Cloning of the gene (*gdcH*) encoding H-protein, a component of the glycine decarboxylase complex of pea (*Pisum sativum* L.)

David MACHEREL, Jacques BOURGUIGNON* and Roland DOUCE

Laboratoire de Physiologie Cellulaire Végétale (Unité Associée au Centre National de la Recherche Scientifique N° 576),
Département de Biologie Moléculaire et Structurale, Centre d'Etudes Nucleaires et Université Joseph Fourier, 85 X,
F-38041 Grenoble-Cédex, France

H-protein is the lipoyl-protein component of the glycine decarboxylase complex, which catalyses, with serine hydroxymethyltransferase, the mitochondrial step of photorespiration in plants. We have isolated and characterized the gene (*gdcH*) encoding the H-protein in pea (*Pisum sativum* L.). The H-protein gene is distributed in a stretch of about 1.55 kbp and contains three introns (75, 64 and 185 bp) located in the coding region. No intervening sequences were detected in the 5' and 3' non-coding regions. This intron–exon structure contrasts with the preliminary H-protein gene structures reported for human and chicken, where these genes (dispersed on 13 and 8 kbp genomic fragments respectively) are composed of five highly conserved exons and are interrupted by four long introns. Two main transcription sites were detected by primer extension of RNA. The first transcriptional initiation site was assigned the +1 position and correlated with a putative TATA box located at position –26. The second transcriptional start site was not correlated with a putative TATA box, but may be regulated by an 'initiator' element described by Smale & Baltimore [(1989) *Cell* (Cambridge, Mass.) 57, 103–113] which contains, within itself, the transcription start site. The presence of two potential promoters may be related to the specialized overexpression pattern of H-protein in leaves, in order to support photorespiration.

INTRODUCTION

Leaf mitochondria from C₃-pathway plants oxidize glycine at extremely high rates (Douce *et al.*, 1977) in order to support the metabolic process of photorespiration. The breakdown of glycine is performed by the glycine decarboxylase multienzyme complex which is located in the mitochondrial matrix space (Bourguignon *et al.*, 1988). The glycine decarboxylase complex consists of four different proteins (Kikuchi, 1973; Walker & Oliver, 1986; Bourguignon *et al.*, 1988). The 100 kDa P-protein binds the α -amino group of glycine through its pyridoxal phosphate cofactor. CO₂ is released, and the remaining methylamine moiety is then transferred to the lipoamide cofactor of the 13.9 kDa H-protein. The H-protein shuttles the methylamine group to the 45 kDa T-protein, where the methylene carbon atom is bound to soluble tetrahydropteroylpolyglutamate, yielding methylenetetrahydropteroylpolyglutamate, and the remaining amino nitrogen is released as NH₃. Finally, the 59 kDa L-protein reoxidizes the lipoamide of H-protein by the sequential reduction of FAD and NAD⁺. The estimated stoichiometry of the complex is 2 P-protein dimers:27 H-protein monomers:9 T-protein monomers:1 L-protein dimer (Oliver *et al.*, 1990). The reaction mechanism implies a pivotal role for H-protein, which interacts with each of the other three proteins in the multienzyme complex. cDNA clones encoding the H-protein have been characterized in pea (Macherel *et al.*, 1990; Kim & Oliver, 1990), chicken (Yamamoto *et al.*, 1991) and human (Fujiwara *et al.*, 1991). A single-copy gene encoding the H-protein was detected in pea (*Pisum sativum* L.) genome (Macherel *et al.*, 1990). Recently, the preliminary structure of the H-protein genes of chicken and human were reported (Yamamoto *et al.*, 1991; Koyata & Hiraga, 1991). In the present paper we report the isolation and the

structural analysis of the H-protein gene (*gdcH*) from pea to explore the molecular mechanisms that control the expression of this critical gene.

EXPERIMENTAL

Gene isolation

Nuclear DNA was prepared from isolated nuclei (Watson & Thompson, 1986). DNA (20 μ g) was digested overnight with restriction endonuclease *Hind*III. After phenol extraction the digested DNA was precipitated and electrophoretically separated though 0.6%-agarose gel in TBE buffer (90 mM-Trizma base/90 mM-boric acid/2 mM-EDTA, pH 8.0). The region of the gel between 5 and 9 kbp was excised, and DNA was electroeluted into a dialysis bag (Maniatis *et al.*, 1982). After isobutanol, phenol and chloroform extraction the DNA was precipitated with ethanol and suspended at 200 ng/ μ l in TE buffer (10 mM-Tris/1 mM-EDTA, pH 8.0). λ -DNA (λ strain NM 1149) was digested with *Hind*III, followed by phenol extraction and ethanol precipitation. A 1 μ g portion of vector DNA was ligated overnight with 400 ng of insert DNA (5–9 kbp nuclear DNA digested with *Hind*III). The entire ligated DNA was packaged using Gigapack Plus extracts (Stratagene). The host, *Escherichia coli* POP 13, was used to grow the recombinant phage (900 000 plaque-forming units). For screening, 60 000 phages were plated on six square (10 cm) plates and grown at 37 °C for several hours. Plaques were blotted on Hybond N⁺ (Amersham International) according to the manufacturer's instructions. Filter hybridization was carried out in 50 % (v/v) formamide/0.2 % (w/v) polyvinylpyrrolidone (*M_r* 40 000)/BSA 0.2 % (w/v)/Ficoll (400 000)/0.1 % (w/v) sodium pyrophosphate/1 M-NaCl/1 % SDS/denatured salmon sperm DNA (100 μ g/ml)/50 mM-Tris/

Abbreviations used: TE buffer, 10 mM-Tris/1 mM-EDTA, pH 8.0; TBE buffer, 90 mM-Trizma base/90 mM-boric acid/2 mM-EDTA, pH 8.0; 1 $\times$ SSC, 0.15 M-NaCl/0.15 M-sodium citrate; poly(A)⁺, polyadenylated.

* To whom correspondence should be sent.

The nucleotide sequence data reported in this paper will appear in the EMBL, Genbank and DDBJ Nucleotide sequence Databases under the accession number X64726.

Fig. 1. Sequence strategy and intron-exon of the H-protein gene

The bottom part shows the *Hind*III genomic fragment pNH and the *Xba*I site used for subcloning (pNH1, pNH2). The region subjected to sequence analysis is represented as a continuous unbroken line and is expanded above. The 5' end of the gene is on the left, and the dark areas correspond to exon structures. Arrows indicate the sequence strategy on both strands.

HCl, pH 7.5. The DNA probe was a 811 bp cDNA encoding the H-protein (Macherel *et al.*, 1990). A 300 ng portion of DNA probe labelled with 32 P (Multiprime labelling system; Amersham International) was used for each hybridization. Hybridization was at 37 °C for 12 h. The filters were washed twice with $2 \times$ SSC/0.1 % SDS at room temperature, followed by $1 \times$ SSC/0.1 % SDS at 65 °C for 30 min ($1 \times$ SSC is 0.15 M-NaCl/0.015 M-sodium citrate). The filters were then autoradiographed for 1 h at -70 °C.

A positive phage was shown to contain an insert of about 5.5 kbp, which was subcloned in pUC 18 as a *Hind*III fragment creating the plasmid pNH. A unique *Xba*I site in the cDNA sequence was used to subclone the insert as two fragments in pUC 18. pNH1 contained a large fragment (about 4.5 kbp) corresponding to the 5' region of the gene up to the *Xba*I site of the cDNA. The plasmid pNH2 contained a 1250 bp fragment corresponding to the region of the gene downstream of the cDNA *Xba*I site. A second *Xba*I site was found 600 bp upstream that was not easily cleaved by the restriction enzyme, thus allowing the subcloning in two fragments only.

DNA sequencing

Plasmid sequencing was performed on both strands by the dideoxy chain-termination method (Sanger *et al.*, 1977) using 2'-deoxyadenosine 5'-[α - 32 S]thio]triphosphate (Amersham International) and Sequenase version 2.0 (United States Biochemical Corp.). pUC reverse and sequencing primers were used to sequence from each end of the inserts. Oligonucleotides specific to the sequence were synthesized with a DNA synthesizer (model 381A; Applied Biosystems) and used as primers.

The nucleotide sequence was managed with PC Gene software (Intelligenetics, Mountain View, CA, U.S.A.).

Primer extension

Poly(A)⁺ RNA was prepared from the leaves of 8-day-old pea plants grown in vermiculite at 25 °C in a growth cabinet with a 12 h light period as described by Macherel *et al.* (1990). A specific 27-mer oligonucleotide (3' TACCGTGAATCCTACACCCGAAGAAGT 5') complementary to the cDNA (nucleotides 95–121; see Macherel *et al.*, 1990) was used as a primer for extension by reverse transcriptase. The primer (5 pmol) was annealed to 5 μ g of pea leaf poly(A)⁺ RNA in a volume of 10 μ l containing 10 mM-MgCl₂, 25 mM-NaCl and 20 mM-Tris/HCl, pH 7.5. After 2 min at 70 °C, the sample was slowly cooled to 30 °C and kept on ice. A 5 μ l portion containing 20 mM-dithiothreitol, 1.5 μ M (each)-dGTP+dCTP+dTTP, 20 μ Ci of [α - 32 S]dATP (> 1000 Ci/mmol) and 20 units of avian-myeloblastosis-virus reverse tran-

scriptase was then added to the sample. After 5 min at 25 °C, 8 μ l of 50 mM-NaCl/180 mM (each)-dGTP+dATP+dCTP+dTTP was added, and incubation was continued at 42 °C for 10 min. The reaction was then stopped by adding 0.1 vol of 3 M-sodium acetate, pH 5.3, and 2 vol. of ethanol. After centrifugation (15000 g, 10 min) the nucleic acids were washed with 70 % (v/v) ethanol and suspended in 1 mM-EDTA/10 mM-Tris/HCl, pH 8.0. The elongation products were analysed by denaturing polyacrylamide (6%, w/v)-gel electrophoresis.

RESULTS AND DISCUSSION

Gene isolation

Since it was previously shown that a single fragment of approx. 6 kb, generated by digestion of pea nuclear DNA with *Hind*III, hybridized to an H-protein cDNA clone (Macherel *et al.*, 1990), we decided to construct a size-selected genomic library. Pea nuclear DNA was digested with *Hind*III and the fragments between 5 and 9 kbp were cloned in λ -phage NM 1149. The resulting library was screened by hybridization to a cDNA probe complementary to the H-protein mRNA (Macherel *et al.*, 1990). A hybridizing phage was characterized in detail and contained a 5.5 kbp *Hind*III restriction fragment which hybridized with the H-protein cDNA (results not shown). For further study, this restriction fragment was cloned in pUC 18, creating plasmid pNH. Taking advantage of a unique *Xba*I restriction site in the cDNA sequence, the inserted fragment of pNH was subcloned in two parts in pUC 18. The plasmid pNH1 contained a 4.5 kbp *Hind*III-*Xba*I fragment corresponding to the gene region upstream of the cDNA *Xba*I site (Fig. 1). A second *Xba*I site was found 560 bp upstream. For unknown reasons it was not easily cleaved by the restriction enzyme. The plasmid pNH2 was shown to contain all the gene sequence downstream of the *Xba*I site in a 1250 bp fragment (Fig. 1).

Gene structure

The sequencing strategy is shown in Fig. 1, and the nucleotide sequence is displayed in Fig. 2. The sequence starts 516 nucleotides upstream of the open reading frame and continues 470 nucleotides beyond the stop codon. Comparison with the cDNA sequence indicates that the gene contains three introns (75, 64 and 185 bp) in the coding region (Figs. 1 and 2). No intervening sequences were detected in the 5' or 3' non-coding regions. Exon 1 contains all of the 5' non-coding sequence, and the coding information for the first 36 amino acids of the protein, including a putative signal peptide. This intron-exon structure sharply

The C (+1) and A (+35) with an asterisk above correspond to the major transcription start sites. Introns are depicted as lower-case letters. The underlined As at positions 1004 and 1152 correspond to the two polyadenylation sites found in the cDNA sequence. Putative TATA boxes revealed by computer analysis are underlined in the upstream region of the gene. The complete cDNA sequence of the H-protein is given in Macherel *et al.* (1990). The overlined sequence is the sequence complementary to the 27-mer oligonucleotide used in the transcript mapping (Fig. 3).

Vol. 286

In order to identify the transcriptional start site, a specific primer complementary to the beginning of the mRNA coding region (see Fig. 2) was used during primer-extension experiments with reverse transcriptase. Two bands were clearly detected (Fig. 3). Such a situation could be interpreted by the presence of two transcriptional initiation sites, by a pausing artifact during reverse transcriptase or by an RNAse-sensitive cleavage site.

However, identical signals were observed when the experimental conditions were modified by changing the annealing temperature, by labelling the primer instead of the elongation products or by changing the temperature and time of reverse transcription (results not shown). Also we could not find any significant secondary structure in this region by computer analysis of the mRNA sequence. The position +1 in the gene nucleotide sequence was tentatively attributed to the nucleotide corresponding to the longer elongation product found in the primer-extension experiment (−63 from the initiation codon). A computer scan for promoter elements in the upstream region of the gene was carried out using the PC Gene program EUKPROM (Intelligenetics), which is based on the work of Bucher (1990). Several TATA boxes were detected (underlined in Fig. 2). The sequence was also scanned for cap signals and the two types of elements were correlated. The TATA box centred at position −26 appeared to be correlated with six Cap signals respectively occurring at positions −7, −5, −3, +2, +4 and +14. The upstream TATA box (−103) is correlated with a cap site at −59, but we could never detect a signal at this position when performing primer extension of mRNA. When considering the results of primer extension and computer analysis of the upstream region of the gene, we suggest that the putative −26 TATA box could direct the initiation of transcription at the observed +1 position. The second transcription start site observed at position +35 (Figs. 2 and 3) most likely cannot be attributed to the same −26 TATA box because the distance between these two elements (61 bases) far exceeds the almost invariant distance of 30 bases (Breathnach & Chambon, 1981; Bucher, 1990). It may be possible that this transcriptional initiation site lacks a TATA-box element or AT-rich element. The sequence TCATTCT surrounding the A (+35) of the second transcriptional start site exactly matches the 'initiator' element described by Smale & Baltimore (1989). These authors described an 'initiator' element, containing the RNA start site, that was sufficient for accurate, basal-level, transcription of a gene in the absence of a TATA box. Such an observation suggests that a similar discrete promoter could direct the transcription of the H-protein gene from the second initiation site. From the primer-extension experiments this second initiation site appears to be predominant (Fig. 3). Surprisingly, the cDNA sequenced previously (Macherel *et al.*, 1990) is longer (15 bases) than the mRNA expected from the first transcription start (+1). This longer cDNA is likely to correspond to a rare species of mRNA, as it never appeared in primer-elongation experiments, even after overexposure. It should be noted that several cDNAs have been shown to exceed the length of the bulk mRNA (for a review, see Kozak, 1987).

Although the expression of H-protein gene in the leaf is stimulated by light, the balance between the two major RNA start sites was similar when dark-grown-leaf RNA was used as a template for primer extension (results not shown). The potential presence of two promoters may be related to the specialized function of glycine decarboxylase complex in plants. In animals the complex is involved in the one-carbon-unit metabolism and is found in low amounts in mitochondria (Motokawa & Kikuchi, 1971). In plants, such a basal expression may occur, but is certainly negligible compared with the very high level of ex-

pression in leaves (Macherel *et al.*, 1990; Kim & Oliver, 1990) required for photorespiration. Indeed, we have demonstrated that the protein components of the glycine-cleavage system represent about one-third of the soluble protein of the matrix of pea leaf mitochondria (Oliver *et al.*, 1990). It should be noted that separate basal-level and tissue-specific expressions were reported for chicken H-protein (Yamamoto *et al.*, 1991). Such a pattern of expression might only be directed by the interaction of several promoters.

Previously we showed that the steady-state level of mRNA for H-protein was higher in light-grown green leaves than in dark-grown leaves. The upstream sequence of the gene (438 bases) did not reveal the conserved light-regulatory elements (GT boxes) found in the −140 region of ribulose biphosphate carboxylase/oxygenase genes in pea, tobacco (*Nicotiana tabacum*) and *Petunia* (Kuhlemeier *et al.*, 1987). A functional promoter analysis will be required to define the eventual light-regulatory and tissue-specific elements. It should be noted that the most common elements, GC and CAAT boxes, were not found in the upstream region of the H-gene. Finally, the availability of this clone will permit informative future studies on the regulation of expression of this gene as well as investigations of structure–function relationships with coding sequences altered by mutagenesis *in vitro*.

REFERENCES

- Bourguignon, J., Neuburger, M. & Douce, R. (1988) *Biochem. J.* **255**, 169–178
- Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383
- Bucher, P. (1990) *J. Mol. Biol.* **212**, 563–578
- Douce, R., Moore, A. L. & Neuburger, M. (1977) *Plant Physiol.* **60**, 625–628
- Fujiwara, K., Okamura-Ikeda, K., Hayasaka, K. & Motokawa, Y. (1991) *Biochem. Biophys. Res. Commun.* **176**, 711–716
- Jackson, I. J. (1991) *Nucleic Acids Res.* **19**, 3795–3798
- Juretic, N., Jaussi, R., Mattes, U. & Christen, P. (1987) *Nucleic Acids Res.* **15**, 10083–10086
- Kikuchi, G. (1973) *Mol. Cell. Biochem.* **1**, 169–187
- Kim, Y. & Oliver, D. J. (1990) *J. Biol. Chem.* **265**, 848–853
- Koyata, H. & Hiraga, K. (1991) *Biochem. Biophys. Res. Commun.* **178**, 1072–1077
- Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125–8148
- Kuhlemeier, C., Green, P. & Chua, N. H. (1987) *Annu. Rev. Plant Physiol.* **38**, 221–257
- Macherel, D., Lebrun, M., Gagnon, J., Neuburger, M. & Douce, R. (1990) *Biochem. J.* **268**, 783–789
- Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: a Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
- Motokawa, Y. & Kikuchi, G. (1971) *Arch. Biochem. Biophys.* **146**, 461–466
- Oliver, D. J., Neuburger, M., Bourguignon, J. & Douce, R. (1990) *Plant Physiol.* **94**, 833–839
- Sanger, F., Nickler, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
- Senapathy, P., Shapiro, M. B. & Harris, N. L. (1990) *Methods Enzymol.* **183**, 252–278
- Smale, S. T. & Baltimore, D. (1989) *Cell* (Cambridge, Mass.) **57**, 103–113
- Walker, J. L. & Oliver, D. J. (1986) *J. Biol. Chem.* **261**, 2214–2221
- Watson, J. C. & Thompson, W. F. (1986) *Methods Enzymol.* **118**, 57–75
- Yamamoto, M., Koyata, H., Matsui, C. & Hiraga, K. (1991) *J. Biol. Chem.* **266**, 3317–3322