conform to the higher-plant splice-junction consensus. Further details on red algal intron structure will be presented elsewhere (Zhou and Ragan 1995).

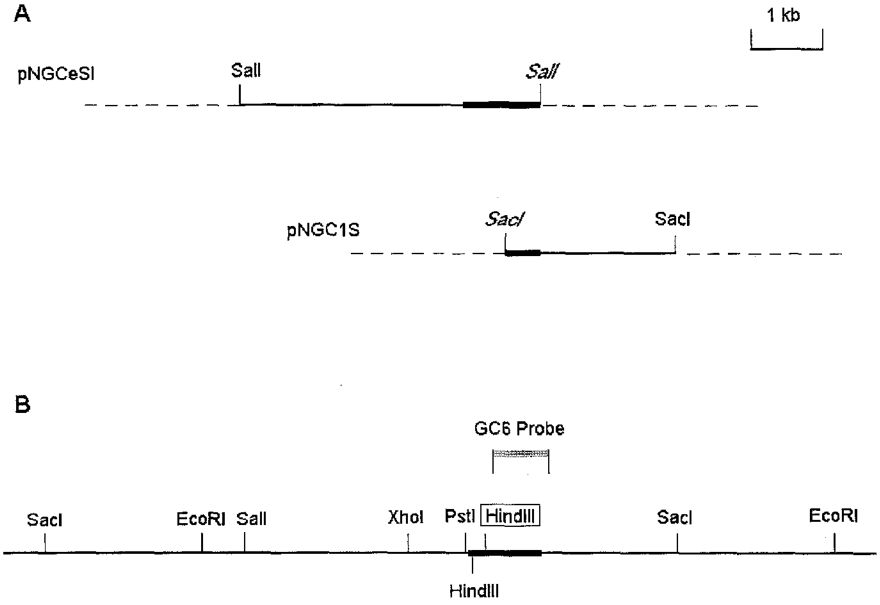

Fig. 3 A Depiction of the *GapC* plasmid clones pGC1S and pGCeSI corresponding to the restriction map above. *Thick lines* represent the *GapC* coding regions (exons and intron); *thin lines* represent the non-coding regions; *dashed* lines represent the pUC18 regions. The restriction-enzyme sites in the vector of the phage clone λNGCe are shown in *italics*. **B** restriction map of the *GapC* gene locus of *G. verrucosa*; the GC6 probe for Southern hybridization is shown corresponding to the restriction map beneath

	6		8		10		12		14		16		18	
	V	G	I	N	G	F	G	R	I	G	R	L	V	L
GapC Gracilaria verrucosa									⇑					
GapC Chlamydomonas reinhardtii					↑									
Gapc Arabidopsis thaliana					↑									
GapC1 Pisum sativum					↑									
GapC1 Zea mays					↑									
GapC4 Zea mays					↑									
Homo sapiens					⇑									
Gallus gallus					⇑									
Schizophyllum commune							↑							
Phanerochaete chrysosporium					⇑		↑							
Agaricus bisporus					⇑		↑							
Ustilago maydis					⇑									
Aspergillus nidulans	Δ												Δ	
Cochliobolus heterostrophus	Δ													
Curvularia lunata	Δ													

Δ phase-0 intron; ↑ phase-1 intron; ⇑ phase-2 intron

Fig. 4 Depiction of intron positions identified in and around the cofactor-I region of GAPDH (in *boldface*). The sequence and the numbering follow the *G. verrucosa* GAPC. Sources of sequences of basidiomycetes: *Ustilago maydis* (Smith and Leong 1990), *Agaricus bisporus, Phanerochaete chrysosporium, Schizophyllum commune* (Harmsen et al 1992); ascomycetes: *Aspergillus nidulans* (Punt et al. 1988), *Cochliobolus heterostrophus* (Van Wert and Yoder 1992), *Curvularia lunata* (Osiewacz and Ridder 1991); animals: *Homo sapiens* (Ercolani et al. 1988); *Gallus gallus* (Stone et al. 1985); *GapC*: *Arabidopsis thaliana* (Shih et al. 1991), *Pisum sativum* (Brinkmann et al. 1989), *Zea mays* (Martinez et al. 1989), *Chlamydomonas reinhardtii* (Kersanach et al. 1994)

Alignment of known GAPDH sequences and positional mapping of introns occurring in or near the conserved cofactor I region (Fig. 4) reveals the *G. verrucosa GapC* intron to occur in a novel position. The closest known introns, phase-one introns in *GapC* of three basidiomycetes, are located two codons (7 bp) upstream in Gly-12. The phase-two introns closest to the *G. verrucosa GapC* intron are conserved between animals and fungi, and are located four amino acids (12 bp) upstream in the Gly-10 codon.

One is impressed by the multiplicity of intron positions in this small, highly conserved region. The novel position of the intron in *G. verrucosa GapC*, and its absence from *GapC* of the phylogenetically related (Ragan et al. 1994) florideophyte *Chondrus crispus* (Liaud et al. 1994) and other eukaryotes (Fig. 4), are readily explained by only one of the three major hypotheses of intron origin, namely specific (late) insertion (Cavalier-Smith 1985; Logsdon and Palmer 1994). Until GAPC genes of closely related species are examined, we cannot assess whether such insertion was specific to *G. verrucosa*, to the monophyletic (Bird et al. 1992) genus *Gracilaria*, to the family Gracilariaceae or to the order Gracilariales. As no other organism is known to have an intron in this position, an ancestral derivation of this intron (Darnell 1978; Doolittle 1978; Gilbert 1978) does not appear likely. An origin from a differently positioned *GapC* intron by intron sliding (Craik et al. 1983; Patthy 1987) is improbable owing to the high degree of sequence conservation, hence functional importance, of the GAPDH cofactor-I region among animals, fungi, green plants and red algae. Even a single amino-acid shift at one boundary of an intron in this region could alter the function of the domain, leading to decreased organismal fitness, unless a compensatory shift occurred simultaneously at the other splice boundary in such as way as not to alter the amino-acid sequence. As this constraint seems severe in a low-copy-number gene, intron sliding through such a highly conservative region would appear to be unlikely.

The position of red algae in the GAPDH tree

Phylogenetic trees were inferred from an alignment of 36 complete glycolytic and plastid-specific GAPDH sequences, representing divergent organismal lineages, by both distance (neighbor-joining) and parsimony methods. The trees were rooted by the method of Iwabe et al. (1989) on the assumption that *GapA/B* and *GapC* arose by duplication of an ancestral GAPDH gene.

Because it is not computationally feasible to conduct an exhaustive search among all possible trees relating 36 species, multiple heuristic searches were conducted based on random input orders of the sequences. Six equally parsimonius trees of 2728 steps each were recovered by unweighted parsimony analysis of the 335-position GAPDH protein alignment. GAPCs of *G. verrucosa* and *C. crispus* group together specifically in all six trees; GAPCs of red algae and green plants appear as sister groups in five of the six trees, and GAPCs of animals, filamentous fungi, green plants and red algae form a monophyletic group in four of the trees (see Fig. 5 of Ragan and Gutell 1995). These results agree with and extend those of Liaud et al. (1994), and support the inclusion of red algae among the late-branching "eukaryote crown taxa" (Ragan and Gutell 1995).

To attempt to correct for possible biases introduced by the high variability among substitution rates at different positions within GAPDH genes (Lawrence et al. 1991), we repeated the analysis with inverse weighting. The resulting tree (Fig. 5) is essentially identical to that produced by unweighted parsimony analysis, although bootstrap support is weak for the red alga-green plant and crown-group clades. The only significant topological difference between the unweighted and weighted parsimony trees involves the GAPDH of *Chlamydomonas reinhardtii*, which diverges basally within the GAPC subtree in the unweighted analysis but appears (albeit with low bootstrap support) as a sister group to red algal GAPC after inverse weighting. GAPDHs of the ascomycetous yeasts *Saccharomyces cerevisiae* and *Zygosaccharomyces rouxii* diverge unexpectedly deeply in both unweighted and weighted parsimony trees (compared with classical

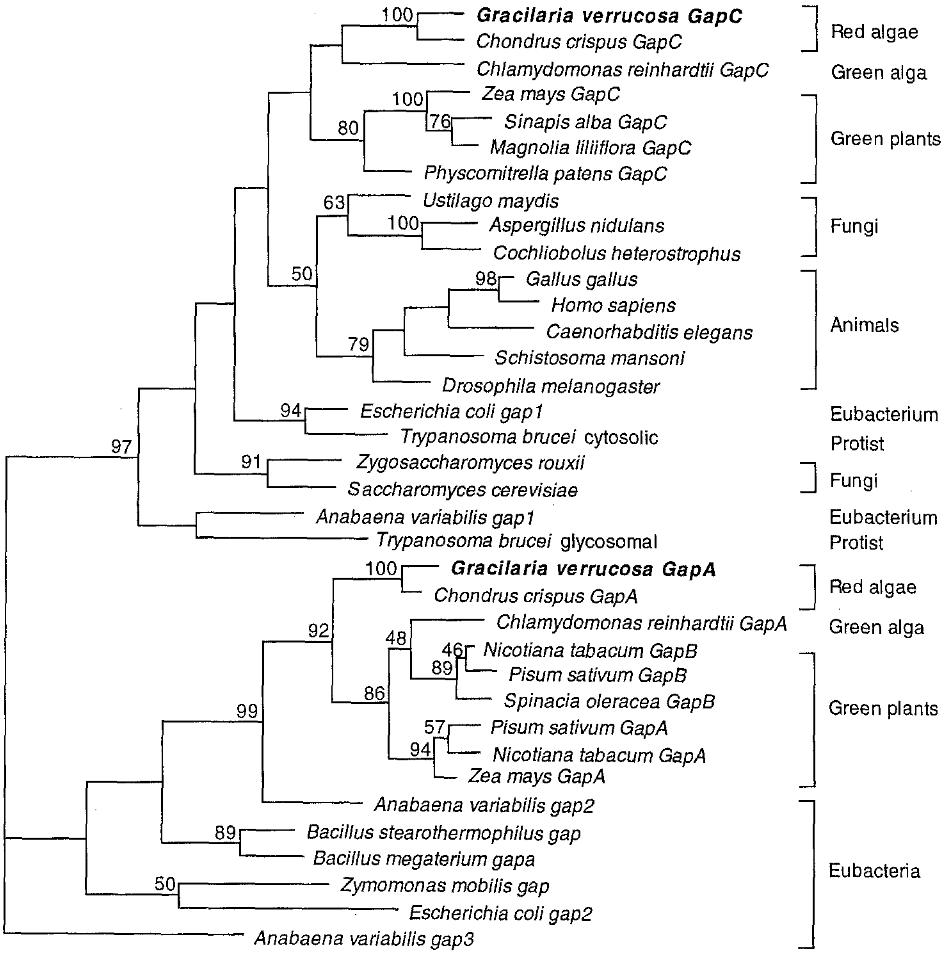

Fig. 5 Inverse-weighted GAPDH-protein maximum parsimony tree with *A. variabilis gap3* as the outgroup. The bootstrap values above 50 (100 replicates) are shown

systematics or small-subunit rRNA-gene trees, e.g., Cavalier-Smith et al. 1994; Ragan and Gutell 1995). We interpret these results as indicating that *C. reinhardtii* GAPC is orthologous with the other GAPCs but has accepted amino-acid replacement at an accelerated rate, whereas the two yeast GAPCs may be paralogous with the GAPCs of other eukaryotes.

The bootstrap is a blunt instrument in respect of phylogenetic inference, highly imprecise (although unbiased) as a measure of the repeatedly of a given result but conservative (although biased) in estimating the probability of correctly inferring a grouping (Hillis and Bull 1993). Thus we examined the costs of alternative topologies among red algae, green plants, animals and (non-yeast) fungi in both GAPDH protein and gene trees. Several alternative branching orders among these taxa cost only one step more than the most-parsimonious tree (Fig. 5). However, forcing the red algal *gapC* to diverge just basally to the other crown taxa increases tree length by four steps, while forcing it to group specifically with the next most-basal branch (*Trypanosoma* cytosolic *gap* and *Escherichia coli gap 1*) costs 11 additional steps. Neighbor-joining inference at both protein and DNA levels (data not shown) yielded poorly resolved trees in which all eukaryote cytosolic GAPC lineages, together with the *E. coli* gapl protein but excluding the perhaps paralogous GAPCs from two yeasts (above), diverge almost simultaneously. With its modest information content and probable paralogous lineages, GAPDH is not an ideal protein for phylogenetic inference; however, based on an analysis of GAPDH genes and proteins, red algae may provisionally be placed among or near the eukaryote crown taxa.

In all trees (e.g., Fig. 5) the plastid-localized GAPA sequences of *G. verrucosa* and *C. crispus* group together and, as reported for *G. verrucosa* GAPA alone (Zhou and Ragan 1993), form a sister group with green-plant plastid-localized GAPA/B sequences. These results are consistent with a single endosymbiotic origin of plastids in red algae and green plants (Zhou and Ragan 1994).

References

Bachmann B, Lüke W, Hunsmann G (1990) Improvement of PCR-amplified DNA sequencing with the aid of detergents. Nucleic Acids Res 18:1309

Bestor TH (1990) DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. Phil Trans R Soc Lond B 326:179–188

Bhattacharya D, Elwood HJ, Goff LJ, Sogin ML (1990) Phylogeny of *Gracilaria lemaneiformis* (Rhodophyta) based on sequence analysis of its small-subunit ribosomal RNA coding region. J Phycol 26:181–186

Bird A (1992) The essentials of DNA methylation. Cell 70:5–8

Bird CJ, Rice EL, Murphy CA, Ragan MA (1992) Phylogenetic relationships in the Gracilariales (Rhodophyta) as determined by 18 rDNA sequences. Phycologia 31:510–522

Boyen C, Leblanc C, Bonnard G, Grienenberger J-M, Kloareg B (1994) Nucleotide sequence of the *cox3* gene from *Chondrus crispus*: evidence that UGA encodes tryptophan and evolutionary implications. Nucleic Acids Res 22:1400–1403

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

Brinkmann H, Cerff R, Salomon M, Soll J (1989) Cloning and sequence analysis of cDNAs encoding the cytosolic precursors of subunits GapA and GapB of chloroplast glyceraldehyde-3-phosphate dehydrogenase from pea and spinach. Plant Mol Biol 13:81–94

Cavalier-Smith T (1985) Selfish DNA and the origin of introns. Nature 315:283–284

Cavalier-Smith T, Allsopp MTEP, Chao EE (1994) Chimeric conundra: are nucleomorphs and chromists monophyletic or polyphyletic? Proc Natl Acad Sci USA 91:11368–11372

Cerff R, Bohnert HJ, Ragan M, Sachs MM (1994) Plant-wide nomenclature of nuclear genes encoding cytosolic and chloroplast glyceraldehyde-3-phosphate dehydrogenases. Pl Mol Biol Rep 12:S36–S37

Corbier C, Clermont S, Billard P, Skarzunski T, Branlant C, Wonacott A, Branlant G (1990) Probing the coenzyme specificity of glyceraldehyde-3-phosphate dehydrogenases by site-directed mutagenesis. Biochemistry 29:7101–7106

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16:10881–10890

Craik CS, Rutter WJ, Fletterick R (1983) Splice junctions:association with variation in protein structure. Science 220: 1125–1129

Darnell JE (1978) Implications of RNA·RNA splicing in evolution of eukaryotic cells. Science 202:1257–1260.

Doolittle RF, Feng DF, Anderson KL, Alberro MR (1990) A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J Mol Evol 31:383–388

Doolittle WF (1978) Genes in pieces: were they ever together? Nature 272:581–582

Douglas SE, Murphy CA, Spencer DF, Gray MW (1991) Cryptomonad algae are evolutionary chimaeras of two phylogenetically-distinct unicellular eukaryotes. Nature 350:148–151

Emanuel JR (1991) Simple and efficient system for synthesis of non-radioactive nucleic acid-hybridization probes using PCR. Nucleic Acids Res 19:2790

Ercolani L, Florence B, Denaro M, Alexander M (1988) Isolation and complete sequence of a functional human glyceraldehyde-3-phosphate dehydrogenase gene. J Biol Chem 263:15335–15341

Felsentein J (1989) PHYLIP – phylogeny inference package (version 3.2). Cladistics 5:164–166

Fothergill-Gilmore LA, Michels PAM (1993) Evolution of glycolysis. Prog Biophys Mol Biol 59:105–235

Gilbert W (1978) Why genes in pieces? Nature 271:501

Hardie DG, Coggins JR (1986) Multidomain proteins – structure and evolution. Elsevier Science Publishers (Biomedical Division), Amsterdam

Harmsen MC, Schuren FHJ, Moukhaa SM, van Zuilen CM, Punt PJ, Wessels JGH (1992) Sequence analysis of the glyceraldehyde-3-phosphate dehydrogenase genes from the basidiomycetes *Schizophyllum commune*, *Phanerochaete chrysosporium* and *Agaricus bisporus*. Curr Genet 22:447–454

Harris JI, Waters M (1976) Glyceraldehyde-3-phosphate dehydrogenase. In: Boyer PD (ed) The enzymes, vol 13. Academic Press, New York, pp 1–14

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biol 42:182–192

Hori H, Osawa S (1987) Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. Mol Biol Evol 4:445–472

Iwabe N, Kuma KI, Hasami M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

Kersanach R, Brinkmann H, Liaud M-F, Zhang DX, Martin W, Cerff R (1994) Five identical intron positions in ancient duplicated genes of eubacterial origin. Nature 367:387–389

Lam E, Chua NH (1989) ASF-2: a factor that binds to the cauliflower mosaic virus 35S promoter and a conserved GATA motif in *Cab* promoters. Plant Cell 1:1147–1156

Lawrence JG, Hartl DL, Ochman H (1991) Molecular considerations in the evolution of bacterial genes. J Mol Evol 33:241–250

Liaud M-F, Valentin C, Martin W, Bouget F-Y, Kloareg B, Cerff R (1994) The evolutionary origin of red algae as deduced from the nuclear genes encoding cytosolic chloroplast glyceraldehyde-3-phosphate dehydrogenases from *Chondrus crispus*. J Mol Evol 38:319–327

Logsdon Jr JM, Palmer JD (1994) Origin of introns – early or late? Nature 369:526

Markos A, Miretsky A, Müller M (1993) A glyceraldehyde-3-phosphate dehydrogenase with eubacterial features in the amitochondriate eukaryote, *Trichomonas vaginalis*. J Mol Evol 37:631–643

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

Martin W, Brinkmann H, Savonna C, Cerff R (1993) Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA 90:8692–8696

Martinez P, Martin W, Cerff R (1989) Structure, evolution and anaerobic regulation of a nuclear gene encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase from maize. J Mol Biol 208:551–565

Michels PAM, Marchand M, Kohl L, Allert S, Wierenga RK, Opperdoes FR (1991) The cytosolic and glycosomal isoenzymes of glyceraldehyde-3-phosphate dehydrogenase in *Trypanosoma brucei* have a distant evolutionary relationship. Eur J Biochem 198:421–428

Osiewacz HD, Ridder R (1991) Genome analysis of imperfect fungi: electrophoretic karyotyping and characterization of the nuclear gene coding for glyceraldehyde-3-phosphate dehydrogenase (*gpd*) of *Curvularia lunata*. Curr Genet 20:151–155

Patthy L (1987) Intron-dependent evolution: preferred types of exons and introns. FEBS Lett 214:1–7

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Perasso R, Baroin A, Qu LH, Bachellerie JP, Adoutte A (1989) Origin of the algae. Nature 339:142–144

Punt PJ, Dingemanse MA, Jacobs-Meijsing BJ, Pouwels PH, van den Hondel CA (1988) Isolation and characterization of the glyceraldehyde-3-phosphate dehydrogenase gene of *Aspergillus nidulans*. Gene 69:49–57

Ragan MA, Gutell RR (1995) Are red algae plants? Bot J Linn Soc (in press)

Ragan MA, Bird CJ, Rice EL, Gutell RR, Murphy CA, Singh RK (1994) A molecular phylogeny of the marine red algae (Rhodophta) based on the nuclear small-subunit rRNA gene. Proc Natl Acad Sci USA 91:7276–7280

Risler JL, Delacroix H, Henaut A (1988) Amino-acid substitution in structurally related proteins (a pattern recognition approach): determination of a new and efficient scoring matrix. J Mol Biol 204:1019–102!9

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Schindler U, Cashmore AR (1990) Photoregulated gene expression may involve ubiquitous DNA-binding proteins. EMBO J 9:3415–3427

Shih M-C, Lazar G, Goodman HM (1986) Evidence in favor of the symbiotic origin of chloroplasts: primary structure and evolution of tobacco glyceraldehyde-3-phosphate dehydrogenase. Cell 47:73–80

Shih M-C, Heinrich P, Goodman HC (1991) Cloning and chromosomal mapping of nuclear genes encoding chloroplast and cytosoloic glyceraldehyde-3-phosphate dehydrogenase from *Arabidopsis thaliana*. Gene 104:133–138

Smith TL (1989) Disparate evolution of yeasts and filamentous fungi indicated by phylogenetic analysis of glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA 86:7063–7066

Smith TL, Leong SA (1990) Isolation and characterization of a *Ustilago maydis* glyceraldehyde-3-phosphate dehydrogenase-encoding gene. Gene 93:111–117

Stone EM, Rothblum KN, Schwartz RJ (1985) Intron-dependent evolution of the chicken glyceraldehyde phosphate dehydrogenase gene. Nature 313:498–500

Swofford DL (1993) PAUP: phylogenetic analysis using parsimony, Version 3.1. (Computer program) The Illinois National History Survey, Champaign, Illinois

Van Wert SL, Yoder OC (1992) Structure of the *Cochliobolus heterostrophus* glyceraldehyde-3-phosphate dehydrogenase. Curr Genet 22:29–35

Zhou Y-H, Ragan MA (1993) cDNA cloning and characterization of the nuclear gene encoding chloroplast glyceraldehyde-3-phosphate dehydrogenase from the marine red alga *Gracilaria verrucosa*. Curr Genet 23:483–489

Zhou Y-H, Ragan MA (1994) Cloning and characterization of the nuclear gene encoding plastid glyceraldehyde-3-phosphate dehydrogenase from the marine red alga *Gracilaria verrucosa*. Curr Genet 26:79–86

Zhou Y-H, Ragan MA (1995) Nuclear-encoded protein-coding genes of the agarophyte *Gracilaria verrucosa* (Hudson) Papenfuss. Proc Intl Seaweed Symp 15 (in press)

Zhu T, Schupp JM, Oliphant A, Keim P (1994) Hypomethylated sequences: characterization of the duplicate soyabean genome. Mol Gen Genet 224:638–645